Heart Disease Prediction

1st Aadit Baxi
School of Computer Engineering.
KIIT Deemed to be University
Bhubaneswar, Odisha, India
2205088@kiit.ac.in

2nd Ananya Sinha
School of Computer Engineering.
KIIT Deemed to be University
Bhubaneswar, Odisha, India
2205100@kiit.ac.in

3rd Ishaan Bhatt
School of Computer Engineering.
KIIT Deemed to be University
Bhubaneswar, Odisha, India
2205119@kiit.ac.in

4th Ananya Sapre
School of Computer Engineering.
KIIT Deemed to be University
Bhubaneswar, Odisha, India
2205875@kiit.ac.in

5th Yerra Prateek
School of Computer Engineering.
KIIT Deemed to be University
Bhubaneswar, Odisha, India
22051215@kiit.ac.in

6th Aditi Prabhat
School of Computer Engineering.
KIIT Deemed to be University
Bhubaneswar, Odisha, India
22053831@kiit.ac.in

***Abstract*—One of the leading causes of death globally is cardiovascular disease (CVD), and successful treatment depends on early identification. This study uses a dataset of 303 patients with 14 clinical parameters to investigate machine learning approaches for heart disease prediction. We apply and analyze various machine learning models, including logistic regression, evaluating their F1-score, accuracy, precision, and recall. The findings support preventive healthcare programs by showing that machine learning offers a practical method for early cardiac disease identification. Heart disease, eXtreme Grading Boosting, RandomForest, K Nearest Neighbor, Machine Learning, Predictive Analytics, Health Informatics, and Logistic Regression are some of the index terms.**

***Index Terms*—Heart Disease, Machine Learning, Predictive Analytics, Health Informatics, Logistic Regression,K Nearest Neighbor, RandomForest, eXtreme Grading Boosting**

I. Introduction

Heart disease remains one of the leading causes of mortality worldwide, affecting millionsofpeople annually [1]. It encompasses a range of cardiovascular conditions, including coronaryartery disease, heart failure, and arrhythmias, which often develop over time and are influencedby various risk factors such as age, lifestyle, genetics, and pre-existing medical conditions. Early detection and timely intervention play a crucial role in reducing the impact of heart disease and improving patient outcomes. With the evolution of healthcare technologies and the exponential growth of electronichealthrecords, there has been a surge in the availability of structured medical data. This has pavedtheway for the application of machine learning (ML) techniques to extract meaningful patternsand support clinical decision-making. ML models can analyze complex datasets, identifyhidden correlations, and predict potential health risks with impressive accuracy. In this project, we aim to harness the power of machine learning to develop a heart diseaseprediction system that can assess a patient's likelihood of having heart disease basedonkeymedical parameters. The model utilizes historical patient data such as age, cholesterol levels, blood pressure, heart rate, and other relevant metrics to generate a predictive score. The primary objectives of this project are: To design and implement a machine learning model capable of accurately predictingthepresence of heart disease based on patient data inputs. To improve diagnostic efficiency by leveraging data-driven insights that aidmedical professionals in making informed decisions. To develop a user-friendly interface that allows healthcare providers or patients to input dataand receive instant predictions in an accessible and interpretable format. To evaluate and optimize the model using various performance metrics and validationtechniques, ensuring reliability and robustness across diverse patient profiles. This report outlines the motivation behind the project, defines the core objectives, describesthe datasets and methodologies employed, and presents the outcomes and potential impactof the solution in real-world healthcare scenarios.

II. Literature Review

A. Machine Learning for Healthcare

Machine learning has revolutionized various sectors, including healthcare. In predictive healthcare analytics, ML models learn from historical patient data to identify patterns and predict potential health risks. The use of ML in cardiology has enabled accurate and faster detection of heart disease risks.

B. Algorithms Implemented

The following machine learning algorithms were implemented in the project:

- **Logistic Regression**: Efficient for binary classification problems.
- **KNN (K-Nearest Neighbour)**: Works well with noisy datasets and adaptable for classification tasks.
- **Random Forest**: Reduces overfitting and improves accuracy by combining multiple decision trees.
- **XGBoost (Extreme Gradient Boosting)**: High-performance boosting algorithm used for large-scale datasets.

