Optimization of Heart Disease Prediction using Machine Learning Model

CAPSTONE PROJECT PHASE-1

Phase – I Report

Submitted by

1. **21BCE10225 Archita Gupta**
2. **21BCE10406 Sonali Raghuwanshi**
3. **21BCE10439 Priyanshi Yadav**
4. **21BCE10669 Tina Chelwani**
5. **21BCE10708 Abhinav Shrivastava**

in partial fulfilment of the requirements for the degree of

Bachelor of Engineering and Technology

VIT®
BHOPAL
www.vitbhopal.ac.in

VIT Bhopal University
Bhopal
Madhya Pradhesh

December, 2024

VIT®
BHOPAL
www.vitbhopal.ac.in

Bonafide Certificate

Certified that this project report titled "Optimization of Heart Disease Prediction using ML Model" is the bonafide work of "21BCE10225 Archita Gupta, 21BCE10406 Sonali Raghuwanshi, 21BCE10439 Priyanshi Yadav, 21BCE10669 Tina Chelwani, 21BCE10708 Abhinav Shrivastava" who carried out the project work under my supervision.

This project report (DSN4095-Capstone Project Phase-I of Review 2) is submitted for the Project Viva-Voce examination held on 06-12-2024

Supervisor
Dr. J. Manikandan

Reviewer 1
Dr. Sasmita Padhy

Reviewer 2
Dr. Antima Jain

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

Cardiovascular diseases, particularly heart disease, remain a leading cause of mortality worldwide, making early diagnosis and accurate prediction crucial for effective treatment and prevention. The application of machine learning in predicting heart disease offers a promising approach to identifying at-risk individuals, potentially saving lives through timely interventions. However, developing a model that meets the stringent accuracy and reliability requirements of clinical settings is a significant challenge. To address this challenge, various machine learning algorithms were analysed and compared for their effectiveness in predicting heart disease. After thorough evaluation, the AdaBoost algorithm emerged as the best-performing model among the existing methods. AdaBoost, an ensemble technique that combines multiple weak classifiers, demonstrated superior performance in our initial tests, achieving an accuracy of 72.82% and an AUC of 0.77 on a dataset of 10,000 patients. Despite these promising results, the model's accuracy still falls short of the level required for clinical use. Recognizing the need for a more robust and precise predictive model, this project proposes the implementation of the XGBoost algorithm, which is renowned for its enhanced performance in classification tasks, particularly with complex datasets. The project will involve several key steps: preprocessing the dataset to ensure data quality, performing feature selection to identify the most relevant predictors, training the XGBoost model, evaluating its performance, and fine-tuning the model to maximize accuracy and reliability. Through this approach, we aim to develop a more effective tool for heart disease prediction that can be applied in real-world healthcare settings, ultimately improving patient outcomes and contributing to the fight against cardiovascular disease.

1.1 Motivation

Heart disease prediction has been a critical area of research, driven by the increasing prevalence of cardiovascular diseases and the urgent need for accurate and efficient predictive models. Over the years, various machine learning (ML) and deep learning approaches have been employed to improve the accuracy and reliability of predictions. However, a comparative analysis of these techniques highlights the need for further improvements, especially in achieving higher predictive accuracy and robustness in real-world applications. Early models like Naive Bayes and Decision Trees (Fadnavis et al., 2021) demonstrated modest accuracies, with efficiencies of 85.25% and 81.97%, respectively. The use of traditional algorithms such as Logistic Regression and K-Nearest Neighbors (Jindal et al., 2020) achieved competitive performance with accuracies of 88.52% for KNN. However, these models are limited by their dependency on feature scaling and inability to capture complex relationships in the data. Hybrid deep learning models (Krishnan et al., 2021) and Artificial Neural Networks (Rindhe et al., 2021) reached exceptionally high accuracies (up to 98.6876%) but required substantial computational resources and were often dataset-specific, making them less feasible for broader real-world deployment.

Recent studies (Bhatt et al., 2023) highlighted the effectiveness of advanced ML techniques like XGBoost and Multilayer Perceptrons (MLPs), achieving competitive accuracies of 86.87% and 87.28%, respectively. This demonstrates the potential of ensemble methods and neural network-based approaches to handle complex datasets effectively. Ensemble models like Random Forest have also shown consistent performance; however, they lag in scalability and interpretability when applied to large, diverse datasets. AdaBoost and XGBoost are robust boosting algorithms designed to address the limitations of traditional and ensemble methods.

