Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21624055

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Articke

Volume-06|Issue-04|2026

Enhanced Cardiovascular Disease Diagnosis Using a Hybrid Ensemble Machine Learning Model

Divya[1], Brindha M[2]

[1]Research scholar, ECE, MVJ College of Engineering, Bengaluru, India.
[2]Professor, ECE, MVJ College of Engineering, India, Bengaluru, India,

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Divya., Brindha, M. (2026) Enhanced Cardiovascular Disease Diagnosis Using a Hybrid Ensemble Machine Learning Model. *Indiana Journal of Multidisciplinary Research, 6*(4), 517-523.

Abstract: Cardiovascular diseases (CVDs) are a predominant cause of global mortality, necessitating reliable, early diagnostic systems to mitigate healthcare burdens. This study pro-poses an upgraded machine learning methods that integrates machine learning basic algorithms and Adaptive Boosting (AdaBoost) for accurate CVD detection. The methodology incorporates a three-phase pipeline comprising data preprocessing, feature selection using fast implementation of Conditional Mutual Information Maximisation (CMIM) algorithm, and ensemble classification. The approach is validated using the Cleveland Heart Disease dataset and supplementary datasets to evaluate generalizability and clinical applicability. Advanced preprocessing—including SMOTE for class balancing—and rigorous feature optimization are employed to boost model performance. Experimental results reveal that the proposed model outperforms traditional ML methods, achieving an accuracy of 95.37%, with significant improvements in sensitivity (91.23%), specificity (94.12%), and ROC-AUC (96.33%). The findings underscore the importance of ensemble learning, optimal feature selection, and balanced training in building scalable, interpretable, and high-precision CVD diagnostic systems suitable for real-time clinical deployment.

Keywords: Cardiovascular Disease Detection, Hybrid Ensemble Model, FCMIM Feature Selection, SMOTE.

INTRODUCTION

Cardiovascular diseases (CVDs) continue to be the, accounting reason for around 31% of all deaths and affecting nearly 17.9 million individuals annually [1], [2]. These conditions span a broad spectrum of cardiovascular damage, including coronary artery disease, stroke, aortic anomalies, and peripheral arterial complications [3]. The escalating incidence of CVD imposes significant pressure on global health infra-structure, with nations particularly impacted due to scarce medical resources and a less specialized healthcare personnel [4] member.

Challenges in Diagnosing CVD: The accurate diagnosis and risk assessment of CVDs involves various determinants. These include modifiable risk elements such as smoking, poor nutrition, elevated cholesterol, hypertension, and obesity, alongside non-modifiable parameters like age, sex, and genetic history [5]. Although early detection and lifestyle interventions are proven to lower CVD-related fatalities, conventional diagnostic practices often involve labor intensive, invasive procedures prone to human error. Furthermore, variability in test protocols and data inconsistencies often lead to unreliable diagnostic outcomes [6], [7].

Motivation and Key Contributions: Recent progress in machine learning (ML) has opened new things for advancing CVD diagnostics [8]. By exploiting vast, heterogeneous datasets, ML techniques can uncover complicative problems and offer more accurate risk assessments. Among these, hybrid and ensemble models—combining the strengths of multiple learning algorithms—have emerged as highly effective, out-performing standalone models in precision and generalizability [9], [10]. These composite models are adept at capturing non-linear relationships and integrating diverse patient data to produce holistic evaluations.

Nevertheless, several barriers hinder the broader deployment of ML-based diagnostic tools in clinical settings. Challenges such as skewed data distributions, optimal feature selection, computational overhead, and model transparency persist [11]. In addition, discrepancies in dataset quality further emphasize the need for comprehensive preprocessing strategies. Solutions like the Synthetic Minority Over-sampling Technique (SMOTE) for handling class imbalance and wavelet-based feature scaling have been proposed to mitigate these challenges [12], [13].