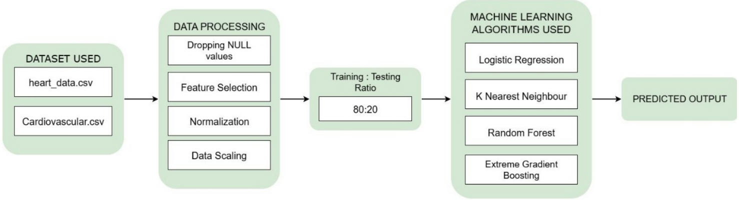

Fig. 1: System Diagram

A machine learning pipeline for predicting heart disease is depicted in the diagram. Betterset.csv and heart data disease.csv are the two datasets utilized. Missing value management, feature selection, normalization, and scaling are all part of data preparation. An 80:20 ratio is used to divide the data into training and testing sets. The processed data is subjected to four machine learning algorithms: Extreme Gradient Boosting (XGBoost), Random Forest, K-Nearest Neighbour (KNN), and Logistic Regression. Each algorithm's result is utilized to produce a projected output that, depending on the input data, most likely indicates the existence or risk of heart-related disorders. Using a variety of models, the procedure guarantees precise and efficient prediction.

III. Implementation

The following procedures were used for the implementation of our project.

A. Methodology

1) Data Collection: Dataset 1 was collected from Kaggel [2].The dataset contains a total of 1025 instances with 14 attributes.

Dataset 2 was collected from Kaggel [3].The dataset contains a total of 918 instances with 12 attributes.

B. Models Used

The following machine learning models were implemented:

1) **Logistic Regression**: Logistic regression Machine learning is a statistical technique used to construct models for machine learning in which the dependent variable is binary, or dichotomous [4]. The relationship between one dependent variable and one or more independent variables is described by data using logistic regression. There are three types of independent variables: nominal, ordinal, and interval. A logistic function that we may use to determine the likelihood of a car breaking down based on the number of years since its last service is shown in the following example.

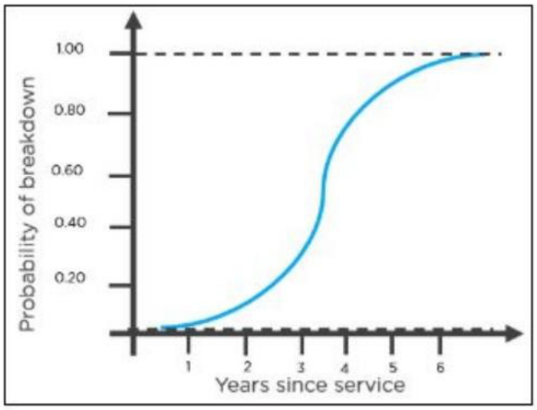

Fig. 2: Logistic Regression Curve

2) **K-Nearest Neighbour (KNN)**: KNN is a straightforward, supervised machine learning (ML) technique that is commonly used in missing value imputation and can be applied to classification or regression tasks. We can categorize unexpected points using the values of the closest existing points since it is predicated on the notion that the observations that are closest to a particular data point are the most "similar" observations in a data set. K allows the user to choose how many close observations the algorithm will use.

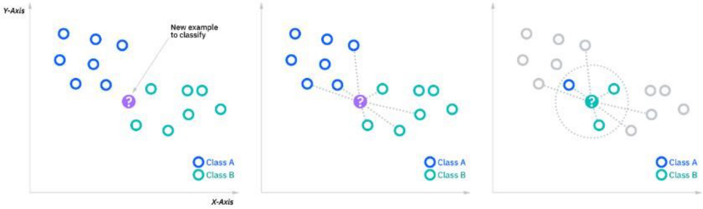

Fig. 3: K-Nearest Neighbour Visualization

3) **Random Forest**: Leo Breiman and Adele Cutler created the popular machine learning method Random Forest, which aggregates the output of several decision trees to produce a single outcome. Because it can handle both classification and regression issues, its popularity has been spurred by its versatility and ease of use as well as its efficacy as a random forest classifier.

