## Optimization of Heart Disease Prediction using Machine Learning Model CAPSTONE PROJECT PHASE-II

**Phase – II Report**

**Submitted by**

**21BCE10225 Archita Gupta 21BCE10406 Sonali Raghuwanshi 21BCE10439 Priyanshi Yadav 21BCE10669 Tina Chelwani 21BCE10708 Abhinav Shrivastava**

**in partial fulfillment of the requirements for the degree of Bachelor of Technology**

****

**VIT Bhopal University Bhopal**

**Madhya Pradesh**

**January, 2025**



## 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-II of Review 1) is submitted for the Project Viva-Voce examination held on 21-01-2025

**Supervisor Dr. J. Manikandan**

**Reviewer 1 Dr. Sasmita Padhy**

**Reviewer 2 Dr. Antima Jain**

# TABLE OF CONTENT

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Topic** | **Page No.** |
| 1.  1.1  1.2 | Introduction  Motivation  Objective | 4  5,6  6 |
| 2. | Existing Work / Literature Review | 7,8 |
| 3. | System Requirement | 9,10 |
| 4.  4.1  4.2  4.3 | Methodology  System Design/Architecture Working Principle  Results and discussions | 11  11-16  17  17-20 |
| 5. | Conclusion | 21 |
| 6. | Individual Contribution by Members | 22,23 |
| 7. | Reference And Publication | 24 |

**CHAPTER 1: INTRODUCTION**

Heart disease remains a leading cause of mortality worldwide, posing a significant challenge to global healthcare systems. Accurate and early detection of heart disease is critical in reducing the risk of severe complications, enabling timely interventions and improving patient outcomes. With advancements in machine learning, predictive models are now capable of analyzing complex medical data to uncover patterns that are often overlooked by traditional statistical methods.

The application of advanced machine learning techniques offers a transformative approach to identifying at-risk individuals, enabling timely interventions and potentially saving lives. However, achieving the high level of accuracy, interpretability, and reliability required for clinical applications remains a significant challenge.

To address this, various machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), were initially explored. While these models demonstrated promising results, their limitations in handling complex, imbalanced datasets and non-linear relationships highlighted the need for more advanced techniques. Subsequently, boosting algorithms like AdaBoost, XGBoost, and Gradient Boosting were analyzed for their effectiveness in heart disease prediction. Among these, Gradient Boosting emerged as the top-performing algorithm, achieving an accuracy of **74.44%** and an AUC-ROC score of **0.81**, followed by AdaBoost and XGBoost with comparable but slightly lower performance metrics.

This project aims to leverage the strengths of boosting algorithms, particularly Gradient Boosting, which optimizes predictions by sequentially correcting errors through gradient descent. The focus will also include XGBoost for its scalability, regularization capabilities, and superior handling of missing data. The project encompasses critical steps, including data preprocessing to ensure quality, feature selection to identify significant predictors like cholesterol, blood pressure, and physical activity, and model fine-tuning to maximize accuracy and robustness.

This project aims to leverage the power of machine learning, focusing on ensemble techniques such as AdaBoost, XGBoost, and Gradient Boosting, to develop an efficient predictive model for heart disease. These algorithms are selected for their ability to handle complex datasets, manage imbalanced data, and provide insights into feature importance, making them particularly suited for this task. By combining clinical measurements and lifestyle factors, this model aspires to offer a robust solution for heart disease prediction.

By developing and comparing these advanced models, the project aspires to create a reliable tool for heart disease prediction that meets the demands of real-world healthcare settings. This tool will empower healthcare professionals with actionable insights, improving patient outcomes and contributing to the global fight against cardiovascular disease.

* 1. **Motivation**

The increasing prevalence of heart disease and its impact on global health have highlighted the need for innovative approaches to early detection and prevention. Traditional diagnostic methods, while effective, often require invasive procedures and may not capture the multifaceted nature of heart disease risk factors. Machine learning, with its ability to process large volumes of data and detect subtle relationships, provides an opportunity to revolutionize cardiovascular disease prediction.

