**Abstract**

This study focuses on comparing the performance of various machine learning algorithms in the context of heart disease diagnosis. The goal is to evaluate different algorithms using specific performance measures and pre-processed data for test prediction. K-Nearest Neighbors (K-NN), Random Forest (RF), and Artificial Neural Network (ANN) models were identified as the most promising for the dataset used in this research.

**Introduction**

Heart failure is a significant cardiovascular condition that necessitates early detection and accurate prediction to improve patient outcomes. Machine learning algorithms have emerged as powerful tools for analyzing medical data and predicting clinical outcomes. This research paper aims to explore the application of different machine learning algorithms for heart failure prediction and compare their performance. By leveraging comprehensive datasets and variables such as demographics, medical history, and biomarkers, these algorithms can identify individuals at high risk of developing heart failure.

Various machine learning algorithms, including decision trees, random forests, support vector machines, and neural networks, have been employed in heart failure prediction studies. Each algorithm possesses distinct strengths and limitations, which must be carefully evaluated. By comparing their performance, this research aims to identify the most effective approach for early detection and risk stratification in heart failure.

Through the use of a comprehensive heart failure dataset, this study applies multiple machine learning algorithms to predict heart failure occurrence. Performance evaluation metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) will be employed. The findings will contribute to understanding the comparative effectiveness of different machine learning algorithms in heart failure prediction and advance personalized and preventive approaches in cardiovascular healthcare.

**Literature review**

The review highlights the significance of using machine learning algorithms for heart failure prediction and summarizes various studies conducted in this area. These studies aim to improve patient outcomes and guide clinical interventions by focusing on early detection and accurate risk stratification.

Different machine learning algorithms are discussed, each with its own strengths and limitations. Decision trees are popular in medical prediction tasks due to their interpretability and ability to partition data based on different variables. Random forests, which are ensemble techniques based on decision trees, can capture complex interactions and improve prediction accuracy. Support vector machines (SVMs) aim to find optimal decision boundaries in high-dimensional feature spaces, while neural networks leverage deep learning architectures to learn intricate patterns in the data.

The review then presents summarized findings from several studies:

1. The HRFLM (Hybrid Random Forest and Linear Method) approach is proposed as an effective method for heart disease prediction. By combining the characteristics of random forest and linear methods, this approach improves the accuracy of predictions.

2. Logistic Regression, Random Forest Classifier, and K-Nearest Neighbors (KNN) achieved an accuracy of 87.5% in heart failure prediction. This suggests that utilizing machine learning techniques in medical databases can lead to more accurate predictions and reduce costs and time associated with manual analysis.

3. The review discusses the limitations of decision trees in small datasets, as they may not always provide accurate results. In contrast, Naïve Bayes is considered more accurate, even with clean and well-maintained input data. It is suggested that combining different algorithms, such as Naïve Bayes and K-means, can lead to greater prediction accuracy by treating variables as individuals.

4. Machine learning and image fusion techniques are highlighted as valuable tools in the medical field for heart disease prediction. The review emphasizes the importance of obtaining reliable datasets with appropriate samples and reliable data to build accurate predictive models. Pre-processing the dataset is also crucial to improve results by preparing the data for use by the machine learning algorithm.

5. The review mentions the positive effects of Artificial Neural Networks (ANN) on heart disease prediction in most models. The utilization of machine learning and image fusion techniques in detecting heart disease is considered essential for both healthcare authorities and patients. However, limited reporting in this area is noted due to challenges associated with accessing patient data. Encouraging hospitals to publish high-quality datasets while ensuring patient privacy would contribute to the improvement of machine learning models in this field.

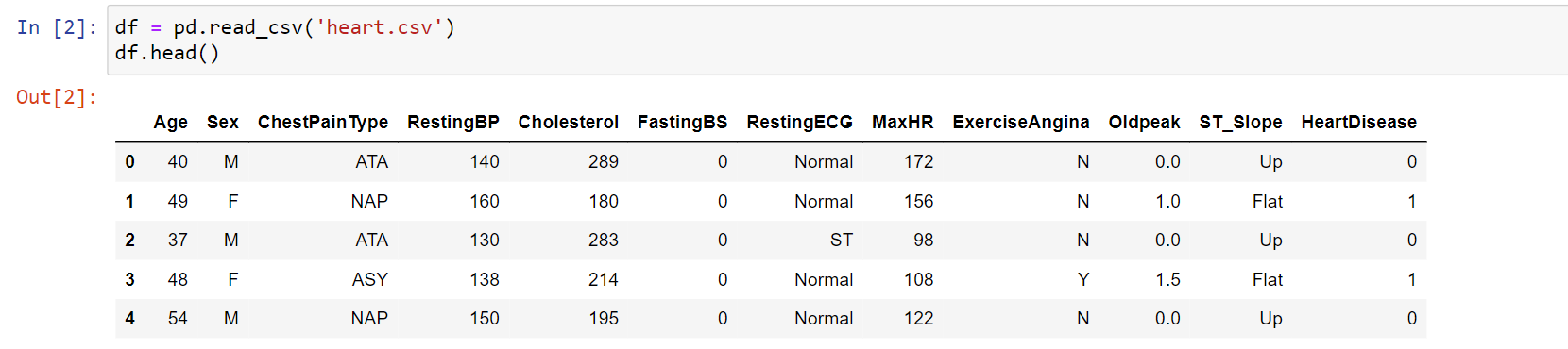
The review concludes by acknowledging the limitations and challenges associated with the application of machine learning algorithms in heart failure prediction. The choice of algorithm depends on various factors, including dataset characteristics, interpretability requirements, and computational resources. Rigorous evaluation and validation are necessary to assess the performance of the algorithms across different patient populations and datasets. Further research and advancements are needed to determine the generalizability, interpretability, and integration of these algorithms into clinical practice for effective heart failure prediction.

