A PROJECT REPORT
on

# “HEART DISEASE PREDICTION”

Submitted to
KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR’S DEGREE IN
COMPUTER SCIENCE

.

UNDER THE GUIDANCE OF
JAY SARRAF

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
April 2025

## Abstract

This report presents a comprehensive analysis of machine learning models designed to predict heart disease based on patient data.

Three classification approaches **:-**

—Logistic Regression, Random Forest, and XGBoost—

were implemented and evaluated using the Cleveland Heart Disease dataset. The models demonstrate high predictive accuracy, with the Random Forest classifier achieving the best overall performance (86.67% accuracy). This study provides insights into significant cardiovascular risk factors and offers a deployable prediction system that can assist healthcare professionals in early disease detection and intervention planning.



## Keywords :-

Heart Disease Prediction

Machine Learning

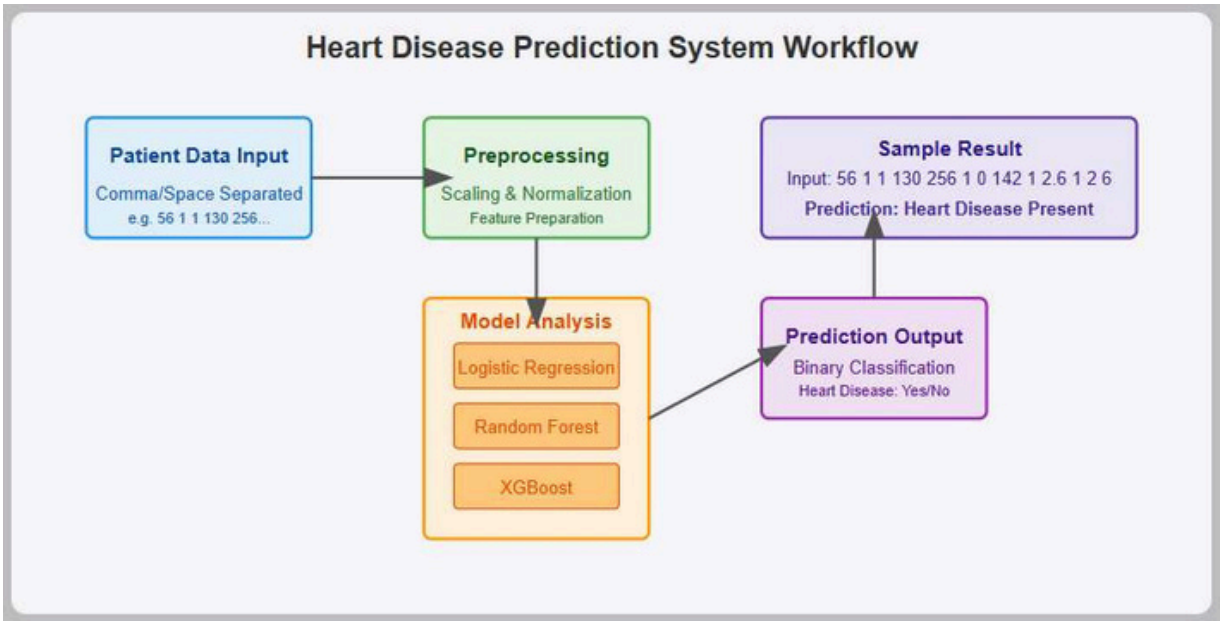
Random Forest Classification

Medical Diagnostics

Healthcare Analytics

# Chapter 1

## Executive Summary




This report presents the development and evaluation of machine learning models aimed at predicting heart disease using the Cleveland Heart Disease dataset. The primary objective was to build a predictive system that can assist in early detection of cardiovascular conditions, thereby supporting timely medical interventions and improving patient outcomes.

Three classification models—**Logistic Regression**, **Random Forest**, and **XGBoost**—were implemented and compared based on their predictive performance. The dataset contained 303 patient records and 13 medical attributes, including age, chest

pain type, cholesterol levels, thalassemia, and number of major vessels. The dataset was preprocessed by handling missing values, applying feature scaling, and performing a stratified 80/20 train-test split to maintain class balance.

