COMPUTOLOGY: Journal of Applied Computer Science and Intelligent Technologies
Volume 3, Issue 2, Jul-Dec 2023, pp. 171-192
DOI: 10.17492/computology.v3i2.2309
https://www.journalpressindia.com/cjacsit

Heart Disease Prediction: A Comparative Analysis of Machine Learning Algorithms

Adarsh Sharma[1], Himanshu Sharma[1], Sudeep Varshney[1] and Nutan Gusain[1]

ABSTRACT

Nowadays, heart disease is one of the biggest concerns. A WHO statistic states that 17.9 million people worldwide die each year, accounting for 32% of all deaths worldwide. It is now very difficult to diagnose and start therapy at an early stage due to population growth. In the healthcare industry, earlier machine learning techniques have been highly successful. The study focuses on predicting cardiac disease using historical data and knowledge. Much greater precision, correctness, and perfection are needed in the analysis and prognosis of cardiac-related issues because, if left undiagnosed, the condition can be fatal. We require a crucial prediction system to address such an issue. This study calculates and determines how accurately machine learning algorithms predict cardiac disease. A variety of machine learning algorithms, including Random Forest Classifier (RF), Neural Network (MLP), Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbour Classifier (KNN), are used to make this prediction. Training and testing of these algorithms are done using the heart dataset. In training, 80% of the dataset is used, and 20% of the dataset is used for testing. The metrics Accuracy, F1-Score, Recall, Precision, and ROC-curve are used for comparison. The results show that RF, MLP, MLP,RF, and RF have the highest accuracy (86.96), F1-score (86.79), recall (0.91), precision (0.82), and ROC-curve (0.93), respectively.

Keywords: *Heart Disease Prediction; Logistic Regression; Machine Learning; Heart Dataset; Performance Evaluation; SVM; KNN; Random Forest; Decision Tree; Logistic Regression; Naïve Bayes; Neural Network.*

1.0 Introduction

Cardiovascular Diseases (CVDs) are the major cause of death world wide. CVDs

[1]*Computer Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India*
**Corresponding author e-mail: himanshugbpuat@gmail.com*

are the collections of disorders which affects individuals' heart, arteries and blood vessels, which comprises of coronary heart disease, aortic disease, rheumatic heart disease and cerebrovascular disease and various conditions, in which there are mainly 4 out of 5 CVDs deaths are caused by strokes and heart attack [1]. According to WHO report, annually 17.9 million humans lose their lives universally, which comprises of 32 percent of global mortality. In America alone, it caused the death of 695,000 people in 2021 which increased to 697,000 deaths per year. In India, one-fifth of the deaths worldwide. India has 28,499 death records in 2021. As per the report by WHO, the deaths by heart attack are increasing annually [11].

Machine learning, a branch of AI research, is one of the most rapidly developing fields in data science. Machine learning algorithms are made so they can carry out a range of processes, including categorization, decision-making, and prediction and so on [9][10].

It can be very difficult to detect and treat heart diseases early because of the scarcity of diagnostic facilities, trained medical experts, and other resources that can affect the correct diagnosis of a cardiac disease, especially in nations with poor infrastructure and limited resources. A system to assist in the early detection of cardiovascular disease is now being developed through the use of machine learning as well as technological advances in computers to medical assistance software, in response to this concern.

The health organisations face a serious difficulty in providing patients with high-quality medical treatments that are affordable. Correct patient diagnosis and identification of appropriate treatments are both necessary for the provision of high-quality care, while avoiding incorrect diagnoses is prohibited. Additionally, early detection of CVD minimises costs and lowers the probability of death rate [6].

Early detection of any ailments related to the cardiac may mitigate the risk of fatalities. The branch of healthcare utilises a variety of ML (machine learning) approaches to figure out the patterns in the data and use them to generalise them. Typically, healthcare data have vast volumes and sophisticated architecture. Big data can be mined by ML algorithms to obtain the necessary information. [10].

We have calculated the accuracy and performed various performance metrics of seven various ML algorithms and based on the performance metrics we will conclude that which is the best ML model among them in this paper.

For this paper, Section 1 gives the introduction of the heart disease data and technology that diagnose the cardiac problems. Section 2 describes the contributions of the work done by researchers. Section 3 of this paper, covers the algorithms applied,

dataset description, proposed workflow of the work and data preprocessing methods. Section 4 is about the result of the research work. And the last, Section 5 tells about the future work and conclusion of the work.

2.0 Literature Survey

Human heart is a vital part of the body organ which helps in pumping and transporting the blood throughout the body. In this survey, the work of several researchers who have contributed in this field has been explored. Machine Learning has become a very effective and useful in diagnosing and treating the heart disease however the size of the dataset has become a serious issue.

KNN is more accurate as compare to the decision tree, support vector machine and linear regression, which gives 87% accuracy in their project [2].Another research have been engaged in the development and application of a security mechanism for safeguarding the records of medical data and predicting the outcome of cardiac disorders in patients using Naïve Bayes and the accuracy of a model is 89.77%. The research reveals that the AES algorithm provides high-security performance evaluation which is 98.2% compared to the PHEA which is 92.21 [3]. By using HRFLM ,model achieved 88.7% accuracy, which is combined form of random forest and linear model [4]. It proves that hybridization of the model gives more and better accuracy. Ali, M. M. et al. have researched on numerous ML and data mining algorithms which is taken from Kaggle and found that KNN, DT and RF have achieved 100% accuracy which other models performs low accuracy [6].

ML techniques are utilised for numerous types of predication of diseases and also, many researchers have worked on project like this analysis on "Heart diseases prediction using a decision support system" in which RF performs 86.6% accuracy after improving performance in feature scaling [7]. Another research used many supervised machine learning algorithms and predicted that RF has the highest accuracy among the machine learning models, which is 95.6% [8]. A Model utilized 4 techniques in which random forest and SVM give the best accuracy; however, the random forest gives the higher accuracy of 99%, whereas SVM gives 98% accuracy [10].Hybridised machine learning algorithm (DT and RF) is used, which has achieved a high accuracy of 88%, while other DT got 79% and RF got 81% [5].

3.0 Methodology

This section clarified the techniques used for the research in this study of the paper. The techniques that we employed to predict cardiac disease have been indicated.

