Received: 24/04/2025, Accepted: 13/07/2025

Research Article

Cardiovascular Disease Prediction through Machine Learning: A Comparative Study of Ensemble Techniques

Nutan Gusain

Department of Computer Science and Engineering, SCSE, Galgotias University, Greater Noida, India

nutan.gusain41@gmail.com

ABSTRACT

Cardiovascular disease is a worldwide health issue that necessitates enhancements in the evaluation of risks and early identification methods. Several lethal heart illnesses have been transmitted to humans through various recognized mechanisms. According to a figure from the World Health Organization (WHO), over 17.9 million individuals in the country perish annually, representing 32% of global mortality. The increasing annual population poses a significant challenge in terms of early-stage diagnosis and treatment. Machine learning approaches have proven highly useful in the realm of healthcare. The study employed the "heart dataset" obtained from Kaggle. The implementation was carried out using the Open editor and the Python programming language. This work aims to calculate and assess the precision of ensemble machine learning algorithms in predicting cardiac disease, along with other performance metrics. The study examined seven proven models, including Gradient Boosting, AdaBoost, XGBoost, CatBoost, Light GBM (boosting techniques), and Random Forest and Extra Tree Classifier (bagging algorithm). These models were tested and trained using the heart dataset. 80% of the dataset will be allocated for training, while the remaining 20% will be used for testing. This study examines machine learning methods to determine the optimal model for predicting cardiac disease. Upon assessing multiple models, we determined that XGBoost achieves the highest accuracy and F1-score, with values of 87.50% and 88.44% respectively, surpassing the performance of other models we employed.

Keywords: *Heart disease prediction, Machine Learning, heart dataset, gradient boosting, Adaboost, Random Forest, XGBoost, CatBoost, Light GBM, Extra tree classifier, Ensemble Learning, Bagging, Boosting*

1. Introduction

Cardiovascular Diseases (CVDs) are the predominant cause of death worldwide. Cardiovascular disease is a broad term used to describe a variety of disorders that impact the circulatory system, blood vessels, and heart. It includes rheumatic heart disease, which is a complication of a streptococcal infection; coronary heart disease, which is caused by narrowed or blocked arteries; cerebrovascular disease, also known as stroke, which results from blood flow disruption to the brain; and congenital heart disease, which is a congenital disability affecting the structure of the heart. This broad category of diseases hinders the body's capacity to circulate blood efficiently, which can result in various health issues. It resulted in 695,000 deaths in America in 2021, a significant increase. The WHO study indicates that 17.9 million people nationwide pass away each year, accounting for 32% of the world's yearly mortality rate of 697,000 deaths. One-fifth of all deaths globally occur in India. As per the report by the WHO, the deaths by heart attack are increasing annually [12].

The lack of diagnostic facilities, qualified medical professionals, and other resources that may influence an accurate diagnosis of a cardiac disease can make it extremely difficult to identify and treat heart disorders in their earliest stages, particularly in nations that lack adequate infrastructure and financial resources. However, traditional diagnostic methods frequently use time-consuming and costly

*Corresponding author: Nutan Gusain, Department of Computer Science and Engineering, SCSE, Galgotias University, Greater Noida, India (nutan.gusain41@gmail.com) procedures like stress examinations and angiography. These operations can also pose dangers for the patient and be intrusive. Moreover, misdiagnosis continues to possess the potential to occur even with these techniques, which could result in inadequate or improper therapy. The problem has led to the development of a system to help with the early diagnosis of cardiovascular illness by the integration of machine learning and technological advancements in computers to provide healthcare tools and applications. Machine learning is a specialised branch of AI, one of data science's fastest-growing areas. The advancement of machine learning represents an innovative approach for addressing the challenges associated with diagnosing cardiovascular disease (CVD). Large and complex datasets comprising patient data, including population size, medical history, test results, and lifestyle factors, can be easily processed by machine learning algorithms. Implementing this research will train machine learning models to detect connections and trends that traditional approaches might neglect. The technique to assess an individual's chance of getting CVD is to look for certain patterns. Machine learning algorithms are designed for multiple jobs, such as classification, decision-making, and predictions [13].

The potential applications of machine learning-based CVD prediction are extensive. ML models can empower healthcare providers to implement preventive actions, which could save lives, by enabling early diagnosis and risk assessment. Additionally, ML can help make better-informed treatment choices, enhancing patient outcomes and lowering medical expenses. Healthcare typically has intricate structures and large volumes of data. Machine learning (ML) methods are capable of mining big data to extract the required information [14]. This research project investigates the performance of seven machine learning algorithms for predicting heart disease. To determine the most effective model, the project will assess accuracy and evaluate performance using metrics like the confusion matrix, F1-score, ROC curve, recall, precision, and accuracy.

