Heart Disease Prediction Using XAI

Utkarsh Singh[1] & Somik Bansal[2]
Utkarsh Singh[1] (BML Munjal University, Computer Science Department, Gurgaon, Haryana, India) Somik Bansal[2] (BML Munjal University, Computer Science Department, Gurgaon, Haryana, India)

Abstract

Heart disease remains a leading cause of mortality worldwide, underscoring the need for accurate and interpretable predictive models. While machine learning models like Neural Networks and Random Forests provide high accuracy, their "black box" nature limits transparency. This project applies Explainable Artificial Intelligence (XAI) techniques, particularly LIME, to make heart disease predictions both accurate and interpretable. Using a dataset of 1,025 records with 14 features (age, cholesterol levels, chest pain type, etc.), a Random Forest Classifier was trained to predict heart disease, yielding strong performance. Key predictors such as cholesterol levels and maximum heart rate were identified. LIME was applied to explain individual predictions, revealing the most influential features in the model's decision-making process. This integration of machine learning and XAI enhances both the predictive power and trust in the model, making it more applicable in clinical settings. Future work could refine model performance and expand interpretability to other models.

Keywords: Heart disease , machine learning, explainable AI, Random forest classifier,LIME(Local Interpretable Model-Agnostic Explanations),

1 Introduction

Explainable Artificial Intelligence (XAI) has become a crucial aspect of machine learning, addressing the need for transparency and interpretability in models that are traditionally viewed as "black boxes." Among the prominent XAI libraries, LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) stand out for their ability to unravel machine learning predictions. In this report, we delve into an end-to-end machine learning process aimed at predicting whether patients have heart disease. By leveraging both the LIME and SHAP libraries, we interpret and explain these predictions, shedding light on the features influencing decision-making. LIME provides instance-specific explanations by approximating complex models with simpler ones, while SHAP assigns importance values to features based on their contribution to a prediction, offering both local and global interpretability. Through this exploration, we aim to showcase the significance of XAI in enhancing model transparency, paving the way for more trustworthy and ethical AI implementations in real-world scenarios.

2 Literature review

The literature review provided elucidates the escalating interest in yoga pose classification, correction, and pose estimation, propelled by advancements in deep learning, human pose estimation (HPE), and Explainable AI (XAI) techniques. Through the amalgamation of Convolutional Neural Networks (CNNs), transfer learning, and pose estimation methodologies, endeavors have been made to surmount the challenges of classifying and rectifying yoga poses, underscoring the intricacy of poses, variations i+n human body structure, and the necessity for real-time feedback. This study consolidates insights from various key research studies, focusing on diverse facets of yoga pose classification and correction, and proposing innovative solutions rooted in deep learning models and explain ability techniques, signifying a concerted effort to enhance interpretability and transparency in machine learning models tailored towards yoga pose analysis.

Unveiling Key Predictors for Early Heart Attack Detection using Machine Learning and XAI Technique using LIME

This paper explores the use of machine learning (ML) and explainable artificial intelligence (XAI) to improve early detection of heart attacks, a leading cause of global mortality. The authors conducted a comparative analysis of several ML classification algorithms, including AdaBoost, Random Forest, Gradient Boosting, and Light Gradient Boosting Machine (LGBM), to predict heart attacks based on clinical data. Among these, LGBM emerged as the top performer, achieving an impressive training accuracy of 99.33%. The study also highlighted the limitations of traditional ML approaches, which often operate as "black boxes," offering high accuracy but limited transparency into how predictions are made. To address this issue, the authors applied XAI techniques, specifically LIME (Local Interpretable Model-Agnostic Explanations), to provide interpretable insights into the model's decisions. LIME identified key features, such as "kcm" (a medical marker) and "troponin" (a protein released during heart muscle damage), as the most critical factors influencing heart attack predictions. The study suggests that integrating XAI with ML models not only enhances prediction accuracy but also provides much-needed transparency, which can be crucial for clinical decision-making. The paper concludes by recommending the adoption of deep learning techniques to further improve predictive capabilities

Heart Attack Prediction using Machine Learning and XAI

This master's thesis examines the application of machine learning (ML) models and explainable AI (XAI) techniques for heart attack prediction. Cardiovascular diseases, particularly heart attacks, are a leading cause of death globally, and early detection can be life-saving. The author uses various ML algorithms, including XGBoost, Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, K-Neighbors Classifier, and Naive Bayes, to predict heart attacks based on a publicly available dataset. The study found that XGBoost outperformed the other models in terms of prediction accuracy, making it the most effective method for heart attack prediction in this study. Additionally, the research explored the use of XAI techniques, specifically SHAP (Shapley Additive Explanations) and LIME, to interpret the model’s predictions. These XAI methods provide transparency by identifying which features most strongly influence the model's predictions, offering both local (instance-specific) and global explanations. This aspect of the research addresses a key issue in healthcare: the need for both accurate predictions and interpretable models that clinicians can trust. The paper concludes that integrating ML with XAI could significantly improve the prediction and management of heart diseases, allowing for more informed clinical decisions

Why Model Why? Assessing the Strengths and Limitations of LIME

This paper critically evaluates the Local Interpretable Model-Agnostic Explanations (LIME) framework, a popular tool used to provide explainability for complex machine learning models, particularly in domains like healthcare, finance, and safety-critical systems such as autonomous vehicles. The authors investigate the performance of LIME in making tabular ML models more interpretable and assess its utility through a series of experiments involving various state-of-the-art machine learning algorithms applied to a tabular dataset. The paper emphasizes that while LIME is a powerful tool for offering local explanations (i.e., explanations specific to individual predictions), it has limitations, particularly when used with models that process tabular data as opposed to image or text data. One of the key findings is that LIME can be highly sensitive to the sampling process it uses to generate explanations, which can sometimes lead to unstable or inconsistent results. Despite these challenges, LIME remains an important tool for improving model transparency and interpretability. The study also includes a usability evaluation, where participants unfamiliar with LIME were asked to interpret its outputs. The results indicate that while LIME can significantly enhance understanding of model predictions, it requires careful application and interpretation. The authors conclude by suggesting improvements to LIME and recommending further research into alternative explainability methods like SHAP and MAPLE that could offer more robust insights into ML models

A Study of LIME and SHAP Model Explainers for Autonomous Disease Prediction

Explainable AI (XAI) addresses the need for transparency in machine learning models, especially in healthcare applications where understanding model predictions is crucial. Traditional ML models, though accurate, often operate as “black boxes” with little insight into their decision-making processes. LIME (Local Interpretable Model-Agnostic Explanations), developed by Ribeiro, Singh, and Guestrin in 2016, provides instance-specific explanations by highlighting features that drive individual predictions. This report explores using LIME in predicting heart disease, emphasizing its value in enhancing model interpretability. By making ML models more transparent, XAI can improve trust and ethical application in healthcare, enabling better-informed clinical decisions.