3.1 Algorithms

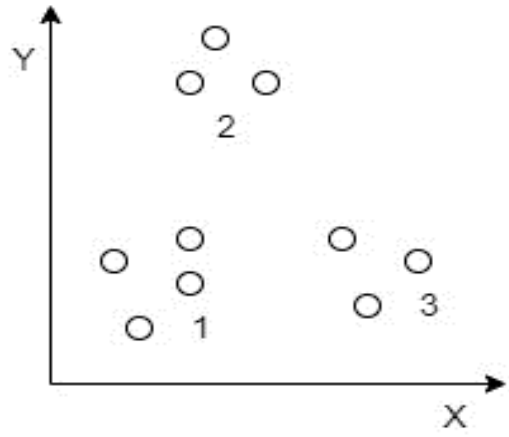
1. **Logistic Regression:** Binary and multiclass classification problems could be solved utilised supervised learning tool reffered to as logistic regression. The classification of categorical data is predicted using probability[22]. A logistic or sigmoid function can be utilized for combining input values linearly in order to predict the result. A mathematical function called the sigmoid function is used to transform predicted values into probabilities [12].

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad \ldots(1)$$

On the other hand, β_0 represents as bias and β_1 represents as the weight in this equation, which has been multiplied by 'X' times.

2. **K-Nearest Neighbour:** K Nearest Neighbor is a supervised machine learning technique which performs well for regression and classification issues. K represents the number of the closest neighbors that were used; it can be computed by simply using the maximum limit that the given value delivers. [13]. Every other set of data is referred to as a neighbor of each other, and the user establishes how many neighbours there are, which is crucial for dataset analysis.

Figure 1: KNN model where K = 3



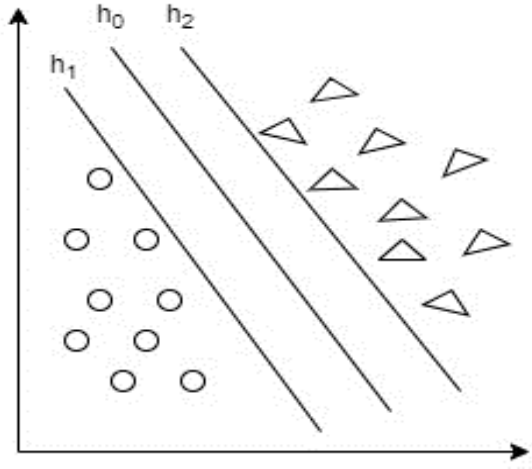
From Figure 1, we can see that there are 3 different neighbouring clusters which is represented in 2D space whose coordinates are (Xi, Yi) in which i = 1, 2, 3, 4, …

3. **Support Vector Machine:** An approach which is supervised to machine learning (using labelled data) is called Support Vector Machine (SVM). With SVM, the

hyper- plane with the broadest possible margin is made to keep identical data of one type at 1 side of the margin and distinct data on the other side, or to separate distinct data. [14].

Here, in Figure 2 we can see that h_0 is a decision boundary, h_1 and h_2 are negative hyperplane and positive hyperplane respectively.

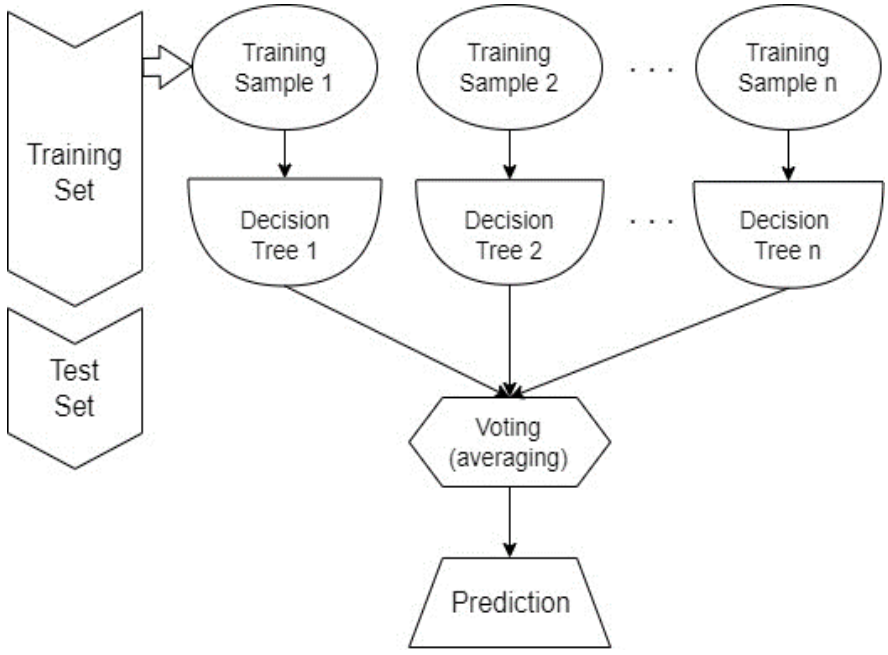
Figure 2: Decision Boundary Separating Two Classes in SVM



4. **Random Forest:** Random Forest is a renowned machine learning algorithm utilised by the supervised learning methodology. It is capable of helping overcome regression and classification-based machine learning problems. Its basis is an understanding of ensemble learning, and it involves the technique of combining a number of classifiers to improve the performance of the model and tackle a challenging. *[15]*.

In Figure 3, we have a decision tree flowchart like structure where each and every internal node is represented as each of the branches of the decision tree indicates an outcome based on the value of a specific attribute, and each leaf node reflects the ultimate prediction. RF employs a bagging technique in which every single subsets of the training sample are 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 outcome produced. Now, the voting outcome will be tested and after that it gives the prediction of the model.

Figure 3: Random Forest Classifier Model



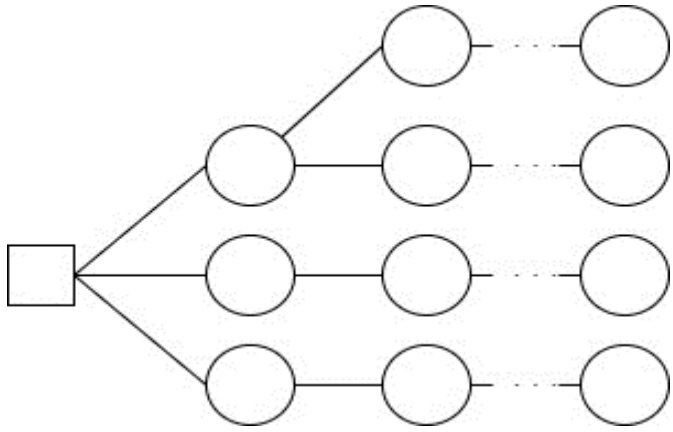
5. **Naïve Bayes:** The Naïve Bayes algorithm, which solves classification problems, is a supervised learning algorithm based on the Bayes theorem. *It is primarily utilized in* spam filtering and disease diagnosing. One of the most straightforward and efficient classification algorithms, it aids in the fast development of machine learning models with rapid prediction capabilities. [16].

$$P(X|Y) = \frac{P(Y|X)\,P(X)}{P(Y)} \quad \ldots(2)$$

6. **Decision Tree:** An effective supervised learning method for identifying and regressing data is a decision tree. Data is split up in a branch-like structure called a decision tree. The root node is divided into sub-branches or further branches according to rules and each attribute's maximal acquisition of information [17][26]. The leaf node will produce the outcome when this process repeats recursively applying the attributes.