2. Literature Survey

Many authors have conducted research and have mentioned and implemented various techniques on cardiac disease using numerous datasets available on Kaggle and the UCI repository. Most of the authors have chosen parameters like age, smoking, diabetes, blood pressure, high body mass index, and poor diet, concluding that these are the reasons which increase the severity of cardiovascular diseases (CVDs). Pal, M. et al. used two ML algorithms, KNN and MLP and observed that only MLP performed best in the accuracy, which is 82.47% and ROC curve 86.41% while KNN performed 73.77% on accuracy but performed better on the ROC curve, that is 86.21%. The paper's authors used the publicly available UCI repository for the dataset [1]. Saboor, A. et al. used nine algorithms, which are MNB, SVM, LR, CART, LDA, AB, RF, ET and XGB, out of which SVM gives the best accuracy of 96.72% in tuning the hyperparameters [2]. Alqahtani et al. used six algorithms, of which four are ML and two are deep learning. RF performs well with an accuracy of 88.65% and DNN gives an accuracy of 87.59%; however, after observing that deep learning did not perform better than ML [3]. Bhatt, C. M. et al. used four ML techniques: DT, RF, XGB and MLP. Among them, the ensemble learning model XGB gives the highest accuracy of 87.02% without using the GridSearchCV method and with the CV method, it got 86.87% [4]. Pradhan, M. R.et al. used five models: LR, SVM, MLP with PCA, DNN and bootstrap aggregation with RF, which gives the best accuracy of 97.67% with the 90-10 split. The authors of this paper have done four different train-test splits, 90-10, 80-20, 70-30, 60-40, for each model they have used, which gives distinct accuracy for every four different splits [5].

Almulihi, A. et al. used five regular ML approaches RF, LR, DT, NB and KNN, two hybrid deep learning approaches CNN-LSTM, CNN-GRU and a proposed model of stacking SVM from which LR got the highest accuracy of 75.6% on full features, from hybrid models CNN-LSTM got the accuracy of 76.64% on full feature set and among all of them stacking SVM got the overall highest accuracy of 78.81% on full feature set [6].

Gupta, C. et al. researched that using six supervised learning models, LR, SVM, NB, DT, KNN and RF, can give better accuracy. Still, among them, only logistic regression gives the overall higher accuracy of 92.30%. The paper's authors used the UCI repository dataset [7]. Roy, R. S. et al. used only two ML classifiers, NB and AdaBoost, which give the same accuracy of 95% for both, but NB has the highest precision of 100%. The research was done using the UCI repository dataset, which they claim to have been thoroughly confirmed by other researchers [8].

Arumugam, K. et al. used three models for their research: NB, SVM and DT, out of which DT claims to have the highest accuracy of nearly 90%. The researchers of this paper used only accuracy as a performance metric, which is surely not enough to get the result. For research work, the authors of this paper have utilised the Cleveland dataset, which is available publicly on the internet [9]. Rahman, M. M. et al. have prepared a web-based system that will indicate the presence of CVD in the patient while using eight ML classifiers, of which only two classifiers give the best accuracy of 99%. The researchers of this paper developed the computer-aided system on Jupyter IDE of Python, a manual diagnosis of the system, in which an individual has to input their data to know their cardiac condition. Researchers have used the UCI repository of datasets [10]. Rani, P. et al. have made a decision support system that predicts cardiac disease using many ML classifiers, NB, SVM, RF, LR and AdaBoost, in which only RF performs with a high accuracy of 86.6%. The authors of this paper used the Cleveland dataset from the UCI repository for their research work [11].

3. Methodology

This section of the paper clarifies the methods employed in cardiac disease prediction using ensemble learning. Before proceeding, it is important to understand ensemble learning. Ensemble learning is a machine learning technique that combines predictions from multiple models to improve overall performance. It leverages the strengths of several classifiers to generate a more accurate and robust prediction than any single model could achieve on its own.

This study explores the use of ensemble learning techniques for predicting heart disease. Ensemble learning combines the strengths of multiple machine learning models to achieve better overall performance than any single model could alone. Two primary ensemble techniques are employed in this research: bagging and boosting. Bagging involves training multiple models on different subsets of the original training data created through bootstrapping. In contrast, boosting train models sequentially, where each subsequent model focuses on improving the errors made by the previous model, leads to a gradual improvement in overall performance. By leveraging these ensemble techniques, the study aims to identify the most effective machine learning model for predicting heart disease.

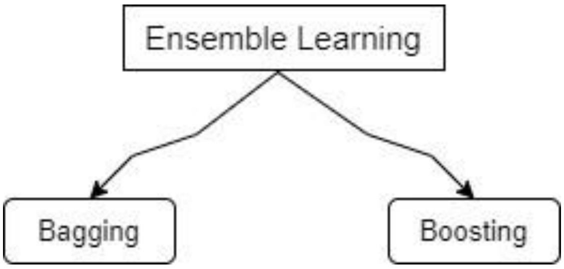


Figure 1: Ensemble learning

3.1. Algorithms

Gradient Boosting: Gradient boosting is an ensemble ML technique that utilizes boosting algorithm which is also the foundation other boosting algorithm which every classifier works upon. It also offers flexibility as it works with every base learner, but is much slower than Light GBM and Xgboost.

Adaptive Boosting: Adaptive boosting is also called “Ada Boost”. A boosting algorithm helps integrate multiple weak learners into a single proficient learner. It acts as a decision tree support system in which every weak learner acts as a stump with 1 root and 2 leaves. The number of iterations of every stump depends upon certain parameters: error rate, learning rate and the complexity of the dataset.