**Existing System**

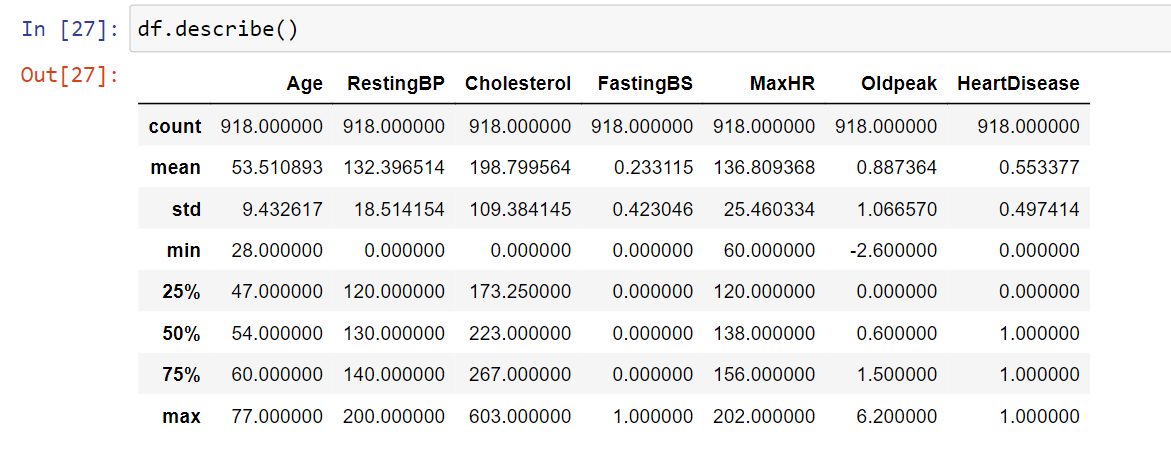
Existing systems for heart failure prediction using machine learning include various machine learning algorithms such as logistic regression, support vector machines, random forests, decision trees, and deep learning. These algorithms learn from historical data to identify patterns associated with heart failure. Clinical decision support systems (CDSS) integrate machine learning algorithms with electronic health records (EHRs) to assist doctors in making better decisions about patient care. Additionally, there are web-based tools available that utilize patient-reported information to predict the risk of heart failure. These systems and tools leverage data and algorithms to provide valuable insights and assist in predicting heart failure.

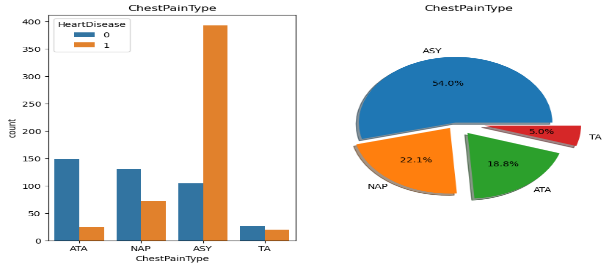
**Dataset Description**

• The dataset contains 11 features that can be used to predict a possible heart disease

• It contains 6 categorical features (Sex, ChestPainType, RestingECG, ST\_Slope, FastingBS, ExerciseAngina) and 5 continuous features (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) along with a label column named as HeartDisease which is to be classified 