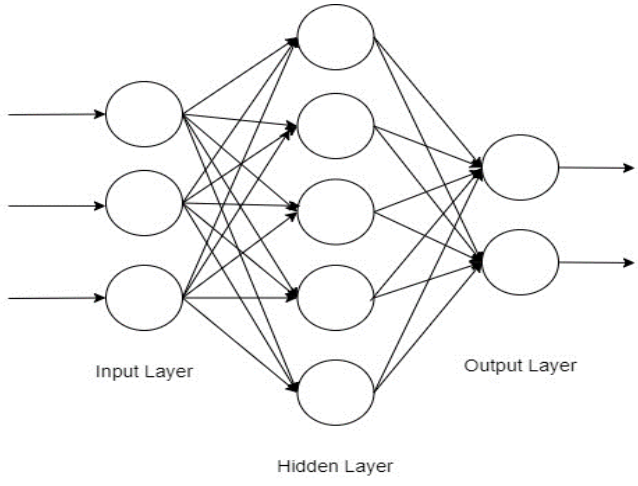
In Figure 4, this model can be visualized as a tree structure from which each and every path from the root node to the leaf node represents a decision rule. When each node splits into a new node, the new decision rules are created for every new node. Leaf node represents a final predicted output.

Figure 4: Decision Tree Model



7. **Multi-layer perceptron:** A supervised learning technique called multi-layer perceptron employs a component of neural networks (ANNs) which consist of over three layers (with the input and output layers included) and several backpropagating hidden layers as opposed to just one hidden layer. MLP is the feed-forward network, meaning that the connections do not form cycles. It functions by acquiring input from other perceptron, establishing a weight to each node, and then moving on to the hidden layer. Additionally, the output layer receives the hidden output. The error value is computed using the determined expected value and the actual output. [18].

Figure 5: Structure of Multilayer Perceptron



In Figure 5, every data is passed from input layer as every node is equal to the equal no. of features. Then, the data is passed through the one or more hidden layer which gives the non-linearity to the model which is now passed through the output layer.

3.2 Dataset description

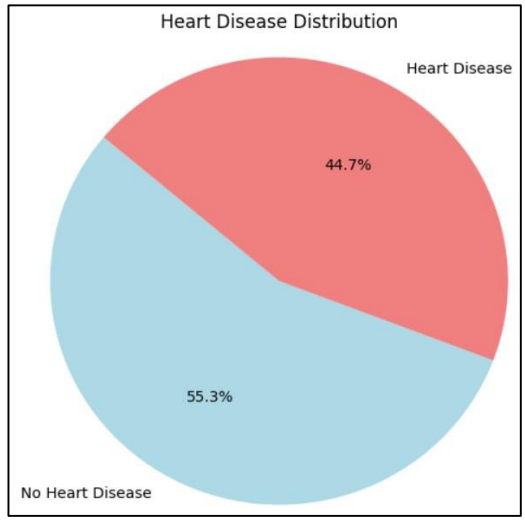
The data used in this research's "heart dataset" was obtained from Kaggle Competitions [21][23]. The availability of this dataset has made it possible for us to work towards developing a more accurate and optimistic model. We will go into detail about the data preparation, model building, assessment measures, and outcomes in the parts that follow in order towards achieving the goal of building more accurate machine learning model. The dataset has been collected from Kaggle website which has 918 observations and 12 attributes has been created by using five distinct datasets which are Cleaveland, Switzerland, Hungarian, Long Beach VA and Stalog (Heart) Data Set. This dataset has also 272 duplicated observations.

Attributes in a dataset represents:

- **Age:** An important feature of the information. It has long been believed that older individuals are more susceptible to cardiac disease.
- **Sex:** Women have the less risk of getting a heart attack in comparison to men.
  0 – Female
  1 – Male
- **Chest pain type:** This feature tells us about the people having which type of chest pain. Four different kind of chest pain are taken here:
  0 – ASY (Asymptomatic)
  1 – ATA (Atypical Angina)
  2 – NAP (Non-Anginal Pain)
  3 – TA (Typical Angina)
- **Resting BP:** It is an essential feature in a data as it can predict that if a person having a high blood pressure is an ideal condition for having a cardiac disease. It can be measured in the units of mm Hg.
- **Cholesterol:** This feature is considered significant as high cholesterol can lead to narrow or nearly blockage of arteries which results in fatalities. This can be measured in mm/dl.
- **Fasting BS:** An Individual having high blood can get heart disease which makes this feature a significant factor.
  1 – if Fasting BS > 120 mg/dl,
  0 – otherwise

- **Resting ECG:** It illustrates electrocardiographic results which has been expressed as:
  0 – Normal: Normal,
  1 – ST: having ST-T wave abnormality,
  2 – LVH: Left Ventricular Hypertrophy
- **Max HR:** It tells us about an individual who achieved the max. heart rate. Numeric value has been taken between 60 and 202.
- **Exercise Angina:** This attribute tells us about the pain induced on the body chest, arms, etc. Values have been taken as:
  1 –Y: Yes,
  0 – N: No
- **Old Peak:** Numerical value measures in depression is ST.
- **ST slope:** It is the peak exercise ST segment slope. Values have been taken as:
  0 – Down (down sloping)
  1 – Flat (flat)
  2 – Up (up sloping)
- **Heart Disease:** This is target class value which represents if someone is experiencing heart disease or not.
  1 – Heart disease
  0 – No Heart disease (Normal)