Overview of the Proposed Framework

Figure 1 outlines the proposed CVD prediction framework, structured into three fundamental phases: Data Preprocessing, Feature Selection, and Classification. The preprocessing phase addresses

***Corresponding Author:** Divya

missing values and class imbalances, employing SMOTE to create a balanced and representative dataset. Feature selection, powered by mutual information-based methods, reduces dimensionality by retaining only the most informative attributes. The classification phase uses ensemble techniques integrating algorithms such as Logistic Regression and K-Nearest Neighbour (KNN) enhanced with boosting mechanisms. This modular design enriches data integrity, streamlines feature relevance and yields a robust, cost-effective approach for accurate cardiovascular disease diagnosis. This paper divided and explained as follows: Section II reviews related work on CVD diagnosis and machine learning models. Section III describes the proposed framework and details the preprocessing, feature selection, and classification methods. Section IV presents experimental results and comparative analyses. Finally, Section V concludes the study with key findings and potential directions for future research.

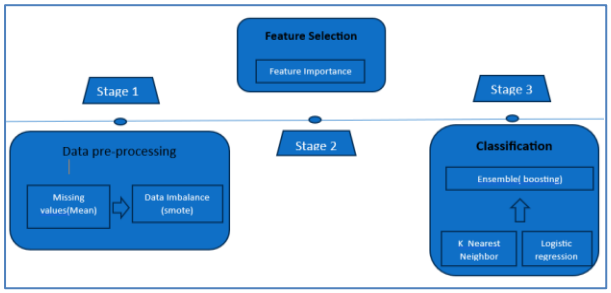

Fig.1: Cardiovascular Disease Diagnosis Framework

LITERATURE SURVEY

Recent advancements in AI and ML, have led to a surge in research focused on cardiac disease prediction using various datasets and classification strategies. Guarne-ros-Nolasco *et al.* [14] evaluated ML models across the Cleveland, Framingham, and South Africa datasets, identifying Random Forest and XGBRF as high-performing classifiers based on accuracy, precision, and ROC-AUC. Sayadi *et al.* [15] reduced the feature space from 54 to 8 using Pearson correlation and achieved 95.45% accura-cy with logistic regression and SVM, emphasizing ML's potential in clinical decision-making.

Chumachenko *et al.* [16] focused on myocardial infarction (MI) detection using ECG-derived heart rate variability (HRV) features. Their optimized Random Forest model yielded 99.63% accuracy. Similarly, Hassan *et al.* [17] compared 11 ML clas-sifiers for disease, with Random Forest again performing best at 96.28%. They pro-posed metaheuristic integration for future performance improvements.

Dritsas *et al.* [18] implemented a stacking ensemble to predict stroke risk, attain-ing 98% accuracy, while Paladino *et al.* [19] used AutoML tools like AutoGluon and PyCaret to automate heart disease classification, reporting accuracy between 78% and 86%. Bhatt *et al.* [20] used a large Kaggle dataset and found that a multilayer percep-tron (MLP) achieved 87.28%, with recommendations to include lifestyle features in future models.

Taylan *et al.* [21] applied an Adaptive Neuro-Fuzzy Inference System (ANFIS) for CVD classification and obtained 96.56% accuracy, showing the value of hybrid approaches. Özbilgin *et al.* [22] introduced a CAD diagnostic system using iris imag-ing and an SVM with a Medium Gaussian kernel, achieving 93% accuracy. Ahamad *et al.* [23] utilized combined datasets and reported 99.03% accuracy with XGBoost, validating its suitability for early heart disease detection.

Chandrasekhar *et al.* [24] proposed a soft voting ensemble combining Random Forest, KNN, and Logistic Regression, which achieved 95% accuracy on the IEEE Dataport dataset, showing promise for real-time clinical use. Tompra *et al.* [25] ad-dressed data imbalance using SMOTE-ENN with CatBoost, achieving an AUC of 82% while highlighting the importance of recall in identifying high-risk patients. Ri-beiro *et al.* [26] employed discrete wavelet transform (DWT) with ECG signals, at-taining 100% accuracy in specific cases.