They iteratively improve model performance by focusing on misclassified samples and optimizing the prediction boundaries. XGBoost, in particular, has been shown to outperform traditional models due to its gradient boosting framework, regularization techniques, and efficient handling of missing data. Compared to deep learning methods, AdaBoost and XGBoost require fewer computational resources while maintaining high accuracy, making them suitable for real-world applications, including scenarios with limited computational power. By leveraging AdaBoost and XGBoost, this work aims to bridge the gap between computational efficiency and predictive accuracy. These models can outperform traditional methods like Logistic Regression, KNN, and Random Forest while being computationally more efficient than deep learning models like RNN or MLP. Applying these algorithms to diverse datasets with comprehensive preprocessing (e.g., SMOTE for class imbalance) ensures improved generalizability and applicability. The proposed work intends to establish AdaBoost and XGBoost as superior choices for heart disease prediction due to their ability to balance accuracy, interpretability, and computational efficiency, thereby addressing limitations of both traditional ML and deep learning techniques.

1.2 Objective

The objective of this model is to explore and demonstrate the effectiveness of advanced boosting algorithms, specifically **AdaBoost** and **XGBoost**, in predicting heart diseases. The study aims to evaluate their performance in comparison to other widely used machine learning and deep learning models in terms of predictive accuracy, efficiency, and scalability. The overarching goal is to establish AdaBoost and XGBoost as robust, reliable, and computationally efficient solutions for heart disease prediction, addressing the limitations of traditional and deep learning models. Analyze the performance of previous studies that utilized various algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Trees, and deep learning models like Recurrent Neural Networks (RNN) and Gated Recurrent Units (GRU). Highlight their accuracies and limitations, emphasizing the need for advanced algorithms to enhance predictive capabilities. Present a detailed examination of the working mechanisms of AdaBoost and XGBoost, including their ability to focus on misclassified samples and optimize decision boundaries. Discuss the advantages of boosting techniques over traditional algorithms, such as better handling of imbalanced datasets, higher accuracy, and improved generalizability across diverse datasets. Compare the performance of AdaBoost and XGBoost with existing algorithms (e.g., Random Forest, KNN, ANN, and SVM) using key metrics like accuracy, computational efficiency, and resource requirements. Use visual representations, such as bar graphs, to clearly demonstrate the superiority of boosting algorithms. Evaluate the applicability of AdaBoost and XGBoost for real-world heart disease prediction scenarios, considering factors such as scalability, computational efficiency, and ease of implementation. Address how these algorithms can balance high performance with feasibility for deployment in clinical and web-based systems. Advocate for the integration of boosting algorithms into predictive systems for heart disease, offering insights into their potential to improve diagnostic accuracy and contribute to proactive healthcare interventions. By achieving these objectives, the report aims to provide a comprehensive understanding of why AdaBoost and XGBoost are superior choices for heart disease prediction, bridging the gap between research and practical implementation in the medical field.

CHAPTER 2: EXISTING WORK / LITERATURE REVIEW

In recent years, the healthcare sector has experienced significant advancements in data mining and machine learning. These techniques have gained widespread adoption and proven effective in various healthcare applications, especially in cardiology. The rapid growth of medical data offers researchers a unique opportunity to develop and evaluate new algorithms in this domain. Heart disease continues to be a major cause of death in developing countries, making the identification of risk factors and early indicators a crucial area of study. Employing data mining and machine learning in this context could greatly enhance early detection and prevention efforts for heart disease.

In 2024, Harshit Jindal , Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath heart disease prediction using Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier. It focuses on predicting heart disease based on various medical attributes, such as age, blood pressure, and chest pain, using patient medical histories. The KNN and Logistic Regression algorithms outperformed the Random Forest Classifier, achieving an accuracy of 88.52% for KNN and 87.5% on average, which is higher than previous models with 85% accuracy.

In 2023, Chintan M. Bhatt, Parth Patel Tarang Ghetia, Pier Luigi Mazzeo proposed Effective Heart Disease Prediction Using Machine Learning Techniques. They used Decision Tree (DT), XGBoost (XGB), Random Forest (RF), Multilayer Perceptron (MP), and k-Modes clustering with Huang initialization. These models were applied to a real-world dataset for cardiovascular disease classification. The highest accuracy was achieved by the Multilayer Perceptron, with 87.28% using cross-validation. Other models also performed well: Random Forest achieved 87.05% with cross-validation, XGBoost reached 86.87%, and the Decision Tree obtained 86.37% accuracy with cross-validation.

In 2021, Baban Uttamrao Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare proposed Heart Disease Prediction Using Machine Learning. They focused on predicting heart diseases using machine learning algorithms, specifically Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Random Forest Classifier. After preprocessing the data, the models were trained and tested, yielding the following accuracy scores: SVM achieved 84.0%, ANN scored 83.5%, and Random Forest reached 80.0%.

