International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **03 | 2026**

Comparative Analysis of Heart Risk Detection Models

Sharon Esther Abreo[1], Renata Reann Pinto[2], Vignesh Vijay Naik[3], Yuvraj Mahesh Hinde[4], Kedar Sawant[5], Shrikrishna Narvekar[6]

[1,2,3,4]Department of Computer Engineering, Agnel Institute of Technology and Design, Goa, India
[5,6] Assistant Professor, Department of Computer Engineering, Agnel Institute of Technology and Design, Goa, India

DOI: 10.5281/zenodo.20630102

ABSTRACT

Cardiovascular Diseases are recognized as a primary cause of morbidity, prevalent in most of the population. Early detection, effective triage and intervention are thus a necessity to provide timely care. Machine learning models can expedite the whole workflow of identifying potential risks, which is especially beneficial for communities that do not have adequate medical services. This study proposes a comparative analysis of 6 ML classification models to ascertain the potential of a heart risk. To develop the machine learning models, two datasets, available publicly, were used: UCI Heart Disease dataset and the Mendeley dataset. Comparative analysis between the "training" and "testing" indicated that the Mendeley dataset achieved higher accuracy and more consistent performance. Additionally, as it is derived from an Indian population and aligns with the target user group, the Mendeley dataset was selected for subsequent model development. Several ML models like XGBoost, SVM, Random Forest, LightGBM, Logistic Regression and Neural Network were tested to determine the optimally performing approach. To determine the most effective approach, model performance was analyzed through accuracy, precision, recall, F1-score, and ROC–AUC metrics. Out of the aforementioned models, XGBoost showed the best performance and highest accuracy with least false negatives. This research conducts a comparative assessment of multiple classification models to demonstrate the practical relevance of ML to change preventive care for CVDs, improving early intervention and aiding decision-making.

***Keyword:** - Machine Learning, Heart Disease, ML classification algorithms, XGBoost, SMOTE, SHAP.*

1. INTRODUCTION

Cardiovascular disease or CVD encompasses a variety of conditions resulting in significant health problems and deaths globally.[1]. Heart diseases occur due to various aspects such as elevated blood pressure, unhealthy dietary habits, obesity, high cholesterol, diabetes, and tobacco use. Genetic factors, limited access to healthcare, delayed identification and intervention often worsen these conditions[2]and individuals are often unaware of the conditions that can be the cause of strokes or failures[3].

While there exist multiple diagnostic approaches to evaluate cardiovascular risks[3], these techniques may be ineffective for timely discovery of the illness and may provide insufficient or incorrect details which lack the required precision in case of complex data[4]. Due to the increasing number of cases, it has fueled the necessity for fast and precise detection techniques which should provide reliable and accurate information[5]. By adopting AI-ML techniques, Support systems can adjust better to new environments and also learn from that data. These systems make use of ML techniques to aid the decision-making process[6].

For early detection of risks, multiple ML methods can be deployed before the patient's condition becomes more severe. Many studies examined that using these methods can reduce the expense of healthcare and detect any risks earlier than would have been identified. Owing to its problem solving capacity, merging machine learning with data analytics has attracted significant attention. "Deep Learning" (DL) and "Machine Learning" (ML) methods attain excellent performance while handling large datasets which helps improve risk classification and he diagnostic accuracy.[7] Thus healthcare professionals can provide improved diagnostic decisions and customized cures for patients. This also helps to extract meaningful insights that infer customized medical approaches improving the results of the treatment along with patient care[1].

Developing a predictive system utilizing ML requires data that is precise and balanced to "train" and "test" the models, which can improve its performance. The dataset features should be relevant and the data should be normalized. Preprocessing is necessary as the existing datasets may be irrelevant and inconsistent. According to research, to provide more reliable and precise predictions, at least 14 distinct features need to be used to train the model[8].

2. LITERATURE REVIEW

Numerous researchers have developed models to assess heart risk, highlighting the ability to greatly enhance patient outcomes. Despite this process, this field is continually evolving, with substantial opportunities for further research and innovation. This section systematically evaluates existing literature to establish the background for our development and benchmarking of six distinct ML models.