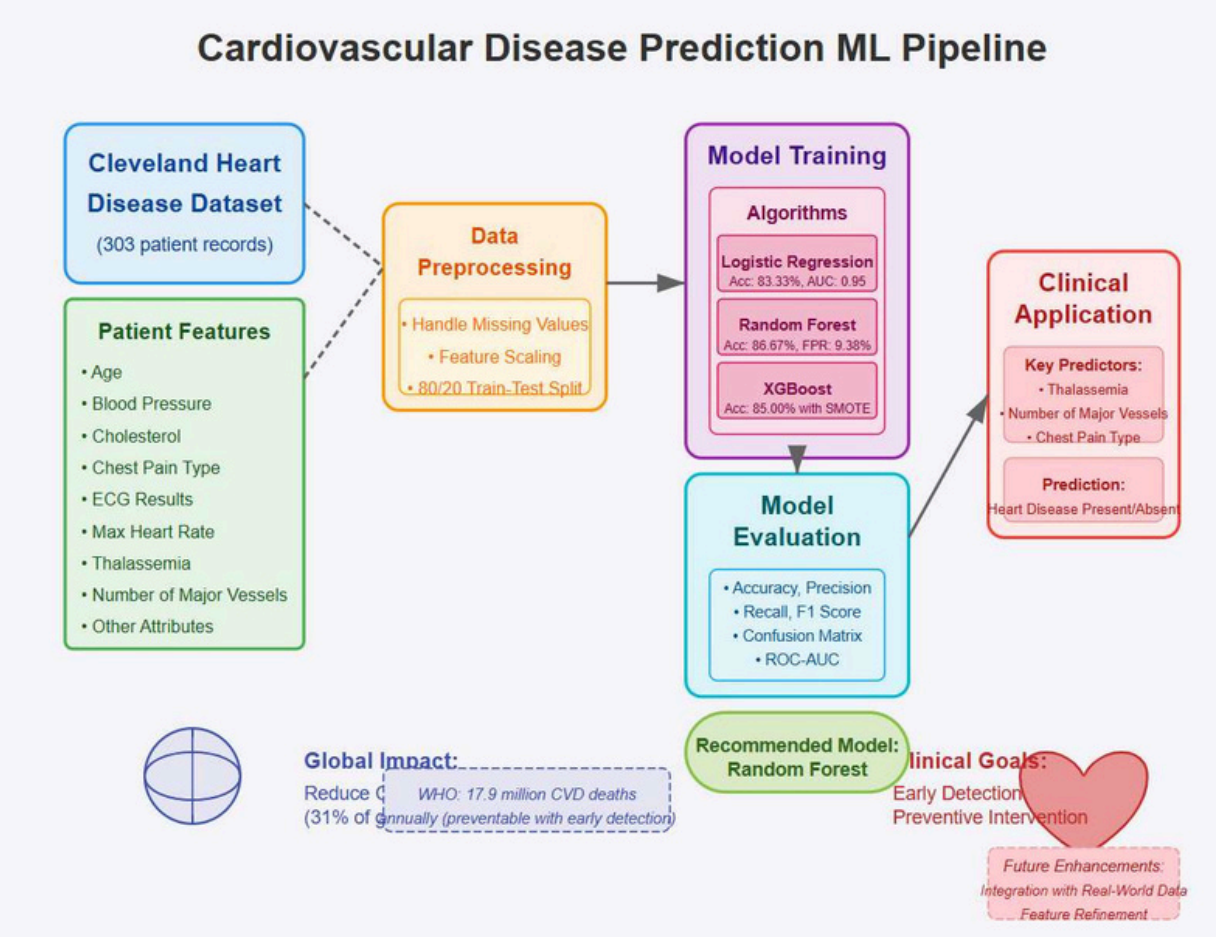
**Logistic Regression** was used as a baseline model due to its simplicity and interpretability. It achieved an accuracy of **83.33%** and the highest ROC-AUC score (**0.9498**), indicating good class discrimination ability. **Random Forest**, an ensemble learning method, achieved the best overall performance with an **accuracy of 86.67%**, and a **false positive rate of just 9.38%**, making it a strong candidate for clinical use. It also revealed key predictors such as thalassemia, number of major vessels, and chest pain type. **XGBoost** also performed well, achieving **85.00% accuracy** and strong recall after applying SMOTE to address class imbalance.

