

# Optimizing heart disease prediction through ensemble and hybrid machine learning techniques

Nomula Nagarjuna Reddy, Lingadally Nipun, MD Uzair Baba, Nyalakanti Rishindra,  
Thoutireddy Shilpa

Department of Computer Science and Engineering, B V Raju Institute of Technology, Narsapur, India

## Article Info

### Article history:

Received Mar 22, 2024

Revised Jul 9, 2024

Accepted Jul 17, 2024

### Keywords:

Artificial intelligence

Ensemble learning

Heart disease

Multilayer perceptron

Random forest

## ABSTRACT

In this modern era, heart diseases have surfaced as the leading factor of fatalities, accounting for around 17.9 million lives annually. Global deaths from heart diseases have surged by 60% over the last 30 years, primarily because of limited human and logistical resources. Early detection is crucial for effective management through counseling and medication. Earlier studies have identified key elements for heart disease diagnosis, including genetic predispositions and lifestyle factors such as age, gender, smoking habits, stress, diastolic blood pressure, troponin levels, and electrocardiogram (ECG). This project aims to develop a model that can identify the best machine learning (ML) algorithm for predicting heart diseases with high accuracy, precision, and the least misclassification. Various ML techniques were evaluated using selected features from the heart disease dataset. Among these techniques, a combination of random forest (RF), multi-layer perceptron (MLP), XGBoost, and LightGBM employing an ensemble method with a stacking classifier, along with logistic regression (LR) as a metamodel, achieved the highest accuracy rate of 95.8%. This surpasses the efficiency of other techniques. The suggested method provides an encouraging framework for early prediction, with the overarching goal of reducing global mortality rates associated with these conditions.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Thoutireddy Shilpa

Department of Computer Science and Engineering, B V Raju Institute of Technology

Narsapur, Medak, Telangana 502313, India

Email: Shilpa.t@bvrit.ac.in, shilpathoutireddy@gmail.com

## 1. INTRODUCTION

The human heart as shown in Figure 1, which weighs approximately 300 grams, is a marvel of biological intricacy and essential for understanding and detecting heart diseases (HD). Cardiovascular disease (CVD) has surfaced as the leading factor of fatalities, accounting for around 17.9 million lives annually. Despite medical advancements, HD remains a major cause of mortality and morbidity, burdening individuals, families, and healthcare systems. Over 80% of these fatalities result from cardiac arrests and strokes, often affecting individuals under 70. HD manifests in various forms, including coronary artery disease, heart failure, arrhythmias, valvular diseases, and congenital defects. Early detection is crucial for timely medical intervention.

In recent years, machine learning (ML) techniques have revolutionized healthcare by offering powerful tools for predictive analytics. ML algorithms can analyze vast medical datasets, detect complex patterns, and create accurate predictive models. Previous studies highlight the importance of genetic predispositions and lifestyle factors [1] such as age, gender, smoking, stress, blood pressure, troponin levels, and electrocardiogram (ECG) readings in assessing HD risk. However, existing HD prediction methods face

challenges in accuracy and efficiency, requiring improvements in reducing misclassification errors and integrating additional factors. By scrutinizing these algorithms against established performance metrics, this study aims to elucidate their effectiveness and comparative performance in predicting heart disease risk. This work contributes to the field by enhancing early detection and intervention strategies, ultimately aiming to reduce the global burden of HD. The following sections will detail the methodologies employed, present the results of our comparative analysis, and discuss the relevance and implications of our findings.

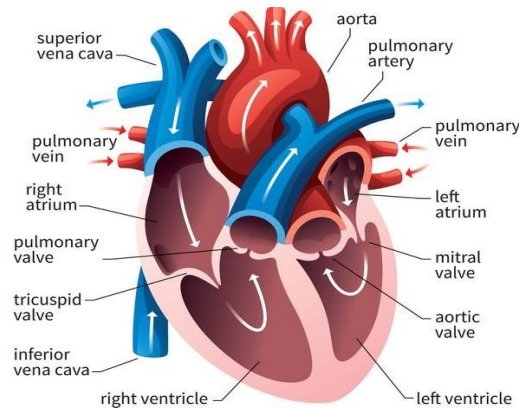


Figure 1. Basic anatomy of the human heart

## 2. LITERATURE REVIEW

HD present a major global health challenge, requiring accurate predictive models for early detection and intervention. Recent studies highlight the efficacy of machine learning (ML) and deep learning (DL) techniques in predicting heart disease. Several studies have demonstrated the effectiveness of various algorithms. One study utilized the classification and regression trees (CART) algorithm, achieving an 87% accuracy in HD prediction [2]. Another research integrated five cardiovascular disease (CVD) datasets, with random forest (RF) outperforming other algorithms after optimization [3]. RF has been shown to demonstrate strong predictive capabilities, though with lower specificity [4]. Significant attributes like age and smoking were identified, achieving a 90% accuracy with RF [5]. Using multi-layer perceptron (MLP) with feature reduction and clustering techniques resulted in an accuracy of 87.28% [6]. Deep neural network (DNN) achieved a remarkable 98.15% accuracy, outperforming prior research [7].

Deep learning (DL) techniques, with accuracies ranging from 90% to 99.1%, were employed using feature fusion and data augmentation [8]. Early-stage identification and treatment were enabled by training static var compensator (SVC), neural network (NN), and random forest classifier (RFC) models [9]. A hybrid system achieved an 86.6% accuracy with RF [10]. The NN algorithm achieved over 93% accuracy [11]. Using logistic regression (LR), k-nearest neighbors (KNN), and RFC, an accuracy of 87.5% was attained [12]. KNN and RF algorithms achieved accuracies of 86.88% and 81.96%, respectively [13].