El-Sofany, Hosam et al. 2024[9] proposed a robust and transparent ML methodology for heart disease prediction at early phases. The methodological framework involved compiling a comprehensive dataset, and applying the "Synthetic Minority Oversampling Technique" (SMOTE) during pre-processing to effectively manage and mitigate class imbalance issues. The core of the approach was a detailed study of 10 supervised classification models, including XGBoost, RF, and SVM). Three unique methods for selection of features, namely, "ANOVA", "Chi-square", and "Mutual Information", utilized to curate optimized feature subsets "SF-1", "SF-2", and "SF-3") designed to maximize predictive accuracy. The ultimate methodology involved selecting the best-performing model (XGBoost achieved 97.57% accuracy on SF-2) and integrating it with methods of "Explainable AI (XAI)" using "SHAP (SHapley Additive exPlanations)" approaches to provide clinical interpretability. [9]

Ogunpola et al. (2024) [7] focused on the issues rooting from datasets which are imbalanced, inherent in CVD detection. The research deployed and evaluated seven distinct ML and DL classifiers. By meticulous hyperparameter optimization and fine-tuning, the final outcomes showed that an optimized XGBoost model stood out as most robust and highly performant, achieving a superior 98.50% accuracy, validating ensemble learning as an effective approach for developing an accurate and reliable clinical decision-support systems for early CVD diagnosis.

Majeed et al. (2024) [10] created a medical chatbot designed for preliminary heart disease identification based on user-provided symptoms. The research focuses on efficient patient triaging and risk assessment prior to clinical consultation. Models were evaluated on the "Cleveland Heart Disease dataset", using "k-best feature selection" and "Grid Search optimization".[10] After analysis, XGBoost achieved highest accuracy (97%).

Doppala et al. (2022) [11] proposed a ML model utilizing ensemble techniques. The core methodology involves using a Voting Ensemble Classifier, which aggregates the predictions from four distinct and powerful base classifiers: NB, RF, SVM, and XGBoost. The model reached accuracies of 96.75%, 93.39%, and 88.24%, on Mendeley, IEEE DataPort, and Cleveland respectively, demonstrating improved reliability and consistency for early CVD detection.

Sourov et al. [2] proposed a ML framework combining classification models for detecting CVD and for risk assessment, using dataset of 1035 cases enhanced via SMOTE. 11 ML classifiers were implemented Random Forest having the best classification accuracy (~97%) while Logistic Regression excelled in risk prediction. Techniques of XAI such as SHAP and LIME were used to improve global and local model understanding.

Chandrasekhar, N., and S. Peddakrishna[12] implemented a supervised ML framework for CVD classification using the Cleveland and IEEE Dataport cardiovascular dataset. After preprocessing, GridSearchCV-based "hyperparameter tuning" and "5-fold cross-validation" were implemented to six classifiers (LR, KNN, RF, NB, GB and AdaBoost) upon which a soft voting ensemble model to aggregate predictions. Optimized models achieved over 90% accuracy, while the ensemble improved performance to 93.44% (Cleveland) and 95% (IEEE), demonstrating enhanced robustness and generalization.

H. F. El-Sofany's 2024 study[13] addresses prediction of heart diseases by using ML to enhance early diagnostic accuracy. Three methods for feature selection ("chi-square", "ANOVA", and "mutual information") generated optimized subsets. 10 ML classifiers like SVM, RF, XGBoost, KNN, and others were evaluated across combined private and public datasets. To mitigate class imbalance, SMOTE oversampling is applied, and SHAP-based "XAI" methods are used to interpret model outcomes. XGBoost on the "SF-2" feature subset recorded superior accuracy and better overall metrics.

Stonier, A. A. et al.[14] trained ML models for cardiac disease risk prediction by analyzing clinical and data from electronic health records to evaluate the likelihood of heart attack occurrence in a patient. The paper compares several ML algorithms and finds that the RF classifier achieved superior predictive accuracy (≈88.52 %) among the tested approaches.