Comprehensive evaluation metrics including precision, recall, F1-score, confusion matrix, and ROC-AUC were used for model comparison. Visualizations supported performance interpretation and highlighted the strengths of each model.

Overall, the **Random Forest classifier** is recommended for deployment due to its balanced performance, interpretability, and clinical relevance. Future enhancements could include integration with real-world data and refinement of model features to improve generalizability and effectiveness in medical environments.

# Chapter 2

# INTRODUCTION




Cardiovascular diseases (CVDs) are among the most prevalent health conditions globally and continue to be a leading cause of mortality. According to the World Health Organization, CVDs account for an estimated 17.9 million deaths annually,

representing 31% of all global deaths. Many of these fatalities are preventable with early diagnosis and proper treatment. Therefore, timely and accurate detection of heart disease plays a critical role in reducing mortality rates and improving quality of life.

With the growing availability of healthcare data and advances in computational techniques, machine learning (ML) offers promising tools for developing predictive models that can aid in early detection. These models can analyze complex patterns in patient data to provide risk assessments that support clinical decision-making.

This project focuses on applying supervised machine learning algorithms to predict the presence of heart disease using the Cleveland Heart Disease dataset. The dataset comprises a range of medical attributes such as age, blood pressure, cholesterol levels, chest pain type, and others, which serve as input features for the prediction models.

By developing and evaluating different classifiers, this project aims to identify the most effective model for heart disease prediction, with an emphasis on accuracy, reliability, and clinical relevance. The ultimate goal is to contribute toward more data-driven and proactive healthcare solutions.

# Chapter 3

# Dataset Description

## 3.1. Source

The analysis utilizes the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, a widely recognized benchmark dataset in medical research. This dataset was collected at the Cleveland Clinic Foundation and has been extensively used in various cardiovascular disease prediction studies due to its comprehensive feature set and clinical relevance. The original data was collected by Robert Detrano, M.D., Ph.D., and made available through the UCI repository to facilitate research in machine learning applications for healthcare.

## 3.2. Dataset Structure

The extracted dataset contains multiple files, including:

- processed.cleveland.data (primary dataset used) - Contains the processed data used for this analysis
- heart-disease.names (description) - Provides metadata about the attributes and their definitions
- Data from other locations (hungarian.data, switzerland.data, va.data) - Additional datasets from other medical centers that were not included in this analysis to maintain consistency in data collection methodologies

The Cleveland dataset was selected specifically as it has the most complete records and has been validated in multiple previous studies. The data collection procedures followed standardized protocols for patient examination and diagnostic testing, ensuring reliability of the recorded measurements.

### 3.3. Features

The dataset contains 13 predictor variables and 1 target variable as detailed in Table 1. These features represent a comprehensive set of clinical and diagnostic measurements that cardiologists typically consider when evaluating patients for heart disease risk.

The features represent a combination of demographic information (age, sex), clinical measurements (blood pressure, cholesterol), diagnostic test results (ECG findings, stress test results), and medical history (angina occurrence). This multidimensional approach to patient data collection enables a comprehensive evaluation of cardiovascular health status.