A hybrid ML model achieved an 88.7% accuracy [14]. NN reported a 93% accuracy [15]. Among six data mining tools, MATLAB's artificial neural network (ANN) model yielded the highest performance [16]. The convolutional neural network (CNN) model achieved a 98.64% accuracy [17]. A hybrid model with RF and support vector machine (SVM) reached a 98.3% accuracy [18]. ML techniques, with SVM exhibiting the highest accuracy of 96%, were analyzed [19]. Using LR and recursive feature elimination (RFE), a 97.35% accuracy in predicting HD risk was achieved [20].

DL algorithms using ECG data attained an 85.6% accuracy [21]. Bayes net and RF had optimal performance in HD prediction in Iraq [22]. The role of ML algorithms in CVD management was emphasized, highlighting advancements in early detection and personalized treatment [23].

In summary, this research contributes to the existing body of literature, affirming the effectiveness of ML and DL algorithms in HD prediction. The importance of feature extraction methods is emphasized, highlighting key physiological features as significant predictors. By synthesizing insights from previous studies, a deeper understanding of HD prediction is achieved, underscoring the capability of ML and DL techniques in advancing early detection and intervention strategies. This research aims to develop accurate predictive models for heart diseases, leveraging ensemble learning, feature extraction, and physiological features to improve diagnostic accuracy and ultimately enhance patient outcomes in cardiovascular healthcare. These findings underscore the significance of ML and DL techniques in healthcare, offering valuable insights for improving patient outcomes and medical care provision.

### 3. METHOD

#### 3.1. Dataset details

The dataset [24] comprises 11 features representing the presence or absence of heart disease, sourced from multiple databases including Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog dataset, totaling 1,190 instances. These features together represent essential patient characteristics for HD prediction. All these features are explained in Table 1 and have been considered for the prediction.

Table 1. Description of features in the dataset

Feature name	Description
ST slope	The incline of the ST segment during peak exercise on an ECG. This reflects the rate of change of the ST segment during exercise.
Chest pain type	The nature of chest discomfort felt by the person. It may include categories such as typical angina (related to heart disease), atypical angina (may not be related to heart disease), non-anginal pain (not related to the heart), or asymptomatic (no chest pain) [25].
Oldpeak	ST depression induced by exercise relative to rest. This measures the extent of ST segment depression on an ECG during exercise compared to resting levels, providing information about myocardial ischemia (insufficient blood flow to the heart).
Resting ECG	Results of the resting ECG test. This test measures the electrical activity of the heart while the individual is at rest, providing information about the heart's rhythm and electrical conduction.
Age	Age of the individual in years. Age is an important demographic factor for CVD risk [26].
Fasting blood sugar	It is measured in mg/dL. Elevated fasting blood sugar levels may indicate insulin resistance or diabetes [27].
Max heart rate	Highest heart rate attained while exercising (bpm). This is reflecting the heart's ability to pump blood efficiently during physical activity.
Sex	Gender of the individual (0 = female, 1 = male). Gender is a factor for CVD risk [28].
Resting BPS	Resting systolic blood pressure (mmHg). This refers to the arterial pressure when the heart contracts and expels blood, providing information about CV health and the risk of conditions like hypertension [29].
Exercise angina	Exercise-induced angina (1 = yes; 0 = no). Angina is discomfort in the chest caused by reduced blood flow to the heart, often triggered by physical stress.
Cholesterol	Serum cholesterol level (mg/dL). High cholesterol levels, especially low density lipoprotein (LDL) cholesterol, heighten the risk of CVD [30].

#### 3.2. Framework

In this study, our objective is to construct an extensive framework as shown in Figure 2 for predicting heart disease utilizing a dataset sourced from various countries. The dataset comprises 1,190 patient records, each featuring 11 distinct attributes alongside a target variable denoting the existence of HD. Our data collection procedure involved compiling information from diverse geographic regions. We meticulously performed data preprocessing tasks, including the removal of null values and the appropriate handling of both numerical and categorical data, to ensure data quality. Moreover, we employed feature extraction techniques aimed at augmenting the predictive capabilities of our models.

For the training and evaluation phases, we selected four prominent ML algorithms: RF, XGBoost (XGB), MLP, and LightGBM (LGBM). These algorithms were picked due to their demonstrated efficacy in handling complex datasets and yielding accurate predictions. To streamline the process of model training and validation, we segregated the observations into separate parts of 80% and 20% respectively. This segregation permitted us to develop our models on one subset while validating their performance on another, ensuring robustness and generalization. Additionally, we employed ensembling using a stacking classifier with LR as the use of a meta-model amplified the predictive capability of this ensemble approach even more, our framework by leveraging the strengths of multiple algorithms.

Following model training, we proceeded with deployment to enable predictions on new, unseen data instances. This deployment phase is critical for real-world applicability, as it simulates the effectiveness of the model in real-world situations. Subsequently, we conducted a thorough performance evaluation to gain insights into the effectiveness and dependability of our heart disease prediction framework as a whole.

To refine our feature selection process, we utilized Kullback-Leibler (KL) divergence and Fischer index techniques. After obtaining the top features from each technique individually, we calculated the average of those features. These averaged features were then used in our predictive model to improve its accuracy and robustness in HD prediction. By adopting a systematic approach that integrates data preprocessing, feature extraction, model training, deployment, and performance evaluation, we aimed to develop a comprehensive framework capable of giving accurate HD prediction. Our endeavor is underpinned by the aspiration to enhance healthcare decision-making and facilitate timely interventions for individuals at risk of cardiovascular ailments.