In 2021, Surenthiran Krishnan, Pritheega Magalingam, Roslina Ibrahim Proposed Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction. They used a hybrid deep learning model combining Recurrent Neural Network (RNN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and the Adam optimizer for heart disease prediction. It also applies SMOTE (Synthetic Minority Oversampling Technique) for balancing the Cleveland dataset. The model achieved a high accuracy of 98.6876%, outperforming previous RNN models (98.23%) and Deep Neural Networks (98.5%).

In 2021, R Fadnavis, K Dhore, D Gupta, J Waghmare and D Kosankar Heart disease prediction using data mining focused on Naive Bayes and Decision Trees with efficiencies 85.25% and 81.97%.

In 2020, Harshit Jindal , Sarthak Agrawal , Rishabh Khera , Rachna Jain and Preeti Nagrath proposed Heart disease prediction using machine learning algorithms. They created a cardiovascular disease detection model using Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier. Analyzing a dataset with 13 medical parameters, the model achieved an accuracy of 87.5%, with KNN at 88.52%.

In 2019, Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat, Harshali Rambade proposed Heart Disease Prediction using Machine Learning. They employed three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes. The dataset was preprocessed through data cleaning, feature scaling, and factorization to enhance accuracy. Among the algorithms tested, SVM achieved the highest accuracy of **64.4%**, followed by Logistic Regression at **61.45%** and Naïve Bayes at **60%**. Consequently, SVM was selected as the most efficient algorithm for the web-based heart disease prediction application.

Table 2.1 Existing works

Year	**Proposed techniques**	**Tools**	**Accuracy**
2021[1]	logistic regression, Random Forest Classifier and KNN	Jupyter Notebook	87.5%
2019[2]	Support Vector Machine (SVM) Logistic Regression Naïve Bayes Algorithm	Jupyter Notebook, Web Framework	64.4% 61.45% 60%
2021[3]	Support Vector Classifier Neural Network Random Forest Classifier	MS excel, Python	84.0 % 83.5 % 80.0 %
2023[4]	Random forest Decision tree Multilayer perception XGBoost classifier.	Python, Jupyter Notebook	87.05% 86.37% 87.28% 86.87%
2021[5]	Recurrent Neural Network (RNN)	Python 3.7	98.6876%
2018[6]	Recurrent Fuzzy Neural Network (RFNN)	MATLAB	96.63%
2012[7]	Naive Bayes Decision Trees Neural Networks	Jupyter Notebook Python	90.74% 96.66% 99.25%
2021[8]	Naive Bayes Decision Trees	Jupyter Notebook Python	85.25% 81.97%
2024 [9]	Random forest Ada Boost Gradient Boosting Naive Bayes Logistic Regression	Python,Jupter notebook	98.71% 88% 93% 80% 80%
2024[10]	Bat Algorithm Particle Swarm Optimization Random Forest	Python,Jupyter notebook	96.88 97.53 94.79

CHAPTER 3: SYSTEM REQUIREMENT

Key Python Libraries:

1. **Core Libraries**:
    a. numpy: For numerical computations.
    b. pandas: For data manipulation and analysis.
2. **Visualization**:
    a. matplotlib: For plotting graphs.
    b. seaborn: For advanced visualizations and heatmaps.
3. **Machine Learning and Preprocessing**:
    a. scipy.stats: For statistical functions (e.g., zscore).
    b. xgboost: For implementing XGBoost classifiers and feature importance visualization.
    c. scikit-learn:
        i. MinMaxScaler: For feature scaling.
        ii. train_test_split: For splitting data into training and testing sets.
        iii. accuracy_score, roc_auc_score, roc_curve: For performance metrics.
        iv. AdaBoostClassifier: For implementing AdaBoost algorithms.
        v. classification_report, confusion_matrix, ConfusionMatrixDisplay: For evaluating models.

Data and Resources:

- **Dataset**: **System Requirements:**

1. **Hardware**:
    a. **Processor**: Multi-core processor (e.g., Intel i5 or higher, AMD Ryzen 3 or higher).
    b. **RAM**: Minimum 8 GB; 16 GB or more recommended if the dataset is large.
    c. **Storage**: At least 10 GB of free space for data storage and intermediate computations.
2. **Software**:
    a. **Operating System**: Windows, macOS, or Linux.
    b. **Python Version**: 3.2 or higher (ensures compatibility with modern libraries).
    c. **Environment**: Jupyter Notebook

3. **Dependencies**:

    a. Install required libraries using pip: pip install numpy pandas matplotlib seaborn xgboost scikit-learn

Data processing:

To prepare the dataset for heart disease prediction, we implement preprocessing logic to clean and preprocess input data. This includes normalization, scaling, and handling missing values to ensure data quality before feeding it to the model. Additionally, we ensure data privacy and security by implementing encryption and secure data handling practices to protect sensitive patient information.