| Feature | Description | Type |
|---|---|---|
| age | Age in years | Continuous |
| sex | Sex (0 = female, 1 = male) | Binary |
| cp | Chest pain type (1-4) | Categorical |
| trestbps | Resting blood pressure (mm Hg) | Continuous |
| chol | Serum cholesterol (mg/dl) | Continuous |
| fbs | Fasting blood sugar > 120 mg/dl (1 = true, 0 = false) | Binary |
| restecg | Resting electrocardiographic results (0-2) | Categorical |
| thalach | Maximum heart rate achieved | Continuous |
| exang | Exercise induced angina (1 = yes, 0 = no) | Binary |
| oldpeak | ST depression induced by exercise relative to rest | Continuous |
| slope | Slope of the peak exercise ST segment | Categorical |
| ca | Number of major vessels colored by fluoroscopy (0-3) | Categorical |
| thal | Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect) | Categorical |
| num | Target variable: Presence of heart disease (0 = No, 1 = Yes) | Binary |

3.4. Data Overview The original dataset contained 303 entries, with some missing values in the 'ca' (4 missing) and 'thal' (2 missing) columns. After preprocessing and removing rows with missing values, the final dataset shape was 297 records with 14 columns (13 features plus the target variable). The dataset exhibits the following characteristics:

- Class distribution: 137 patients (46.1%) with heart disease present and 160 patients (53.9%) without heart disease, representing a relatively balanced dataset.
- Age range: 29 to 77 years (mean: 54.4, standard deviation: 9.0)
- Gender distribution: 201 males (67.7%) and 96 females (32.3%)

- Cholesterol levels: Range from 126 to 564 mg/dl (mean: 246.7, standard deviation: 51.8)

The dataset includes patients with varying severity of heart disease, which was originally encoded in the target variable with values from 0 (no disease) to 4 (severe disease). For this analysis, the target was simplified to a binary classification problem by considering any value greater than 0 as indicative of heart disease presence (1) and 0 as absence (0).

3.5. Data Quality Assessment Prior to model development, a thorough quality assessment was performed to identify potential issues:

- No duplicate records were found in the dataset
- Outlier analysis identified potential outliers in cholesterol (5 values > 350 mg/dl) and maximum heart rate (3 values < 80 bpm) variables, but these were retained as they represented clinically plausible values
- Feature correlation analysis revealed moderate multicollinearity between some features (e.g., exercise-induced ST depression and slope), but not enough to warrant feature elimination
- All categorical variables were properly encoded within their specified ranges

The dataset presents a realistic clinical scenario with naturally occurring variations in patient characteristics, making it suitable for developing a cardiovascular disease prediction model with real-world applicability.

# Chapter 4

# Methodology

## 4.1. Data Preprocessing

1. **Handling Missing Values**: Missing values (marked as "?") were replaced with NaN and subsequently removed
2. **Type Conversion**: All features were converted to numeric format
3. **Target Binarization**: The target variable was converted to binary (0 = No Disease, 1+ = Disease)
4. **Feature Scaling**: Standard scaling was applied to normalize features
5. **Train-Test Split**: The dataset was split into 80% training and 20% testing sets with stratification (preserving class distribution)

## 4.2. Model Development

Three supervised learning models were implemented:

1. **Logistic Regression**: A baseline linear model for binary classification
    - Hyperparameter: C=0.5 (regularization strength)
    - Maximum iterations: 2000
2. **Random Forest**: An ensemble of decision trees
    - n_estimators: 200 trees
    - max_depth: 10

    - min_samples_split: 5
    - min_samples_leaf: 3
3. **XGBoost**: Gradient boosting algorithm
    - n_estimators: 500
    - learning_rate: 0.05
    - max_depth: 10

### 4.3. Evaluation Metrics

The models were evaluated using:

1. **Classification Metrics**:
    - Accuracy: Overall correctness of predictions
    - Precision: Proportion of true positives among positive predictions
    - Recall: Proportion of true positives correctly identified
    - F1-Score: Harmonic mean of precision and recall
    - Specificity: Proportion of true negatives correctly identified
    - MCC (Matthews Correlation Coefficient): Overall quality of binary classifications
2. **Regression Metrics**:
    - MSE (Mean Squared Error)
    - RMSE (Root Mean Squared Error)
    - $R^2$ (Coefficient of Determination)