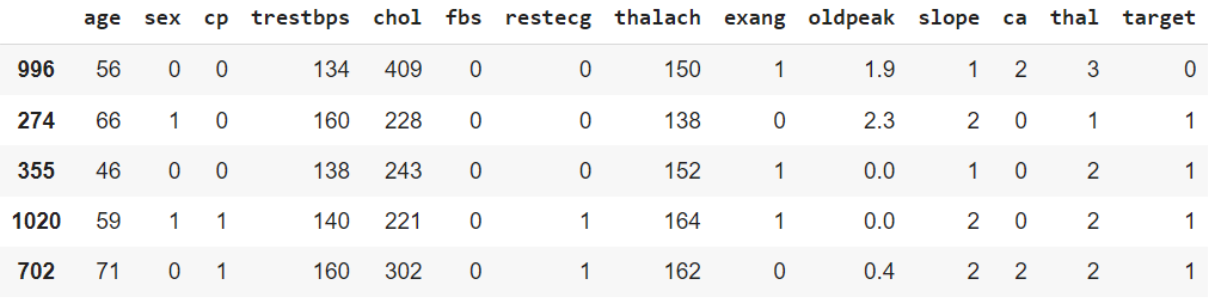
“Table 1: Overview of the Literature Review”

Ref NO.	Title	Techniques used	Pros	Cons	Limitations	Dataset used
1	A Study of LIME and SHAP Model Explainers for Autonomous Disease Prediction(**BASE PAPER**)	**LIME** and **SHAP** applied on a Naive Bayes classifier model to enhance interpretability.	Enhances transparency, trust, and flexibility in AI-driven healthcare predictions.	High computation, limited global insight, and complex interpretation for non-experts.	Limited scalability, accuracy, and consistent interpretability with complex models.	• Diabetes prediction • Heart disease prediction • Breast cancer prediction
2	Unveiling Key Predictors for Early Heart Attack Detection using ML and XAI (LIME)	AdaBoost, Random Forest, Gradient Boosting, LightGBM, LIME	High accuracy (LGBM 99.33%), identifies key predictors like kcm and troponin	AdaBoost performed poorly compared to others	Limited generalizability due to dataset size (Heart Attack dataset with 8 features)	Heart Attack dataset with 1319 samples
3	Heart Attack Prediction using Machine Learning and XAI (SHAP, LIME)	XGBoost, Logistic Regression, Stochastic Gradient Descent, SVM, KNN, Naive Bayes, LIME, SHAP	XGBoost showed best accuracy, feature importance visualized using SHAP and LIME	High complexity of models, SHAP and LIME may be difficult to implement for non-experts	Relies on black-box models and complex XAI methods	UCI Heart Disease Dataset
4	Why Model Why? Assessing the strengths and limitations of LIME	LIME applied on Decision Trees, Random Forest, Logistic Regression, XGBoost	LIME provides local explanations for model predictions, enhances model interpretability	Limited to local explanations, lacks global model behavior insights	Limited effectiveness when used for models trained on non-image or complex, high-dimensional data	Various tabular datasets, e.g., Australian Rain dataset

3 Methodology

3.1 Dataset collection and Importing

The dataset was obtained from a publicly available source and uploaded for analysis. It included various features relevant to predicting heart disease, such as age, gender, blood pressure, cholesterol levels, and biochemical markers. The data was then imported into the environment for further processing.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
996	56	0	0	134	409	0	0	150	1	1.9	1	2	3	0
274	66	1	0	160	228	0	0	138	0	2.3	2	0	1	1
355	46	0	0	138	243	0	0	152	1	0.0	1	0	2	1
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
702	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1

“Figure 1: Sample of the dataset

3.2 Data Description

- **Age**: The age of the patient in years.

- **Sex**: The gender of the patient (1 = male, 0 = female).

- **CP (Chest Pain Type):** Indicates the type of chest pain experienced by the patient:
  0: Typical angina
  1: Atypical angina
  2: Non-anginal pain
  3: Asymptomatic]

- **Trestbps (Resting Blood Pressure):** The patient’s resting blood pressure (in mm Hg) on admission to the hospital

- **Chol (Cholesterol):** Serum cholesterol in mg/dl.

- **FBS (Fasting Blood Sugar):** Whether the patient's fasting blood sugar is > 120 mg/dl (1 = true, 0 = false).

- **Restecg (Resting Electrocardiographic Results):** Results of the resting electrocardiogram:
  0: Normal
  1: Having ST-T wave abnormality (such as T wave inversions or ST elevation or depression of > 0.05 mV)
  2: Showing probable or definite left ventricular hypertrophy.

- **Thalach (Maximum Heart Rate Achieved):** The maximum heart rate achieved during a stress test.

- **Exang (Exercise Induced Angina):** Whether the patient experienced angina during exercise (1 = yes, 0 = no).

- **oldpeak:** ST depression induced by exercise relative to rest, measuring heart stress.

- **Slope (Slope of the Peak Exercise ST Segment):** The slope of the peak exercise ST segment:
  0: Upsloping
  1: Flat
  2: Downsloping

- **CA (Number of Major Vessels Colored by Fluoroscopy):** The number of major vessels (0-3) colored by fluoroscopy (a type of imaging test).

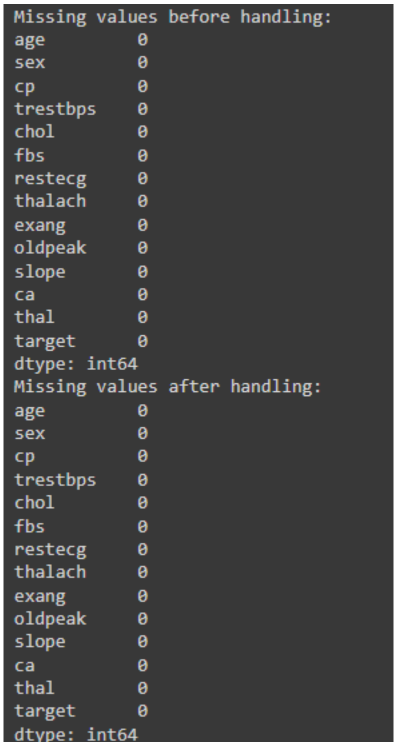
- **Thal (Thalassemia):** A blood disorder that affects oxygen-carrying proteins:
  1: Fixed defect (no reversible blood flow)
  2: Normal
  3: Reversible defect (reversible blood flow)

- **Target:** The diagnosis of heart disease (0 = no heart disease, 1 = heart disease).