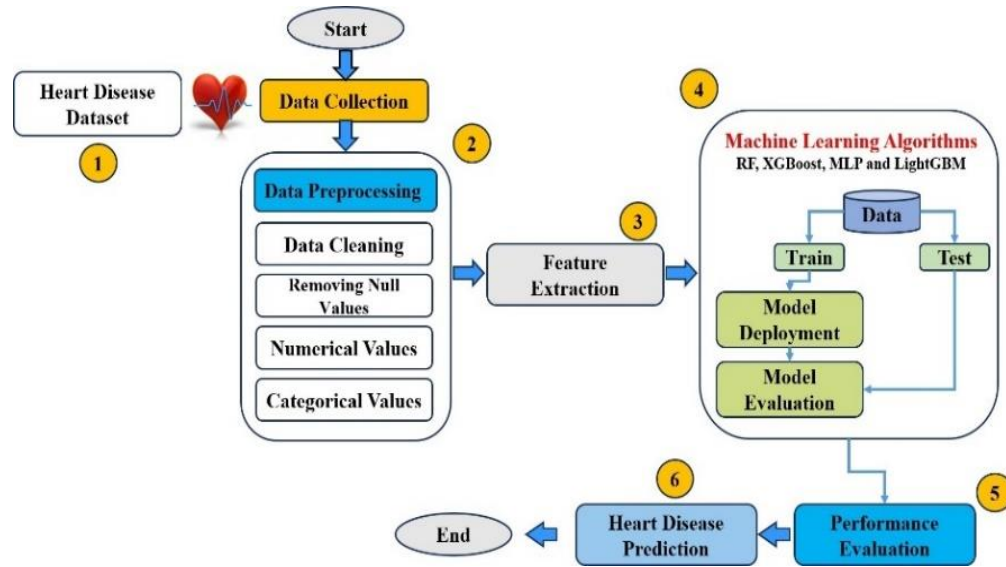


Figure 2. System framework

### 3.3. Data preprocessing

It is a crucial initial step in preparing raw data for ML, essential for enhancing both the accuracy and efficiency of predictive models. Real-world datasets often contain issues such as noise, missing values, and formats that are unsuitable for direct use in ML models. Thus, preprocessing is necessary to clean and format the data, making it suitable for modeling.

The preprocessing phase addresses missing values, which can significantly impact model performance. These values are either removed or imputed to ensure data completeness and consistency. Additionally, categorical variables must be encoded into numerical formats required by ML algorithms. This step ensures that the models can interpret and utilize all features effectively.

Moreover, feature scaling guarantees uniform contribution from all features in the learning phase, preventing any individual feature from overshadowing others due to its scale. Essential stages in data preprocessing encompass dataset importation and cleansing, addressing missing data, converting categorical variables into numerical formats, dividing the dataset into training and testing sets, and standardizing features to maintain consistency during model training. Typically, techniques like z-score normalization or min-max scaling are used for standardization. These preprocessing steps are integral to preparing the dataset for effective ML model training, ensuring that the models perform optimally and provide reliable predictions.

### 3.4. Feature selection

It is crucial in improving heart disease risk assessment models by identifying key attributes from datasets. In this context, sophisticated techniques such as KL divergence and the Fisher Index are employed to pinpoint the most informative features among the dataset's 11 attributes. Utilizing two distinct feature extraction methods and combining their results offer several advantages. Different methods provide complementary insights, offering varied perspectives on feature importance, which helps achieve a more comprehensive understanding. Additionally, this combination enhances the robustness of the feature selection process by reducing reliance on a single technique, thereby improving model performance through a broader capture of relevant features. By extracting the top features from each method and averaging them, the model's accuracy and robustness are further enhanced, integrating diverse and valuable insights from both techniques.

#### 3.4.1. Fisher index

It measures the significance of features in classification tasks, identifying relevant features and reducing dimensionality before classification. It calculates the discriminative power of each feature, guiding the inclusion or exclusion of features in models. The Fischer index ( $F(i)$ ) formula is given by:

$$F(i) = \frac{(\mu_{i1} - \mu_{i2})^2}{s_{i1}^2 + s_{i2}^2}$$

Here:

- $F(i)$  is the Fischer index for feature  $i$ .
- $\mu_{i1}$  and  $\mu_{i2}$  are the means of feature  $i$  for both the  $+ve$  and  $-ve$  samples respectively.
- $s_{i1}^2$  and  $s_{i2}^2$  are the variances of feature  $i$  for both the  $+ve$  and  $-ve$  samples respectively.

### 3.4.2. KL-divergence

It evaluates differences between probability distributions, identifying features vital for discriminating between normal and abnormal heart conditions. It compares distributions  $M$  and  $N$ , highlighting extra information needed to represent  $M$  using  $N$ 's code. The Kullback-Leibler (KL) divergence formula is given by:

$$D_{KL}(M||N) = \sum_{a \in X} M(a) \log \left( \frac{M(a)}{N(a)} \right)$$

Here:

- $D_{KL}(M||N)$  is the KL divergence from distribution  $N$  to distribution  $M$ .
- $M(a)$  and  $N(a)$  are probability distributions over the sample space  $X$ .
- The sum is taken over all possible outcomes ' $a$ ' in the sample space  $X$ .