• The description of the 918 rows and 12 columns is as follows



**Dataset Analysis**

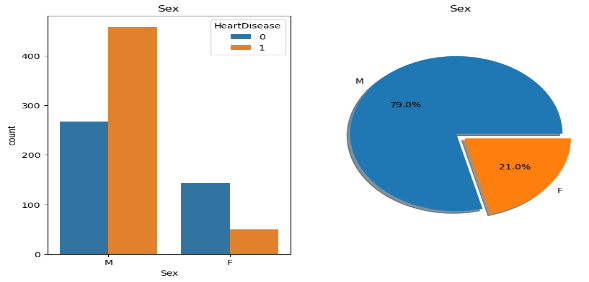
****

Figure 2: Chest Pain

Figure 1: Sex

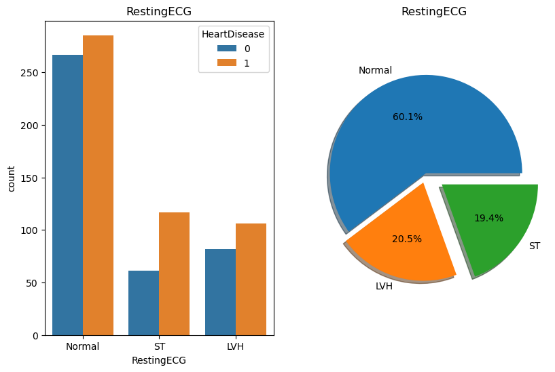
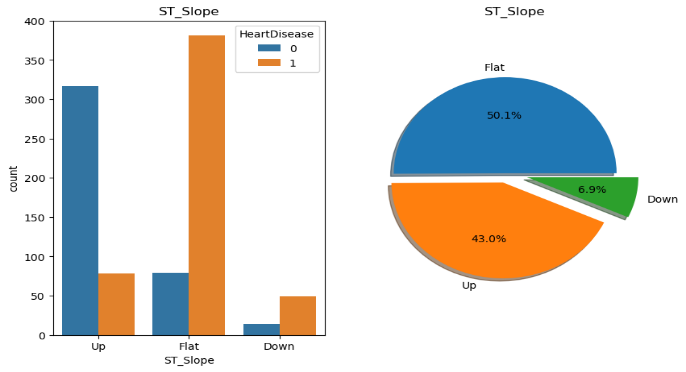
 

Figure 3: RestingECG Figure 4: ST slope

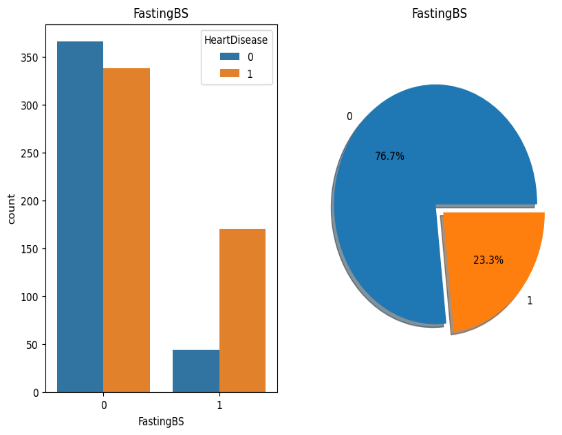
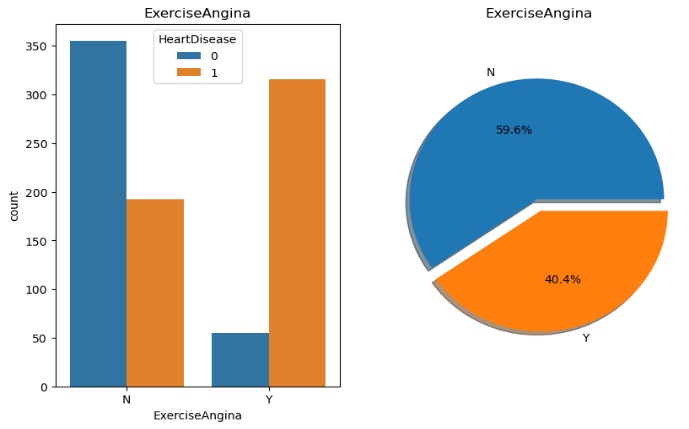
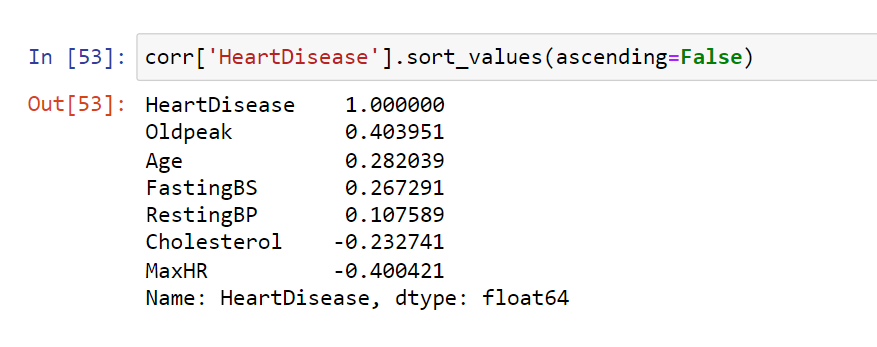
**** ****

Figure 5: fastingBS Figure 6: ExerciseAngina

• From the plots of Sex column, we can see that the values female, have more chances of no heart disease, whereas the values male, have a higher chance of suffering from heart disease.

• Similarly, ChestPainType ASY has a high chance of being suffering with heart disease, and ST\_slope with value flat has higher chance of being suffering with a heart disease.