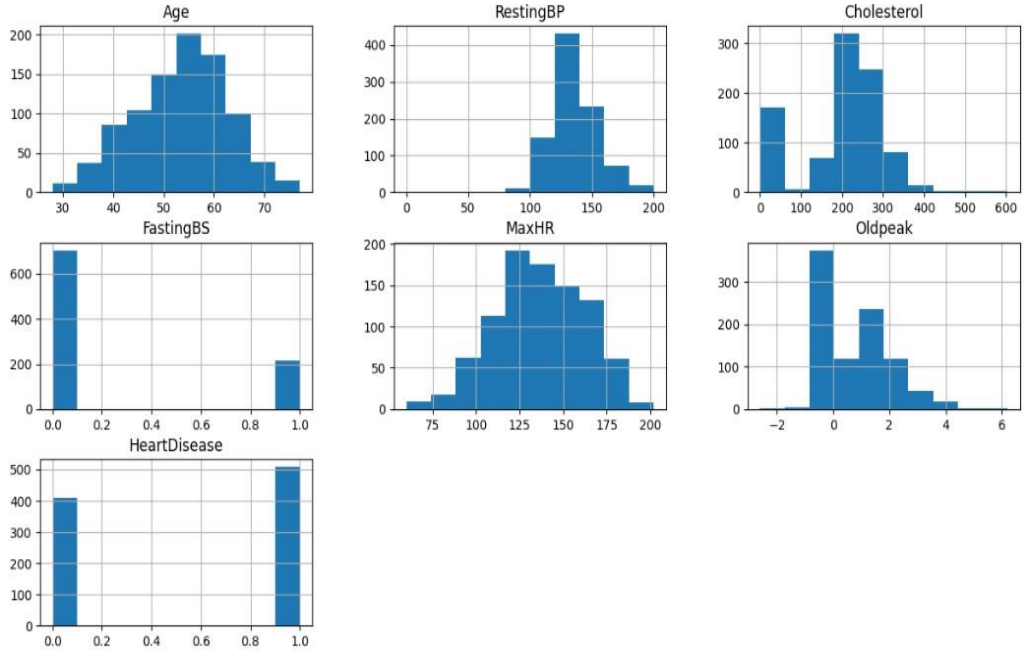
Figure 6: Data Visualization in Heart Disease in Heart Dataset



In Figure 6 data balancing is a necessary thing in machine learning so that it doesn't affect accuracy and give bias data. From the dataset we chose, we observed that 44.7% people suffered from heart disease and 55.3% people have not suffered from heart disease. In Figure 7 histogram of attributes show the graphical representation of the features of the dataset which includes age, resting bp, cholesterol, fasting bs, max hr, old peak and target class heart disease.

In Age, it is graphical representation of no. of people (y-axis) and age of people (x-axis). It showed us that people whose age is between 40 and 70 suffered the most heart disease problems. In FastingBS, 1 indicates that no. of people having high heart rate and 0 indicates that people having normal heart rate. In Heart Disease, 1 refers that people have heart disease and 0 refers that no heart disease.

Figure 7: Histogram of Attributes

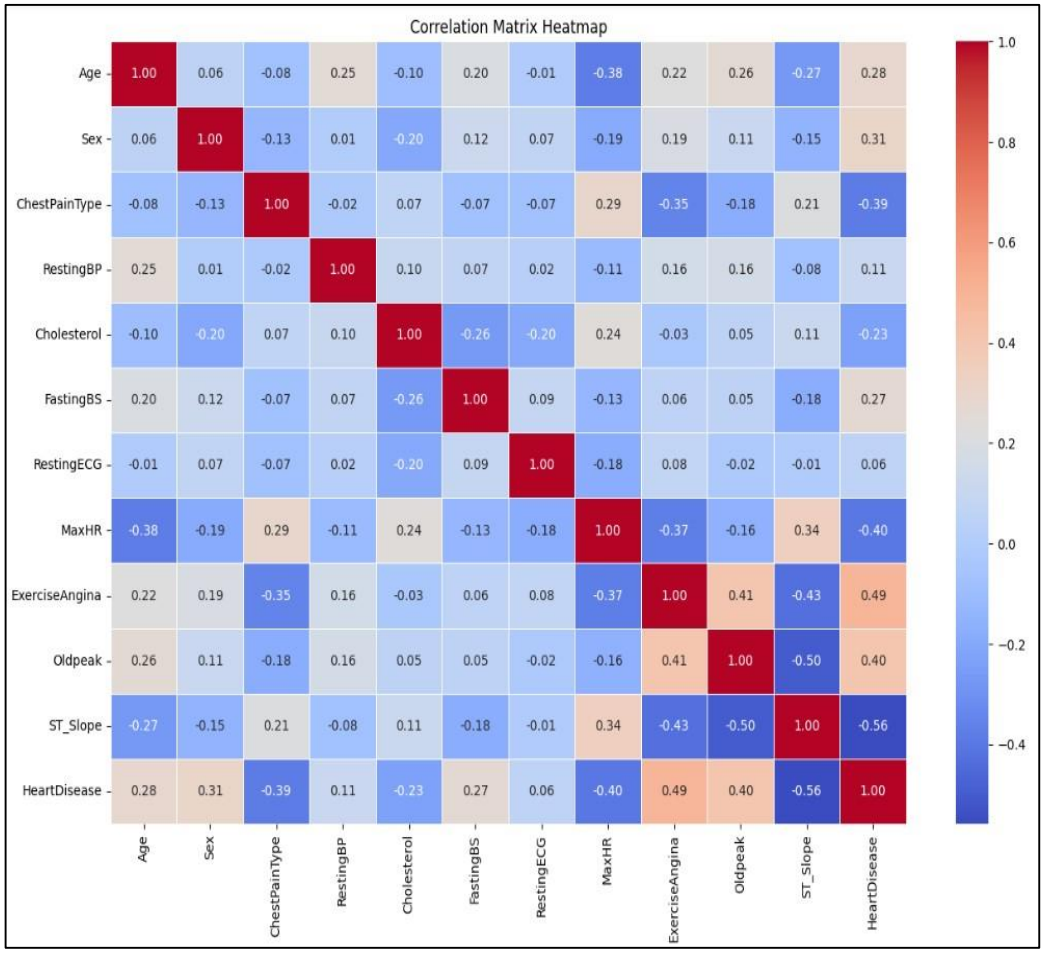


3.3 Analysis of feature correlation

A technique that can be used to comprehend the fundamental connections between various data features within a dataset are known as feature correlations. Feature correlation can be useful in two applications: one is to determine how each feature

influences the output feature and the other is to determine the interdependencies between the data features [19]. We were able to ascertain the correlation values between the data features by calculating the correlation coefficients of the feature matrix M of size p × q, which may be represented as $M = [v1, v2, \ldots, vq]$, where $v1, v2, \ldots, vq$ are the vectors representing q number of characteristics. One full medical procedure at a specific time for each vector defines the vector's length, which is denoted by p. The computed correlation values between the target disease and different medical variables for each dataset are shown in Figure 8.