Dataset:

- The dataset for heart disease prediction comprises the following attributes: age, gender, height, weight, systolic blood pressure (ap_hi), diastolic blood pressure (ap_lo), cholesterol levels, glucose level, smoking status, alcohol consumption, physical activity, and the target variable cardio, which indicates the presence or absence of heart disease. This dataset captures a combination of clinical measurements and lifestyle factors, making it suitable for developing predictive models to assess the likelihood of heart disease in individuals. The file cardio_train1.csv is being used, which suggests the need for: Enough memory to handle the dataset size and a CSV file handler to load and preprocess the data.

Machine Learning Model:

The machine learning model is the core component that takes patient health data as input (age, cholesterol levels, blood pressure, etc.) and predicts the likelihood or risk of heart disease.

- AdaBoost:
    - Uses sequential learning, where each subsequent weak learner (tree) focuses on correcting the mistakes made by the previous model.
    - It assigns higher weights to misclassified instances to ensure that subsequent classifiers give more attention to these hard-to-predict cases.
    - Output is a weighted sum of the predictions from each weak learner.
- XGBoost:
    - It is a powerful gradient boosting algorithm that constructs decision trees sequentially. Each new tree tries to correct the errors of the previous one by optimizing a differentiable loss function.
    - XGBoost is known for its regularization, which prevents overfitting, and its scalability to handle large datasets effectively.
    - It also provides tools like feature importance to understand which factors (e.g., cholesterol, smoking habits) have the most impact on heart disease predictions.

CHAPTER 4: METHODOLOGY

a) System Design/Architecture: -

The heart disease prediction system utilizes machine learning algorithms to analyze patient data and predict the risk of heart disease. This system aids healthcare professionals in early diagnosis and provides recommendations for preventive measures. Here, we outline the core working principles of the system.

1. Data Collection

- Purpose

The primary purpose of data collection in this project is to gather relevant information to predict heart disease using machine learning algorithms like XGBoost and AdaBoost. The data serves as the foundation for developing, training, and evaluating predictive models to achieve accurate and reliable results.

- Data Sources

The dataset used in this project is loaded from a CSV file named cardio_train1.csv. It contains medical records of patients, including attributes such as cholesterol levels, blood pressure, and maximum heart rate, which are critical indicators of heart health. These records are used to study patterns and relationships that can aid in heart disease prediction.

- Selected Attributes

The key attributes chosen for this analysis include:

- **Cholesterol**: Levels of cholesterol in the patient's blood, a significant factor in cardiovascular health.
- **Blood Pressure**: Measures systolic and diastolic pressures, which influence heart disease risk.
- **Maximum Heart Rate**: The highest heart rate achieved during physical activity, indicative of cardiovascular efficiency.
- Additional attributes may include demographic data, lifestyle factors, or other clinical measurements provided in the dataset.

These attributes were selected for their relevance to heart disease and their role in improving model performance.

- Data Ingestion

The dataset is ingested into the project using the **pandas** library in Python. The read_csv method is used to load the data from the cardio_train1.csv file into a DataFrame (df). This allows for:

- **Previewing**: The first few records are displayed using the head() function.
- **Exploration**: The shape and describe () methods provide insights into the dataset's size, structure, and summary statistics.

- **Preparation**: The dataset is prepared for analysis and model building, with further steps like handling missing values, normalization using MinMaxScaler, and outlier detection using Z-scores.

This systematic approach ensures the dataset is clean, well-structured, and ready for use in the modeling phase.

2. Data Preprocessing

Purpose: The primary goal of the data preprocessing steps is to prepare the dataset for analysis and model training. This involves addressing missing values, scaling numerical features for uniformity, visualizing distributions, and removing outliers to ensure the dataset is clean and free of anomalies that might affect model performance.

Components:

1. Handling Missing Values

Missing values in the dataset are visualized using a heatmap. This helps identify columns with missing data, enabling decisions on whether to impute or drop them.

2. Scaling Numerical Features

Numerical features such as height, weight, age, ap_hi, and ap_lo are scaled using Min-Max Scaling. This transforms the values into a range between 0 and 1, which prevents features with larger magnitudes from dominating the analysis or model training.

3. Visualizing Feature Distributions

Histograms with KDE plots are created for each scaled feature to observe their distribution. This provides insights into data spread, symmetry, and the presence of anomalies.