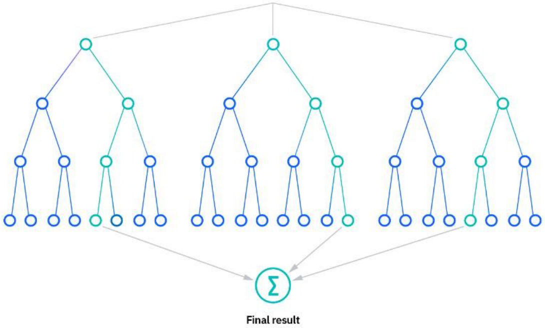

Fig. 4: Random Forest Model Visualization

4) **XGBoost (Extreme Gradient Boosting)**: The distributed, open-source machine learning package XG-

Boost (eXtreme Gradient Boosting) employs gradient boosted decision trees, a supervised learning boosting approach that leverages gradient descent. It is renowned for being quick, effective, and scalable when working with big datasets. [5]

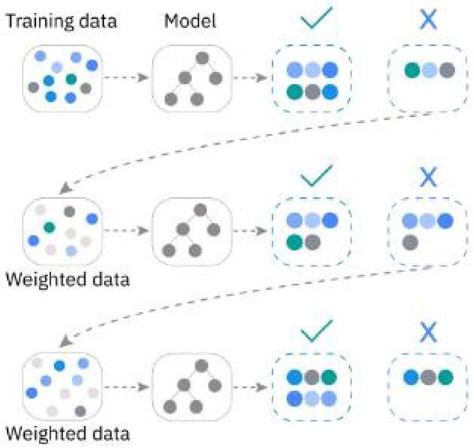

Fig. 5: XGBoost Visualization

IV. Results

TABLE I: Performance Metrics - Dataset 1

Performance Metric	Logistic Regression	K Nearest Neighbour	Random Forest	XGBoost
Accuracy	80.33%	75.41%	83.61%	83.61%
Balanced Accuracy	79.38%	75.11%	83.77%	83.77%
Precision	76.92%	74.29%	87.10%	87.10%
Average Precision	74.85%	70.19%	81.10%	81.10%
Recall	90.91%	81.25%	81.82%	81.82%
F1 Score	83.33%	77.61%	84.38%	84.38%

TABLE II: Performance Metrics - Dataset 2

Performance Metric	Logistic Regression	K Nearest Neighbour	Random Forest	XGBoost
Accuracy	96.00%	84.00%	97.00%	96.00%
Balanced Accuracy	96.06%	83.70%	96.59%	95.89%
Precision	97.37%	86.96%	95.83%	96.55%
Average Precision	97.37%	82.82%	95.51%	95.22%
Recall	95.69%	85.47%	99.14%	96.55%
F1 Score	96.52%	86.21%	97.46%	96.55%

V. RESULT ANALYSIS

A. DATASET 1:

1) Data Correlation Heatmap:

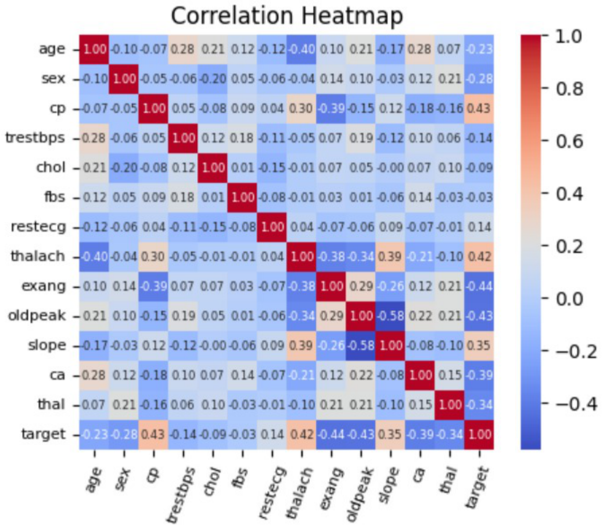

Fig. 5: Dataset 1 - Data Correlation Heatmap

With a few noteworthy patterns, the heatmap shows primarily modest relationships between attributes [6]. The type of chest discomfort and maximum heart rate are positively correlated with heart disease (the goal), while major vessels, oldpeak, and exercise-induced angina are negatively correlated. The relationship between age and maximal heart rate is somewhat inverse. While the majority of other factors show little interrelation, these variables stand out as significant predictors.