• Hence, we can say that these features may have a high importance while doing the classification



Firstly, we found the correlations of the features with respect to target column and the results are as follows – We can conclude that the features ExerciseAngina, Oldpeak and Sex are the most correlated among all the features. After this, we encoded all the categorical columns with the help of LabelEncoder()

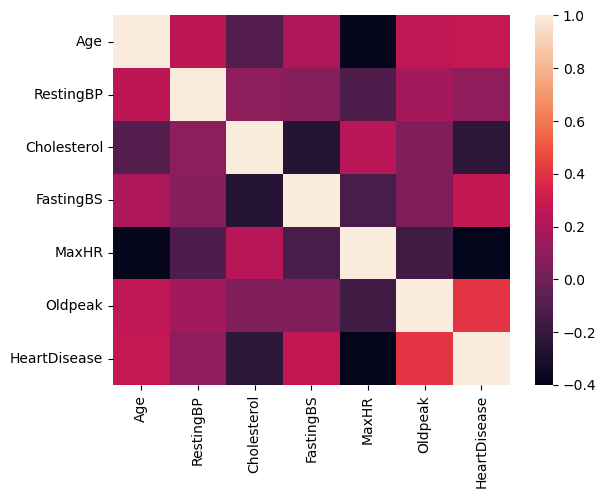
****

Figure 7: Heat Map

Then we applied PCA to figure out the principal components that can store up to 99.9 percent of variance, and the following was the result –

• The data got reduced to 4 features

• 1 feature showed up to 90 percent variance, 2 features showed up to 95 percent variance, 3 features showed up to 99 percent variance

• The accuracy score was about 71 % on the Random Forest classifier with precision scores also about 70% The plot clearly shows that the data is not linearly separable, neither it gives any information so that we can apply the KMeans clustering and the chances are high that clustering also may not give appropriate results.

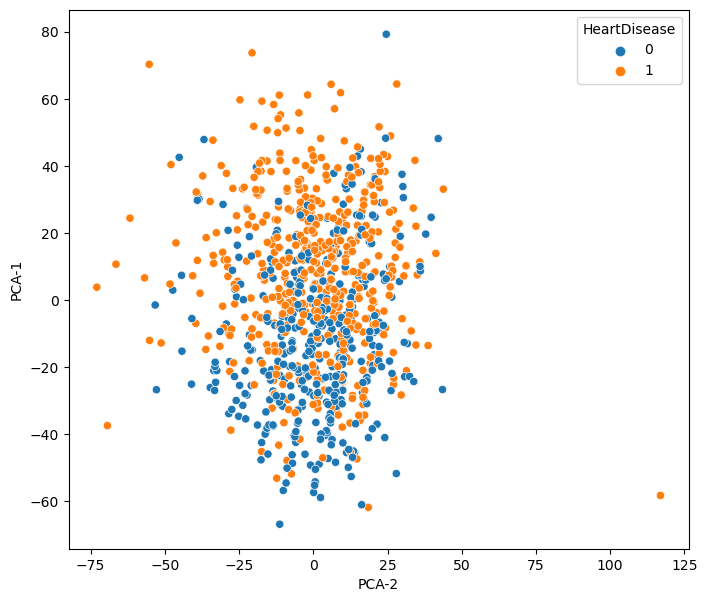
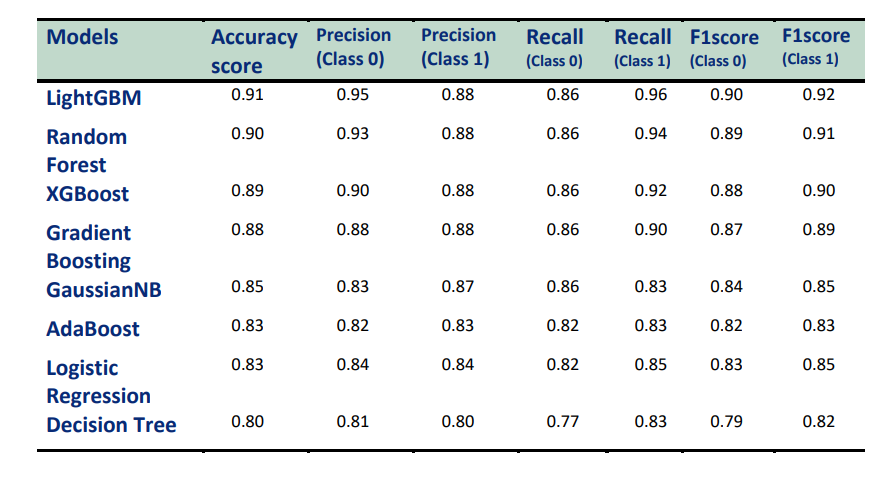


Figure 8: PCA

**Methodology**

• We applied the following models on the training dataset and then tested the scores on validation set –

****

Considering the above table, we have decided the best 5 models based on the accuracy score and F1 score as LightGBM, Random Forest, XGBoost, Gradient Boosting and Gaussian Naïve Bayes.