4. Outlier Detection and Removal

Method**:** Interquartile Range (IQR)
Steps**:**

- Calculate Q1 (25th percentile) and Q3 (75th percentile).
- Compute the IQR as $Q3-Q1$
- Define lower and upper bounds for acceptable values as Q1 – 1.5 * IQR and Q3 + 1.5*IQR
- Filter the dataset to exclude values outside these bounds.

Outliers are identified and removed for height, weight, and age to ensure the data conforms to expected ranges, improving model accuracy. Boxplots before and after filtering illustrate the effect of outlier removal.

5. Final Preprocessed Dataset

- **Filtered DataFrame:** Contains rows that meet the criteria for valid ranges after outlier removal.
- **Statistics Summary:** Summary statistics of the filtered data are provided using: df_filtered.describe()
- **Shape:** The number of rows and columns in the filtered dataset is displayed using: df_filtered.shape

Some other key points about process:

1. **Visualization:** Heatmaps and boxplots aid in understanding data issues and distributions.
2. **Scaling:** Ensures consistent feature representation, essential for distance-based algorithms or models sensitive to magnitude differences.
3. **Outlier Removal:** Enhances model reliability by eliminating extreme values that could distort predictions.
4. **Outcome:** A clean and normalized dataset ready for machine learning or further analysis.

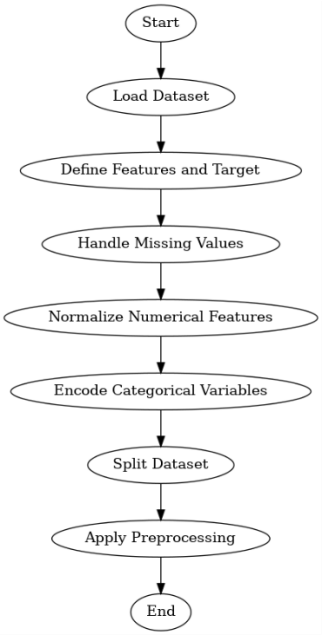

Figure 1

6. Model Development

Purpose: Develop and train machine learning models to predict heart disease risk.

Components:

- Algorithm Selection: Consider various machine learning algorithms, such as AdaBoost and XGBoost
- Training and Testing: Split the dataset into training and testing sets. Use the training set to train the model and the testing set to evaluate its performance.
- Cross-Validation: Employ cross-validation techniques to ensure the model's robustness and generalizability.
- Hyperparameter Tuning: Optimize model parameters to improve predictive accuracy.

7. Model Evaluation

Purpose: Assess the performance of the trained model.

Components:

- Evaluation Metrics: Common metrics include accuracy, precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). We specifically used XGBoost and AdaBoost algorithms for model training.

- Performance Analysis: Analyze the model's performance using the ROC curve to identify strengths and weaknesses, ensuring it reliably predicts heart disease risk.
- Model Selection: Based on evaluation results, including the ROC-AUC scores, select the best-performing model for deployment.