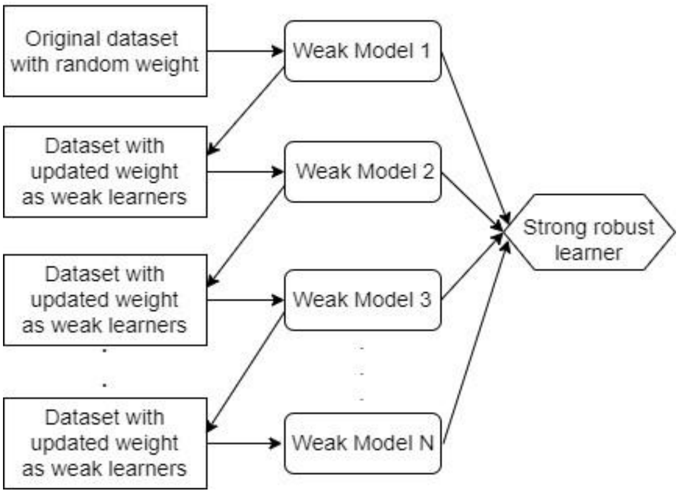

Original dataset with random weight
Weak Model 1
Dataset with updated weight as weak learners
Weak Model 2
Dataset with updated weight as weak learners
Weak Model 3
Dataset with updated weight as weak learners
Weak Model N
Strong robust learner

Figure 2: Adaptive boosting model

Extreme Boosting: Extreme boosting is also called “Xg Boost”. XGBoost is an ensemble machine learning algorithm that uses the gradient descent framework to optimise the loss function during training, leading to a more precise prediction. This model combines all weak learners to produce an accurate result. This boosting algorithm is known for its scalability, which handles large datasets and efficient memory resources, and has a reduced training time.

Cat Boost: Also known as an ordered and symmetric tree. The category boosting algorithm is an ensemble machine learning algorithm that uses the gradient boosting framework. It mainly focuses on helping with categorical data without any preprocessing. It is scalable, handles large datasets, and is robust enough to overfit, which acts as a valuable option for various tasks across several domains. It builds the sequence of weak learners iteratively, where every weak model focuses on correcting the errors of the previous weak models.

Light GBM: Light gradient boosting machine is an ensemble ML classifier which uses a histogram-based algorithm for decision tree splitting, which makes training time faster. It focuses on efficiency and high accuracy. It offers flexibility, scalability and low memory usage, making it a valuable option for NLP and other domains. It balances accuracy, efficiency and robustness for various classification tasks.

Extra Tree Classifier: Extra Tree Classifier is an ensemble ML model that utilises no bagging techniques, i.e., the whole training dataset is used to train every tree. It introduces randomness in the tree-building process, combining every tree model at the end and predicting the model.

Random Forest: Random Forest is a renowned ensemble machine learning algorithm that utilises the bagging technique. As the name signifies, an algorithm that generates a forest with a sizeable no. of trees. It combines several classifiers to improve the model's performance and tackle a challenge. [15].

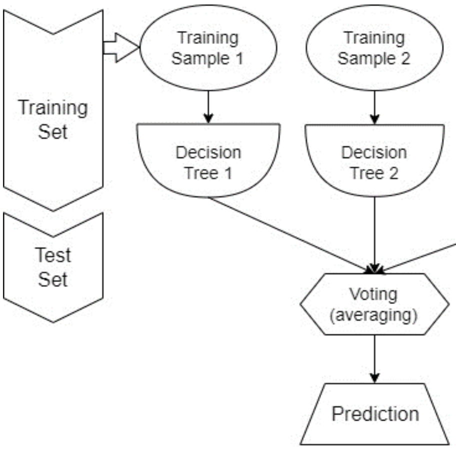


Figure 3: Random Forest Classifier Model

In Figure 3, the research utilises a decision tree flowchart structure. Each internal node within the tree represents a decision point based on the value of a specific attribute. The branches stemming from these nodes indicate the possible outcomes based on the chosen attribute value. Finally, the leaf nodes at the end of each branch represent the final prediction for a particular data point. RF employs a bagging technique in which every subset of the training sample is made by random sampling, after training the samples in every decision tree node. Now, in the voting part, every outcome of the decision tree is merged and gives the average of the outcomes produced. Now, the voting outcome will be tested, and after that, it will provide the model's prediction.

3.2. Dataset Description

This research leverages the "heart dataset" from Kaggle Competitions [16]. The accessibility of this dataset facilitates the development of a more precise and optimistic model for heart disease prediction.

The dataset contains 918 patients’ data and 12 test attributes with no null values. This dataset includes categorical and non-categorical data, as shown in Table 1. It consists of patients aged between 28 and 77. Cleaveland, Switzerland, Hungarian, Long Beach, VA, and the Stalog (Heart) Data Set are the five datasets used in the compilation. Moreover, 272 duplicate observations have been captured in this dataset.

Table 1. Attributes of the dataset

S. No.	*Attribute*	*Description*	*Value Range*
1	Age	Patient's age in years	28 - 77
2	Sex	Patient’s gender (0: Female, 1: Male)	0, 1
3	Cp	Chest Pain Type	0, 1, 2, 3
4	Resting BP	Resting Blood Pressure	0 - 200
5	Cholesterol	Serum cholesterol [mm/dl]	0 - 603
6	Fasting BS	Fasting blood sugar [mg/dl] (If sugar>120 mg/dl, then 1, else 0)	0, 1
7	Resting ECG	Resting Electrocardiographic Result	0, 1, 2

		(0: Normal, 1: ST, 2: LVH)	
8	Max HR	Max. Heart Rate	60 – 202
9	Exercise Angina	Exercise-induced angina (0: No, 1: Yes)	0, 1
10	Old Peak	ST depression induced by exercise	-2.6 – 6.2
11	ST slope	Slope of exercise ST segment (0: Down, 1: Flat, 2: Up)	0, 1, 2
12	Heart Disease	Target class (0: Normal, 1: Heart disease)	0, 1

Figure 4 suggests that males exhibit a higher risk of heart disease compared to females. A portion of the black bar shown in the graph shows that people have no heart disease, and the blue bar shows people with heart disease. Compared to males, females have less risk of contracting cardiac disease.