• The feature importance from one of the models is shown below and as we have predicted from the data visualization that ST\_slope, Sex and ChestPainType will play a major role in predicting the heart failure chances –

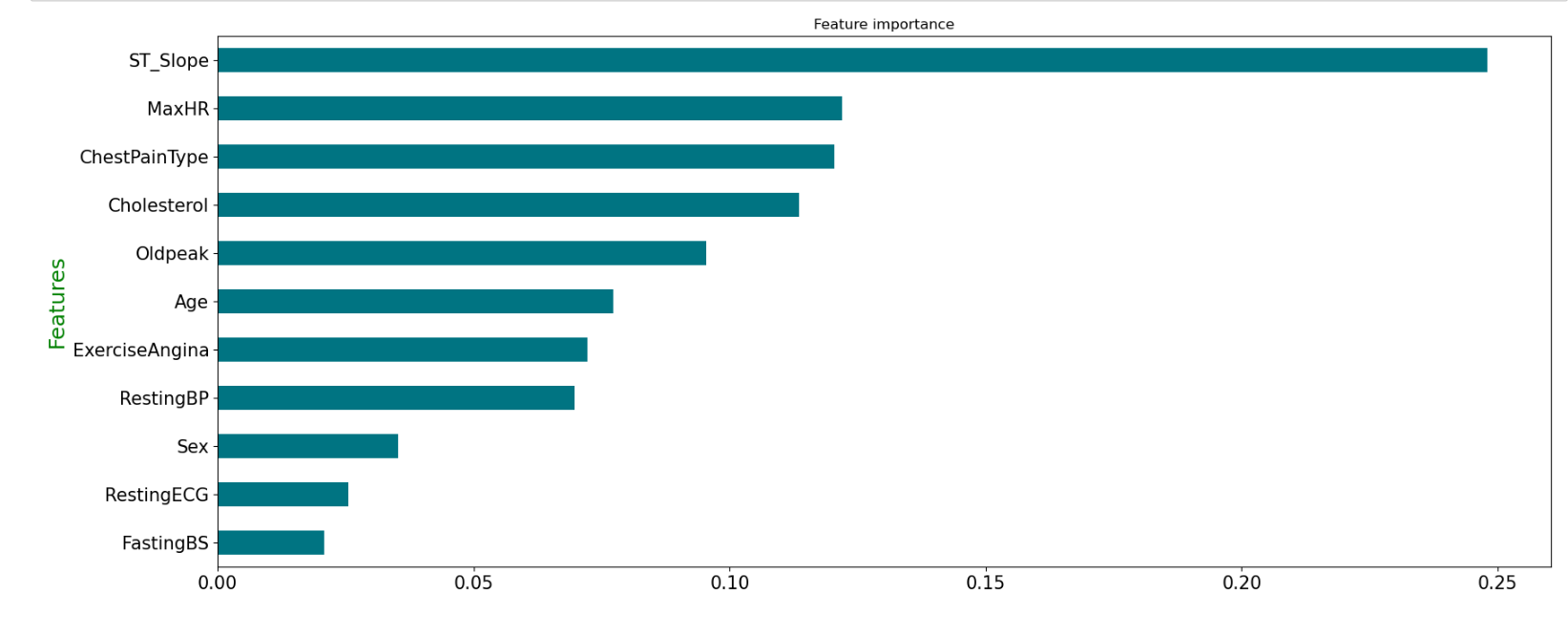
****

Figure 9:   
Feature importance

Next, we concatenated both the train and validation data to create a new training data and implemented a class based on this training data and tested the scores on the testing data. The functioning of the implemented class is as follows –

• Input to the class List of models (best 5 models selected above), metric (any of the metrics such as accuracy score, precision, f1-score etc.)

• Methods implemented inside the class

1. Fit () – Used to train all the models which are given as input and then it is storing the input metrics, class (0,1) predictions and predicted probabilities.

2. Mode\_accuracy() – It will consider the predictions of all the models and then calculates the mode of the predictions.

For eg: If 3 models predict the value as 1, and 2 models predict the value as 0, then it will consider the final prediction based on the majority and hence will store the final prediction as 1 in this case.

3. Mean\_proba\_accuracy() – Calculates the mean of the probabilities of all 5 models and then makes the prediction on the basis of the final calculated probability.

For eg: If the predicted probabilities from the 5 models are [p1, p2, p3,, p4, p5] then the final probability will be –

𝒑 = 𝒑𝟏 + 𝒑𝟐 + 𝒑𝟑 + 𝒑𝟒 + 𝒑𝟓 𝟓

Then this value will get rounded to find the prediction as class 0 or class 1, and then calculates the input metrics based on the test data.

4. Weighted\_mean\_accuracy() – Taking the prediction array of all the models as input and then calculating the weighted mean of the predictions by taking the weights as the accuracy score of that model.