Singh, Amrit et al. [15] proposes a novel framework for early CVD detection to address overfitting issues commonly observed in existing models. Using supervised ML techniques like DT, RF, SVM and PCA the system evaluates performance of the model on imbalanced clinical data and compares accuracy using graphical analysis. The model

performs reliably in both development and evaluation stages, demonstrating the usefulness of ML in enabling timely identification of heart disease along with supporting better health outcomes at lower costs.
Singh et al. [16], authors apply techniques like “Pearson correlation” and “Chi-Square” tests to select significant clinical attributes and evaluate classifiers including LR, RF, NB, and SVM. Their findings show that feature selection enhances the performance of prediction and reduces computational complexity, with LR performing best among the tested models.
Victor P. Chang et al. [17] proposes an AI driven system for CVD detection with ML to enhance early diagnosis and predictive accuracy in cardiovascular healthcare. Healthcare data were preprocessed in Python, with LR applied for initial analysis and a Random Forest classifier achieving ~83% training accuracy. The study outlines key stages including database collection, data preprocessing, classifier development, and evaluation, highlighting how AI-driven methods can augment traditional diagnostic approaches.
Bhatt et al. [18] develop a ML framework for CVD prediction that integrates feature selection, preprocessing, and classifier optimization to enhance diagnostic accuracy. Using a 13-feature dataset, features were ranked and normalized, and classifiers like “RF”, “GB”, “SVM”, “KNN”, and “LR” were evaluated. Gradient Boosting achieved the highest accuracy (~91.2%), followed by Random Forest (~89.8%), demonstrating that feature selection enhances efficiency and captures nonlinear clinical relationships.

3. METHODOLOGY

Models like “XGBoost”, “Random Forest”, “LightGBM”, “SVM”, “Neural Network” and “Logistic Regression” were tested for identifying the best algorithm for the prediction system. The first phase of the methodology was dataset preparation for the Cardiovascular Risk Triage System, by cleaning the data, eliminating null values and redundant values. To develop and assess the models, the dataset was separated into training and validation sets. Each model was then trained accordingly and subsequently evaluated..

3.1 Dataset

The algorithms were tested on two datasets : the ‘Mendeley’ and the ’UCI Heart Disease’ dataset. The latter is composed of 920 patient records derived from four geographic locations: Hungary, Cleveland, Long Beach and Switzerland. It includes 726 male entries and 194 female entries. The features include resting blood pressure, age, chest pain type, serum cholesterol, sex and fasting blood sugar. The Mendeley dataset contains 1000 records with 12 features. It includes 765 male entries and 235 female entries. The features are serum cholesterol, age, resting blood pressure, gender, max heart rate, serum cholesterol, resting heart rate, resting ecg results, chest pain type and classification, oldpeak ST, exercise induced angina, number of major vessels. This dataset is suitable for ML and data mining tasks as well as for CVD detection. This data was derived from the Indian population, collected from one of the multispeciality hospitals in India therefore it was more suitable and was chosen for further development.

3.2 Preprocessing Techniques

The dataset had many missing or incomplete values due to inconsistent data entry and unavailability of certain data. The median was utilized for filling up numerical values that were incomplete and mode for categorical values. Elimination of duplicate values was done by the drop_duplicate() function. Label encoding. implemented to translate the categorical values to quantitative or numerical values as portrayed in Table 1. Outlying values were identified using the interquartile range (IQR) criterion, which were then removed to make sure the data was within the realistic range. Since the medical records often exhibit an extensive numerical distribution, this can create problems during analysis, especially for algorithms which are sensitive to these differences such as LR and SVM. To standardize the data Z-Score Normalization was used. The target attribute contained a disease severity value in the range 0 - 5, since we decided the model classification in binary, the values were converted to 0 for no risk and 1 for risk present. The distribution of targets categorized as “at risk” and “not at risk” is illustrated in Figure 1.
It was split randomly into a “training set” comprising 70% of data and a “testing set” comprising the remaining 30%. After cleaning and transforming the dataset six models were trained on it. Standard metrics were utilised to compare their results and find the model exhibiting the superior performance for the Cardiovascular Risk Triage System. Model performance was evaluated using “confusion matrix”, “recall”, “accuracy”, “F1-score” and “precision”. Collectively, these metrics demonstrate the efficiency and performance, helping choose the one that works best with new data.