## 3.5. Proposed ensemble method

### 3.5.1. Random forest

It is a flexible algorithm adept at handling both classification and regression tasks. It enhances model performance and addresses complex problems by integrating multiple decision trees (DTs) as shown in Figure 3. This approach effectively reduces the limitations of single decision trees, improving accuracy and minimizing overfitting. RF offers several advantages, including fast training times suitable for large datasets, strong predictive capabilities even in the presence of missing data, and the ability to rank feature importance, which helps in understanding data trends. Random forest, functioning as an ensemble learning (EL) technique, constructs multiple DTs during training. In classification tasks, it yields the most frequent class as the output, whereas in regression tasks, it offers the average prediction from the individual trees.

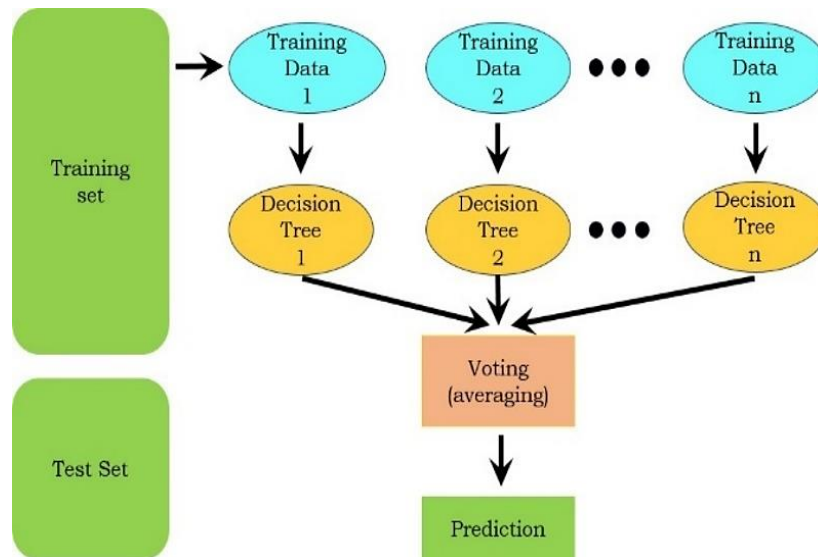


Figure 3. The working of a random forest

### 3.5.2. Multilayer perceptron

It is an advanced artificial neural network designed with multiple interconnected layers, forming a feedforward structure. It effectively tackles non-linear problems and complex datasets by identifying intricate relationships within the data. An MLP typically consists of an input layer along with one or multiple hidden layers, and an output layer, each composed of nodes (neurons) connected by adjustable weights as shown in Figure 4. During training, these weights are optimized to minimize errors. The network employs activation

functions at each node to introduce non-linearity, enabling the modeling of complex functions. A significant feature of MLPs is backpropagation, an algorithm that iteratively updates the weights by propagating errors backward through the network, improving model accuracy over time. This continuous weight optimization makes MLPs highly effective for predictive modeling across various applications.

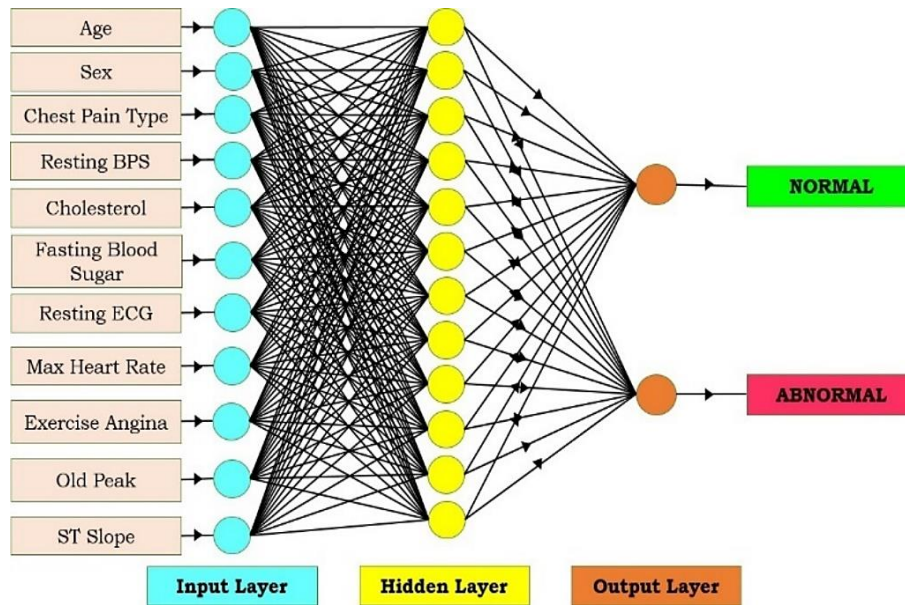


Figure 4. The functioning of a multilayer perceptron

**3.5.3. XGBoost**

It stands out as a powerful algorithm rooted in gradient-boosting decision trees, designed for speed, ease of use, and effectiveness, especially with large datasets [31]. It operates by creating a series of models as shown in Figure 5 and combining them to enhance overall accuracy. Its advantages include exceptional efficiency and scalability, effortlessly handling vast amounts of data, while also offering interpretability and resilience against overfitting. By iteratively adding new models to rectify errors of previous ones, XGBoost crafts a sequence of models that collectively improve predictive performance. This implementation of gradient-boosting machines ensures a potent tool for data analysis and prediction tasks, providing users with both accuracy and scalability in their endeavors.