Boosting algorithms, specifically AdaBoost, XGBoost, and Gradient Boosting, stand out for their accuracy, robustness, and interpretability. Their capacity to focus on misclassified samples and manage class imbalances makes them ideal for medical applications where data quality and precision are paramount. This project is motivated by the potential of these algorithms to enhance healthcare outcomes by enabling data-driven, non-invasive, and scalable solutions for heart disease prediction.

Heart disease is a significant global health issue, affecting millions of people annually. Early prediction of heart disease can greatly improve outcomes by facilitating timely intervention and treatment. Machine learning offers powerful tools for analyzing large and complex datasets, uncovering patterns that traditional statistical methods may overlook. Among these tools, boosting algorithms—such as AdaBoost, XGBoost, and Gradient Boosting—have demonstrated exceptional performance in classification tasks due to their ability to combine weak learners into strong predictive models. These algorithms are particularly well-suited for medical applications, where precision and reliability are critical.

Boosting algorithms offer several advantages, making them ideal for heart disease prediction. Their accuracy and robustness stem from their iterative approach to improving predictions by focusing on difficult-to-classify samples. This results in highly accurate models capable of handling complex datasets. In heart disease prediction, where small inaccuracies can have significant consequences, these algorithms help minimize errors, making them highly dependable for clinical applications. Furthermore, boosting algorithms provide insights into feature importance, allowing healthcare professionals to identify critical risk factors, such as blood pressure, cholesterol, and glucose levels, that contribute to heart disease.

Handling imbalanced datasets is another strength of boosting algorithms. Heart disease datasets often exhibit class imbalances, with fewer cases of disease compared to healthy individuals. Boosting methods effectively address this challenge by adjusting weights to focus on minority classes, ensuring balanced and reliable predictions. Additionally, the comparative performance of different boosting algorithms highlights their unique strengths. For example, AdaBoost is simple and effective for smaller datasets, although it may struggle with noisy data. XGBoost, optimized for speed and performance with advanced regularization techniques, is particularly suited for large-scale datasets. Gradient Boosting, while versatile and incrementally building strong models, may require longer training times compared to XGBoost.

In practical healthcare applications, boosting algorithms stand out for their interpretability and reliability. Their transparency makes them invaluable for clinical decision-making, where understanding model predictions is essential. By leveraging the strengths of these algorithms, this project sets out to evaluate the effectiveness of AdaBoost, XGBoost, and Gradient Boosting in predicting heart disease. The study will compare their performance based on metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Additionally, it will analyze computational efficiency, focusing on training and inference times for real-time healthcare applications. The project will also identify key clinical and lifestyle factors influencing heart disease through feature importance analysis.

Ultimately, this research aims to advance the development of machine learning-based tools for heart disease prediction. By improving healthcare outcomes and contributing to the growing body of medical AI research, the project aspires to create reliable, interpretable, and efficient predictive models that can significantly impact patient care.

* 1. **Objective**

Heart disease significantly affects the structure and function of the heart, impacting blood vessels, rhythm, or muscle performance. It remains the leading cause of death globally, encompassing various conditions such as Coronary Artery Disease (CAD), which involves the narrowing or blockage of blood vessels; arrhythmias, characterized by irregular heartbeats; heart failure, where the heart pumps blood ineffectively; congenital heart defects, which are structural abnormalities present from birth; cardiomyopathy, involving diseases of the heart muscle; and heart valve diseases, which result from valve dysfunction. Early detection and intervention are critical to mitigating the devastating impact of these conditions.

Machine learning offers a transformative approach to heart disease prediction by processing large medical datasets to uncover hidden patterns and risk factors. This capability enables early and accurate detection of heart disease, empowering healthcare professionals to make data-driven decisions. Machine learning models are non-invasive, scalable, cost-effective, and precise, supporting timely interventions that improve patient outcomes. These attributes make machine learning an invaluable tool in addressing the global burden of heart disease.