Figure 8: Dataset Features Heatmap



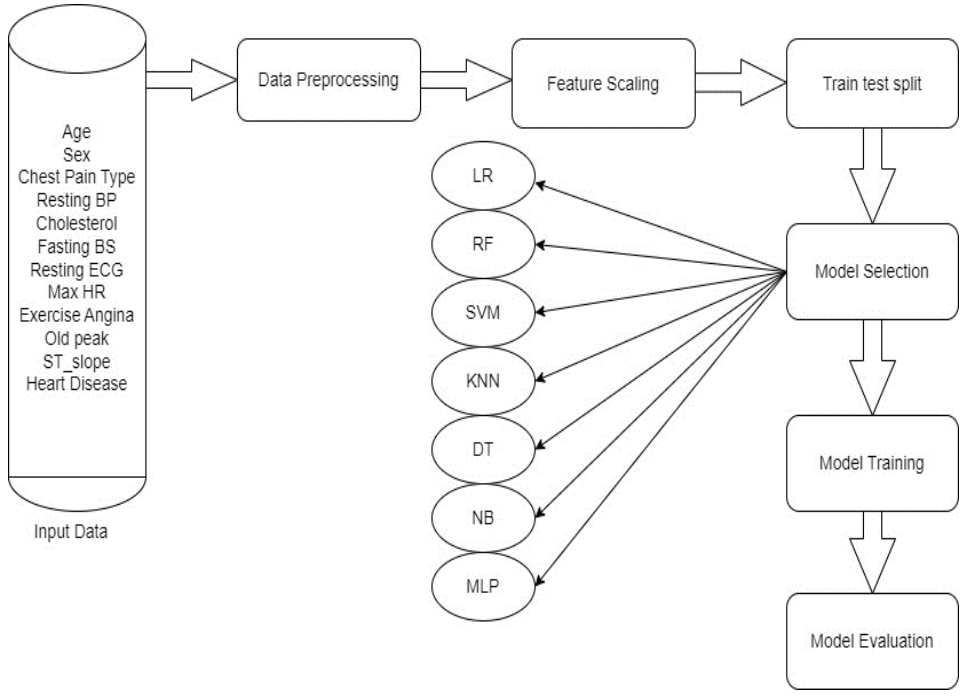
From Figure 8, we can observe that p and q are no. of rows and columns where p is equal to q which indicates they both have the same no. of rows and columns. p and q having 12 no. of features.

Our objective is to determine the correlation between multiple variables and thereafter structure the information we've gathered. In this figure, a matrix data structure is utilized to store the information. This figure provides us with much detail on a feature-by-feature basis. First of all, with correlations of 0.49, 0.40, and 0.31, respectively, {ExerciseAngina, Oldpeak, and Sex} are the three features exhibiting the greatest class feature dependence. The second fact represents the feature-feature correlation with correlation values of 0.49, 0.41, 0.40, and 0.34 that are displayed in HeartDisease – ExerciseAngina, Oldpeak –ExerciseAngina, HeartDisease –Oldpeak, and ST_Slope – MaxHR, respectively. Furthermore, Age, Sex, ChestPainType, RestingBP, Cholesterol, RestingBS and RestingECG have the lowest correlation with the target.

3.4 Proposed workflow

We discussed the model we applied on for our experiment in this section. Figure 9 below illustrates the model's architecture.

Figure 9: Model Architecture



For further discussion, we have divided the section into 3 more sections under this section. In Section 1, we have explained the architectural structure of the Figure 9 step by step. In Section 2, we have explained the process of data preprocessing techniques that we have applied in our experiment. And furthermore, in the last section, we have discussed about the five different performance metrics that will evaluate the performance of the model.

3.4.1 Working of the architectural model

- Initially, we will be gathering dataset from the Kaggle and perform Exploratory Data Analysis (EDA) to learn about the dataset and then refine it using Data Preprocessing techniques like removing duplicates, data normalization, detecting outliers, etc.
- Then, we will be applying feature scaling method which provide a fixed range for every independent feature that are available in the data. It is done during the data pre-processing which manages the drastically varied magnitudes, units and values i.e., between 0,1 and -1. Converting every value to decimal form.
- Now, we will perform train test split in which 80% data will be in training and 20% data will be in testing.
- Then, we will be using model selection technique in which we will be selecting appropriate model like SVM, KNN, DT, RF, LR, NB and MLP.
- The dataset is subjected to distinct implementations of several machine learning techniques, which are then combined for data assessment.
- Now, we will be importing important libraries and models will be trained.
- Using performance measures to conduct the evaluation be the final step of the trained model by using ROC curve, F1 score, recall, precision, accuracy and confusion matrix.
- Afterwards, results are analysed and one of the best algorithms in performance is found.

3.4.2 Data preprocessing

- **Data examination and quality assessment:** In the preprocessing phase, the integrity and purity of the dataset are first assessed. It reveals there are no null values in the dataset. Analysing our data closely to gain a sense of its overall quality, applicability to our project, and consistency. Missing values, outliers, mismatched data types, mismatched data values, and other characteristics are examples of factors to watch out for.

- **Data Cleansing:** The technique for filling in missing data, fixing, deleting, or fixing inaccurate or redundant information from a dataset, as well as outlier identification.
- **Feature Scaling:** Feature Scaling is a technique to provide a fixed range for every independent feature that are available in the data. It is done during the data pre-processing which manages the drastically varied magnitudes, units and values i.e., between 1 and -1.

3.4.3 Evaluation metrics

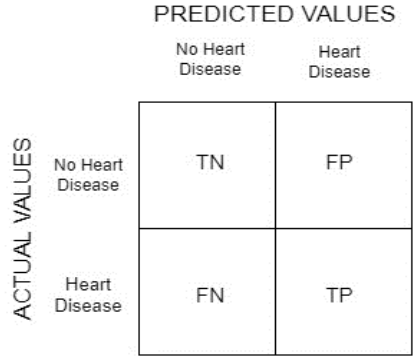
A particular classification algorithm's performance can be evaluated using a variety of metrics. Consequently, the kind of problem we are going to tackle will impact which parameters are best. Certain scenarios, like this one, may call for accuracy to be the best option, while others may call for recall or precision. We could select our classifier using recall (sensitivity) as a performance parameter because we are working with medical cases.

We have assessed the performance of ML classification models using five frequently employed performance evaluation metrics: accuracy, F1-score, ROC, recall, and precision. [20][24][25].

Table 1: 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

Figure 10: Skeleton of Confusion Matrix



- *Confusion matrix:* An expected table structure that makes it possible to observe how well the algorithm for supervised learning is working is called a confusion matrix. In the N x N matrix, each row represents an instance that occurs in an actual class, and each column shows occurrences in a predicted class. A representation of a confusion matrix for a binary classification—from which another terminology or metric could possibly be derived—is shown in the Table 1. The following discusses a few of the metrics.
- *Accuracy:* It can be expressed as the proportion of accurate predictions which is divided by the total no. of predictions.