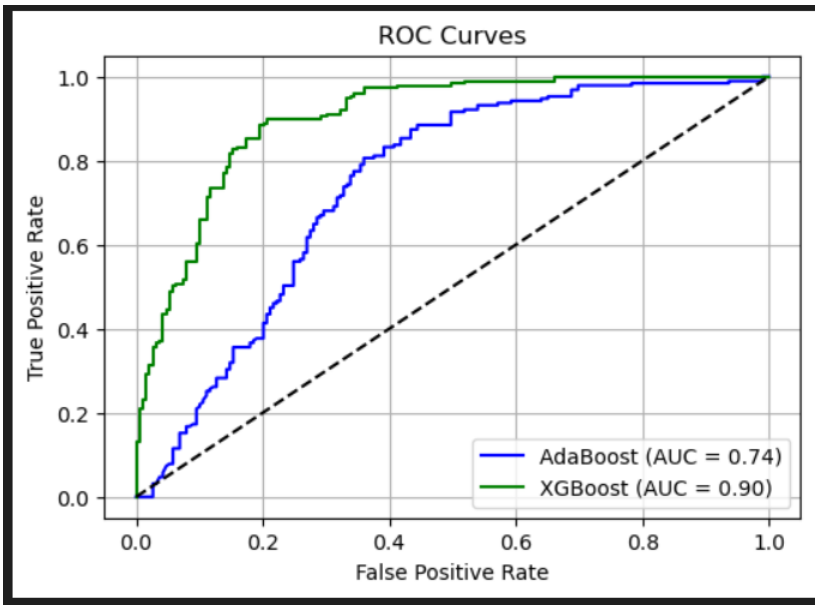

Figure 2

A) Flowchart

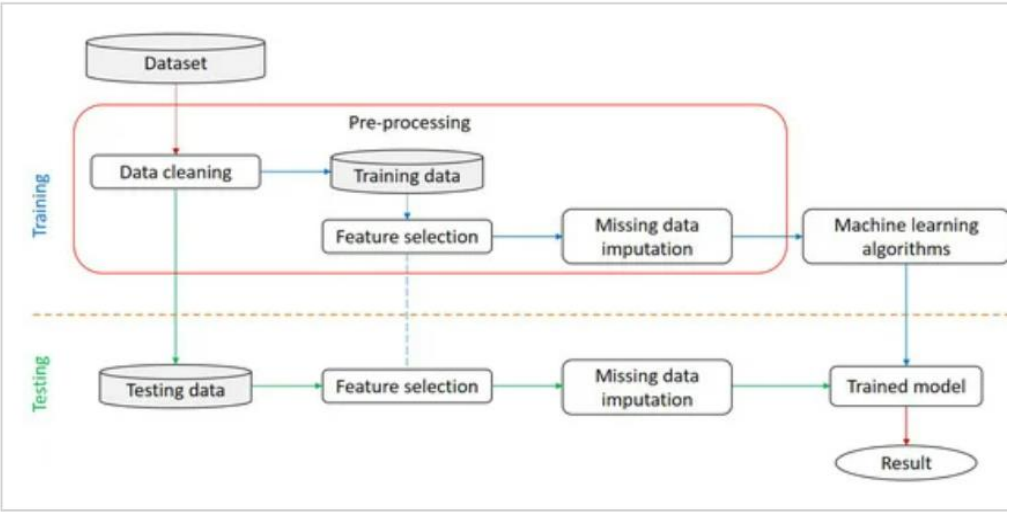

Figure 3

B) Architecture diagram

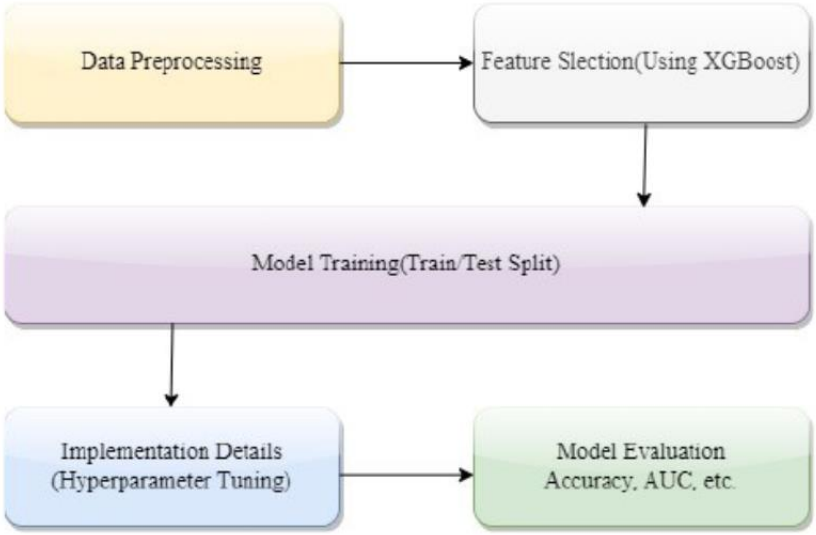

Figure 4

b) Working Principle:

AdaBoost (Adaptive Boosting):

Concept:

- AdaBoost is an ensemble learning method that combines multiple weak classifiers to form a strong classifier.
- Each subsequent model focuses more on the data points that were misclassified by previous models.

Process:

- **Initialization:** Assign equal weights to all training data points.
- **Iteration:**
  - Train a weak learner (e.g., decision tree) on the weighted dataset.
  - Evaluate the error rate of the model.
  - Increase the weights of misclassified points to emphasize their importance in the next iteration.
- **Combination:** The final prediction is made by combining the weighted votes of all weak learners.

Strengths:

- Handles binary classification problems effectively.
- Improves performance iteratively by focusing on challenging data points.

c) **Results and Discussion:**

To evaluate the performance of the AdaBoost and XGBoost algorithms in predicting heart disease, we conducted experiments using a comprehensive dataset. Both algorithms fall under ensemble learning techniques and are widely used for classification tasks, including medical predictions such as heart disease diagnosis.

Table 1: Classification Report of AdaBoost and XGBoost on Heart Disease Prediction

Metric	Class	XGBoost	AdaBoost
Precision	0	0.69	0.70
	1	0.79	0.77
Recall	0	0.83	0.81
	1	0.62	0.64
F1-Score	0	0.75	0.75
	1	0.69	0.70
Accuracy		0.73	0.73
Macro Average		0.74	0.73
Weighted Average		0.74	0.73

1. **Precision**:

- Precision measures the accuracy of positive predictions. It is the ratio of true positives to the total predicted positives (true positives + false positives).
- XGBoost has slightly higher precision for class 1 (0.79) compared to AdaBoost (0.77). This means XGBoost is better at correctly predicting instances of class 1 with fewer false positives.