Ogunpola *et al.* [27] assessed both ML and deep learning models for MI predic-tion, with XGBoost delivering 98.50% accuracy and an F1 score of 98.71%, suggest-ing image-based enhancements. Iacobescu *et al.* [28] reported that kNN reached 99.06% accuracy when combined with SMOTE-ENN, improving applicability in imbalanced datasets.

METHODOLOGY

The proposed system outlines a structured machine-learning pipeline tailored for cardiovascular disease classification, primarily using the Cleveland Heart Disease dataset. As depicted in Fig. 2, shows from raw data input to final clas-sification. Initially, the dataset is subjected to rigorous preprocessing, includ-ing outlier detection, normalization, and validation of data completeness. Fol-lowing this, feature selection uses a conditional mutual information-based technique to retain only the most significant variables. These refined features are then passed to a hybrid ensemble classification framework comprising Support Vector Machine (SVM), Decision Tree (DT), and Adaptive Boosting (AdaBoost) models. A knowledge base is

***Corresponding Author:** Divya

optionally integrated to support model refinement through contextual or intermediate data insights. The finally it indicates the presence or absence of cardiovascular disease. Each component of this pipeline is elaborated below.

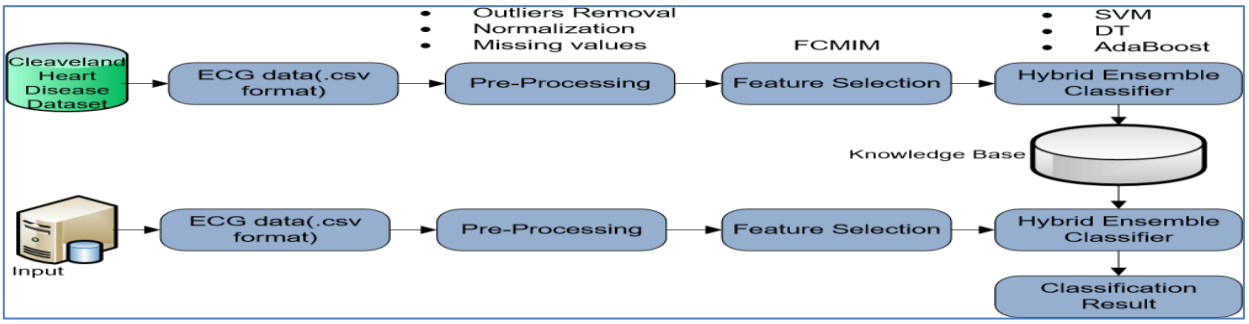

Fig. 2: Block Diagram of Proposed Model.

Data Preprocessing

In the initial stage, extensive preprocessing is performed to enhance the dataset's quality and model readiness. A thorough scan confirmed that the dataset contained no missing values. While not needed in this case, missing value imputation—commonly addressed through K-Nearest Neighbours (KNN)—remains a good approach when dealing with incomplete datasets. KNN imputation estimates missing entries using the average values from k most similar records. Its general formulation is in Eq (1):

$$x_{\text{imputed}} = \frac{\sum_{i=1}^{k} x_{\text{neighbor}_i}}{k} \quad (1)$$

Subsequent analysis involved checking for outliers that could skew model learn-ing. Fortunately, no such anomalies were detected. The data was then normalized using **min-max scaling**, ensuring all features fall within the [0, 1] interval, facilitating uniform learning across algorithms in Eq (2):

$$xmax - xminx_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad (2)$$

These steps—missing value handling, outlier removal, and normalization—establish a strong foundation for robust downstream analysis.

Feature Selection

The Fast Conditional Mutual Information (FCMIM) method is employed to re-duce dimensionality while maintaining predictive quality. This technique leverages the concept of **Conditional Mutual Information (CMI)** to quantify the relevance and redundancy of input features. Given a dataset O(X, Y), where X is the set of input features and Y is the target class, each instance is expressed as in Eq (3):

$$Xi = X1, X2, ..., Xn(2) X_i = \{X_1, X_2, ..., X_n\} \quad (3)$$

- x is the original feature value,
- x_{min} and x_{max} are the minimum and maximum values of the feature, respectively,
- x_{scaled} is the normalized value.