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \ldots(3)$$

- *Recall:* It determines the rate of actual positives, or the fraction of actual positive cases that are found and divided by the no. of cases that the model estimates to be positive.

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \quad \ldots(4)$$

- *Precision:* The number of positive predictions that turned out to be actually accurate is determined by the precision.

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \quad \ldots(5)$$

- *F1-Score:* It is represented as a harmonic average of the model's recall and precision, integrating both of these parameters.

$$F1 - Score = 2\frac{\text{recall}*\text{precision}}{\text{recall+precision}} \quad \ldots(6)$$

- *ROC curves:* The receiver operating characteristic curve is a graphical diagram that shows the true positive rate and false positive rate as a function of a binary classification algorithm's performance.

4.0 Result

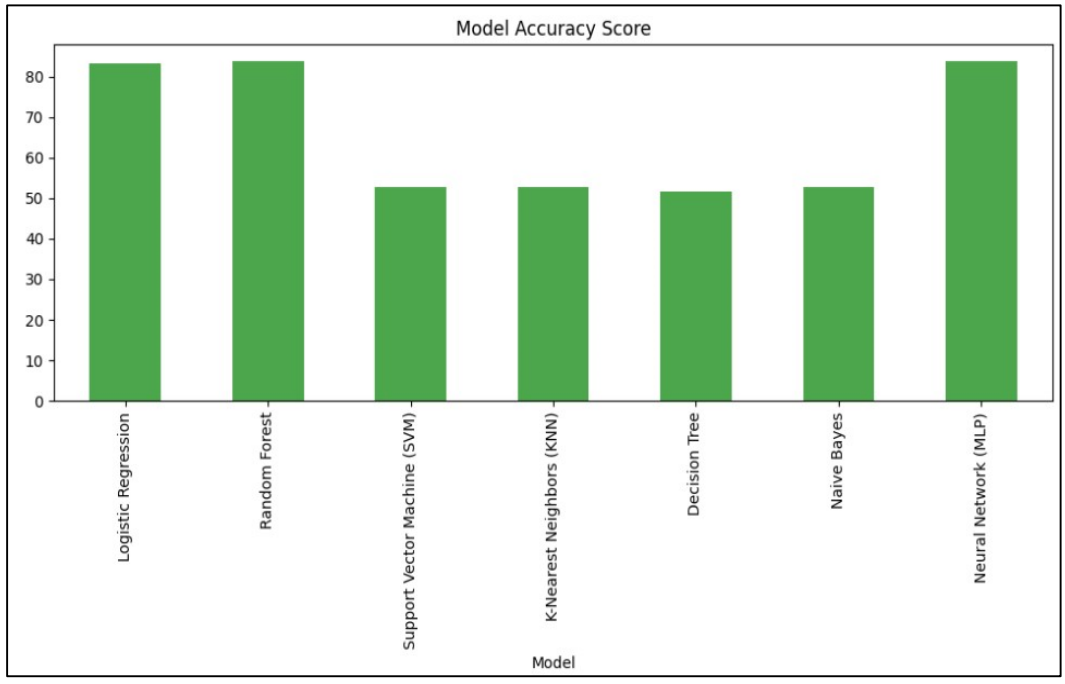
We discover that the random forest algorithm's accuracy becomes substantially more effective than existing algorithms following the training and testing of the machine learning approach. When measuring accuracy, one should consider the confusion matrix, F1 score, precision, and recall of each algorithm. After measuring the outcomes, it is determined that random forest has the highest accuracy of all of them, at 86.96%. Table 2 displays the comparison. Unfortunately, we were unable to come across any additional algorithms that worked well with our data.

Table 2: Comparative Performance of Heart Disease Prediction Models

Model	Model Accuracy (%)	F1 Score (%)	Recall	Precision	ROC curve
RF	86.96	86.57	0.88	0.82	0.93
MLP	84.78	86.79	0.91	0.72	0.91
LR	82.61	84.00	0.85	0.77	0.91
NB	63.59	71.97	1.00	0.53	0.88
SVM	59.24	69.04	1.00	0.53	0.92
KNN	52.72	69.04	1.00	0.53	0.50
DT	47.28	71.70	0.87	0.51	0.47

In Figure 11 we can observe that RF model performs the best model accuracy among several ML models which gives us the accuracy of 86.96%.

Figure 11: Comparison of Algorithm in Terms of Model Accuracy



In Figure 12 Neural Network (MLP) model got the highest performance metrics in F1 score which is 86.79%.

Figure 12: Comparison of Algorithm in Terms of Performance Metrics (F1 Score)