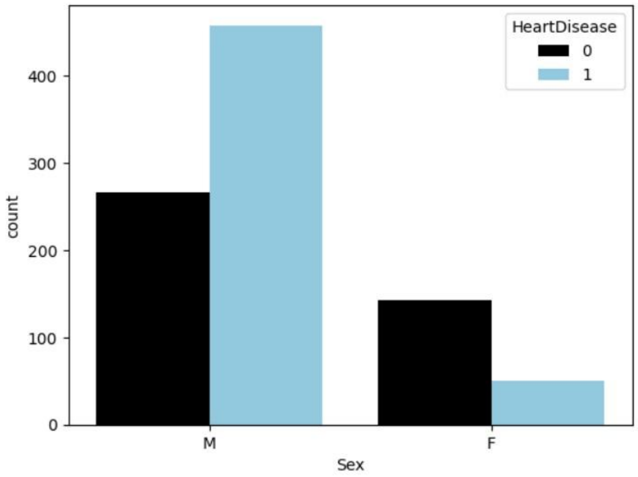


Figure 4: Data Visualization in Heart Disease

In Figure 5, an increased risk of contracting cardiac disease between the ages of 40 and 70 has been observed. Normally, the resting blood pressure of an individual is considered normal below 120. Still, in this graph, most individuals have more than 120 bp, which is regarded as lethal and increases the chance of cardiac disease.

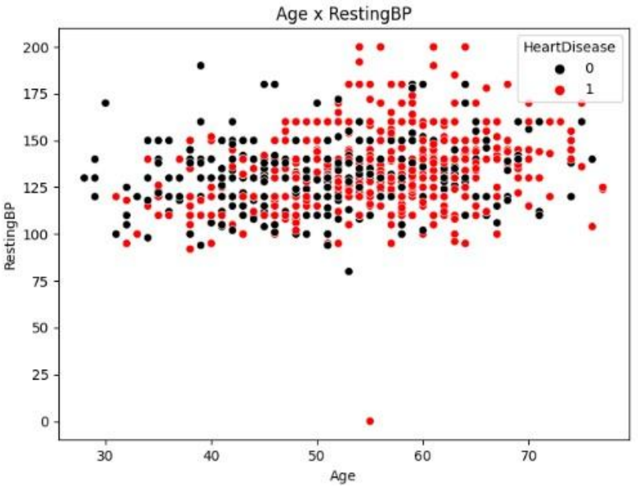


Figure 5: Scatterplot between Resting BP and Age

In Fig. 6, the individuals aged between 35 and 77 with high cholesterol (200 mg/dL) have heart disease, and this cholesterol level can increase the risk of heart attack and cardiac disease. Furthermore, the presence of heart disease in individuals with lower cholesterol levels highlights the need to analyse other parameters.

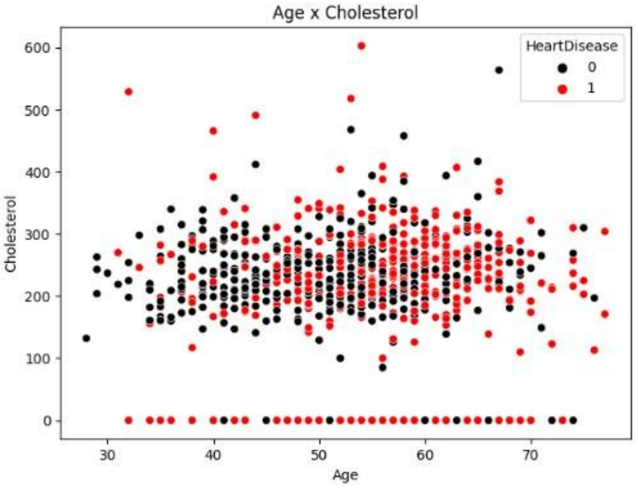


Figure 6: Scatterplot between Age and Cholesterol

Figure 7 shows that individuals aged between 35 and 70 with heart disease also tend to have a high heart rate. It's important to note that normal heart rate ranges vary across ages (children vs. adults). For individuals, a heart rate exceeding 100 beats per minute can be life-threatening.

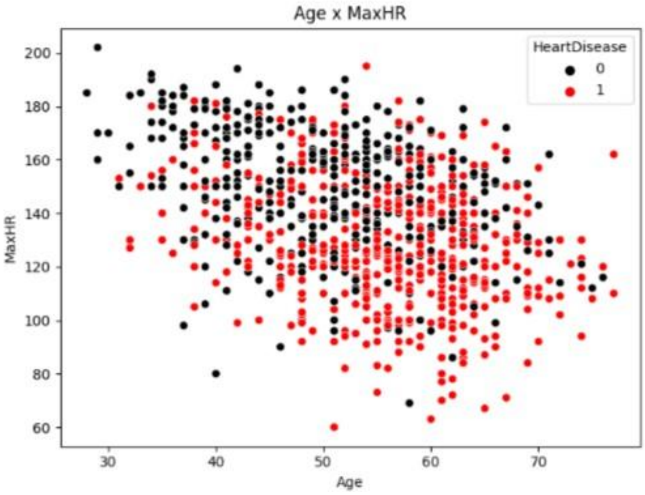


Figure 7: Scatterplot between Age and Max Heart Rate

In Figure 8. Individuals below point 2 have less risk of contracting cardiac disease, which means values which are low or negative signify a dip in depression. Individuals who are above pt. 2 have a high probability of getting cardiac-related diseases.