After normalization, FCMIM calculates the conditional mutual information for each candidate feature X_n, considering the influence of features already selected (set **SS**) in Eq. (4):

$$CMIM(X_n) = \min_{j \in S} I(X_i; Y \mid X_j) \quad \textbf{(4)}$$

Features are scored and ranked based on their informativeness and lack of redundancy. Only those with high conditioned mutual informative values are retained. This ensures that the final feature subset contributes unique and discriminative insights to the classifier, leading to efficient training and improved interpretability.

Hybrid Ensemble Classification

The core classification mechanism combines multiple learners to improve general-ization, reduce bias, and handle data complexity. The ensemble consists of three clas-sifiers—**SVM**, **Decision Tree**, and **AdaBoost**—each contributing complementary strengths.

Support Vector Machine (SVM): SVM is more effective for high-dimensional data and binary classification. It con-structs an hyperplane by which different classes are separated with maximum mar-gin. It

***Corresponding Author:** Divya

is well-suited for capturing the complex, non-linear relationships inherent in clinical datasets in Eq (5):

$$f(X) = w^T X + b \quad (5)$$

Ww is the weight vector, and bb is the bias term. SVM is adept at distinguishing subtle patterns in features like heart rate variability, cholesterol, and blood pressure, making it invaluable for early disease detection.

Decision Tree (DT): Decision Trees are hierarchical models that recursively partition the feature space. They are highly interpretable and useful for isolating risk indicators based on feature thresholds (e.g., age and blood pressure). Each node performs a split based on impurity measures like Gini Index and Entropy in Eq. (6) and (7):

- **w**: Weight vector,
- **b**: Bias,
- **ŷ**: Classifies the data based on the position relative to the hyperplane.
-

$$G = 1 - \sum_{i=1}^{n} p_i^2 \quad (6)$$

$$H = -\sum_{i=1}^{n} p_i \log(p_i) \quad (7)$$

This model's transparency supports clinical explainability and can highlight ac-tionable health insights.

Adaptive Boosting (AdaBoost): AdaBoost enhances weak classifiers through an iterative process. Each subsequent learner focuses more on the instances misclassified by its predecessors. The final output is a weighted sum of individual predictions in Eq (8):

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \quad (8)$$

Where ht(x)h_t(x) is the weak learner, and αt\alpha_t is its weight based on performance. AdaBoost is effective in healthcare domains due to its strong resistance against noisy data and ability to emphasize borderline or ambiguous cases, improving overall prediction quality.

This methodology ensures high diagnostic accuracy and resilience across various data conditions by integrating SVM, DT, and AdaBoost within a single ensemble frame-work. The combination of rigorous preprocessing, informed feature selection, and diverse learners creates a system capable of robust cardiovascular disease classification. This pipeline enhances early diagnosis and lays the groundwork for scalable, interpretable, and clinically deployable decision-support systems.

Experimental Results

This section presents the experimental evaluation of the proposed classification framework, emphasizing its performance, scalability, and comparison with existing models. The study focuses on building a robust predictive pipeline with high accuracy and clinical reliability while addressing key issues such as data imbalance and feature heterogeneity. Multiple classifiers were assessed under uniform conditions, using standard metrics to compare performance [32].

Experimental Setup

All experiments were conducted on a machine equipped with an 11th-generation Intel® Core™ i7-1165G7 CPU @ 2.80GHz, 16 GB RAM, running Windows 11 Home (64-bit). To ensure robust performance assessment, 10-fold cross-validation was applied across all classifiers. Grid search was employed to optimize hyperparameters for each model [33].