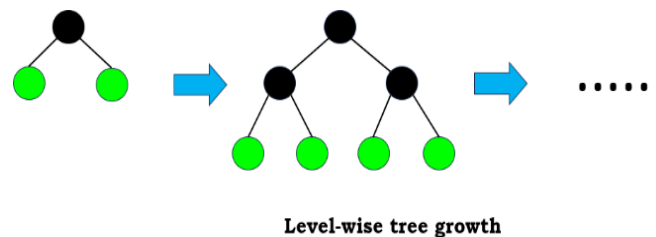


Figure 5. The working mechanism of XGBoost

**3.5.4. LightGBM**

It stands out as an efficient and swift gradient-boosting framework, leveraging decision trees and advanced techniques like gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) to bolster efficiency while curbing memory usage. Its leaf-wise splitting strategy as shown in Figure 6, which prioritizes selecting the leaf with the highest loss, elevates both accuracy and efficiency. Notably, LightGBM boasts superior performance on large datasets, even outperforming XGBoost in certain scenarios. This achievement is underpinned by its adept memory management, thanks to optimization techniques woven into its architecture. By utilizing algorithms based on tree learning principles and prioritizing distributed and

effective training methods, LightGBM optimizes its operations for scalability and effectiveness. Its approach to leaf-wise splitting, aiming for the leaf with the highest delta loss for expansion, further underscores its commitment to minimizing loss compared to traditional level-wise methods.

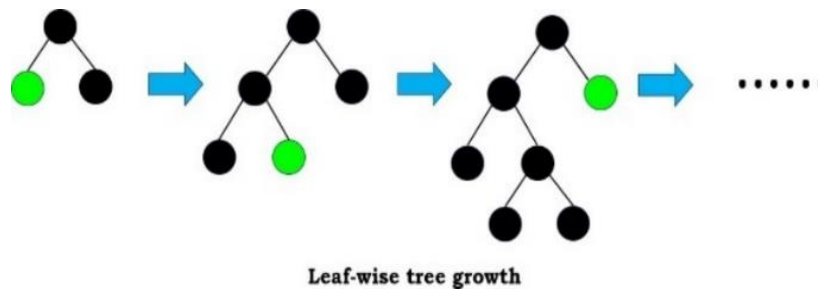


Figure 6. The working mechanism of LightGBM

### 3.6. Ensemble learning

It is a potent tool in machine learning, amalgamating multiple models to enhance predictive accuracy, robustness, and generalization [32]. It tackles issues like high variance or bias in individual models and the uncertainty surrounding the optimal model choice. By amalgamating predictions, ensemble methods mitigate inconsistencies and errors, yielding more dependable outcomes. The crux of ensemble learning (EL) lies in extracting diverse perspectives from various models and merging them into a cohesive prediction. This diversity can arise from employing different algorithms, and data subsets, or introducing randomness during training.

In our study, we utilized RF, MLP, XGB, and LGBM. While RF may struggle with complex relationships, XGB and LGBM can be prone to overfitting, and MLP is sensitive to hyperparameters and smaller datasets. Hence, a comprehensive ensemble approach merging RF, XGB, MLP, and LGBM is proposed. Each model contributes unique strengths: RF stabilizes against overfitting, XGB and LGBM capture intricate patterns, and MLP introduces flexibility with nonlinear relationships. By integrating these algorithms, the ensemble harnesses their collective strengths to improve predictive accuracy and reliability, thereby advancing the efficacy of ML in practical applications.

#### 3.6.1. Stacking classifier

The proposed ensemble method employs the stacking classifier, which combines predictions from various base models like MLP, RF, XGB, and LGBM to enhance accuracy and reliability. Each base model is trained independently with specific hyperparameters. The stacking classifier aggregates these predictions using logistic regression as a meta-estimator, capturing diverse aspects of the data. By leveraging the strengths of multiple models and correcting individual weaknesses through the meta-model, this approach improves predictive power and generalization to new data. LR is chosen for its simplicity, efficiency, and interpretability, striking a balance between complexity and performance. Its linear nature facilitates efficient learning and adjustment of weights for combining predictions, while directly outputting probabilities aids interpretation, making it suitable for classification tasks.

### 3.7. Performance analysis

The evaluation of classification algorithms for HD prediction aims to increase diagnostic accuracy and patient outcomes. The following metrics provide insights into model performance, balancing false positives and negatives, guiding algorithm selection, and optimization to enhance heart disease diagnosis and patient care.

#### 3.7.1. Accuracy

It assesses the model's overall correctness by determining the proportion of correctly predicted instances among the total instances. Accuracy measures the model's overall performance by calculating the ratio of correctly predicted instances to the total number of instances. It serves as a key indicator of how well the model performs across the entire dataset. A higher accuracy value signifies better model performance, indicating that the model is making fewer errors in its predictions.

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

### 3.7.2. Precision

It evaluates the model's capacity to identify positive instances among all correctly predicted positives, playing a critical role in minimizing false positives (FP). Precision evaluates the model's ability to correctly identify positive instances among all predicted positives. It is crucial for reducing the number of false positives (FP) in the model's predictions. High precision indicates that the model has a high accuracy in its positive predictions, ensuring reliable identification of positive instances.