This project aims to develop a robust predictive model using advanced boosting algorithms, including AdaBoost, XGBoost, and Gradient Boosting, for heart disease detection. By comparing and analyzing the performance of these algorithms using evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, the study will identify the most effective model for this critical application. Additionally, it seeks to identify and rank the most significant features contributing to heart disease prediction, such as cholesterol levels, blood pressure, and glucose levels. The insights gained from this analysis will provide valuable guidance to healthcare professionals, aiding in early detection and prevention strategies to combat heart disease effectively.

# 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%.

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Proposed techniques** | **Tools** | **Accuracy** |
| 2021[[1]](https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf) | logistic regression, Random Forest  Classifier and KNN | Jupyter Notebook | 87.5% |
| 2019[[2]](https://www.ijresm.com/Vol.2_2019/Vol2_Iss2_February19/IJRESM_V2_I2_89.pdf) | Support Vector Machine (SVM) Logistic Regression  Naïve Bayes Algorithm | Jupyter Notebook, Web Framework | 64.4%  61.45%  60% |
| 2021[[3]](https://www.researchgate.net/publication/351545128_Heart_Disease_Prediction_Using_Machine_Learning) | Support Vector Classifier Neural Network  Random Forest Classifier | MS excel, Python | 84.0 %  83.5 %  80.0 % |
| 2023[[4]](https://www.mdpi.com/1999-4893/16/2/88) | Random forest Decision tree  Multilayer perception XGBoost classifier. | Python, Jupyter Notebook | 87.05%  86.37%  87.28%  86.87% |
| 2021[[5]](https://ijece.iaescore.com/index.php/IJECE/article/view/24402/15276) | Recurrent Neural Network (RNN) | Python 3.7 | 98.6876% |
| 2018[[6]](https://www.jetir.org/papers/JETIR1811A41.pdf) | Recurrent Fuzzy Neural Network (RFNN) | MATLAB | 96.63% |
| 2012[[7]](https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f4de0213b4a5777ff39d5a94cd574713799ca221) | Naive Bayes  Decision Trees Neural Networks | Jupyter Notebook Python | 90.74%  96.66%  99.25% |
| 2021[[8]](https://iopscience.iop.org/article/10.1088/1742-6596/1913/1/012099/pdf) | Naive Bayes Decision Trees | Jupyter Notebook Python | 85.25%  81.97% |
| 2024 [[9]](https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf) | Random forest Ada Boost  Gradient Boosting Naive Bayes  Logistic Regression | Python,Jupter notebook | 98.71%  88%  93%  80%  80% |
| 2024[[10]](https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf) | Bat Algorithm  Particle Swarm Optimization Random Forest | Python,Jupyter notebook | 96.88  97.53  94.79 |

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

# CHAPTER 3: SYSTEM REQUIREMENT

## Key Python Libraries:

* + 1. **Core Libraries**:
       1. numpy: For numerical computations.
       2. pandas: For data manipulation and analysis
    2. **Visualization**:
       1. matplotlib: For plotting graphs.
       2. seaborn: For advanced visualizations and heatmaps.
    3. **Machine Learning and Preprocessing**:
       1. scipy.stats: For statistical functions (e.g., zscore).
       2. xgboost: For implementing XGBoost classifiers and feature importance visualization.
       3. scikit-learn:
          1. MinMaxScaler: For feature scaling.
          2. train\_test\_split: For splitting data into training and testing sets.
          3. accuracy\_score, roc\_auc\_score, roc\_curve: For performance metrics.
          4. AdaBoostClassifier: For implementing AdaBoost algorithms.
          5. GradientBoostingClassifier: For implementing Gradient Boosting algorithms.
          6. classification\_report, confusion\_matrix, ConfusionMatrixDisplay: For evaluating models.