2) Logistic regression confusion matrix:

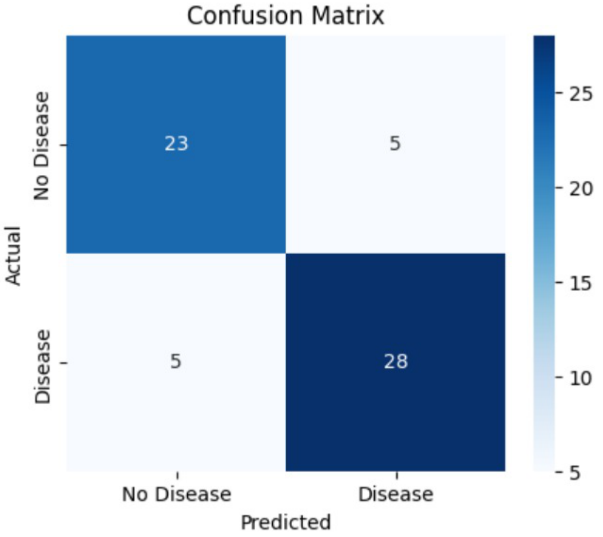

Fig. 5: Dataset 1 - Logistic Regression Result

The performance of the Logistic Regression model is seen in the confusion matrix: 28 cases of ”Disease” and 23 cases of ”No Disease” were accurately predicted. Five false negatives and five false positives were found. This suggests a rather successful model with balanced error kinds, with an overall accuracy of 83.6

3) KNN confusion matrix:

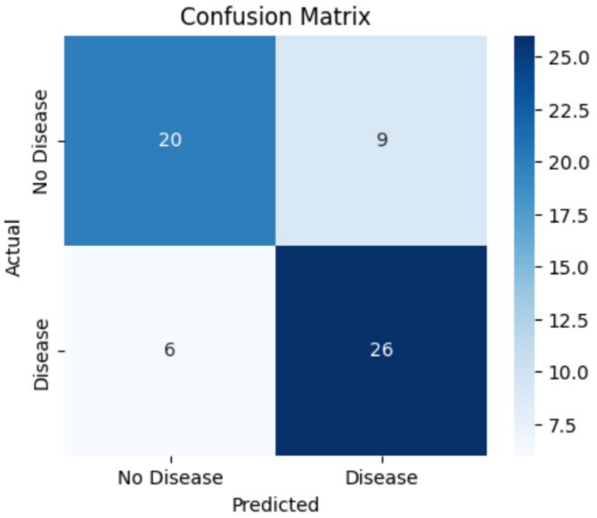

Fig. 5: Dataset 1 - Knn Result

Twenty true negatives and twenty-six true positives were obtained with the KNN model. Nevertheless, it produced six false negatives, failing to detect ”Disease” cases, and nine false positives, misclassifying ”No Disease” as ”Disease.” This yields a 75.4

4) Random Forest confusion matrix:

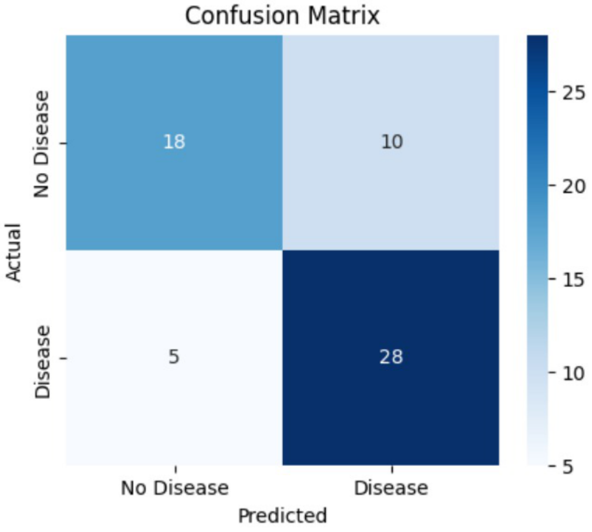

Fig. 5: Dataset 1 - Random Forest Result

The Random Forest model’s ability to forecast the presence of disease is displayed in the confusion matrix. Twenty-eight instances were accurately classified as ”Disease” (true positives) and eighteen as ”No Disease” (true negatives). Five false negatives and ten false positives were found. The findings show that overall accuracy is strong, with somewhat higher predictions for ”No Disease”.

5) XG Boost confusion matrix:

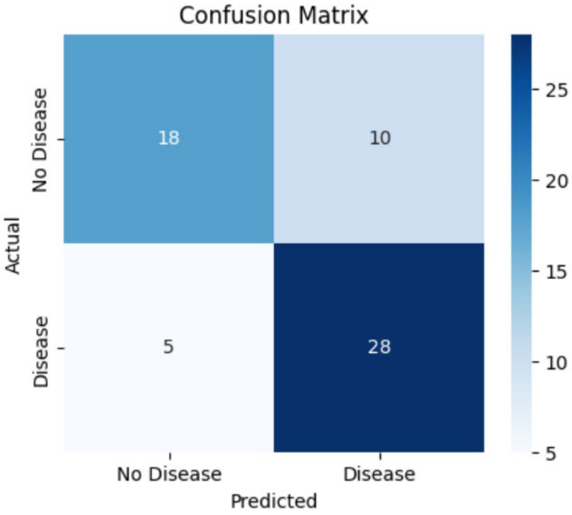

Fig. 5: Dataset 1 - XGBoost Result

The XG Boost model is evaluated by the confusion matrix. It accurately classified 28 individuals as having ”Disease” and 18 as having ”No Disease.” Nevertheless, there were five false negatives (missing genuine disease) and ten false positives (predicting disease when absent). Although the model performs well, it is somewhat more effective at identifying illness than at ruling it out.

6) Comparision Of All 4 Models:

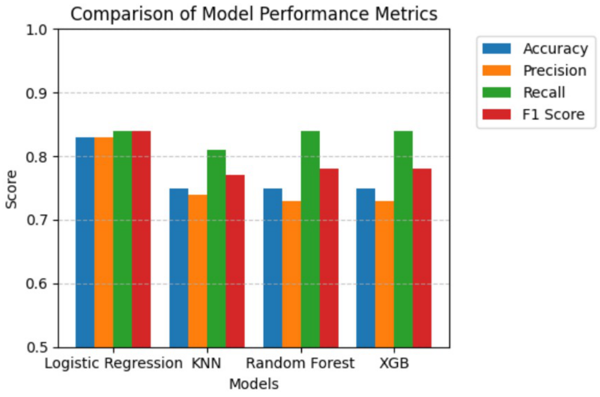

Fig. 5: Dataset 1 - Comparison Of all models

Four classification models—Logistic Regression, KNN, Random Forest, and XGB—are contrasted in the bar chart based on their accuracy, precision, recall, and F1 score. Overall, logistic regression performs best, consistently receiving the highest ratings across all measures. With noticeably higher recall than precision, KNN, Random Forest, and XGB perform similarly, indicating that they are more adept at spotting good situations. Based on these findings, Logistic Regression emerges as the most well-rounded and successful model overall [7].