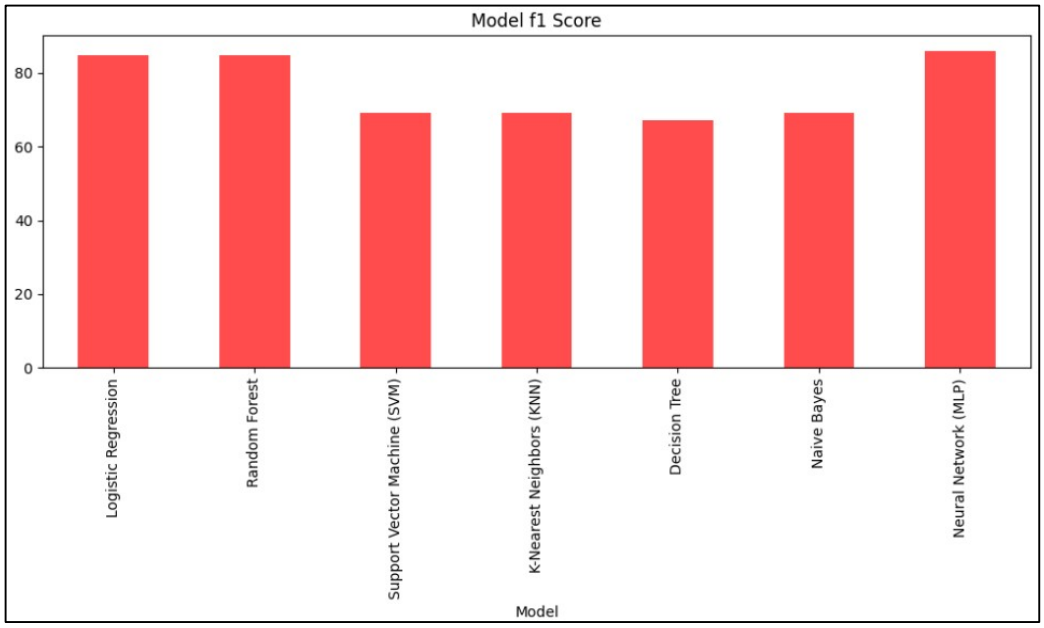
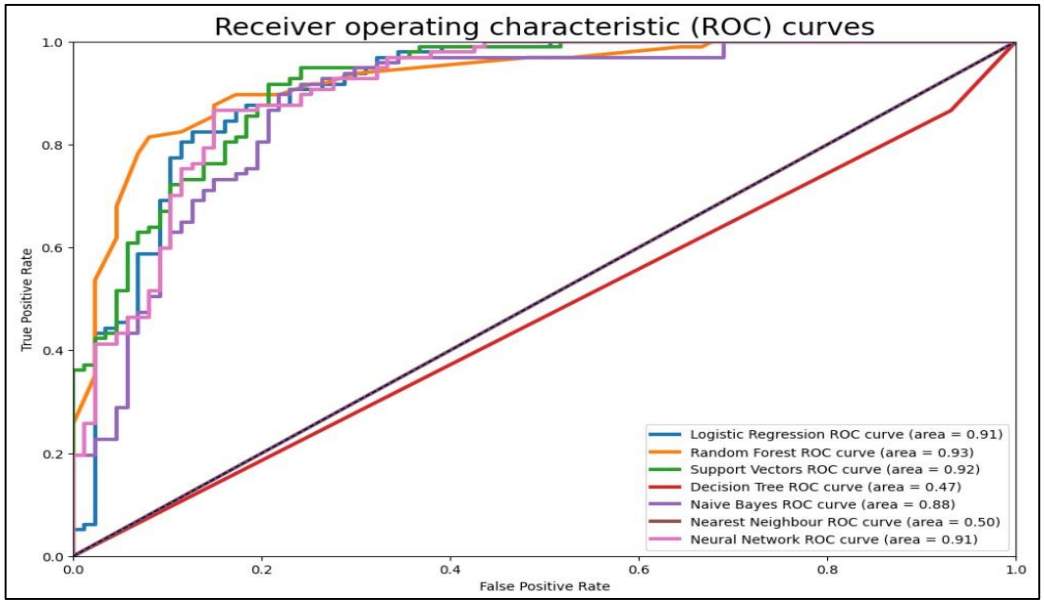


Figure 13: Comparison in Terms of Performance Metrics in Receiver Operating Characteristic (ROC) Curves



In Figure 13 each and every algorithm have performed well in ROC curve except Decision Tree which has fall below the threshold which has the value of 0.47. Figure 14 gives the highest accuracy of 86.96% in RF. From this figure we obtained 88 as true positive,72 as true negative, 8 as false negative and 15 as false positive.

Figure 14: Confusion Matrix for Random Forest

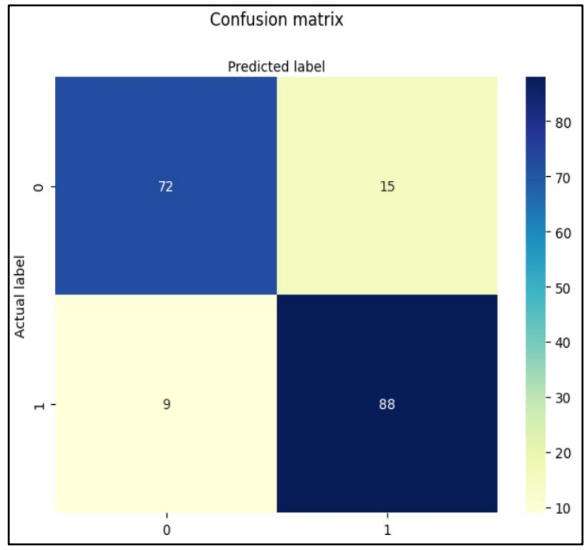


Figure 15: Confusion Matrix for Neural Network (MLP)

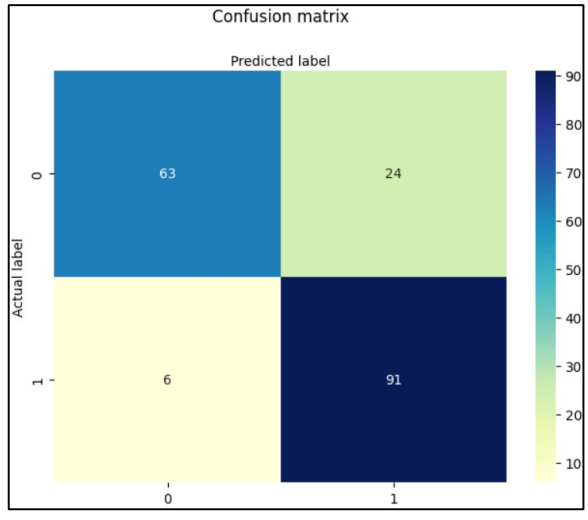


Figure 15 illustrates the highest F1 score in heart disease prediction for the Neural Network (MLP). From this figure we obtained 91 as true positive, 63 as true negative, 6 as false negative and 24 as false positive. The graph in Figure 16 and Figure 17 illustrates the model accuracy and model loss respectively.