3.3 Data Cleaning

To ensure data quality, duplicate records were identified and removed. Structural errors in categorical columns were corrected by verifying that values matched predefined valid categories (e.g., specific values for gender, chest pain type, etc.). This step also included filtering out unwanted outliers from numerical columns (e.g., age, systolic blood pressure) using the Interquartile Range (IQR) method to improve data consistency.

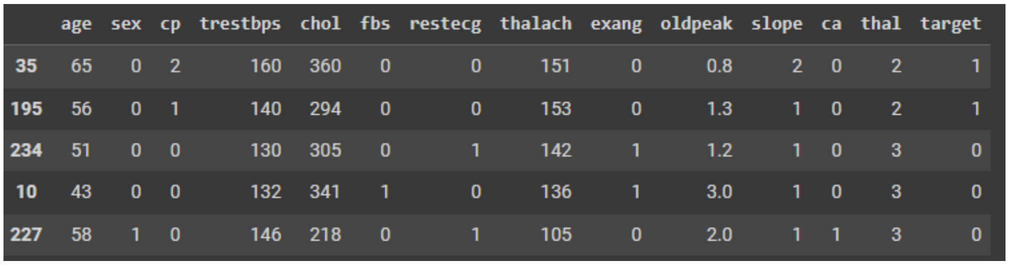
- Removing Duplicate Values : 723 duplicate values are removes
- Fixing Structural Errors
- Filtering Unwanted Outliers
- Handling Missing Data



```
Missing values before handling:
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
Missing values after handling:
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

- Validation and Quality Assurance

3.4 Loading Cleaned Dataset



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
35	65	0	2	160	360	0	0	151	0	0.8	2	0	2	1
195	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
234	51	0	0	130	305	0	1	142	1	1.2	1	0	3	0
10	43	0	0	132	341	1	0	136	1	3.0	1	0	3	0
227	58	1	0	146	218	0	1	105	0	2.0	1	1	3	0

3.5 Data Visualization

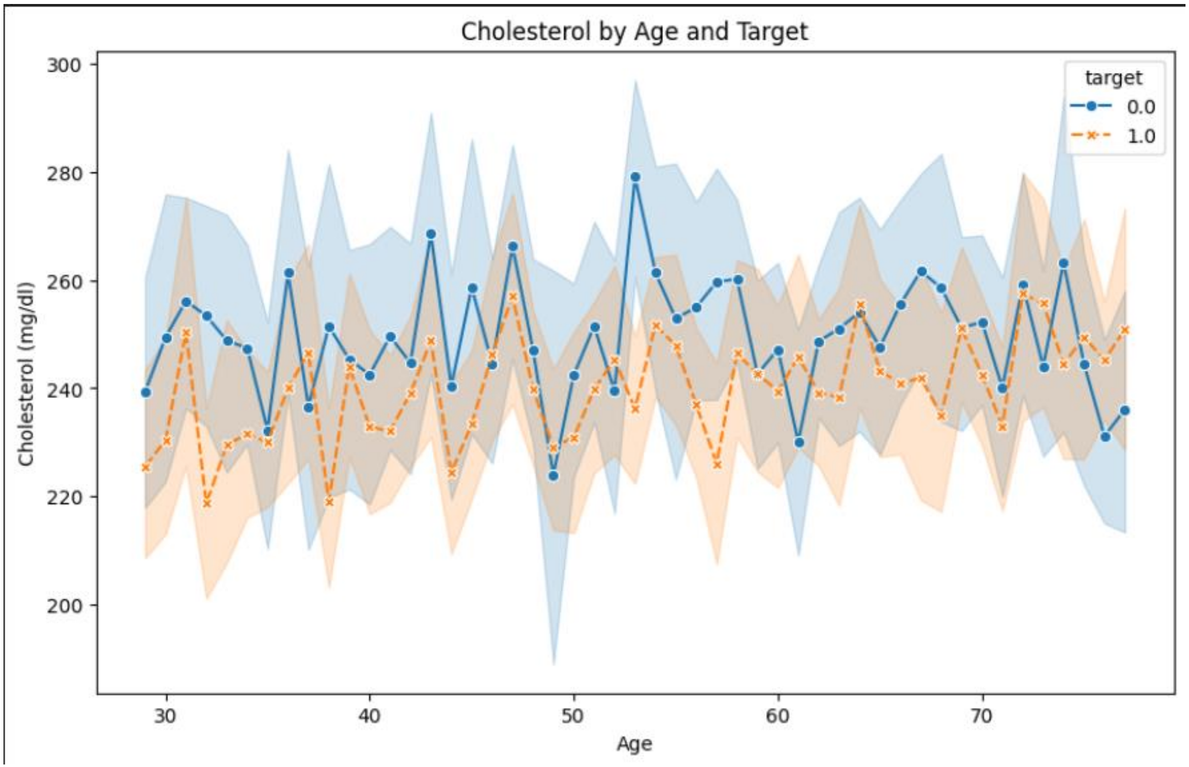


Fig 2 : Cholesterol by Age and Target

3.6 EDA (Exploratory Data Analysis)

- Distribution of variables

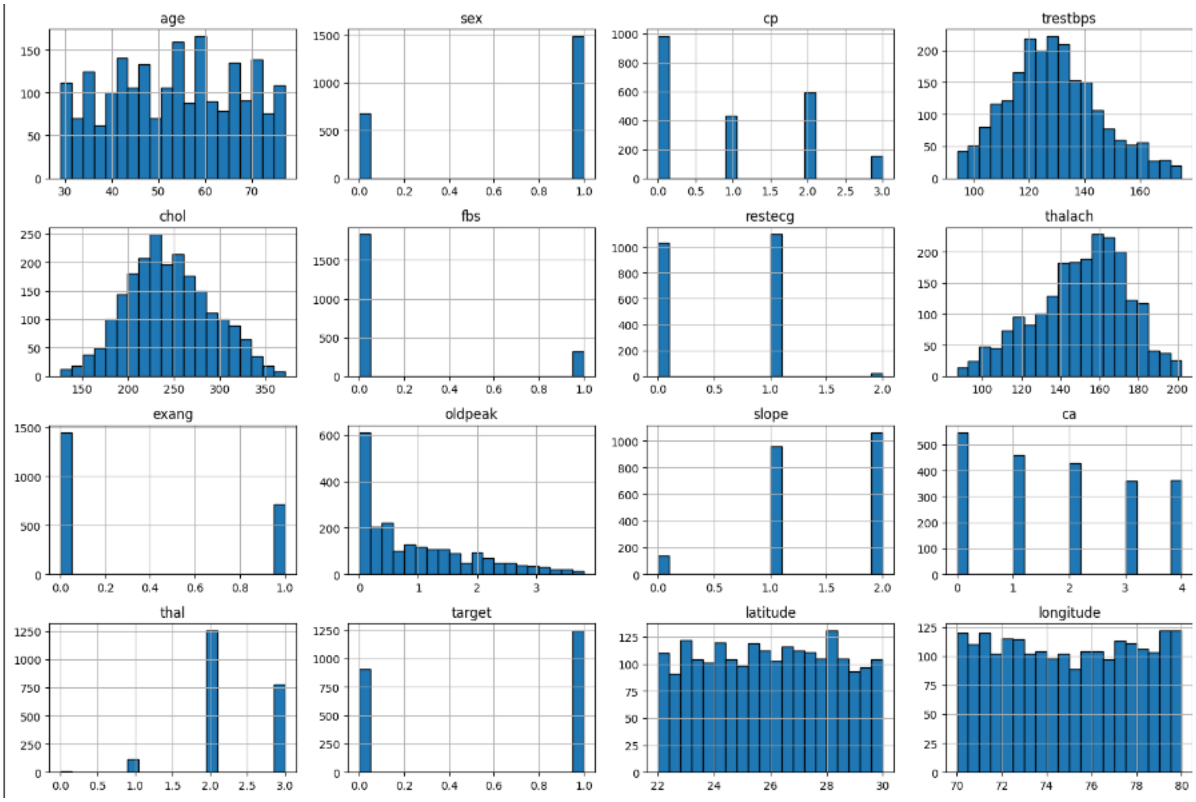
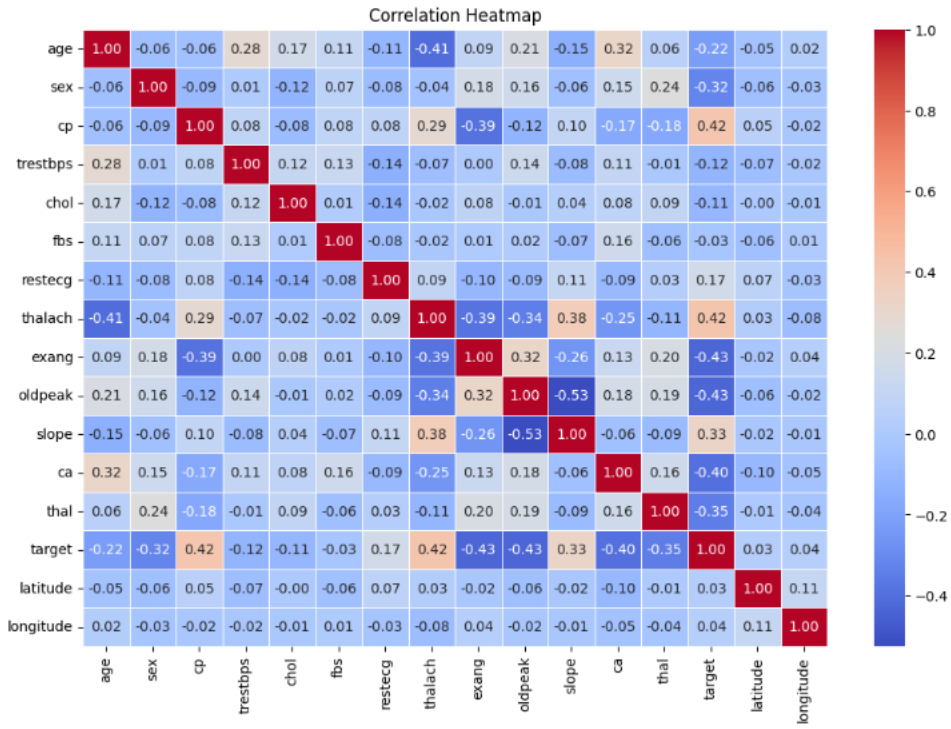


Fig 3 :

- Correlation heat map



- Class Imbalance

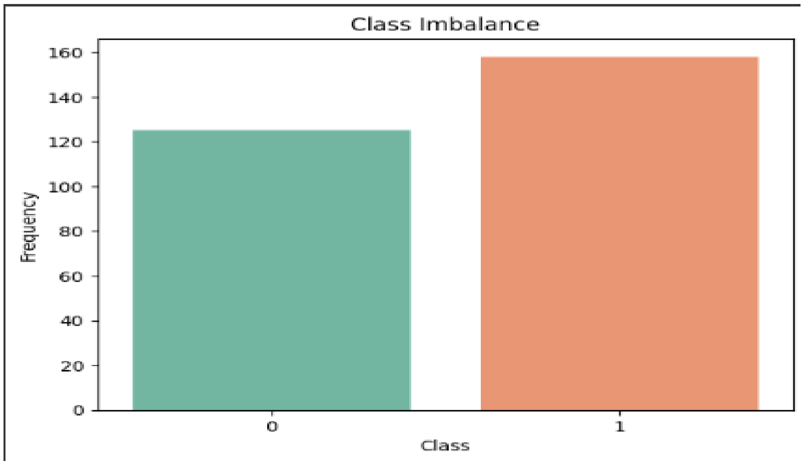


Fig 5

3.7 Model building

- **Random Forest Classifier :** A Random Forest Classifier can be used in heart disease prediction by analyzing patterns in patient data (e.g., age, blood pressure, cholesterol) to classify whether a patient is at risk. It combines multiple decision trees to improve accuracy and reduce the risk of overfitting, making it effective for complex medical datasets.

- **Logistic Regression**: Logistic Regression can be used for heart disease prediction by modeling the probability of a patient having the disease based on features like age, cholesterol, and blood pressure. It is effective for binary classification tasks and helps identify the impact of each feature on the risk.