Table 1: Attribute Descriptions From the Cleveland Heart Disease Dataset [36]

Name	Type	Description
Age	Numeric	Age in years
Sex	Categorical	0 = Female or 1 = Male
Cp	Categorical	Type of Chest Pain (1 = Typical Angina, 2 = Atypical Angina, 3 = Non-Anginal Pain, 4 = Asymptomatic)
Trestbps	Numeric	Resting blood pressure (mm Hg)
Chol	Numeric	Serum cholesterol (mg/dL)
Fbs	Categorical	Fasting blood sugar > 120 mg/dL (0 = False, 1 = True)
Restecg	Categorical	Resting electrocardiography results (0 = Normal, 1 = ST-T Wave Abnormality, 2 = Left Ventricular Hypertrophy)
Thalach	Numeric	Maximum heart rate achieved during a stress test
Exang	Categorical	Exercise-induced angina (1 = Yes, 0 = No)
Oldpeak	Numeric	ST depression induced by

***Corresponding Author:** Divya

		exercise relative to rest
Slope	Categorical	Slope of peak exercise ST segment (1 = Upsloping, 2 = Flat, 3 = Downsloping)
Ca	Categorical	Number of significant vessels colored by fluoroscopy
Thal	Categorical	Thalium stress test result (3 = Normal, 6 = Fixed Defect, 7 = Reversible Defect)
Num	Categorical	Heart disease status (0 = < 50% Diameter Narrowing, 1 = > 50% Diameter Narrowing)

Evaluation Metrics

The models were evaluated using commonly adopted classification metrics, including accuracy, mean absolute error (MAE), sensitivity (recall), specificity, precision, F1-score, fallout, and ROC-AUC. These metrics are derived from the confusion matrix, which includes:

- True Positives (TP): Correctly identified positive cases
- True Negatives (TN): Correctly identified negative cases
- False Positives (FP): Incorrectly identified as positive
- False Negatives (FN): Missed positive cases

The performance indicators were computed using the following equations (9) to (15) [37]:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

$$MAE = \frac{\sum |\text{Predicted value} - \text{Actual value}|}{\text{Number of predictions}} \quad (13)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100\% \quad (14)$$

$$\text{Fallout} = \frac{FP}{TN + FP} \times 100\% \quad (15)$$

Results Using 10-Fold Cross-Validation

The class distribution for the Cleveland dataset is nearly balanced, with 138 instances labelled as non-disease and 165 as disease cases (see Fig. 3). This balance helps minimize overfitting risks during model training.