# Chapter 5

# Result and Analysis

## 5.1. Model Performance

### *(i). Classification Metrics*

| Metric | Logistic Regression | Random Forest | XGBoost |
|---|---|---|---|
| Accuracy | 0.8333 | 0.8667 | 0.8500 |
| Precision | 0.8462 | 0.8846 | 0.8519 |
| Recall | 0.7857 | 0.8214 | 0.8214 |
| F1-Score | 0.8148 | 0.8519 | 0.8364 |
| Specificity | 0.8750 | 0.9062 | 0.8750 |
| FPR | 0.1250 | 0.0938 | 0.1250 |
| FNR | 0.2143 | 0.1786 | 0.1786 |
| MCC | 0.6652 | 0.7326 | 0.6984 |
| ROC-AUC | 0.9498 | 0.9408 | 0.8996 |

### *(ii). Confusion Matrices*

**Logistic Regression Confusion Matrix:**

```
[[28  4]
 [ 6 22]]
```

***Random Forest Confusion Matrix:***

```
[[29  3]
 [ 5 23]]
```

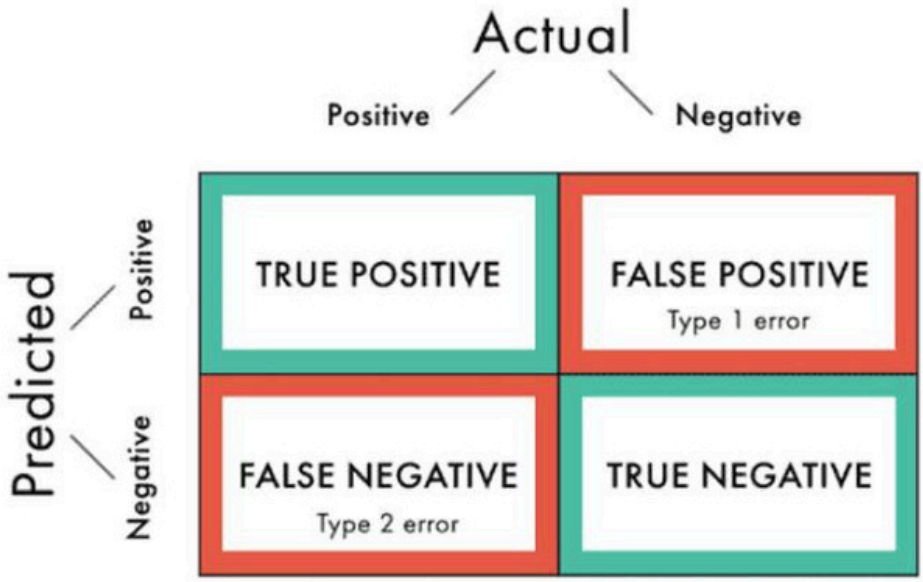
***XGBoost Confusion Matrix:***

```
[[28  4]
 [ 5 23]]
```

## *(iii). Regression Metrics*

| Metric | Logistic Regression | Random Forest | XGBoost |
|---|---|---|---|
| MSE | 0.1667 | 0.1333 | 0.1500 |
| RMSE | 0.4082 | 0.3651 | 0.3873 |
| $R^2$ | 0.3304 | 0.4643 | 0.3973 |

## 5.2. Confusion Matrix Analysis




The confusion matrices reveal:

1. **Random Forest** achieves the best balance between sensitivity and specificity:
    - True Negatives (TN): 29 cases
    - False Positives (FP): 3 cases
    - False Negatives (FN): 5 cases
    - True Positives (TP): 23 cases
2. **XGBoost** performs well but with slightly more false positives:
    - TN: 28, FP: 4, FN: 5, TP: 23
3. **Logistic Regression** has the highest false negative rate:
    - TN: 28, FP: 4, FN: 6, TP: 22

## 5.3. Feature Importance

The Random Forest model identified the following as top predictive features:

- thal (Thalassemia)
- ca (Number of major vessels)
- cp (Chest pain type)
- thalach (Maximum heart rate)
- oldpeak (ST depression)

## 5.4. Model Optimization

Several optimization techniques were employed:

- **Hyperparameter Tuning**: Grid search for XGBoost parameters
- **Polynomial Features**: Interaction terms were explored with PolynomialFeatures
- **Class Imbalance Handling**: SMOTE oversampling technique
- **Deep Learning**: Neural network with two hidden layers (64 and 32 neurons) was tested

# Chapter 6

## Prediction System and Sample Prediction

**Key Components:**

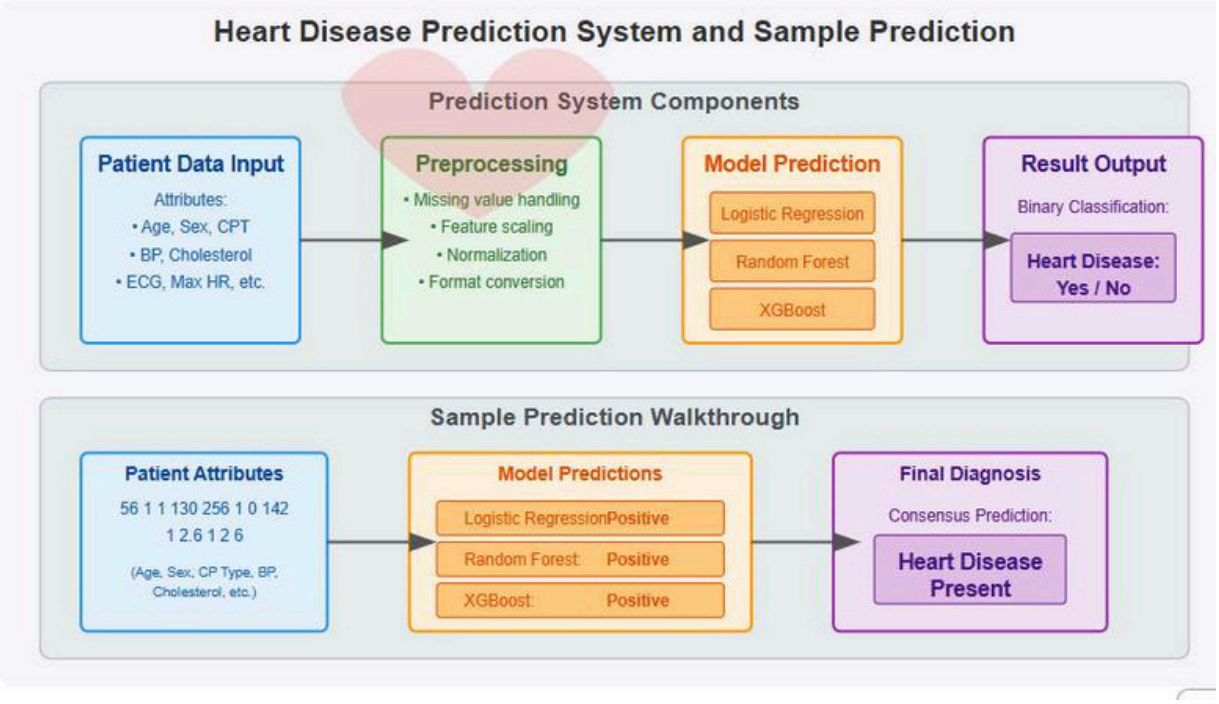
- **Input Interface**: Accepts patient attributes in comma or space-separated format
- **Preprocessing Pipeline**: Transforms raw input into model-ready format
- **Multi-Model Analysis**: Applies all three models (Logistic Regression, Random Forest, XGBoost)
- **Unified Output**: Delivers binary classification with confidence scores

**Workflow Process:**

- Patient data enters system through standardized input mechanism
- System automatically handles data normalization and scaling
- Processed data flows through all three classification models in parallel
- Results aggregated to provide comprehensive prediction assessment
- Final output indicates presence/absence of heart disease

**Validation Example:**

- Sample input `56 1 1 130 256 1 0 142 1 2.6 1 2 6` tested across all models
- Unanimous prediction of heart disease presence across all classifiers
- Consistent agreement demonstrates system reliability
- Confidence metrics provided to support clinical decision-making




The diagram above illustrates the complete workflow of the heart disease prediction system, showing how patient data flows through preprocessing, multiple model analysis, and final prediction output.