For eg: If the predicted accuracy from the 5 models are [a1, a2, a3, a4, a5] and the predicted classes are [p1, p2, p3, p4, p5] then the final probability will be –

𝒑 = 𝒑𝟏(𝒂𝟏) + 𝒑𝟐(𝒂𝟐) + 𝒑𝟑(𝒂𝟑) + 𝒑𝟒(𝒂𝟒) + 𝒑𝟓(𝒂𝟓) 𝟓

Then this will be considered as the final probability to calculate the class 0 or class 1 prediction.

5. Weighted\_proba\_accuracy() – Taking the probability prediction array of all the models as input and then calculating the weighted mean of the probabilities by taking the weights as the accuracy score of that model.

For eg : If the predicted accuracy from the 5 models are [a1, a2, a3, a4, a5] and the predicted probas are [p1, p2, p3, p4, p5] then the final probability will be –

𝒑 = (𝒑𝟏(𝒂𝟏) + 𝒑𝟐(𝒂𝟐) + 𝒑𝟑(𝒂𝟑) + 𝒑𝟒(𝒂𝟒) + 𝒑𝟓(𝒂𝟓)) (𝒂𝟏 + 𝒂𝟐 + 𝒂𝟑 + 𝒂𝟒 + 𝒂𝟓 )

Then this will be considered as the final probability to calculate the class 0 or class 1 prediction.

**Algorithms**

For the project of heart failure prediction using machine learning, there are several existing systems or approaches that have been utilized in research and practice. Some commonly used existing systems for heart failure prediction include:

1. Logistic Regression: Logistic regression is a widely used statistical modeling technique that can be applied to predict heart failure. It analyzes the relationship between input variables (such as age, gender, medical history, and vital signs) and the likelihood of heart failure occurrence.

2. Random Forest: Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It has been applied to heart failure prediction by considering various features and their importance in determining the risk of heart failure.

3. Support Vector Machines (SVM): SVM is a machine learning algorithm that can be used for classification tasks, including heart failure prediction. It works by finding an optimal hyperplane that separates different classes based on input features.

4. Artificial Neural Networks (ANN): ANN is a deep learning technique that mimics the structure and functionality of the human brain. It has been utilized for heart failure prediction by training neural networks with a large dataset containing patient information and predicting the likelihood of heart failure.

5. Gradient Boosting Algorithms (e.g., XGBoost, LightGBM): Gradient boosting algorithms are powerful machine learning techniques that sequentially build a strong predictive model by combining weak models. They have shown promising results in heart failure prediction by leveraging the strengths of multiple weak models.

**Results and Analysis**

The results we got by implementing the above class on the best 5 selected models are –

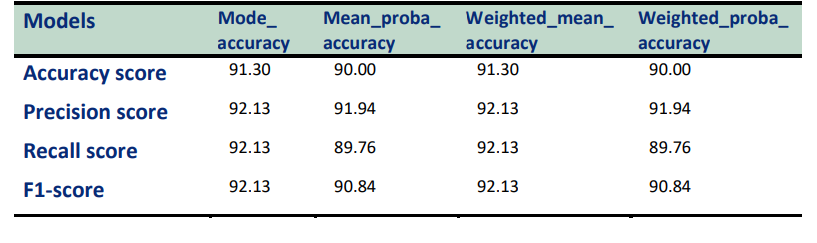
****

Figure 10: Result

As we can see that, the final accuracy score increased by a slight amount, as in the previous table the max accuracy was about 91 percent and here the max accuracy is about 91.30 percent.

**ROC plots and 10-fold cross validation scores –**

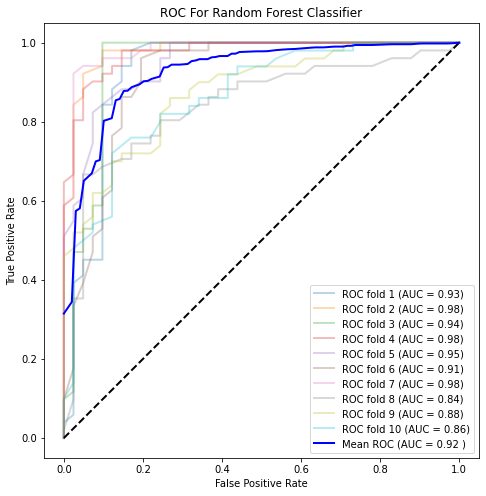
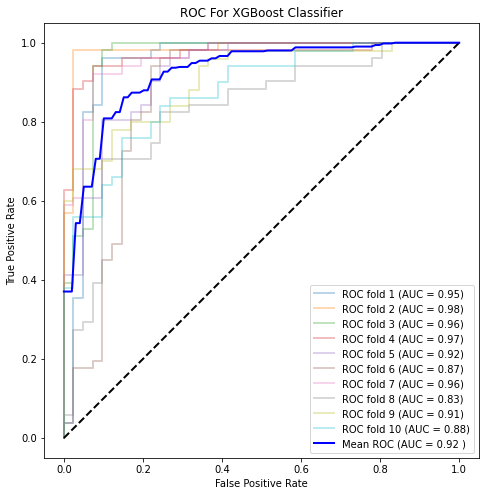
 