Table -1: Label Encoding Table

Feature	Mapping
age_category	0 : Young Adult, 1 : Middle-Aged, 2 : Senior
gender	0 : Female, 1 : Male
chestpain	1 : Typical Angina, 2 : Atypical Angina, 3 : Non-Anginal Pain,4 : Asymptomatic

restingBP_category	0 : Normal, 1 : Elevated, 2 : Hypertension Stage 1, 3 : Hypertension Stage 2, 4 : Hypertensive Crisis
serumcholestrol_category	0 : Desirable, 1: Borderline High, 2: High
fastingbloodsugar	0 : <= 120 mg/dl, 1 : > 120 mg/dl
restingelectro	0 : Normal, 1 : ST-T Wave Abnormality, 2 : Left Ventricular Hypertrophy
maxheartrate_category	0 : Zero Response, 1: Low Response, 2 : High Response
exerciseangia	0 : No, 1: Yes
oldpeak_category	0 : Normal, 1 : Moderate, 2 : High
slope	1 : Upsloping, 2 : Flat, 3 : Downsloping
Noofmajorvessels	0 : Normal, 1 : One vessel, 2 : Two vessels, 3 : Three vessels
target	0 : No Heart Disease, 1 : Heart Disease

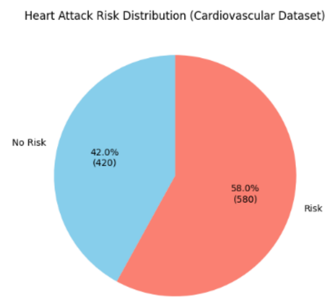

Fig -1: Percentage of risk and no risk from 'Mendeley dataset' after preprocessing

3.3 Models Used

XGBoost
XGBoost is a tree-based ensemble learning method built on gradient boosting principles. It incrementally constructs decision trees, with each iteration optimizing the residual errors of prior models. Through built-in regularization and efficient parallel computation, it achieves strong predictive performance while maintaining computational efficiency on high-dimensional datasets. [10]

SVM
SVM is a classification, regression model, it is a supervised(labelled) machine learning algorithm. Its goal is to determine a boundary that clearly separates distinct classes, giving the model the ability to accurately assign new data instances to the correct category. It focuses on finding a decision function that creates the maximum margin between classes. It finds the support vectors,the extreme points to create hyperplanes [10].

Logistic regression
Logistic regression, one of the supervised machine learning models, utilized to classify data points according to their features. It uses a collection of independent variables to predict categorical classification of a dependent variable[19]. The input parameters were mapped to a continuous range from 0 to 1 using a sigmoid activation function which were then mapped to a discrete value (True/False) [10].

LightGBM
LightGBM is a fast and efficient ML model that builds trees step-by-step to improve predictions.It follows the "gradient boosting" techniques, which can combine many small models to make one strong model. It speeds up training by grouping data into bins and growing trees leaf by leaf, making it work well on large datasets with less memory.[20]

Random forest
Random Forest constructs an ensemble of decision trees using bootstrapped samples and randomly selected subsets of features. It is used for categorizing data and regression analysis. This method uses the votes of several decision trees to determine the final result. This algorithm also works well with large datasets[18].

Neural Networks

Neural networks consist of multiple linked levels of neurons that learn complex patterns through forward propagation and backpropagation. They adjust weights iteratively to minimize prediction errors, enabling modeling of nonlinear relationships in data[15].

4. ANALYSIS AND RESULTS

Model evaluation was conducted through precision, recall, accuracy, F1-score, and comparison of the confusion matrix. The outcome of ROC-AUC curves and the confusion matrix are shown from Figure 2 to Figure 13.