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

### 3.7.3. Recall

Recall, also known as sensitivity, assesses the model's capacity to accurately detect true positive cases among all actual positive instances, crucial for minimizing false negatives (FN). Recall, also known as sensitivity, measures the model's effectiveness in correctly identifying true positive cases among all actual positive instances. It is essential for minimizing false negatives (FN), as it ensures that the model detects as many positive instances as possible. High recall is particularly important in applications where missing positive instances can have significant consequences.

$$Recall = \frac{TP}{FN + TP}$$

### 3.7.4. F1-Score

The F1-Score offers a balanced assessment of the model's performance by taking into account both precision and recall, making it particularly valuable for evaluating models on datasets with class imbalances. The F1-Score provides a balanced evaluation of the model's performance by considering both precision and recall. It is particularly valuable for assessing models on datasets with class imbalances, where relying on accuracy alone might be misleading. A higher F1-Score indicates a better balance between precision and recall, making it a robust metric for model evaluation.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Together, these metrics show the model's performance, ensuring strong and dependable predictions for heart disease.

## 4. RESULTS AND DISCUSSION

The performance metrics of different ML models for HD prediction exhibit notable variations in accuracy, precision, recall, and F1-scores, as depicted in Figure 7. RF demonstrated high precision (96%) and recall (92%), resulting in accuracy (94.5%) and an F1-score (94%), indicating its robust predictive capability. Similarly, XGB showed strong performance with an accuracy of 92.8% and an F1-score of 92%, while MLP achieved an accuracy of 91.2% and an F1-score of 90%. Combining RF, XGB, MLP, and LGBM resulted in the highest overall accuracy (95.8%) and F1-score (96.2%), underscoring the ensemble methods' effectiveness. These findings are consistent with previous research that underscores the effectiveness of ensemble methods in medical predictions. Combining multiple algorithms significantly improves predictive accuracy and robustness, as evidenced by the superior performance of the RF, XGB, MLP, and LGBM ensemble. This aligns with existing literature, confirming that integrating different models can capture a wider range of patterns in complex datasets. However, the study had limitations, such as the dataset not fully incorporating comprehensive medical records and social factors, which are crucial for a holistic risk assessment. Additionally, the exclusion of unstructured data like text and images represents a limitation that, if addressed, could further enhance model accuracy.

This research aimed to evaluate the efficacy of multiple ML algorithms in HD prediction. The findings demonstrate that ensemble methods, particularly the combination of RF, XGBoost, MLP, and LGBM, offer superior predictive performance. These results emphasize the importance of robust ML techniques in early detection and intervention strategies for heart disease. Future research should focus on incorporating a broader range of patient data, including lifestyle factors and unstructured data, to heighten the accuracy and reliability of predictive models. Additionally, exploring advanced preprocessing techniques and more sophisticated ML algorithms could further improve the efficacy of heart disease prediction, ultimately contributing to better healthcare outcomes and reducing the global CVD burden.



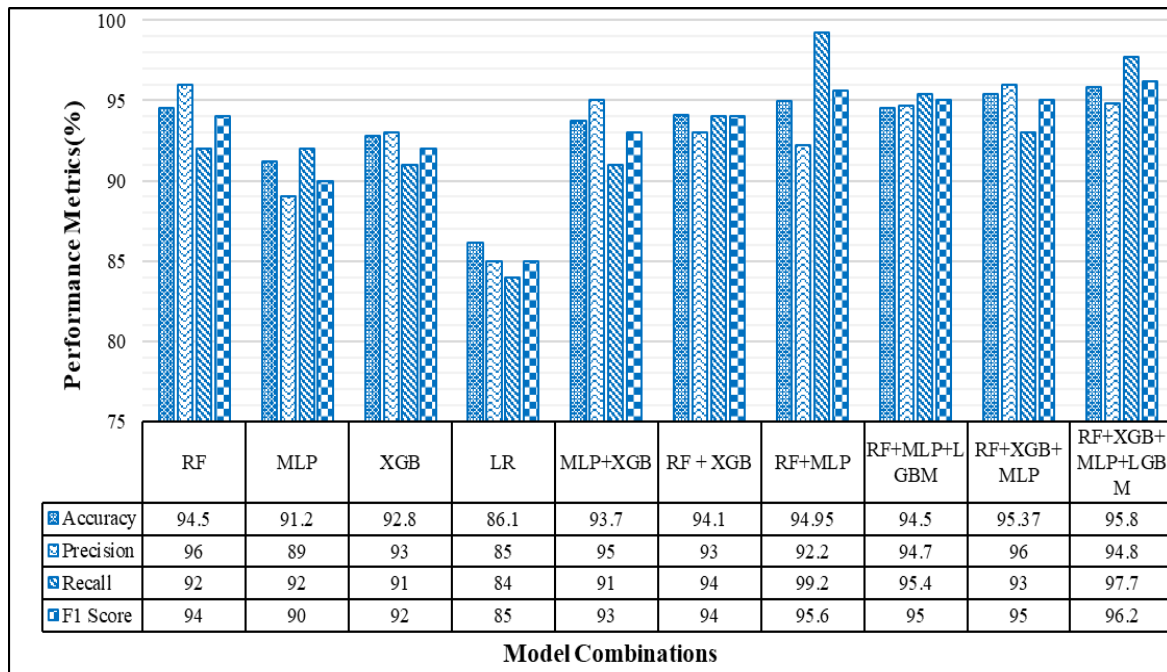


Figure 7. Performance analysis

## 5. CONCLUSION

In summary, this research delved into the application of ML algorithms for heart disease prediction, leveraging a dataset of 1190 patient records with 11 distinct features. Through the utilization of various ML techniques, including RF, MLP, XGBoost, and LightGBM, we evaluated their effectiveness in predicting CVD. Notably, ensemble methods, such as stacking classifiers with logistic regression as a meta-model, yielded a commendable accuracy rate of 95.8%. The amalgamation of RF, XGBoost, MLP, and LightGBM showcased a preceding accuracy of 95.37%, underscoring the potency of ensemble techniques in enhancing predictive performance.