B. DATASET 2:

1) Data Correlation Heatmap:

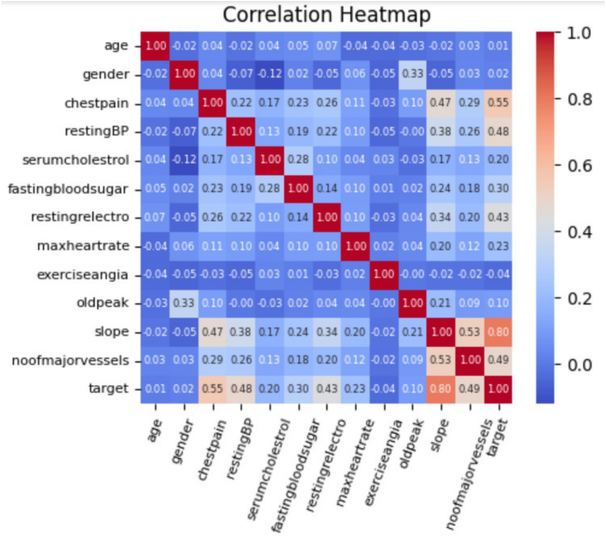

Fig. 5: Dataset 2 - Data Correlation Heatmap

The heart disease dataset's correlation heatmap primarily displays weak feature correlations. The number of main vessels (0.49), slope (0.80), and chest discomfort (0.55), on the other hand, show larger positive relationships with the target variable. These imply that exercise slope patterns and the type of chest discomfort are significant determinants. Relationships between other variables, such as age, cholesterol, and resting blood pressure, are less strong.

2) Logistic regression confusion matrix:

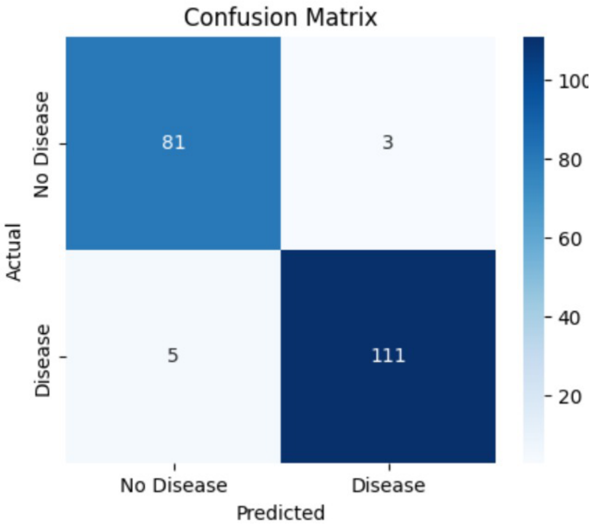

Fig. X: Dataset 2 - Logistic Regression Result

This confusion matrix demonstrates how accurately the Logistic Regression model predicts cardiac disease. 111 illness cases and 81 no-disease cases out of 200 are appropriately classified. There were just eight misclassifications (3 false positives and 5 false negatives). With somewhat fewer errors in predicting the existence of the disease than its absence, the model exhibits high precision and recall. The performance is outstanding.

3) KNN confusion matrix:

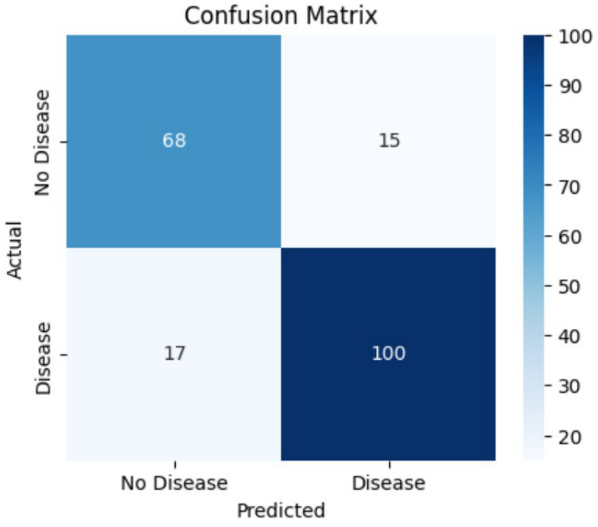

Fig. X: Dataset 2 - KNN Result

The predictions of the KNN model for heart illness are displayed in this confusion matrix. It accurately identified 100 cases of disease and 68 cases of no disease out of 200. 15 false positives and 17 false negatives are examples of misclassifications. Although the model's overall accuracy is strong, errors are distributed evenly across classes. To cut down on misclassifications, both sensitivity and specificity could be improved.