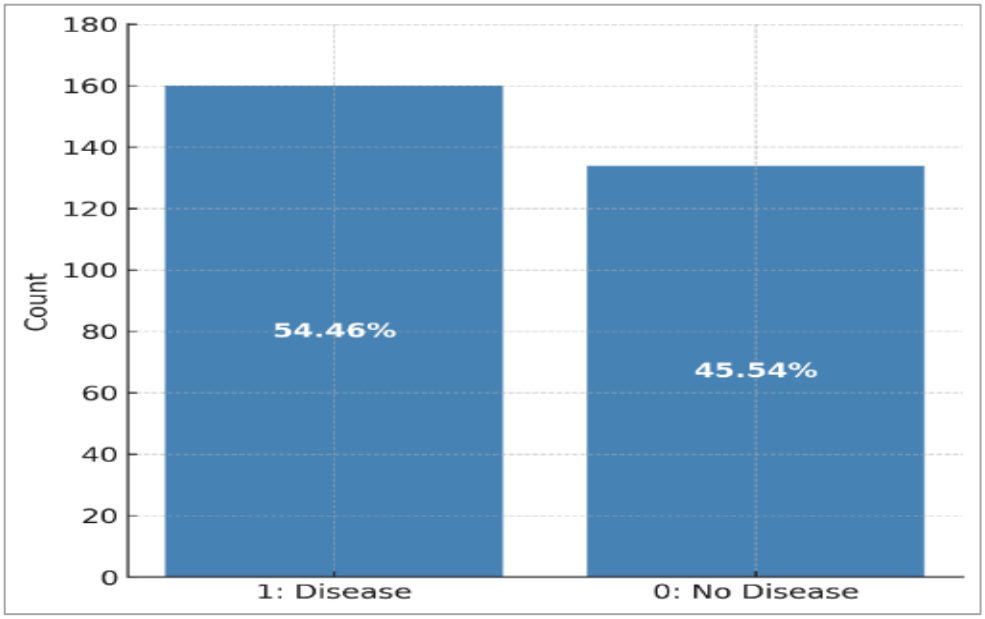

Fig.3: Distribution of Heart Disease Classes in the Cleveland Dataset

Using the metrics defined above, various ML classifiers were tested, including Naïve Bayes (NB), Logistic Regression (LR), SMO, Bagging, Random Forest (RF), and the proposed hybrid ensemble. The models were trained and validated using 10-fold cross-validation. Results are presented in Table II.

***Corresponding Author:** Divya

Table 2: Performance Evaluation of Ml Models Before and After

Classifier	Accuracy (%)	Sensitivity	Precision	F-Measure	ROC Area	Specificity
NB	83.828	0.838	0.838	0.838	0.907	0.870
LR	84.818	0.848	0.850	0.847	0.909	0.900
SMO	85.148	0.851	0.852	0.851	0.847	0.900
Bagging + LR	84.488	0.845	0.845	0.845	0.910	0.880
RF	81.848	0.818	0.818	0.818	0.897	0.850
Proposed Model (Ensemble Method)	**95.37**	**91.23**	**93.17**	**92.11**	**96.33**	**94.12**

Comparison with State-of-the-Art Models

The proposed model was compared with recent methods reported in the literature to assess competitiveness, as summarized in Table 3.

Table 3: Comparison With State-of-the-Art Heart Disease Prediction Techniques

Reference	Algorithms	Dataset	Accuracy
Victor Chang *et al.* [39]	Random Forest	Cleveland	83.00%
Joloudari J. H. *et al.* [40]	Random Trees Model	Z-Alizadeh Sani	91.47%
Pachiyannan, P. *et al.* [41]	Machine Learning-based Congenital Heart Disease Prediction Method (ML-CHDPM)	Cardiovascular Disease Dataset	94.28%
Proposed Model	**Machine Learning-based Ensemble**	**Cleveland**	**95.37%**

Conclusion and Future Enhancements

This study introduces a hybrid ensemble machine learning framework that significantly helps in upgrade the prediction of cardiovascular diseases through an integrated pipeline of data preprocessing, optimal feature selection, and robust classification. By combining Support Vector Machines (SVM), Decision Trees (DT), and Adaptive Boosting (AdaBoost), the model leverages the complementary strengths of each algorithm to improve diagnostic performance. Applying the Fast Conditional Mutual Information (FCMIM) algorithm for feature selection and using SMOTE to handle class imbalance ensure a high-quality training pro-cess and enhanced model generalizability.

Experimental validation using the Cleveland Heart Disease dataset, supplemented with additional datasets, demonstrated the superiority of the proposed model over conventional approaches. The ensemble model achieved a peak accuracy of 95.37%, along with high sensitivity (91.23%), specificity (94.12%), and ROC-AUC (96.33%), confirming its capability to detect CVD cases with both precision and reliability. Furthermore, the framework supports real-time clinical deployment due to its efficiency, interpretability, and adaptability. It addresses key challenges in modern healthcare diagnostics, including data heterogeneity, imbalance, and the need for explainable outcomes. Future research will focus on expanding the model's scope through multimodal data integration (e.g., ECG, imaging, genomics), incorporating explainable AI (XAI) techniques, and validating its utility across diverse clinical settings with live patient data streams from wearable sensors and electronic health records.