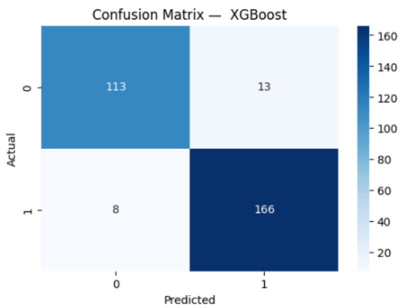

Fig -2: XGBoost Confusion Matrix

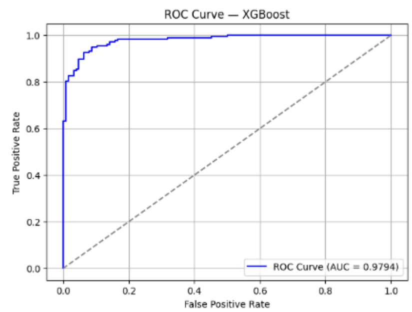

Fig -3: XGBoost ROC Curve

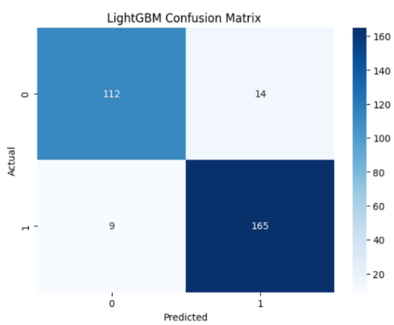

Fig -4: LightGBM Confusion Matrix

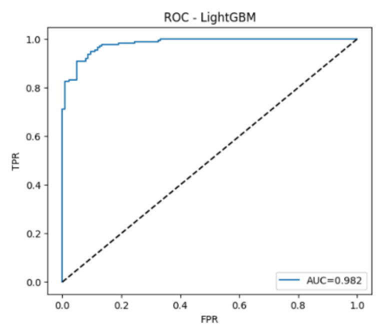

Fig -5: LightGBM ROC Curve

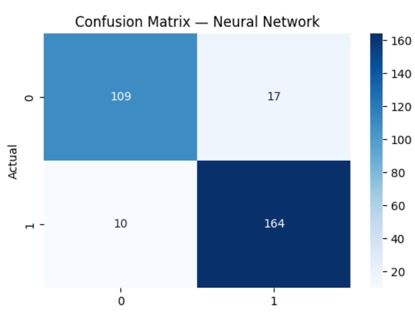

Fig -6: Neural Networks Confusion Matrix

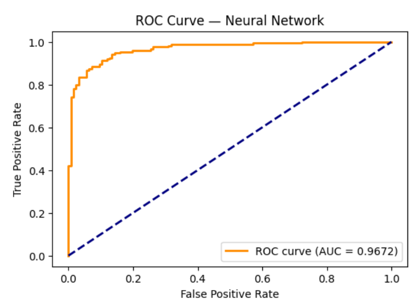

Fig -7: Neural Network ROC Curve

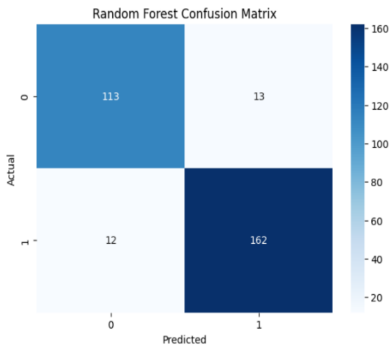

Fig -8: Random Forest Confusion Matrix

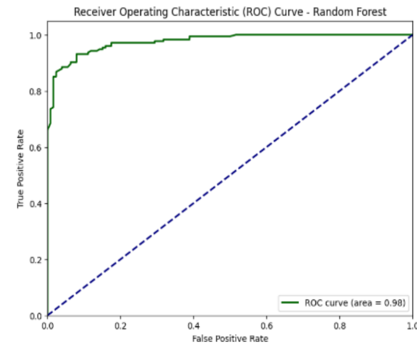

Fig -9: Random Forest ROC Curve

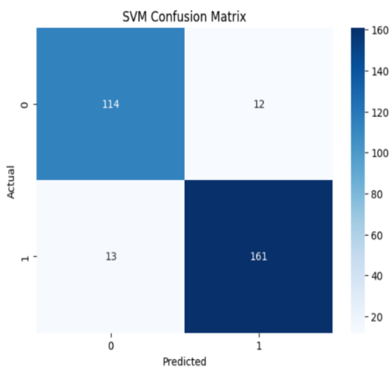

Fig -10: SVM Confusion Matrix

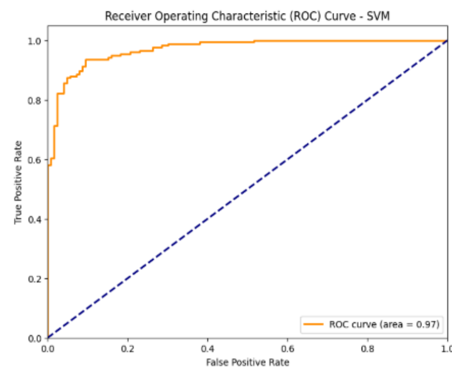

Fig -11: SVM ROC Curve

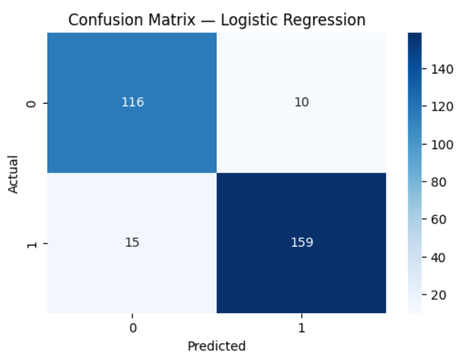

Fig -12: Logistic Regression Confusion Matrix

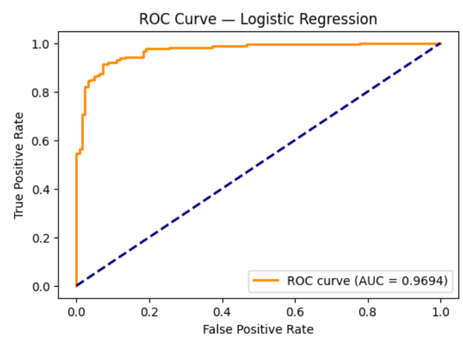

Fig -13: Logistic Regression ROC Curve

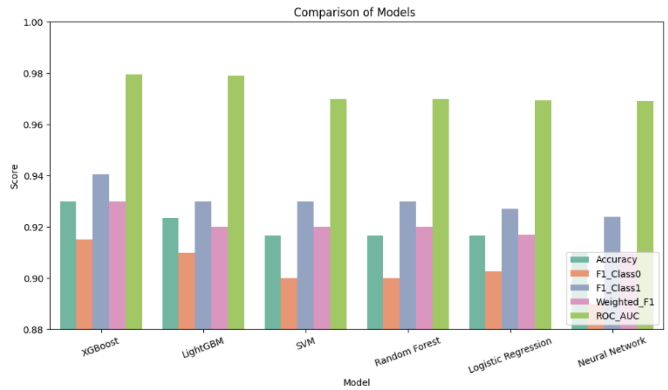

Chart-1: Comparison Graph of model performance

Table -2: Comparison of Performance metrics of models

Metric	XGBoost	LightGBM	SVM	Random Forest	Logistic Regression	Neural Network
Accuracy	0.93	0.9233	0.9167	0.9167	0.9167	0.91
F1-score (Class 0)	0.915	0.91	0.9	0.9	0.9027	0.8898
F1-score (Class 1)	0.9405	0.93	0.93	0.93	0.9271	0.9239
Weighted Avg F1-score	0.9298	0.92	0.92	0.92	0.9169	0.9096
True Positives (TP)	166	165	161	162	159	164
True Negatives (TN)	113	112	114	113	116	109
False Positives (FP)	13	14	12	13	10	17
False Negatives (FN)	8	9	13	12	15	10
ROC AUC Score	0.9794	0.979	0.97	0.97	0.9694	0.969

The experimental results revealed that XGBoost achieved the superior performance metrics, attaining 95% accuracy and a ROC-AUC value of 0.9794. This model also exhibited the lowest number of false negatives, quintessential in heart risk prediction to avoid missed diagnoses and delayed interventions. The comparative analysis, shown in Chart 1 and Table 2, indicates that XGBoost delivered the highest performance across all considered evaluation metrics

5. CONCLUSIONS

Cardiovascular diseases stand as a prominent cause of death all across the globe, and various strategies and technologies have been developed by the Medical and IT Industry. Attending heart health as early as possible is a very effective method to decrease the mortality rates due to various CVDs. Machine Learning has a very high potential to be at the forefront of proactive healthcare, by detecting early any signs of Cardiovascular risks. Using medically verified datasets to train various ML models and testing them on industry benchmarks can tell us the best possible way to know the risks at a very early stage which let us make better healthcare decisions and increase personal awareness. Detection of CVDs at early phases can lead more people to opt for preventive actions to reduce their risk exposure or take requisite measures to mitigate them. ML can help in making preventive care more accessible.