4) Random Forest confusion matrix:

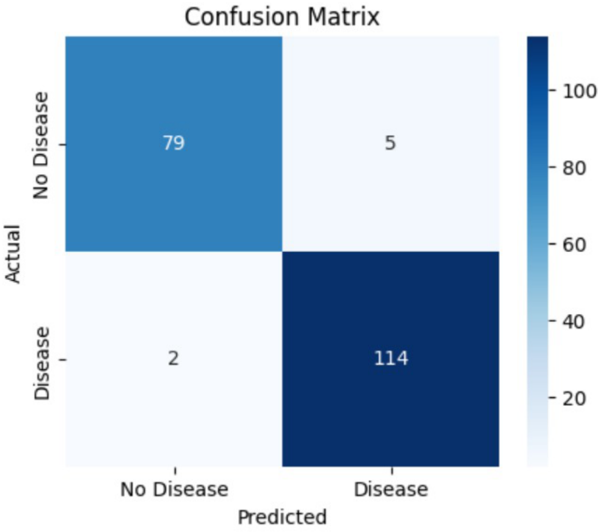

Fig. X: Dataset 2 - Random Forest Result

The Random Forest model performs exceptionally well in this confusion matrix. 114 illness cases and 79 no-disease cases out of 200 are appropriately classified. There are just seven misclassifications (5 false positives and 2 false negatives). The model is very dependable for predicting the presence or absence of cardiac disease since it exhibits high accuracy, sensitivity, and specificity with very few errors.

5) XG Boost confusion matrix:

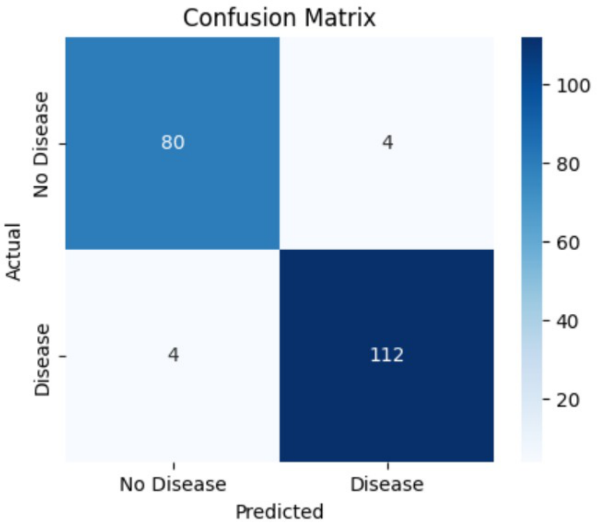

Fig. X: Dataset 2 - XGBoost Result

This confusion matrix shows excellent classification performance for XG Boost model. Out of 200 cases, 80 no-disease and 112 disease cases are predicted correctly. Only 8 misclassifications occurred: 4 false positives and 4 false negatives. The model has balanced sensitivity and specificity, with high overall accuracy. Predictions for both classes are strong, indicating a highly reliable heart disease prediction model.

6) Comparision Of All 4 Models:

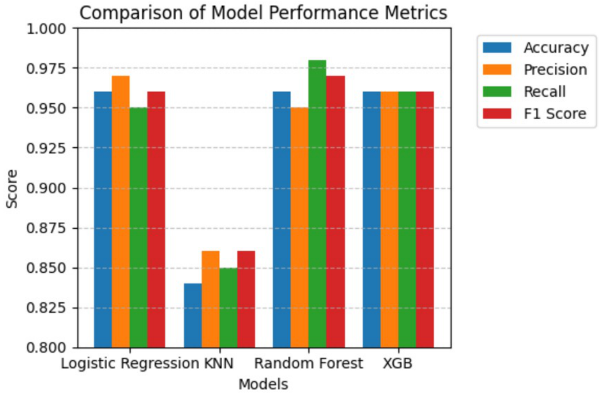

Fig. X: Comparison Of all models

The bar chart compares four models across accuracy, precision, recall, and F1 score. Random Forest performs best overall, with the highest recall and F1 scores near 0.98. Logistic Regression and XGBoost also perform strongly around 0.96. KNN performs weakest, especially in accuracy and recall. Random Forest is the most consistent and balanced among all evaluated metrics.