**Data and Resources:**

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

1. **Hardware**:
   1. **Processor**: Multi-core processor (e.g., Intel i5 or higher, AMD Ryzen 3 or higher).
   2. **RAM**: Minimum 8 GB; 16 GB or more recommended if the dataset is large.
   3. **Storage**: At least 10 GB of free space for data storage and intermediate computations.
2. **Software**:
   1. **Operating System**: Windows, macOS, or Linux.
   2. **Python Version**: 3.2 or higher (ensures compatibility with modern libraries).
   3. **Environment**: Jupyter Notebook
3. **Dependencies**:
   1. 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.
  + Gradient Boosting
    - Gradient Boosting is an ensemble learning algorithm that builds decision trees sequentially, correcting previous errors by optimizing a loss function. It prevents overfitting through regularization and efficiently scales to handle large datasets.
    - The algorithm also provides feature importance, helping identify key factors—such as cholesterol and blood pressure—that significantly impact heart disease predictions.

# CHAPTER 4: METHODOLOGY

1. 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, AdaBoost and Gradient Boosting algorithm. 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.

1. **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.

1. 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.

1. 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.

1. 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.

1. 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.

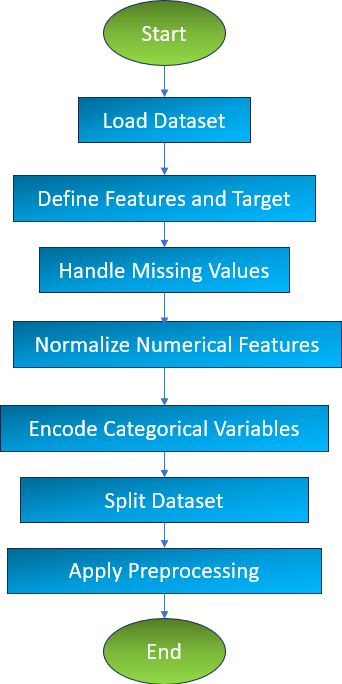


Figure 1

1. **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, Gradient.
  + 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.

1. **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, AdaBoost, Gradient 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.