- **Decision Tree Classifier**: A Decision Tree Classifier predicts heart disease by splitting patient data into branches based on features (e.g., age, heart rate) to classify risk. It is simple, interpretable, and can capture non-linear patterns in medical data.

- **XGBoost Classifier:** The XGBoost (Extreme Gradient Boosting) Classifier predicts heart disease by building an ensemble of decision trees through a gradient boosting framework. It iteratively improves prediction accuracy by optimizing a loss function and correcting errors made by previous trees. XGBoost is highly efficient, handles missing data, and is well-suited for large and complex datasets. It can capture both linear and non-linear relationships in medical data, making it a robust choice for heart disease prediction.

3.8 LIME (Local Interpretable Model-agnostic Explanations)

In our project, LIME (Local Interpretable Model-Agnostic Explanations) was employed to enhance the interpretability of the XGBoost classifier used to predict heart disease. LIME provides a way to explain individual predictions made by any machine learning model, making it an ideal tool to understand complex models.

Key Features of LIME in Our Project:

- Model-Agnostic: LIME works with any machine learning model, regardless of its complexity, allowing us to use it with our XGBoost classifier.

- Local Interpretability: LIME generates explanations for specific predictions by creating a simpler, linear model that approximates the behavior of the more complex XGBoost model for each instance.

- Feature Importance: LIME identifies which features (e.g., age, cholesterol, exercise-induced angina) had the most significant impact on a particular prediction, helping us understand what drives the model's decision.

Application in Our Heart Disease Prediction:

By using LIME, we were able to explain how individual predictions of heart disease risk were made. For example, it helped us pinpoint whether features like cholesterol levels, thalach (maximum heart rate achieved), or restecg (electrocardiographic results) were the main factors contributing to a positive prediction for heart disease.