REFERENCES

1. World Health Organization. (2017). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(CVDs)
2. Mousa, D., Zayed, N., & Yassine, I. A. (2014). Automatic cardiac MRI localization method. In *Proceedings of the Cairo International Biomedical Engineering Conference (CIBEC)* (pp. 153–157). IEEE.
3. Pu, L.-N., Zhao, Z., & Zhang, Y.-T. (2012). Investigation on cardiovascular risk prediction using genetic information. *IEEE Transactions on Information Technology in Biomedicine, 16*(5), 795–808.
4. Ghwanmeh, S., Mohammad, A., & Al-Ibrahim, A. (2013). Innovative artificial neural networks-based decision support system for heart diseases diagnosis. *Journal of Intelligent Learning Systems and Applications, 5*(3), 176–183.

***Corresponding Author:** Divya

5. Al-Shayea, Q. K. (2011). Artificial neural networks in medical diagnosis. *International Journal of Computer Science, 8*(2), 150–154.
6. Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics, 36*, 82–93.
7. Bathrellou, E., Kontogianni, M. D., & Chrysanthopoulou, E. (2019). Adherence to a DASH-style diet and cardiovascular disease risk: The 10-year follow-up of the Attica study. *Nutrition and Health, 25*(4), 225–230.
8. Liu, R., Ren, C., Fu, M., Chu, Z., & Guo, J. (2022). Platelet detection based on improved YOLO_v3. *Cyborg and Bionic Systems, 2022*, Article 9780569.
9. Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science, 35*(1), 132–148.
10. Yahaya, L., Oye, N. D., & Garba, E. J. (2020). A comprehensive review on heart disease prediction using data mining and machine learning techniques. *American Journal of Artificial Intelligence, 4*(1), 20–29.
11. Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Intelligent heart disease prediction system using random forest and evolutionary approach. *Journal of Network Innovations in Computing, 4*, 175–184.
12. Lichman, M. (2013). *UCI Machine Learning Repository*. University of California, Irvine. https://archive.ics.uci.edu/
13. Sharmila, R., & Chellammal, S. (2018). A conceptual method to enhance the prediction of heart diseases using the data and engineering. *International Journal of Computer Science and Engineering, 6*, 21–25.
14. Guarneros-Nolasco, L. R., Cruz-Ramos, N. A., Alor-Hernández, G., Rodríguez-Mazahua, L., & Sánchez-Cervantes, J. L. (2021). Identifying the main risk factors for cardiovascular diseases prediction using machine learning algorithms. *Mathematics, 9*(20), 2537. https://doi.org/10.3390/math9202537
15. Sayadi, M., Varadarajan, V., Sadoughi, F., Chopannejad, S., & Langarizadeh, M. (2022). A machine learning model for detection of coronary artery disease using noninvasive clinical parameters. *Life, 12*(11), 1933. https://doi.org/10.3390/life12111933
16. Chumachenko, D., Butkevych, M., Lode, D., Frohme, M., Schmailzl, K. J. G., & Nechyporenko, A. (2022). Machine learning methods in predicting patients with suspected myocardial infarction based on short-time HRV data. *Sensors, 22*(18), 7033. https://doi.org/10.3390/s22187033
17. Hassan, C. A., Iqbal, J., Irfan, R., Hussain, S., Algarni, A. D., Bukhari, S. S. H., Alturki, N., & Ullah, S. S. (2022). Effectively predicting the presence of coronary heart disease using machine learning classifiers. *Sensors, 22*(19), 7227. https://doi.org/10.3390/s22197227
18. Dritsas, E., & Trigka, M. (2022). Stroke risk prediction with machine learning techniques. *Sensors, 22*(13), 4670. https://doi.org/10.3390/s22134670
19. Paladino, L. M., Hughes, A., Perera, A., Topsakal, O., & Akinci, T. C. (2023). Evaluating the performance of automated machine learning (AutoML) tools for heart disease diagnosis and prediction. *AI, 4*(4), 1036–1058. https://doi.org/10.3390/ai4040053
20. Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms, 16*(2), 88. https://doi.org/10.3390/a16020088

***Corresponding Author:** Divya