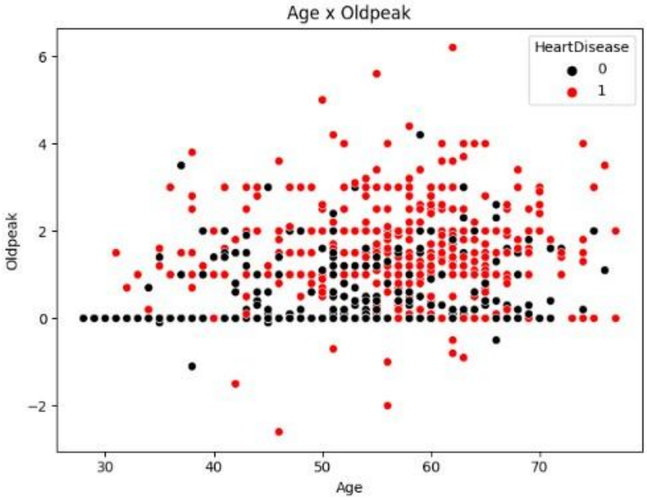


Figure 8: Scatterplot between Age and Oldpeak

4. Proposed Workflow

The following section delves into the model utilized for the experiment. Figure 9 provides a visual representation of the model's architecture. To facilitate a deeper understanding, this section is divided into three sections. Section 1 meticulously dissects the architectural structure depicted in Figure 9. Section 2 explores the data preprocessing techniques employed during the experiment. Finally, the last section discusses the five-performance metrics used to assess the model's effectiveness.

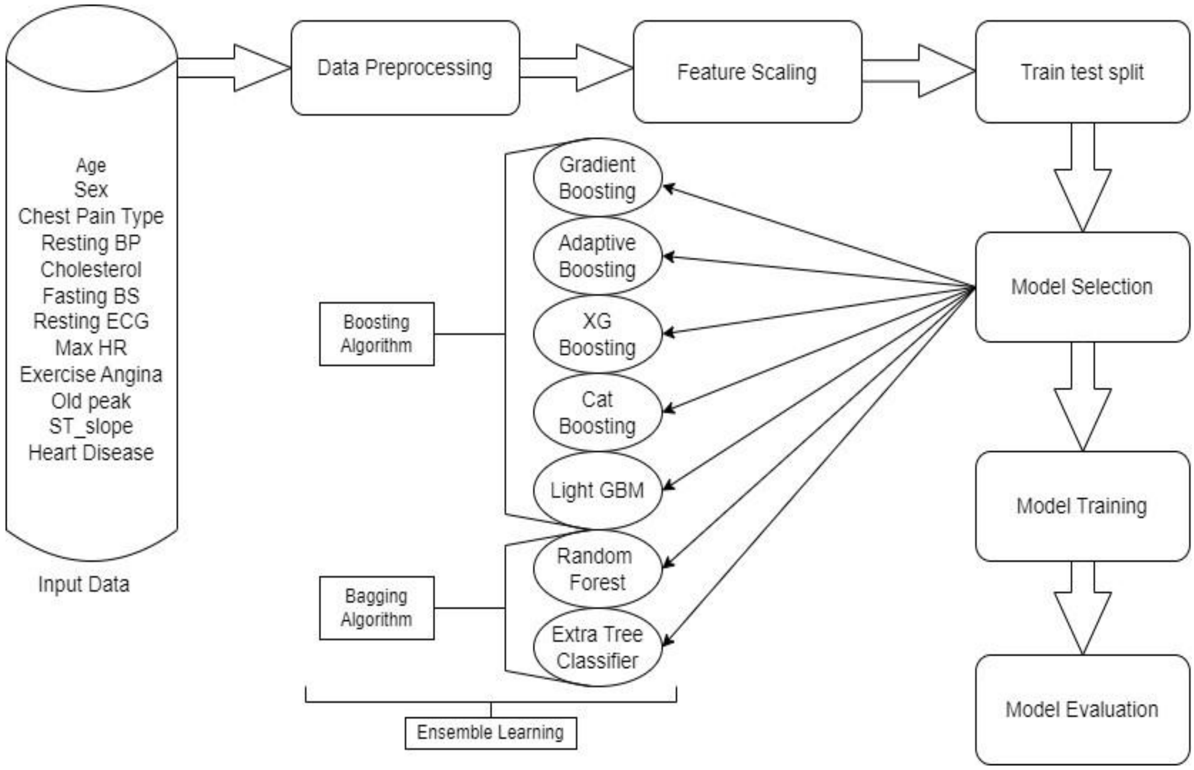


Figure 9: Model Architecture

4.1. Working of the architectural model

- The research begins by collecting and downloading the dataset from Kaggle. Exploratory Data Analysis (EDA) is then performed to understand the dataset. Subsequently, data preprocessing methods are applied to prepare the data for modelling. These methods may include outlier detection, data normalisation, and duplicate values removal.
- After applying the feature scaling method, every independent feature within the data is provided with a fixed range. This is carried out as part of the pre-processing phase of the data, which addresses the greatly variable magnitudes, units, and values, ranging from 1 to -1, putting all values in a decimal system.
- Then, a train-test split is done, where 80% of the data will be used for training and 20% for testing.
- After that, a model selection approach is utilised to identify the best model, such as an Extra Tree classifier, Random Forest, Gradient Boost, AdaBoost, CatBoost, XGBoost, and Light GBM.
- At this point, essential libraries will be imported, installed, and models trained.
- The trained model's evaluation uses performance measures, including the ROC curve, F1 score, recall, precision, accuracy, and confusion matrix.
- One of the most effective algorithms is then identified after the results are analysed.