This interpretability was crucial in validating the model's trustworthiness, especially in a healthcare context where understanding why a decision was made is as important as the decision itself. By leveraging LIME, we ensured transparency in the decision-making process, fostering greater trust in the model's predictions.

3.8 SHAP (Shapley Addictive exPlanations)

In our project, SHAP was employed to enhance the interpretability of the XGBoost classifier used to predict heart disease. SHAP assigns each feature an importance value for a particular prediction, providing a clear explanation of how each feature contributes to the model's output.

Key Features of SHAP in Our Project:

- Model-Specific and Model-Agnostic: SHAP can be tailored for specific models like XGBoost using TreeExplainer while also offering model-agnostic capabilities, making it a versatile tool for interpretability.
- Global and Local Interpretability: SHAP provides:
  - **Global insight**s: By summarizing feature importance across all predictions (e.g., summary plots).
  - **Local explanations**: By detailing the contribution of each feature for individual predictions (e.g., force plots).
- **Feature Importance and Interaction**: SHAP not only ranks feature importance (e.g., cholesterol, age, exercise-induced angina) but also explores feature interactions, such as the relationship between thalach (maximum heart rate achieved) and ca (number of major vessels colored by fluoroscopy).

Application in Our Heart Disease Prediction:

Using SHAP, we were able to interpret both global and individual predictions made by the XGBoost classifier:

Global Explanations:

Summary plots revealed that features like thalach, chol, and oldpeak were the most influential in predicting heart disease risk.

Local Explanations:

- Force plots illustrated how specific features influenced the prediction for individual patients. For example, a high thalach value might reduce the risk score, while a high oldpeak value could increase it.
- This ability to explain predictions helped us validate the model's decisions, fostering trust and reliability, especially critical in healthcare contexts where the "why" behind predictions is as important as the predictions themselves.

4. Results & Discussion