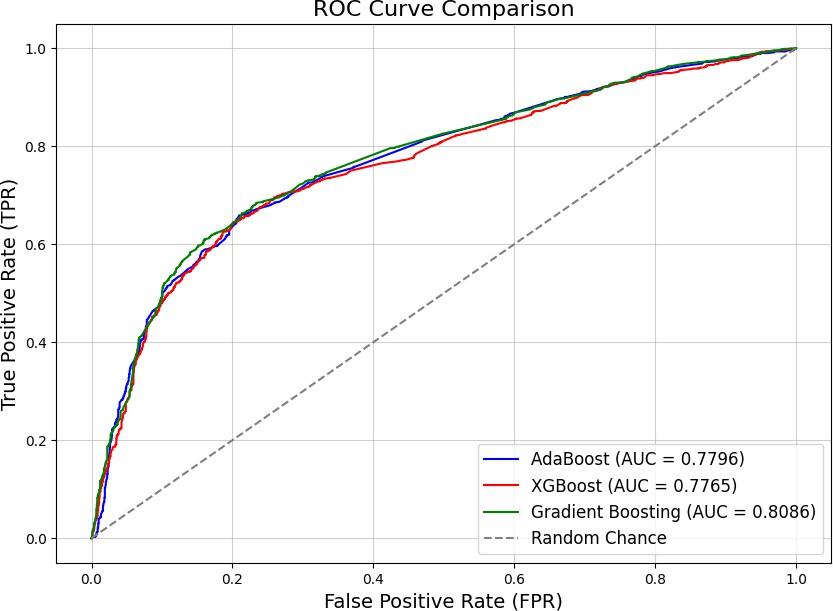


Figure 2

1. **Flowchart**

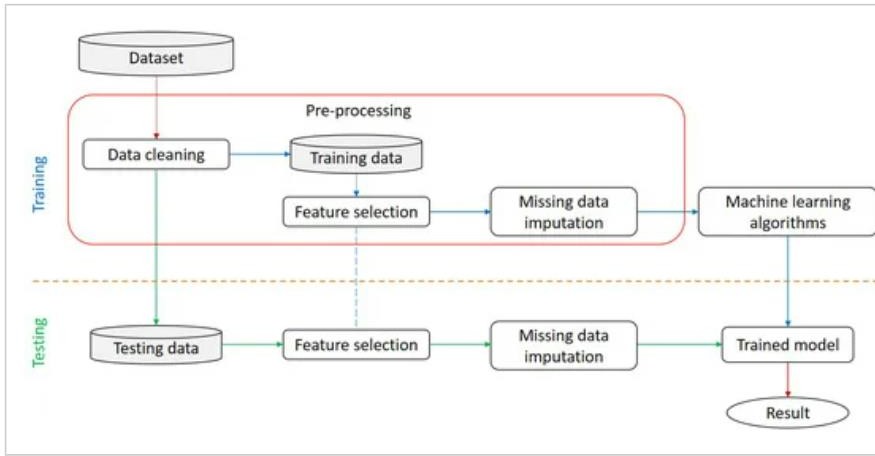
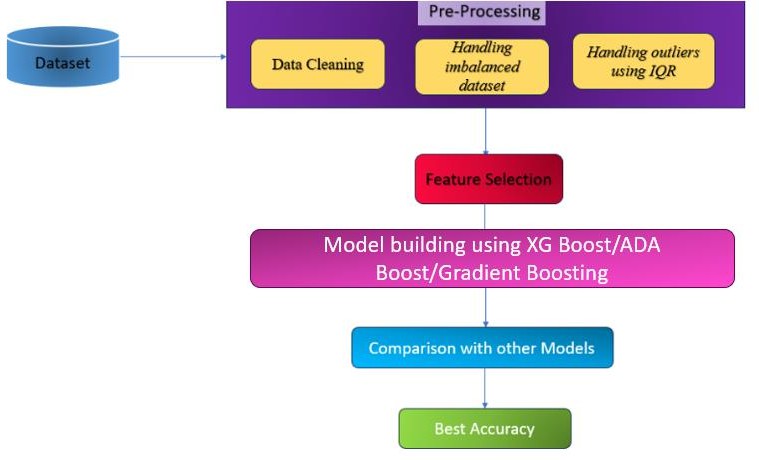


Figure 3

Architecture diagram

**Figure 4**



1. **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.