6. REFERENCES

[1]. Gunjal A, Judgi T (2025) A comprehensive survey of artificial intelligence methods for cardiovascular disease detection: Recent advances and future challenges. MethodsX 15:103678. https://doi.org/10.1016/j.mex.2025.103678
[2]. Sourov MEA, Hossen MS, Shaha P, Hossain MM, Iqbal MS (2025) An explainable AI-enhanced machine learning approach for cardiovascular disease detection and risk assessment. arXiv preprint arXiv:2507.11185
[3]. Mohapatra S, Maneesha S, Mohanty S, Patra PK, Bhoi SK, Sahoo KS, Gandomi AH (2023) A stacking classifiers model for detecting heart irregularities and predicting Cardiovascular Disease. Healthc Anal 3:100133. https://doi.org/10.1016/j.health.2022.100133

[4]. Al-Mahdi IS, Darwish SM, Madbouly MM (2025) Heart disease prediction model using feature selection and ensemble deep learning with optimized weight. Comput Model Eng Sci 143:875–909. https://doi.org/10.32604/cmes.2025.061623
[5]. Saleh Alghamdi N, Zakariah M, Shankar A, Viriyasitavat W (2024) Heart disease prediction using autoencoder and DenseNet architecture. Egypt Inform J 28:100559. https://doi.org/10.1016/j.eij.2024.100559
[6]. Anooj PK (2012) Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. J King Saud Univ Comput Inf Sci 24:27–40. https://doi.org/10.1016/j.jksuci.2011.09.002
[7]. Ogunpola A, Saeed F, Basurra S, Albarrak AM, Qasem SN (2024) Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics 14:144. https://doi.org/10.3390/diagnostics14020144
[8]. Kiran S, Reddy GR, Girija SP, et al (2023) A Gradient Boosted Decision Tree with Binary Spotted Hyena Optimizer for cardiovascular disease detection and classification. Healthc Anal 3:100173. https://doi.org/10.1016/j.health.2023.100173
[9]. El-Sofany H, Bouallegue B, El-Latif YMA (2024) A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Sci Rep 14:23277. https://doi.org/10.1038/s41598-024-74656-2
[10]. Majeed BA, Hardan AY, Hardan BY, Munaf DF (2024) Accurate AI-based Chatbot to diagnose heart diseases pre-human doctor consultation. Rev Intell Artif 38(1):213–220. https://doi.org/10.18280/ria.380121
[11]. Doppala BP, Bhattacharyya D, Chakkravarthy M, Thirunavukkarasu U (2022) A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques. J Healthc Eng 2022:2585235. https://doi.org/10.1155/2022/2585235
[12]. Chandrasekhar N, Peddakrishna S (2023) Enhancing heart disease prediction accuracy through machine learning techniques and optimization. Processes 11(4):1210. https://doi.org/10.3390/pr11041210
[13]. El-Sofany HF (2024). Predicting heart diseases using machine learning and different data classification techniques. IEEE Access 12:106146–106160. https://doi.org/10.1109/ACCESS.2024.3437181
[14]. Stonier AA, Gorantla RK, Manoj K (2023) Cardiac disease risk prediction using machine learning algorithms. Healthc Technol Lett 11(4):213–217. https://doi.org/10.1049/htl2.12053
[15]. Singh A, Mahapatra H, Biswal AK, et al (2024) Heart disease detection using machine learning models. Procedia Comput Sci 235:937–947. https://doi.org/10.1016/j.procs.2024.04.089
[16]. Singh P, Pal GK, Gangwar S (2022) Prediction of cardiovascular disease using feature selection techniques. Int J Comput Theory Eng 14(3):101–107. https://doi.org/10.7763/IJCTE.2022.V14.1316
[17]. Chang V, Bhavani VR, Xu AQ, Hossain MA (2022) An artificial intelligence model for heart disease detection using machine learning algorithms. Healthc Anal 2:100016. https://doi.org/10.1016/j.health.2022.100016
[18]. Bhatt CM, Patel P, Ghetia T, Mazzeo PL (2023) Effective heart disease prediction using machine learning techniques. Algorithms 16(2):88. https://doi.org/10.3390/a16020088
[19]. Ullah T, Ullah SI, Ullah K, Ishaq M, Khan A, Ghadi YY, Algarni A (2024) Machine learning-based cardiovascular disease detection using optimal feature selection. IEEE Access 12:16431–16446. https://doi.org/10.1109/ACCESS.2024.3359910
[20]. Yang H, Chen Z, Yang H, Tian M (2023) Predicting coronary heart disease using an improved LightGBM model: performance analysis and comparison. IEEE Access 11:23366–23380. https://doi.org/10.1109/ACCESS.2023.3253885