4.2. Data Preprocessing

- Data examination and quality assessment: The validity and quality of the dataset are evaluated first during the preprocessing stage. The data must be analysed thoroughly to ensure quality, alignment with the research objective, and internal consistency.
- Data Cleansing: Identifying outliers in a dataset, updating missing data, and removing, editing, or correcting erroneous or redundant information.

- Feature Scaling: Feature scaling is an approach that gives each independent feature in the dataset a fixed range. It is carried out at the pre-processing phase of the data, which addresses the greatly varied magnitudes, units and values, i.e., between 1 and -1.

4.3. Evaluation Metrics

Many performance metrics can be used to assess the performance of a specific classification technique. Different parameters' effectiveness depends on the nature of the problem being tackled. In certain situations, like this, accuracy might be the ideal choice; in others, precision or recall might be necessary. In healthcare applications, recall (sensitivity) frequently takes centre stage as a critical performance metric when choosing classification algorithms.

Five commonly used performance evaluation metrics have been used to evaluate the performance of ensemble learning classification models: accuracy, F1-score, ROC, recall, and precision [15].

Confusion Matrix: In this area of research, the most useful tool for assessing heart disease prediction is the confusion matrix. It is employed to observe how various classifiers perform [7]. A confusion matrix is a table structure which enables monitoring the performance of the supervised learning technique. Each row in an N x N matrix depicts an event in an actual class, and every column depicts occurrences in a class that is predicted. The table below is an example of a confusion matrix for a binary classification, based on which another term or metric might be put together. A few of the metrics are covered below.

Table 2: Description of the Confusion Matrix

Term	Full form	Descriptions
TP	True Positive	+ve cases which are predicted as +ve
FP	False Positive	-ve cases which are predicted as +ve
TN	True Negative	-ve cases which are predicted as -ve
FN	False Negative	+ve cases which are predicted as -ve

Accuracy: A performance metric assesses how much the system can predict the true predictions.

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{1}$$

Recall: A performance metric assesses the system's capability for accurate positive prediction.

$$\boldsymbol{Recall = \frac{TP}{TP+FN}} \tag{2}$$

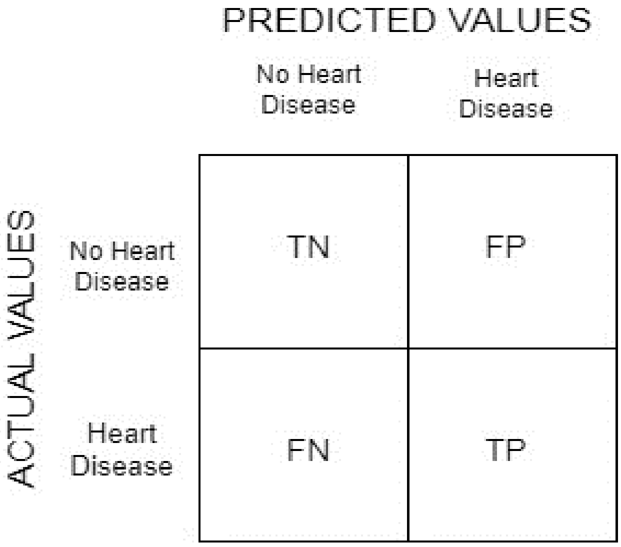
Precision: It assesses the ability of a system to provide only the most relevant outcomes.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{3}$$

F1-Score: It is a performance metric that uses the harmonic mean to combine the results of sensitivity and precision.

$$F1 - Score = 2\frac{\mathrm{recall*precision}}{\mathrm{recall+precision}} \tag{4}$$

ROC curves: The receiver operating characteristic curve is a graphical representation of the true positive rate (TPR) and false positive rate (FPR) as a function of the performance of a binary classification system. By analysing TPR to FPR over numerous threshold values, the ROC curve effectively distinguishes the signal from the noise [8].



		PREDICTED VALUES	
		No Heart Disease	Heart Disease
ACTUAL VALUES	No Heart Disease	TN	FP
	Heart Disease	FN	TP

Figure 10: Confusion Matrix

The theory section should expand on the foundational context introduced in the Introduction, serving as a bridge to support deeper exploration in the study. It should lay out the theoretical underpinnings that form the basis for the research, without repeating basic background information. On the other hand, the Calculation section should concentrate on the practical implementations and developments that emerge from the established theoretical framework. Rather than covering basic definitions or widely known theories, the focus should remain on the specific theoretical concepts and their direct relevance to the current work.

5. Results and Discussion

The results indicate that after training and testing the machine learning approach, the XGBoost algorithm achieves significantly higher accuracy than existing algorithms. Each algorithm's precision, recall, F1 score, and confusion matrix should be considered while calculating accuracy. Measuring the findings, it concludes that, at 87.50%, the Xgboost algorithm has the highest accuracy. The comparison is shown in Table 3. The research did not identify other algorithms that performed comparably well on the data.