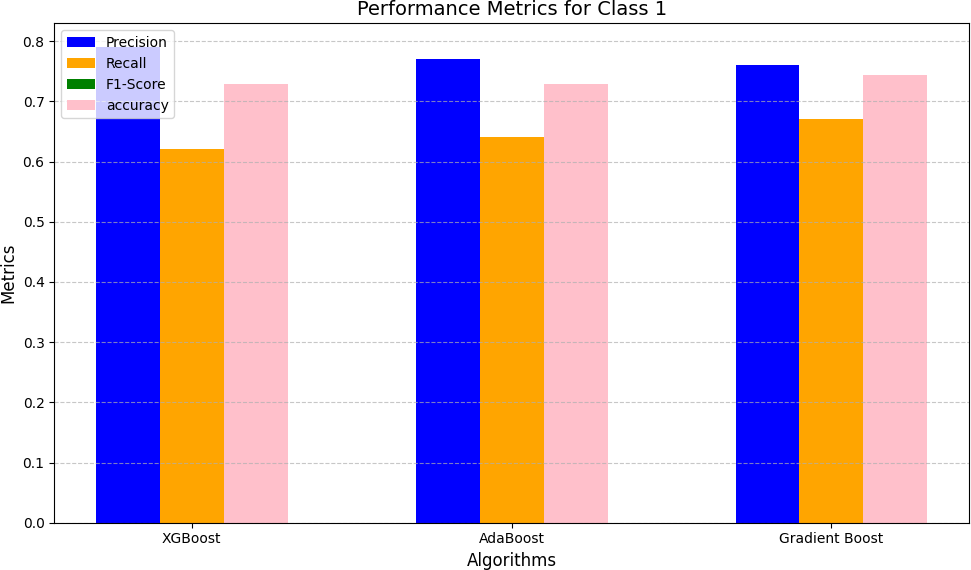
1. **Results and Discussion:**

To evaluate the performance of the AdaBoost, XGBoost and Gradient 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 , XGBoost and Gradient on Heart Disease Prediction**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Class | XGBoost | AdaBoost | Gradient Boost |
| Precision | 0 | 0.69 | 0.70 | 0.70 |
|  | 1 | 0.79 | 0.77 | 0.76 |
| Recall | 0 | 0.83 | 0.81 | 0.79 |
|  | 1 | 0.62 | 0.64 | 0.67 |
| F1-Score | 0 | 0.75 | 0.75 | 0.74 |
|  | 1 | 0.69 | 0.70 | 0.71 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy |  | 0.73 | 0.73 | 0.73 |
| Macro Average |  | 0.74 | 0.73 | 0.73 |
| Weighted Average |  | 0.74 | 0.73 | 0.73 |

****

1. **Precision**:
   * Gradient Boosting's precision (0.76) is slightly lower than XGBoost (0.79) but higher than AdaBoost (0.77).
   * Gradient Boosting is better suited for scenarios requiring balanced precision and recall, given its strong overall metrics.
   * Gradient Boosting's classification report further highlights its balanced performance, achieving moderate precision (0.76) and recall (0.67) for class 1, with an overall accuracy of 73%.
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, XGBoost achieves the highest recall (0.83), followed by AdaBoost (0.81) and Gradient Boosting (0.79).
   * For class 1, Gradient Boosting outperforms both models with a recall of 0.67, indicating it is the best at identifying true positives with fewer false negatives, followed by AdaBoost (0.64) and XGBoost (0.62).
3. **F1-score:**
   * For class 1, Gradient Boosting achieves the highest F1-score (0.71), followed by AdaBoost (0.70) and XGBoost (0.69).
   * Gradient Boosting provides the most balanced performance, excelling in harmonizing precision and recall for both classes, making it the best option when both metrics are equally important.
4. **Accuracy:**
   * Accuracy is the ratio of correctly predicted instances (true positives and true negatives) to the total instances.
   * The Gradient Boosting model achieves an accuracy of 73%, indicating it classifies instances correctly at the same rate overall. Both AdaBoost and XGBoost also achieve this accuracy, with each using unique techniques—Gradient Boosting focuses on combining trees, AdaBoost assigns weights to misclassified instances, and XGBoost incorporates advanced optimization for improved performance.
5. **Macro Average:**
   * Calculates precision, recall, and F1-score for each class independently and averages them.
   * Does not account for class imbalance.
   * In Gradient Boosting, AdaBoost, and XGBoost, the macro average values for precision, recall, and F1-score are nearly identical.
   * All models maintain balanced performance across both classes.
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.
   * **Gradient Boosting**: The precision for class 1 and recall for class 0 are slightly improved compared to AdaBoost.
   * It balances performance well across both classes, with consistent F1-scores and higher recall for class 0 compared to AdaBoost.
   * Similar to AdaBoost, Gradient Boosting also performs slightly better in handling class imbalance while maintaining overall accuracy.

# 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 three powerful ensemble algorithms, XGBoost, AdaBoost, and Gradient Boosting, for model training and evaluation. XGBoost achieved an accuracy of **72.82%**,

AdaBoost achieved **72.86%**, and Gradient Boosting achieved **74.44%**, showcasing its superior performance among the three models.

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, Support Vector

Machines (SVM), K-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), and 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, AdaBoost, and Gradient Boosting. Compared to the earlier models, these algorithms have demonstrated better efficiency and performance. They 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 provide 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. In this research, the dataset was inherently balanced with 5,030 cases classified as negative and 4,969 as positive, eliminating the need for oversampling or undersampling techniques like SMOTE (Synthetic Minority Over-sampling Technique). Following the

training with 30%dataset used for testing and 70% used for training, I achieved an accuracy of 72.86% and an AUC of 77%, highlighting the model's strong predictive capability.

Additionally, I studied a 2023 paper by Juan Carlos Yang on heart disease prediction using XGBoost, which provided valuable insights for refining my approach.

* + **Abhinav Shrivastava**

**(Registration Number – 21BCE10708)**

I identified the base research paper from 2023 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 removed outliers with Z-scores, 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 72.82%. 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).