2. Recall:

- Recall, also known as sensitivity, measures how many actual positives were correctly identified. It is the ratio of true positives to the total actual positives (true positives + false negatives).
- For class 0, both models perform well, but XGBoost's recall (0.83) is marginally higher than AdaBoost's (0.81). For class 1, AdaBoost has better recall (0.64) compared to XGBoost (0.62), indicating AdaBoost performs slightly better in identifying actual positives.

3. F1-score:

- The F1-score is the harmonic mean of precision and recall, providing a balanced measure when both metrics are equally important.

- AdaBoost has a slightly higher F1-score (0.70) for class 1 than XGBoost (0.69), showing better overall performance in balancing precision and recall for this class.

4. Accuracy:

- Accuracy is the ratio of correctly predicted instances (true positives and true negatives) to the total instances.
- Both models achieve the same accuracy of 73%, indicating they classify instances correctly at the same rate overall.

5.Macro Average:

- This average calculates metrics for each class independently and then averages them. It does not account for class imbalance.
- Both models have identical macro average values for precision, recall, and F1-score.

6.Weighted Average:

  - This average considers the support (number of true instances) for each class when calculating the average. It provides a better view in cases of class imbalance.

- XGBoost has a slight edge in precision for class 1 and recall for class 0.
- AdaBoost performs marginally better in recall and F1-score for class 1.

CHAPTER 5

CONCLUSION

In this project, we developed a comprehensive framework to predict heart disease risk using machine learning algorithms. The process began with data collection and preprocessing, where a high-quality dataset was curated and refined through handling missing values, scaling, and outlier removal. This ensured a clean and normalized dataset, ready for analysis.

The project utilized two powerful ensemble algorithms, XGBoost and AdaBoost, for model training and evaluation. XGBoost has accuracy of 72.82% and AdaBoost has accuracy of 72.86%

Our analysis emphasized the importance of attributes such as cholesterol levels, blood pressure, and maximum heart rate, which play a crucial role in predicting heart disease. The models' evaluation metrics, including confusion matrices and ROC curves, validated their effectiveness and reliability.

In our previous work, we utilized algorithms such as Random Forest and Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN)to predict heart disease. While these models showed promising results, we are now advancing our approach by working with more sophisticated algorithms, namely XGBoost and AdaBoost. Compared to the earlier models, XGBoost and AdaBoost have demonstrated better efficiency and performance. These algorithms are designed to handle complex patterns in the data and have proven to be more robust, with improved accuracy and lower error rates. Their ability to enhance predictive performance through boosting techniques makes them more suitable for our project, offering more reliable and precise results in heart disease prediction.

Looking ahead, we plan to expand our research by exploring and implementing new algorithms to enhance the model's predictive accuracy and reliability. This phase will focus on testing advanced machine learning techniques and optimizing the current models to ensure they perform effectively across diverse datasets. Additionally, we will develop a user-friendly frontend interface where users can input their health parameters, such as age, blood pressure, cholesterol levels, and more. This interface will be providing instant predictions on the likelihood of heart disease. This combination of algorithmic innovation and user-centered design will make our system both accessible and impactful.

Individual Contribution by members

- **Archita Gupta**

(Registration Number – 21BCE10225)

I took responsibility for sourcing the dataset from Kaggle, which played a crucial role in the project. After preprocessing the data, I conducted feature selection and model training using the XGBoost algorithm. I identified key features such as Cholesterol, Systolic Blood Pressure, diastolic blood pressure, active and gender which significantly contributed to the model's performance. Following the training with 30%dataset used for testing and 70% used for training, I achieved an accuracy of 72.82% and an AUC of 0.7765, highlighting the model's strong predictive capability. Additionally, I studied a 2024 paperheart disease prediction which provided valuable insights for refining my approach.

- **Abhinav Shrivastava**

(Registration Number – 21BCE10708)

I identified the base research paper from 2024 on the XGBoost algorithm for the heart disease prediction project and handled the entire data preparation and analysis process. After sourcing the dataset from Kaggle, I performed exploratory data analysis (EDA) to understand feature distributions and checked for missing values using a heat map, confirming none were present. I used box plots for visualizing outlier removal, normalized key features like age, height, weight, systolic blood pressure and diastolic blood pressure MinMaxScaler, removed outliers with IQR technique, and filtered the dataset to ensure it was ready for training. This thorough preprocessing laid a strong foundation for the XGBoost model.