Our findings emphasize the transformative potential of machine learning-based approaches in revolutionizing heart disease diagnosis, equipping clinicians with powerful tools for early detection and intervention, thereby leading to improved patient outcomes and healthcare efficiency. By integrating multiple algorithms, we demonstrated the ability to capture a broader spectrum of patterns in complex datasets, which is essential for accurate CVD prediction. While achieving substantial accuracy, it is crucial to acknowledge the computational complexity and hyperparameter sensitivity as limitations. Future research endeavors should explore alternative ensemble techniques, conduct further hyperparameter optimization, and integrate supplementary data sources, including lifestyle factors and unstructured data, to bolster predictive performance. These enhancements hold promise for improving the predictive model's accuracy and reliability, ultimately contributing to enhanced healthcare outcomes and a reduction in the global burden of CVD. In summary, this study not only highlights the efficacy of machine learning algorithms in predicting heart disease but also sets the stage for future progress in this field. By addressing identified challenges and limitations, researchers can chart a path toward continued improvements in heart disease management, thereby enhancing patient care and mitigating mortality rates associated with cardiovascular disorders.

## REFERENCES




- [1] B. H. Huang, M. J. Duncan, P. A. Cistulli, N. Nassar, M. Hamer, and E. Stamatakis, "Sleep and physical activity in relation to all-cause, cardiovascular disease and cancer mortality risk," *British Journal of Sports Medicine*, vol. 56, no. 13, pp. 718–724, Jul. 2022, doi: 10.1136/bjsports-2021-104046.
- [2] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2022.100130.
- [3] M. A. Kadhim and A. M. Radhi, "Heart disease classification using optimized machine learning algorithms," *Iraqi Journal for Computer Science and Mathematics*, vol. 4, no. 2, pp. 31–42, Feb. 2023, doi: 10.52866/ijcsm.2023.02.02.004.
- [4] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A novel study on machine learning algorithm-based cardiovascular disease prediction," *Health and Social Care in the Community*, vol. 2023, pp. 1–10, Feb. 2023, doi: 10.1155/2023/1406060.

- [5] M. I. Hossain *et al.*, "Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison," *Iran Journal of Computer Science*, vol. 6, no. 4, pp. 397–417, 2023, doi: 10.1007/s42044-023-00148-7.
- [6] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, Feb. 2023, doi: 10.3390/a16020088.
- [7] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary artery heart disease prediction: a comparative study of computational intelligence techniques," *IETE Journal of Research*, vol. 68, no. 4, pp. 2488–2507, Jul. 2022, doi: 10.1080/03772063.2020.1713916.
- [8] T. M. A. Monisha Sharean and G. Johncy, "Deep learning models on heart disease estimation - a review," *Journal of Artificial Intelligence and Capsule Networks*, vol. 4, no. 2, pp. 122–130, Jul. 2022, doi: 10.36548/jaicn.2022.2.004.
- [9] A. Gupta, V. Misra, K. Chauhan, and K. Manoj, "Heart disease prediction using machine learning," in *2023 5th International Conference on Advancing in Computing, Communication Control and Networking (ICAC3N)*, Dec. 2023, pp. 108–112, doi: 10.1109/ICAC3N60023.2023.10541622.
- [10] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, Sep. 2021, doi: 10.1007/s40860-021-00133-6.
- [11] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution," *Future Science OA*, vol. 7, no. 6, Jul. 2021, doi: 10.2144/fsoa-2020-0206.
- [12] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012072.
- [13] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012046.
- [14] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Jan. 2021, pp. 1329–1333, doi: 10.1109/ICICT50816.2021.9358597.
- [15] D. E. Salhi, A. Tari, and M. T. Kechadi, "Using machine learning for heart disease prediction," *Lecture Notes in Networks and Systems*, vol. 199 LNNS, pp. 70–81, 2021, doi: 10.1007/978-3-030-69418-0\_7.
- [16] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," *Health and Technology*, vol. 10, no. 5, pp. 1137–1144, 2020, doi: 10.1007/s12553-020-00438-1.
- [17] T. Suresh, T. A. Assegie, S. Rajkumar, and N. K. Kumar, "A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1831–1838, 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.
- [18] R. R. K. Al-Taie, B. J. Saleh, A. Y. F. Saedi, and L. A. Salman, "Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5229–5239, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5229-5239.
- [19] B. S. Shukur and M. M. Mijwil, "Involving machine learning techniques in heart disease diagnosis: a performance analysis," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 2177–2185, Apr. 2023, doi: 10.11591/ijece.v13i2.pp2177-2185.
- [20] A. O. Salau, T. A. Assegie, E. D. Markus, J. N. Neneh, and T. I. Ozue, "Prediction of the risk of developing heart disease using logistic regression," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, pp. 1809–1815, Apr. 2024, doi: 10.11591/ijece.v14i2.pp1809-1815.
- [21] A. Naizagarayeva *et al.*, "Detection of heart pathology using deep learning methods," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 6, pp. 6673–6680, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6673-6680.
- [22] M. Almatari *et al.*, "Cardiovascular disease risk factors prediction using deep learning convolutional neural networks," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 4, pp. 4471–4487, 2024, doi: 10.11591/ijece.v14i4.pp4471-4487.
- [23] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges," *Algorithms*, vol. 17, no. 2, p. 78, Feb. 2024, doi: 10.3390/a17020078.
- [24] M. Siddhartha, "Heart disease dataset (comprehensive)," *IEEE-dataport.org*, pp. 23–25, 2020. Accessed: Feb. 05, 2024. [Online]. Available: <https://ieee-dataport.org/authors/manu-siddhartha>
- [25] C. Lenfant, "Chest pain of cardiac and noncardiac origin," *Metabolism: Clinical and Experimental*, vol. 59, pp. 41–46, Oct. 2010, doi: 10.1016/j.metabol.2010.07.014.
- [26] C. Wang *et al.*, "Association of age of onset of hypertension with cardiovascular diseases and mortality," *Journal of the American College of Cardiology*, vol. 75, no. 23, pp. 2921–2930, Jun. 2020, doi: 10.1016/j.jacc.2020.04.038.
- [27] R. M. Al-Amer, M. M. Sobeh, A. A. Zayed, and H. A. Al-Domi, "Depression among adults with diabetes in Jordan: risk factors and relationship to blood sugar control," *Journal of Diabetes and its Complications*, vol. 25, no. 4, pp. 247–252, Jul. 2011, doi: 10.1016/j.jdiacomp.2011.03.001.
- [28] A. H. E. M. Maas and Y. E. A. Appelman, "Gender differences in coronary heart disease," *Netherlands Heart Journal*, vol. 18, no. 12, pp. 598–603, Nov. 2010, doi: 10.1007/s12471-010-0841-y.
- [29] F. D. Fuchs and P. K. Whelton, "High blood pressure and cardiovascular disease," *Hypertension*, vol. 75, no. 2, pp. 285–292, Feb. 2020, doi: 10.1161/HYPERTENSIONAHA.119.14240.
- [30] J. A. S. Carson *et al.*, "Dietary cholesterol and cardiovascular risk: a science advisory from the american heart association," *Circulation*, vol. 141, no. 3, pp. E39–E53, Jan. 2020, doi: 10.1161/CIR.0000000000000743.
- [31] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [32] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, Apr. 2020, doi: 10.1007/s11704-019-8208-z.