Table 3: Description of the Confusion Matrix

Model	Model Accuracy (%)	F1 Score (%)	Recall	Precision	ROC curve
ADB	84.78	86.27	0.91	0.84	0.91
XGB	87.50	88.44	0.91	0.86	0.93
CatB	86.96	87.88	0.90	0.86	0.94
GB	86.96	87.76	0.89	0.87	0.94
LGBM	84.78	86.14	0.90	0.83	0.92
RF	85.33	85.57	0.89	0.84	0.92
ETC	84.78	87.13	0.91	0.84	0.93

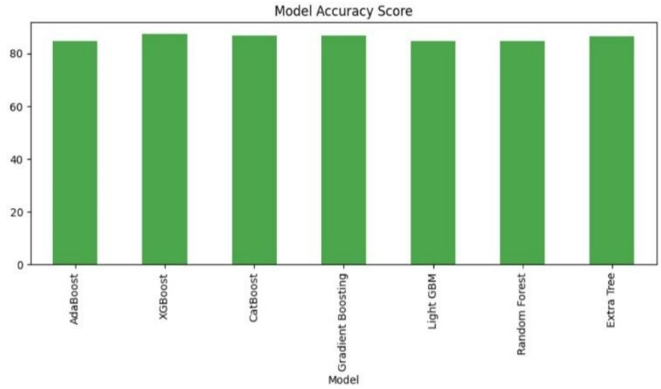


Figure 11: Comparison of algorithms in terms of model accuracy

As shown in Fig. 11, the Xgboost model achieves an accuracy of 87.50%, which is the highest achievable model accuracy out of multiple ensemble ML models.

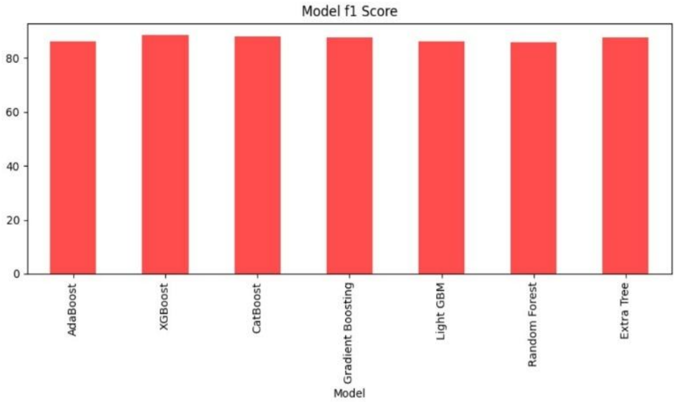


Figure 12: Comparison of algorithms in terms of performance metrics (F1 score)

As shown in Figure 12, the Xgboost model achieved the highest performance metrics in the F1 score of 88.44%.

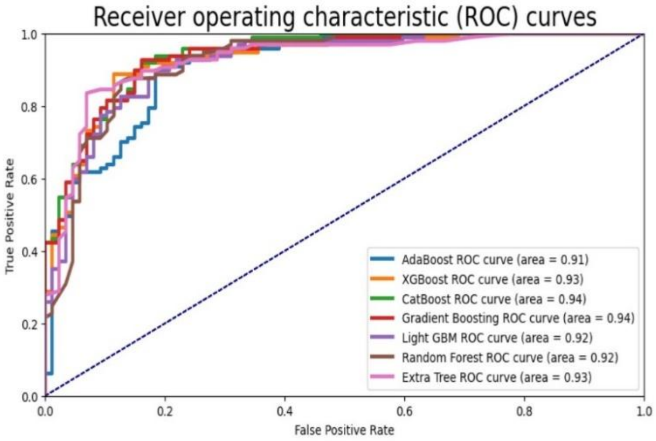


Figure 13: Comparison in terms of performance metrics in Receiver Operating Characteristic (ROC) Curves

As shown in Figure 13, each of the algorithms used in this research has ROC curves that perform well; however, CatBoost and Gradient Boosting have the best ROC values, both of which are 0.94.

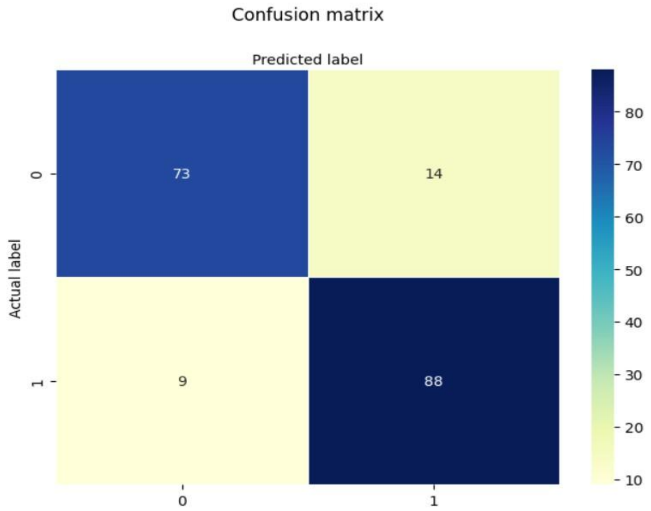


Figure 14: Confusion matrix for XG Boost

The greatest accuracy in Xgboost is 87.50%, as shown in Fig. 14. It is determined that 88 true positives, 73 true negatives, 9 false negatives, and 14 false positives are from this number.