VI. Standards Adopted

A. Design Standards

To ensure clarity and maintainability, the project follows these design standards:

- **IEEE 830-1998:** Defines system requirements specifications.
- **ISO/IEC 25010:** Ensures software quality and maintainability.
- **UML Diagrams:** Includes use case, class, and sequence diagrams for proper system modeling.
- **Database Design:** Normalization techniques are applied to ensure efficiency and reduce redundancy.

B. Coding Standards

The project adheres to coding best practices for maintainability and readability:

- **PEP 8:** Ensures consistency in Python code style.
- **Modular Code:** Functions and classes are designed to perform single tasks for reusability.
- **Code Structure:** Proper indentation, meaningful comments, and logical segmentation are maintained.
- **Error Handling:** Use of `try-except` blocks to prevent runtime failures.

C. Testing Standards

To ensure reliability and accuracy, the project follows these testing standards:

- **IEEE 829-2008:** Guides test documentation.
- **ISO/IEC/IEEE 29119:** Provides a standardized testing framework.
- **Testing Strategies:**
  - Unit Testing: Validates individual functions.
  - Integration Testing: Ensures components work together correctly.
  - Performance Testing: Evaluates model efficiency.
  - Cross-validation: Uses K-Fold to assess accuracy.
  - User Testing: Evaluates usability of the Gradio interface.

VII. Conclusion and Future Scope

A. Conclusion

An effective and precise technique for early diagnosis was made possible by the project's successful development of a machine learning-based heart disease prediction system. Reliable predicting capabilities were shown by the application of algorithms like Random Forest, KNN, XGBoost, and Logistic Regression. Both users and medical professionals can utilize Gradio thanks to its user-friendly design. All things considered, this system can lower the risks of heart disease and support proactive healthcare decision-making.

B. Future Scope

Future upgrades could improve the system's usability and efficacy:

- **Integration with Real-time Data:** Connecting wearable medical equipment to the system to enable ongoing surveillance.
- **Expanded Dataset:** To enhance model generalization, additional patient data from various demographics should be included.

- **Advanced Machine Learning Models:** Using deep learning methods to improve prediction accuracy even more.
- **Cloud-based Deployment:** For scalability and remote accessibility, the model can be deployed on cloud platforms.
- **Mobile Application Development:** Making a version that is optimized for mobile devices to make it more accessible to a larger user base.

By enhancing these features, the system can develop into a more complete predictive healthcare tool th

By expanding on these aspects, the system can become a more comprehensive tool in predictive healthcare, aiding in early disease detection and prevention strategies.

References

[1] C. Boukhatem, H. Y. Youssef, and A. B. Nassif, "Heart disease prediction using machine learning," in *2022 Advances in Science and Engineering Technology International Conferences (ASET)*.

[2] J. Smith, "Heart disease dataset," https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset, 2023.

[3] F. Soriano, "Heart failure prediction dataset," https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction, 2023.

[4] D. Dua and C. Graff, "Uci heart disease dataset," https://archive.ics.uci.edu/dataset/45/heart+disease.

[5] A. N. (replace with actual names from the paper), "Heart disease prediction using machine learning algorithms." [Online]. Available: https://www.mdpi.com/1999-4893/16/2/88

[6] A. N. (replace with actual authors from the paper), "Title of the paper (replace with actual title)." [Online]. Available: https://arxiv.org/pdf/2505.09969

[7] ——, "Comparative analysis of heart disease prediction using logistic regression, svm, knn and random forest with cross-validation for improved accuracy." [Online]. Available: https://www.researchgate.net/publication/390924883