- For the XGBoost algorithm, an accuracy of 94.68% was achieved, with precision, recall, and F1-score values of 94.40%, 96.33%, and 95.35%, respectively.

- For the Random Forest algorithm, an accuracy of 93.75% was achieved, with precision, recall, and F1-score values of 92.91%, 96.33%, and 94.59%, respectively.

- The Logistic Regression algorithm demonstrated lower performance with an accuracy of 82.87%, precision of 82.02%, recall of 89.39%, and an F1-score of 85.55%.

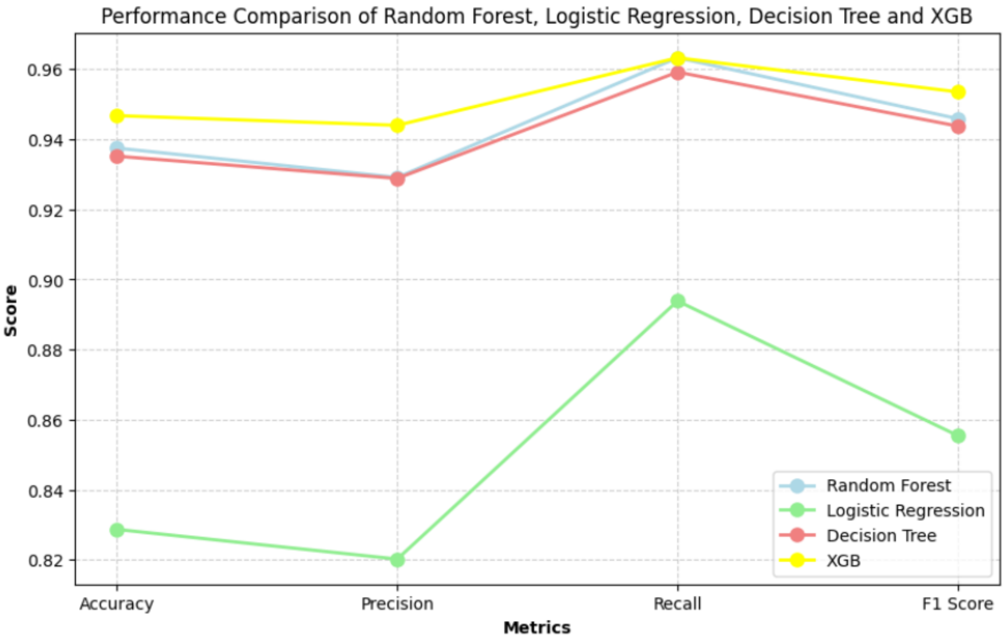
- The Decision Tree classifier achieved an accuracy of 93.52%, with precision, recall, and F1-score values of 92.89%, 95.92%, and 94.38%, respectively.

Model	Accuracy	F1 Score	Precision	Recall
XGBboost	0.9468	0.9535	0.9440	0.9633
Random Forest	0.9375	0.9459	0.9291	0.9633
Decision Tree	0.9352	0.9438	0.9289	0.9592
Logistic Regression	0.8287	0.8555	0.8202	0.8939

Table 2 : Model Differentiation

- Overall, the XGBoost algorithm showed the best performance across all evaluation metrics, closely followed by the Random Forest and Decision Tree classifiers.

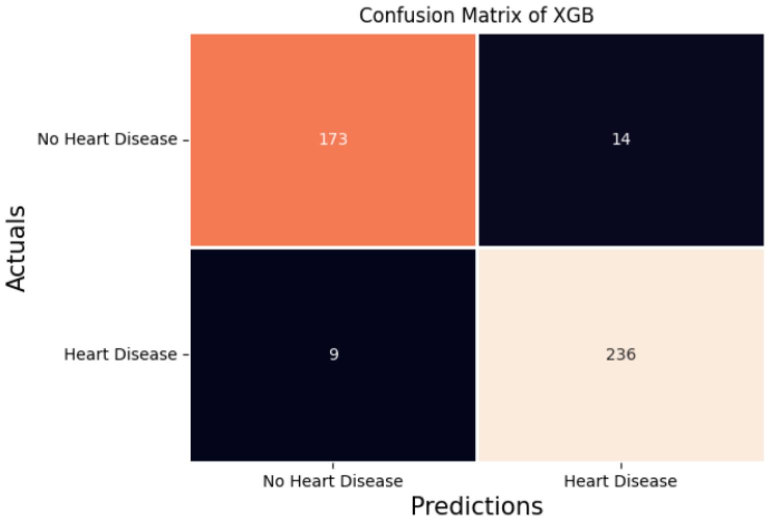
- The models' accuracy was tested using cross-validation, revealing variations in performance based on the dataset. XGBoost Classifier demonstrated the best results across all metrics, followed by Random Forest, Decision Tree, and Logistic Regression.



“Figure 6: Comparison graph

- **XgBoost Classifier: Confusion Matrix Analysis**
  The confusion matrix for the XGBoost Classifier breaks down the model's performance into four key outcomes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). In heart disease prediction, recall is particularly important because it measures how well the model identifies positive cases (patients with heart disease).

- **Importance of High Recall**
  A high recall is essential in medical diagnostics since it reduces false negatives, where heart disease is present but the model fails to detect it. Missing these cases can have serious consequences, as patients might not receive the necessary treatment on time.