6. Conclusions

After conducting extensive research, it has been shown that unaddressed risk factors can significantly elevate the chances of developing serious heart disease. The research has identified several key factors that can increase a person's risk of heart disease. This study investigates the application of various ensemble learning algorithms to the heart disease prediction task. The research leverages a heart disease dataset to assess the effectiveness of these algorithms in predicting the presence of heart disease. This research investigated machine learning algorithms for predicting cardiovascular illness outcomes. The XGBoost algorithm demonstrated exceptional performance on the utilised dataset, suggesting its potential for accurate cardiovascular disease prediction. The Xgboost Classifier has achieved the most favourable results regarding accuracy and other evaluation metrics. Compared to other researchers, Bhatt, C. M. et al., who got the accuracy of 87.02% on the XGBoost model, the model used in this paper performed better and exceptionally well. Hybrid Machine learning and stacking models are currently used throughout all fields for future work to accomplish better results. Future research directions include exploring the application of more machine learning techniques, such as stacking models and hybrid deep learning approaches, to improve the prediction of cardiovascular disease outcomes.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, “Risk prediction of cardiovascular disease using Machine Learning Classifiers,” Open Medicine, vol. 17, no. 1, pp. 1100–1113, Jan. 2022. doi:10.1515/med-2022-0508

[2] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, “A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms,” Mobile Information Systems, vol. 2022, no. 1, pp. 1–9, Mar. 2022, doi: https://doi.org/10.1155/2022/1410169.

[3] A. Srivastava, Sandhya Umrao, and S. Biswas, “Exploring Forest Transformation by Analyzing Spatial-temporal Attributes of Vegetation using Vegetation Indices,” International

Journal of Advanced Computer Science and Applications, vol. 14, no. 5, Jan. 2023, doi: https://doi.org/10.14569/ijacsa.2023.01405114.

[4] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” IEEE Access, vol. 7, no. 1, pp. 81542–81554, 2019, doi: https://doi.org/10.1109/access.2019.2923707.

[5] D. Swain, B. Parmar, H. Shah, A. Gandhi, M. R. Pradhan, H. Kaur, & B. Acharya, “Cardiovascular Disease Prediction using Various Machine Learning Algorithms,” Journal of Computer Science, vol. 18, no. 10, pp. 993–1004, Oct. 2022, doi: https://doi.org/10.3844/jcssp.2022.993.1004.

[6] H. Sharma, P. Kumar, and K. Sharma, “Identification of Device Type Using Transformers in Heterogeneous Internet of Things Traffic,” Lecture notes in networks and systems, pp. 471–481, Jan. 2023, doi: https://doi.org/10.1007/978-981-99-3010-4_40.

[7] C. Gupta, A. Saha, N. V. Subba Reddy, and U. Dinesh Acharya, “Cardiac Disease Prediction using Supervised Machine Learning Techniques.,” Journal of Physics: Conference Series, vol. 2161, no. 1, p. 012013, Jan. 2022, doi: https://doi.org/10.1088/1742-6596/2161/1/012013.

[8] A. Srivastava and P. Ahmad, “A Probabilistic Gossip-based Secure Protocol for Unstructured P2P Networks,” Procedia Computer Science, vol. 78, pp. 595–602, 2016, doi: https://doi.org/10.1016/j.procs.2016.02.122.

[9] K. S. Kaswan, J. S. Dhatterwal, H. Sharma, and K. Sood, “Big Data in Insurance Innovation,” Big Data: A Game Changer for Insurance Industry, pp. 117–136, Jul. 2022, doi: https://doi.org/10.1108/978-1-80262-605-620221008.

[10] M.M. Rahman, M.R. Rana, M. Nur-A-Alam, M.S.I. Khan, K.M.M. Uddin, “A web-based heart disease prediction system using machine learning algorithms,” Network Biology, pp. 64–80, June 2022.

[11] 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, Jan. 2021, doi: https://doi.org/10.1007/s40860-021-00133-6.

[12] Nutan Gusain and Himanshu Sharma, “Communication-Efficient Federated Learning in Industrial IoT — A Framework for Real-Time Threat Detection and Secure Device Coordination,” International Journal on Computational Modelling Applications, vol. 2, no. 2, pp. 18–29, May 2025, doi: https://doi.org/10.63503/j.ijcma.2025.115.

[13] P. Sapra, Divya Paikaray, Nutan Gusain, M. Abrol, S. Ramesh, and S. Bhardwaj, “Evaluation of soft computing in methodology for calculating information protection from parameters of its distribution in social networks,” Soft Computing, Jun. 2023, doi: https://doi.org/10.1007/s00500-023-08633-8.

[14] Anubhava Srivastava, Rakesh Dubey & Susham Biswas, “Comparison of Sentinel and Landsat Data Sets over Lucknow Region Using Gradient Tree Boost Supervised Classifier,” In International Conference on Emerging Trends and Technologies on Intelligent Systems (pp. 221-232), 2023. Singapore: Springer Nature Singapore.

[15] S. Dev, F. M. Savoy, Y. H. Lee, and S. Winkler, “Nighttime sky/cloud image segmentation,” in: Proc. IEEE International Conference on Image Processing (ICIP), IEEE Xplore, pp. 345–349, Sep. 01, 2017.

[16] H. Sharma, P. Kumar, and K. Sharma, “Recurrent Neural Network based Incremental model for Intrusion Detection System in IoT,” Scalable Computing: Practice and Experience, vol. 25, no. 5, pp. 3778–3795, Aug. 2024, doi: https://doi.org/10.12694/scpe.v25i5.3004.