Figure 16: Graph of MLP for Model Accuracy

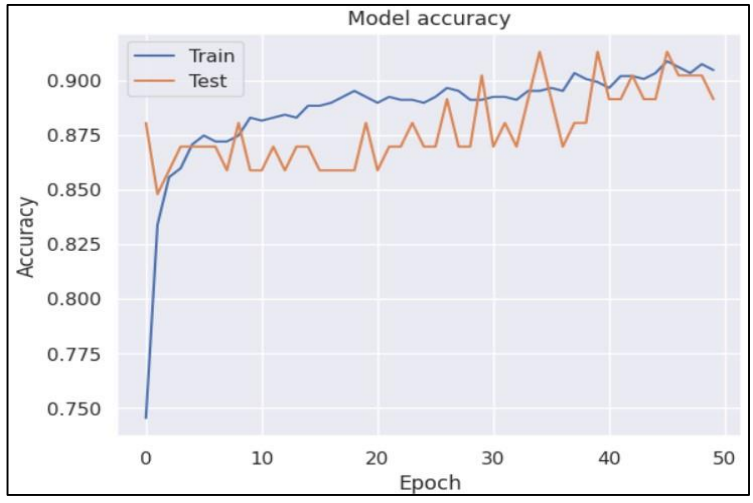
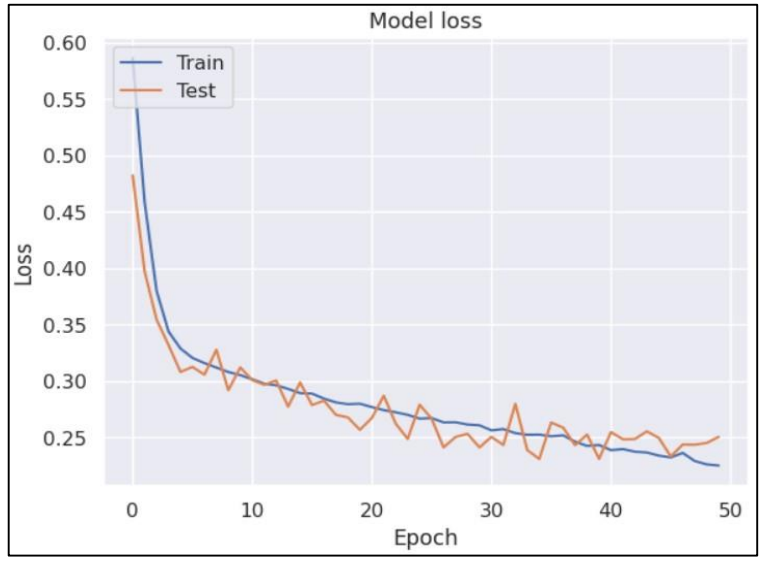


Figure 17: Graph of MLP for Model Loss



5.0 Conclusion

Combining the findings of this study with those of others has helped bring to light the fact that the dangers of ignoring the possibility of getting heart disease are real. A more complete picture of the many risk factors for cardiovascular disease is painted. This study includes a comparison of various machine learning methods that use the cardiac dataset to forecast the occurrence of heart disease. After searching for an algorithm that could help improve the accuracy and quality of cardiovascular disease outcome predictions, RF was determined to have the greatest performance on the dataset that was used. The Random Forest Classifier outperformed the others with ROC-Curve, Precision, and Accuracy values of 0.93, 0.82, and 86.96 correspondingly. In order to get better results in future work, learning techniques like ensemble learning and deep learning are currently being used across all industries. To further enhance Random Forest's outputs, we want to incorporate more machine learning algorithms that leverage deep learning and ensemble learning in the near future.

References

[1] Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics, 2*, 100060.

[2] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. *In 2020 International Conference on Electrical and Electronics Engineering (ICE3)* (pp. 452- 457). IEEE.

[3] Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using naives Bayesian. *In 2019 3rd International conference on trends in electronics and informatics (ICOEI)* (pp. 292-297). IEEE.

[4] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access, 7*, 81542-81554.

[5] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. *In 2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333). IEEE.

[6] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine, 136*, 104672.

[7] Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments, 7*(3), 263-275.

[8] Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: a comparative study and analysis. *Health and Technology, 11*, 87-97.

[9] El-Hasnony, I. M., Elzeki, O. M., Alshehri, A., & Salem, H. (2022). Multi-label active learning-based machine learning model for heart disease prediction. *Sensors, 22*(3), 1184.

[10] Sharma, V., Yadav, S., & Gupta, M. (2020, December). Heart disease prediction using machine learning techniques. *In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 177-181). IEEE.

[11] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[12] Rani, P., Kumar, R., Jain, A. & Lamba, R. (2020) Taxonomy of machine learning algorithms and its applications. *Journal of Computational and Theroretical Nanoscience, 17*(6):2509–2514.

[13] Luo, X., Lin, F., Chen, Y., Zhu, S., Xu, Z., Huo, Z., ... & Peng, J. (2019). Coupling logistic model tree and random subspace to predict the landslide susceptibility areas with considering the uncertainty of environmental features. *Scientific Reports, 9*(1), 15369.

[14] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–97.

[15] Jabbar, M. A., Deekshatulu, B. L. & Chandra, P. (2016). Prediction of heart disease using random forest and feature subset selection. *In Innovations in Bio-Inspired Computing and Applications*, pp 187–196. Springer: Cham.

[16] Dulhare, U. N. (2018). Prediction system for heart disease using Naive Bayes and particle swarm optimization. *Biomedical Research, 29*(12), 2646–2649.

[17] Karthiga, A. S., Mary, M. S., & Yogasini, M. (2017). Early prediction of heart disease using decision tree algorithm. *International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST), 3*(3), 1-17.

[18] Dangare, C. & Apte, S. (2012). A data mining approach for prediction of heart disease using neural networks. *International Journal of Computer Engineering and Technology (IJCET), 3*(3), 30-40.

[19] Gopika, N., & ME, A. M. K. (2018, October). Correlation based feature selection algorithm for machine learning. *In 2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, (pp. 692-695). IEEE.

[20] Dev, S., Savoy, F. M., Lee, Y. H., & Winkler, S. (2017, September). Nighttime sky/cloud image segmentation. *In 2017 IEEE International Conference on Image Processing (ICIP)* (pp. 345-349). IEEE.

[21] Fedesoriano (September 2021). Heart failure prediction dataset. Retrieved from https://www.kaggle.com/fedesoriano/heart-failure-prediction

[22] Sharma, H., Kumar, P., & Sharma, K. (2023, February). Identification of device type using transformers in heterogeneous internet of things traffic. *In International Conference on Innovative Computing and Communication* (pp. 471-481). Singapore: Springer Nature Singapore.

[23] Srivastava, A. & Ahmad, P. (2016). A probabilistic gossip-based secure protocol for unstructured P2P networks. *Procedia Computer Science, 78*, 595-602. Retrieved from 10.1016/j.procs.2016.02.122.

[24] Dubey, R., Bharadwaj, S., Zafar, I. & Biswas, S. (2021). GIS mapping of short-term noisy event of diwali night in Lucknow city. *ISPRS International Journal of Geo-information, 11*(1), 25.

[25] Srivastava, A., Umrao, S., Biswas, S. & Dubey, R. (2023). FCCC: Forest cover change calculator user interface for identifying fire incidents in forest region using satellite data. *International Journal of Advanced Computer Science and Applications, 14*(7), 948–959.

[26] Srivastava, A., Shruti, B., Dubey, R. & Sharma, V. B. (2022). Mapping vegetation and measuring the performance of machine learning algorithm in lulc classification in the large area using Sentinel-2 and Landsat-8 datasets of dehradun as a test case. Retrieved from DOI: 10.5194/isprs-archives-XLIII-B3-2022-529-2022.