- In our model, XGBoost achieved a recall of 96.33%, meaning it correctly identified over 96% of actual heart disease cases. This exceptionally high recall minimizes missed diagnoses, thereby enhancing patient safety and making the model highly reliable for detecting heart disease. Furthermore, XGBoost maintains a strong balance between recall and precision, reducing both false negatives and false positives, ensuring more effective and trustworthy predictions.



"Figure 7: Confusion Matrix

➢ **LIME EXPLANATION FOR XgBoost MODEL**

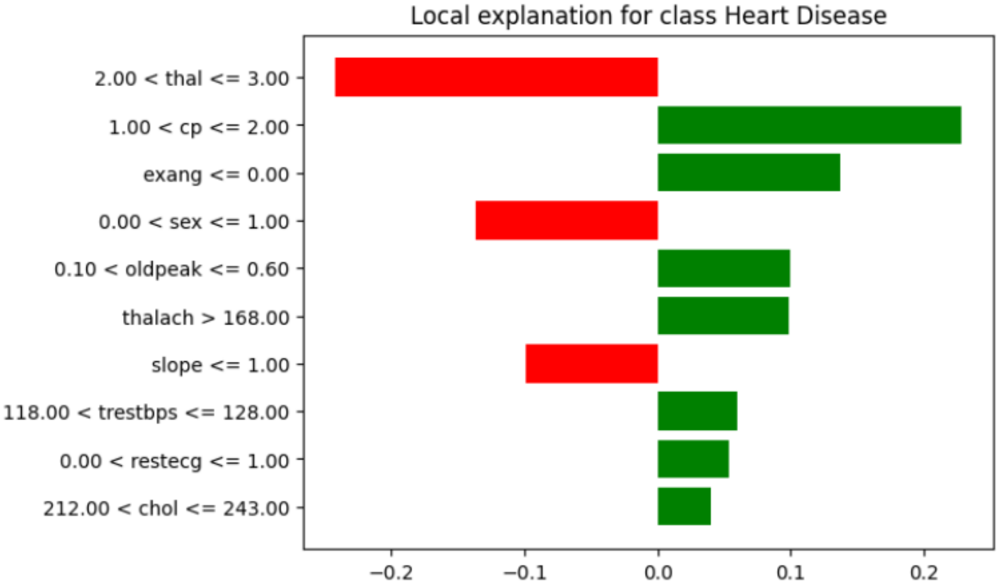
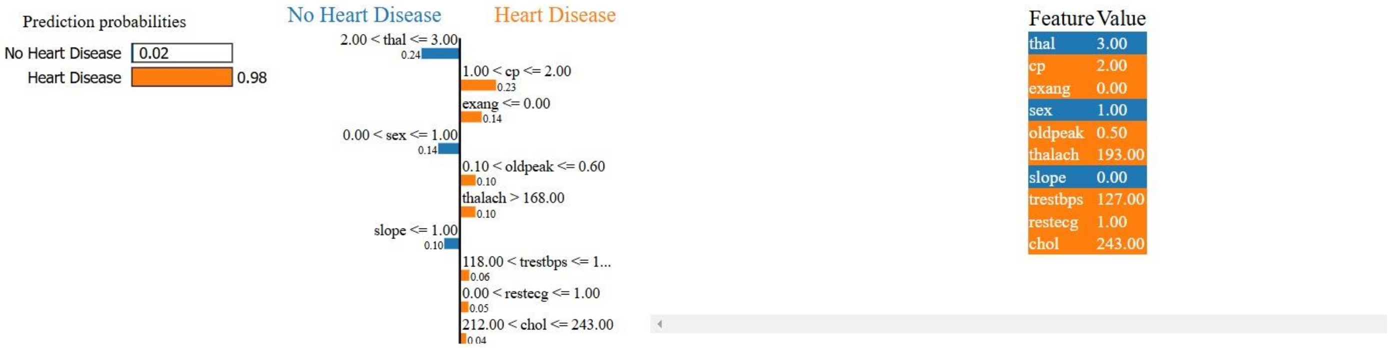


Fig 8 :Local explanation for class Heart Disease



Feature	Value
thal	3.00
cp	2.00
exang	0.00
sex	1.00
oldpeak	0.50
thalach	193.00
slope	0.00
trestbps	127.00
restecg	1.00
chol	243.00

FIG 9: No Heart Disease (blue color) and Heart Disease (orange color)

Features pushing the prediction towards Heart Disease (Orange color):

- **cp (Chest Pain Type**): A value of 2.00 (non-anginal pain) moderately pushes the prediction towards heart disease, as certain chest pain types are associated with cardiac issues.
- **exang (Exercise-Induced Angina**): A value of 0.00 (no exercise-induced angina) slightly contributes to heart disease, as the absence of angina during exercise could still indicate underlying cardiac risks.
- **oldpeak:** An ST depression value of 0.50, caused by exercise, contributes to the prediction of heart disease, as even mild ST depression indicates stress on the heart.
- **thalach (Maximum Heart Rate Achieved):** A high value of 193.00 slightly contributes to heart disease, as extremely high heart rates may indicate abnormal stress responses or cardiac conditions.
- **trestbps (Resting Blood Pressure):** A resting blood pressure value of 127.00 contributes slightly towards heart disease. While in a near-normal range, it can still indicate mild cardiac stress.
- **restecg (Resting Electrocardiographic Results**): A value of 1.00 (ST-T wave abnormality) moderately pushes the prediction towards heart disease, as it often reflects irregularities in heart activity.
- **chol (Cholesterol):** A cholesterol level of 243.00 contributes to the prediction of heart disease, as high cholesterol is a known risk factor for cardiovascular conditions.

➢ SHAP(SHaply Addictive exPlanations)

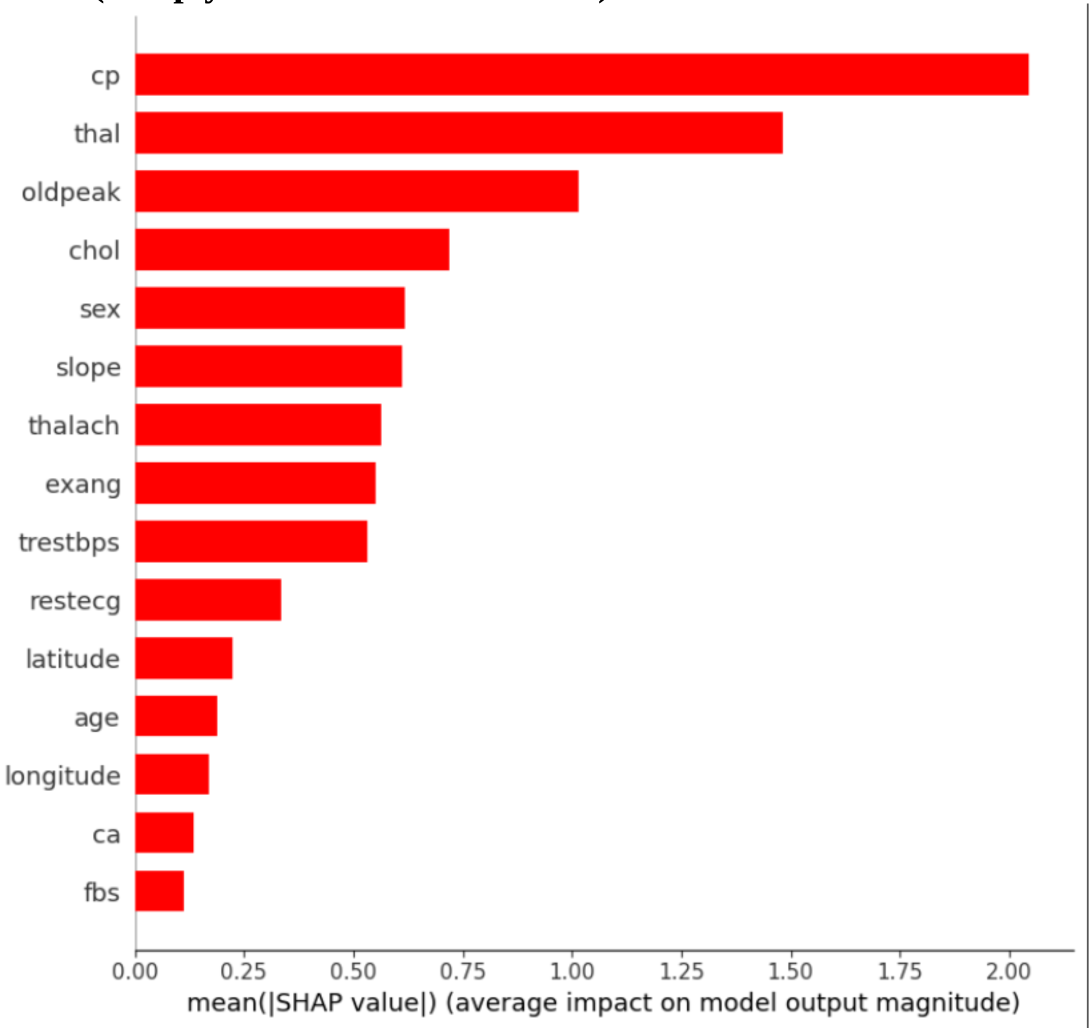
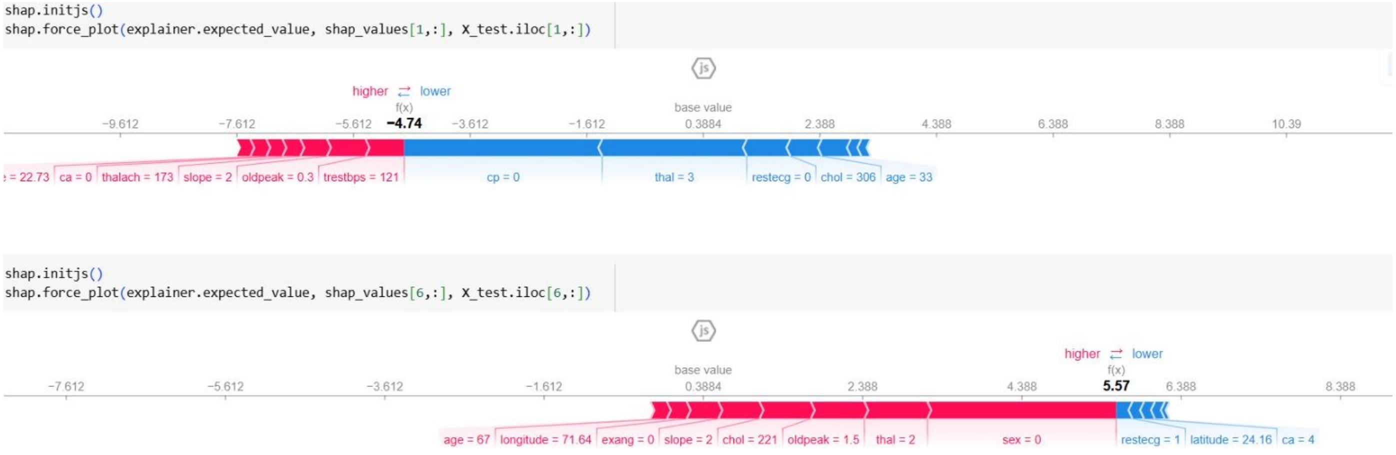


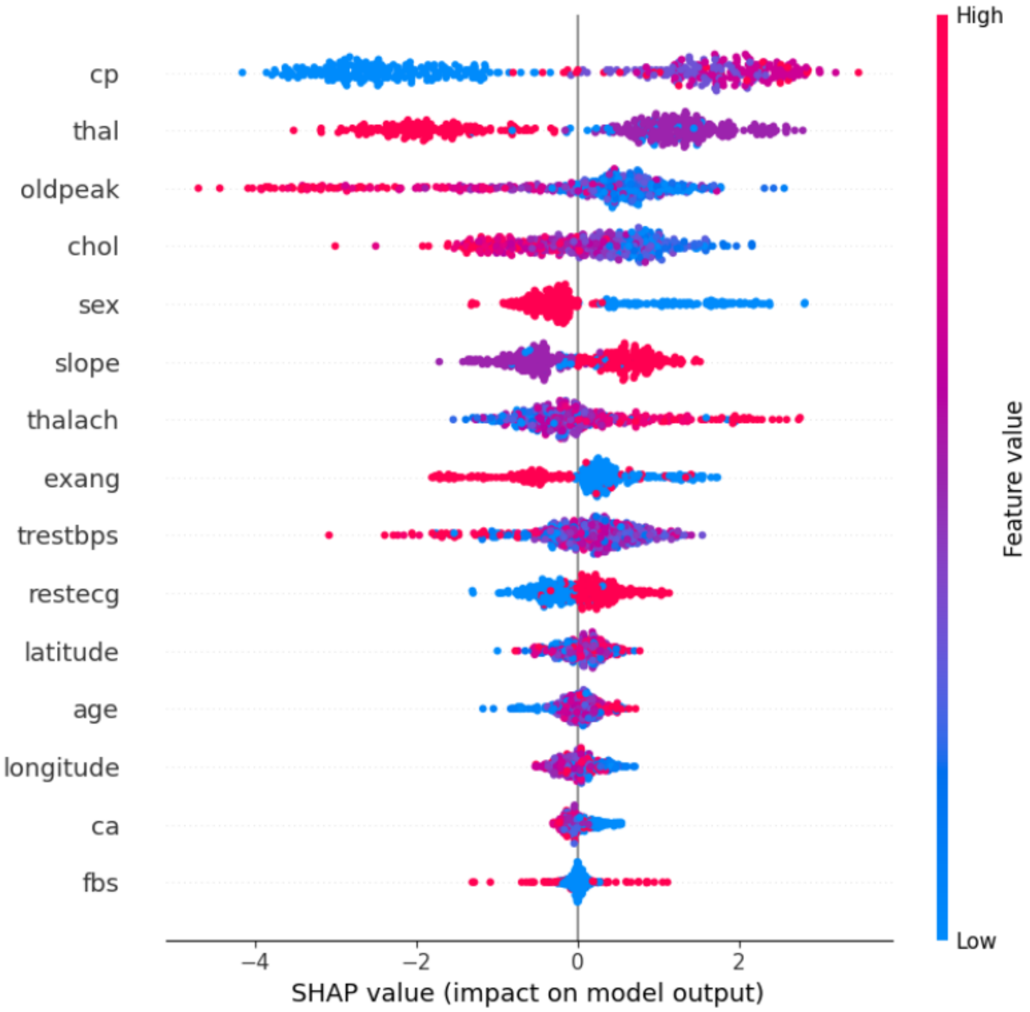
FIG 10