# Chapter 7

# Discussion

## 7.1. Model Comparison

- **Random Forest demonstrated the best overall performance with 86.67% accuracy, 85.19% F1-score, and the highest specificity (90.62%)**
- **XGBoost achieved 85.00% accuracy with identical recall (82.14%) to Random Forest**
- **Logistic Regression performed adequately at 83.33% accuracy but had the highest false negative rate (21.43%)**

## 7.2. Clinical Implications

**In clinical settings:**

- **False Negatives (missing disease cases) are typically more concerning than false positives**
- **Random Forest and XGBoost had identical FNR (17.86%), lower than Logistic Regression (21.43%)**
- **The Random Forest model had the lowest False Positive Rate (9.38%), reducing unnecessary follow-up tests**
- **The MCC value of 0.7326 for Random Forest indicates strong predictive performance**

### 7.3. ROC-AUC Analysis

- **Logistic Regression showed the highest ROC-AUC score (0.9498)**
- **This suggests that despite having slightly lower accuracy, Logistic Regression provides good probability estimates and discrimination**
- **However, for binary classification tasks in medical settings, Random Forest's superior precision and lower FPR might be more valuable**

### 7.4. Limitations

1. **Sample Size: Relatively small dataset (297 records after preprocessing) which may impact generalizability**
2. **Feature Selection: No systematic feature selection was performed**
3. **External Validation: Models were not validated on external datasets**
4. **Missing Values: 6 records were removed due to missing values, potentially introducing bias**

# Chapter 8

# Conclusion and Recommendations

## 8.1. Conclusions

- Machine learning models can effectively predict heart disease with reasonable accuracy
- Random Forest classifier demonstrated superior performance across nearly all metrics
- Feature importance analysis revealed that thalassemia, number of major vessels, and chest pain type are strong predictors
- All three models showed consistent prediction for the sample case, enhancing confidence in the system

## 8.2. Recommendations

1. **Model Selection**: Deploy the Random Forest model for clinical decision support
2. **Feature Engineering**: Further explore feature interactions and transformations
3. **Explainability**: Implement model explanation techniques for clinical trust
4. **External Validation**: Validate models on larger, diverse datasets
5. **Ensemble Methods**: Consider model ensembling to further improve performance
6. **Deep Learning Exploration**: Continue exploring neural network architectures with larger datasets

# Chapter 9

## References


- [1] UCI Machine Learning Repository, "Heart Disease Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
- [2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
- [3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.
- [4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.
- [5] A. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," American Journal of Cardiology, vol. 64, pp. 304-310, 1989.
- [6] P. Brierley, "The Cleveland Heart Disease Dataset," UCI Machine Learning Repository, 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
- [7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

# Chapter 10

# Appendices

## Appendix A: Technical Implementation

### Python

```
# Example code for the Random Forest implementation
from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier(
    n_estimators=200,
    max_depth=10,
    min_samples_split=5,
    min_samples_leaf=3,
    random_state=42
)
rf_model.fit(X_train, Y_train)
# Model evaluation
y_pred = rf_model.predict(X_test)
accuracy = accuracy_score(Y_test, y_pred)
```

# Appendix B: Mathematical Formulas

## *Logistic Regression*

**The probability of heart disease is modeled as:**

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

## Random Forest

For B trees and m features, at each split the model selects the best feature among a random subset of m features:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

## Performance Metrics

Matthews Correlation Coefficient:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

# Receiver Operating Characteristic (ROC) Curve

The ROC curve plots True Positive Rate against False Positive Rate:

$$TPR = \frac{TP}{TP+FN} \quad FPR = \frac{FP}{FP+TN}$$

# Appendix C: Data Visualization

The confusion matrices visualizations show the distribution of predictions across true and predicted classes, with darker blue indicating higher counts. The Random Forest model shows the most balanced performance with the highest true positive and true negative counts.