**BIOGRAPHIES OF AUTHORS**

**Nomula Nagarjuna Reddy**    is a budding scholar whose academic journey is marked by determination and a quest for knowledge. He graduated with distinction from Alphores High School, Karimnagar, Telangana, India, 505451, completing his matriculation in 2020 and senior secondary in 2022. Currently, he is pursuing a bachelor of technology in computer science engineering at B V Raju Institute of Technology, Narsapur, Medak, Telangana, India, 502313. His research interests include artificial intelligence and machine learning, data science, and full stack development, reflecting his extensive knowledge and enthusiasm for innovation in the field of computer science. He can be contacted at email: [nagarjunareddynomula2@gmail.com](mailto:nagarjunareddynomula2@gmail.com).






**Lingadally Nipun**    is a talented student determined to make a significant impact through his knowledge. He graduated from Vikas the Concept School, Bachupally, Hyderabad, Telangana, and completed his Senior Secondary at Excellencia Junior College, Madhapur, Hyderabad, Telangana, in 2020 and 2022, respectively. Currently, he is pursuing a bachelor of technology in computer science engineering at B V Raju Institute of Technology, Narsapur, Medak, Telangana, India, 502313. His research interests include cyber security, web development, artificial intelligence, data science, and app development. He can be contacted at email: [nipun.lingadally@gmail.com](mailto:nipun.lingadally@gmail.com).






**MD Uzair Baba**    is a committed student with a keen interest in technology. He completed his matriculation at Telangana Minorities Residential Educational Institutions Society in Devarakadra, Mahabubnagar, Telangana, India, and his Polytechnic at Government Polytechnic College, Gadwal, Telangana, India. Currently, he is pursuing a bachelor of technology in computer science and engineering at B V Raju Institute of Technology, Narsapur, Medak, Telangana, India, 502313. His interests include artificial intelligence, machine learning, and full stack development. He can be contacted at email: [uzairmohd2026u@gmail.com](mailto:uzairmohd2026u@gmail.com).



**Nyalakanti Rishindra**    is a dedicated student with a passion for technology. He completed his matriculation at Dr. K K R's Gowtham School, Gandimaisamma (Rd, Borampet (Vill)), Hyderabad, Telangana, India, in 2020, and his Senior Secondary School at NSR Impulse College, Pragathinagar, Hyderabad, Telangana, India, in 2022. Currently, he is pursuing a bachelor of technology in computer science and engineering at B V Raju Institute of Technology, Narsapur, Medak, Telangana, India, 502313. His interests include artificial intelligence, machine learning, and data science. He can be contacted at email: [nyalakantyrishi11@gmail.com](mailto:nyalakantyrishi11@gmail.com).



**Thoutireddy Shilpa**    received her B.Tech. degree in computer science and engineering from Vaagdevi College of Engineering, JNTUH, Warangal (D), India, in 2006 and a Postgraduate (M.Tech.) degree in computer science and engineering from Amina Institute of Technology, JNTUH, RangaReddy (D), India, in 2012. Currently, she is working as an assistant professor in the Department of B V Raju Institute of Technology, Narsapur, Medak, Telangana, India, 502313, and a research scholar at, the Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India. She can be contacted at email: [shilpathoutireddy@gmail.com](mailto:shilpathoutireddy@gmail.com).