Figure 11: Random Forest Figure 12: XGBoost

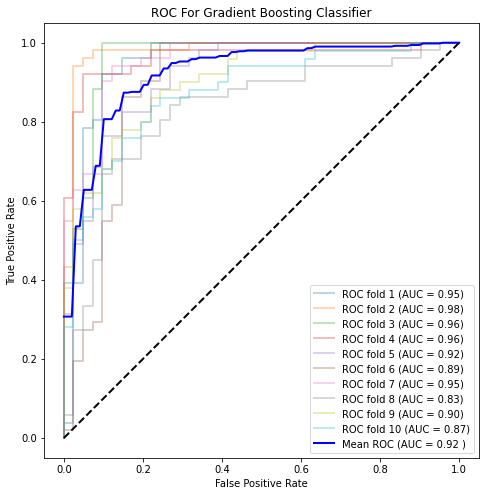
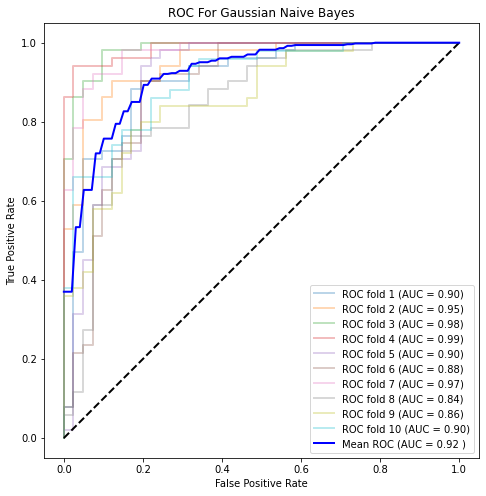
 