```
shap.initjs()
shap.force_plot(explainer.expected_value, shap_values[1,:], X_test.iloc[1,:])
```

```
shap.initjs()
shap.force_plot(explainer.expected_value, shap_values[6,:], X_test.iloc[6,:])
```

Fig11 : SHAP Explainer Plot


cp
thal
oldpeak
chol
sex
slope
thalach
exang
trestbps
restecg
latitude
age
longitude
ca
fbs
−4
−2
0
2
SHAP value (impact on model output)
High
Low
Feature value

➢

FIG 12

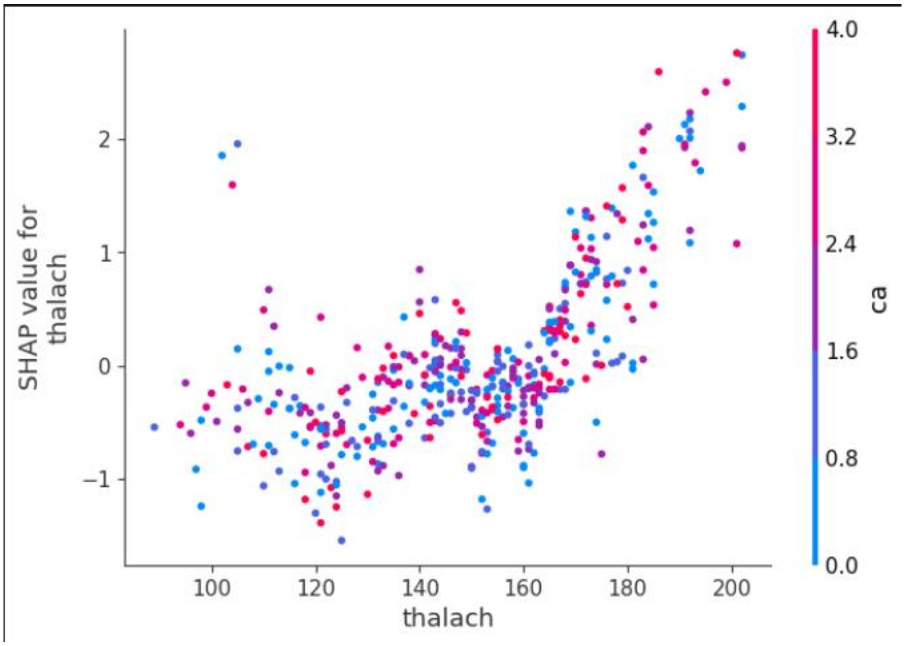

SHAP value for thalach
2
1
0
−1
100
120
140
160
180
200
thalach
4.0
3.2
2.4
1.6
0.8
0.0
ca

FIG 13

5 Conclusion

In medical diagnostics, especially in the context of heart disease prediction, minimizing the risk of false negatives—where an actual patient is misclassified as healthy—is critical. Therefore, selecting a model that offers high recall, which indicates the model's ability to correctly identify true positive cases, is of utmost importance. Precision, recall, and overall accuracy are key factors in assessing the performance of machine learning models, but in healthcare applications, recall often takes precedence to ensure that patients with the disease are not overlooked.

After comparing the performance of Logistic Regression, Decision Tree, Random Forest, and XGBoost classifiers, it was observed that the XGBoost model consistently achieved superior results. It not only showed the highest recall but also outperformed the other models in terms of precision, F1-score, and accuracy. Most importantly, the XGBoost model demonstrated the lowest occurrence of false negatives, making it highly effective for the early detection and treatment of heart disease. The Random Forest and Decision Tree classifiers also showed strong performance, with Random Forest coming close to XGBoost in recall and precision. Logistic Regression, while performing adequately, lagged behind the other models in most metrics.

In conclusion, the XGBoost model proved to be the most reliable and robust for predicting heart disease, making it an ideal choice for applications where identifying at-risk patients is critical. This analysis highlights the significance of evaluating machine learning models based on multiple metrics, ensuring that the chosen model is optimized for both predictive accuracy and medical relevance.