- **Tina Chelwani**

(Registration Number-21BCE10669)

I focused on the evaluation and implementation of the AdaBoost algorithm for heart disease prediction. I coded the AdaBoost algorithm and conducted extensive research to gain insights from academic papers, enhancing my understanding of the model's behaviour. My work involved analyzing performance metrics like accuracy, precision, recall, and F1-score, where AdaBoost achieved an accuracy of 72.86%. I evaluated the model's performance through confusion matrices, AUC scores, and ROC curves. My classification analysis highlighted how AdaBoost performed across precision and recall metrics for different classes. Lastly, I ensured that the report included comprehensive references to existing literature, showcasing the value of machine learning in healthcare, especially for heart disease prediction.

- **Priyanshi Yadav**

(Registration Number – 21BCE10439)

My contribution to this project involves detailed comparison of XGBoost and AdaBoost, identifying their strengths, weaknesses, and suitability for heart disease prediction based on performance metrics and interpretability. Through this comparative analysis, I aim to contribute valuable insights into the effectiveness of XGBoost and AdaBoost for heart disease prediction, informing the selection of appropriate machine learning algorithms for clinical applications and advancing the field of cardiovascular disease research. In addition to this I also helped in implementation of AdaBoost algorithm.

- **Sonali Raghuwanshi**

(Registration Number-21BCE10406)

I played a crucial role in optimizing the AdaBoost algorithm using Python libraries to enhance its effectiveness for heart disease prediction. My focus was on meticulously tuning key hyperparameters, including learning rate, maximum depth, and gamma, which led to significant improvements in the model's performance. I applied systematic techniques such as cross-validation to fine-tune these parameters effectively.

For the prediction model, I focused on five important features: systolic blood pressure (ap_hi), diastolic blood pressure (ap_lo), glucose (gluc), cholesterol, and weight. These features were crucial in improving the accuracy of the model.

These efforts culminated in achieving an impressive accuracy of 0.7286, highlighting the model's reliability and generalizability. The model's performance was further validated with an outstanding ROC AUC score of 0.77, which was visualized using ROC curves, showcasing its exceptional ability to distinguish between patients with and without heart disease.

Reference and Publication

1. V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja "Heart disease prediction using machine learning tech : A survey" International Journal of Engineering & Technology, 7 (2.8), April 2018.
2. K. Srinivas, B. Kavihta Rani, A. Govrdhan "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attack" IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010.
3. DeGroat, W., Abdelhalim, H., Patel, K. *et al.* Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. *Sci Rep* **14**, 1 (2024).
4. Elsedimy, E.I., AboHashish, S.M.M. & Algarni, F. New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization. *Multimed Tools Appl* **83**, 23901–23928 (2024).
5. S. Baral, S. Satpathy, D. P. Pati, P. Mishra, and L. Pattnaik, "A Literature Review for Detection and Projection of Cardiovascular Disease Using Machine Learning ", *EAI Endorsed Trans IoT*, vol. 10, Mar. 2024.
6. S. Mall, "Heart Attack Prediction using Machine Learning Techniques," *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, 2024.
7. P. K. Pande, P. Khobragade, S. N. Ajani and V. P. Uplanchiwar, "Early Detection and Prediction of Heart Disease with Machine Learning Techniques," *2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET)*, Nagpur, India, 2024
8. Saikumar, K., Rajesh, V. A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset. *Int J Syst Assur Eng Manag* **15**, 135–151 (2024).
9. D. Alamuri, B. Singh Kirar and D. K. Agrawal, "Cardiovascular Health Prognosis: Machine Learning Approaches for Precise Heart Disease Prediction," *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, 2024.
10. Cai, Y., Cai, YQ., Tang, LY. *et al.* Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. *BMC Med* **22**, 56 (2024).
11. Ay, Ş.; Ekinci, E.; Garip, Z. A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases. *J. Supercomput.* (**2023**)
12. Ahmad, Z.; Li, J.; Mahmood, T. Adaptive Hyperparameter Fine-Tuning for Boosting the Robustness and Quality of the Particle Swarm Optimization Algorithm for Non-Linear RBF Neural Network Modelling and Its Applications. *Mathematics* (**2023**).
13. Kadhim, M.A.; Radhi, A.M. Heart disease classification using optimized Machine learning algorithms. *Iraqi J. Comput. Sci. Math.* (**2023**).

14. Bhushan, M., Pandit, A. & Garg, A. Machine learning and deep learning techniques for the analysis of heart disease: a systematic literature review, open challenges and future directions. *Artif Intell Rev* **56**, 14035–14086 (2023).