Figure 13: Gradient Boosting Figure 14: Naïve Bayes

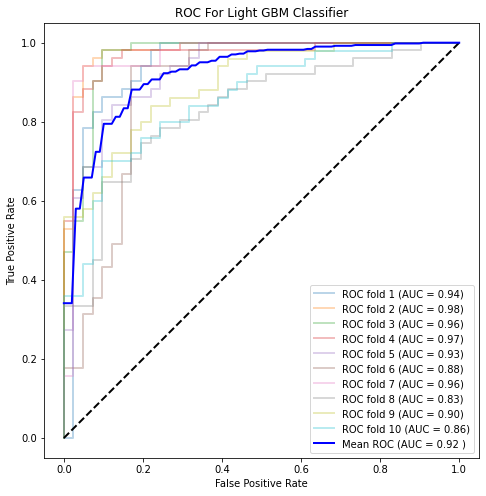
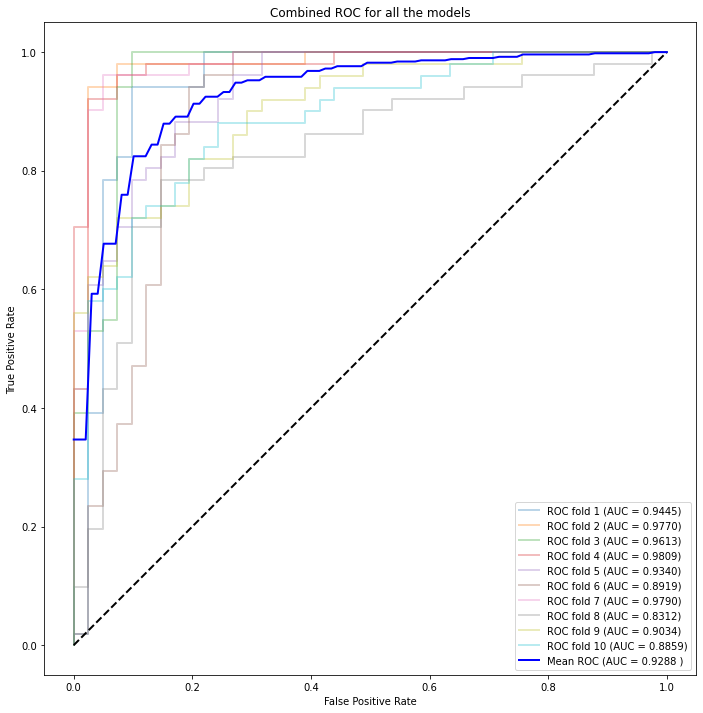
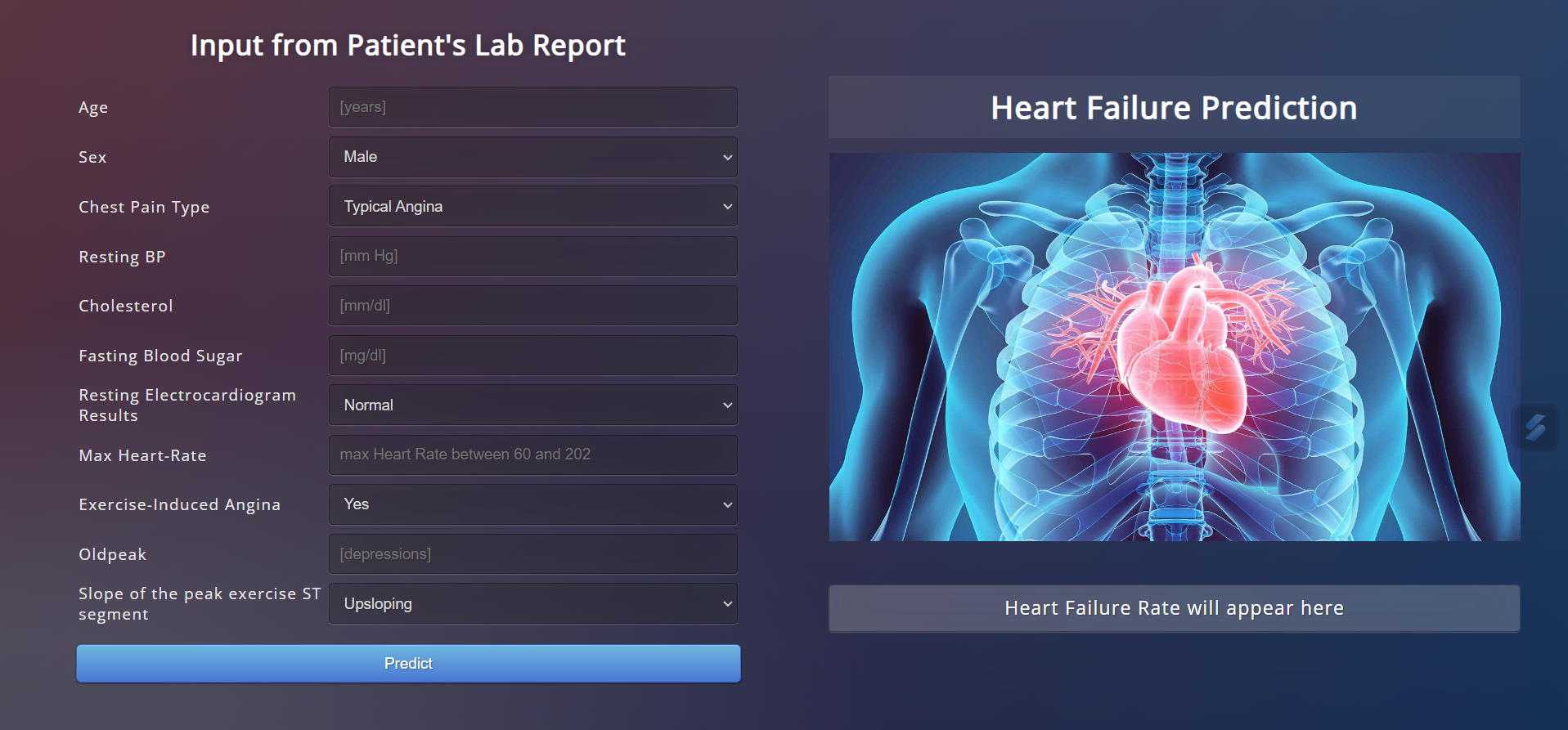
 

Figure 15: Light GBM Figure 16: ALL models

**Web page**

****

**Conclusion**

In conclusion, the project on heart failure prediction using machine learning aims to address the crucial need for early identification and personalized care in the field of cardiovascular health. By comparing and utilizing existing models and employing an ensemble approach, the project achieves higher accuracy in predicting heart failure risk. The proposed solution offers several unique features, including personalized risk assessment, optimized resource allocation, and improved healthcare decision making. It presents a significant value proposition with its potential for cost savings, scalability, and commercial viability. Moreover, the solution contributes to environmental sustainability by reducing healthcare waste and promoting preventive care. Overall, this project represents an innovative application of machine learning in healthcare, showcasing the potential to enhance patient outcomes and optimize healthcare practices.

**Reference**

[1] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE access* 7 (2019): 81542-81554.

[2] Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." *IOP conference series: materials science and engineering*. Vol. 1022. No. 1. IOP Publishing, 2021.

[3] Yadav, Anupama, Levish Gediya, and Adnanuddin Kazi. "Heart disease prediction using machine learning." *International Research Journal of Engineering and Technology (IRJET* 8.09 (2021).

[4] Diwakar, Manoj, et al. "Latest trends on heart disease prediction using machine learning and image fusion." *Materials Today: Proceedings* 37 (2021): 3213-3218.

[5] Khourdifi, Youness, and Mohamed Baha. "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization." *International journal of Intelligent engineering & systems* 12.1 (2019).