# Heart Disease Prediction Lab Report

L-55+56

S. Muskan - 23BCE7305
SK. Salma - 23BCE20344

## 1 Introduction

### Problem Statement

Focus on the diagnosis and prediction of heart disease using machine learning models.

### Brief Description of Demand

Growing need for reliable and efficient diagnostic tools in healthcare to aid in early detection of heart diseases.

### Objective

Predict the likelihood of heart disease using patient health data.

### Dataset

Utilizes medical parameters such as age, sex, blood pressure, cholesterol, etc.

### ML Algorithms

Implements models like Naive Bayes, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN).

### Benefits

Enables early detection of heart disease for timely intervention.

### Workflow

Involves data preprocessing, feature selection, model training, and evaluation using metrics like accuracy and precision.

## Outcome

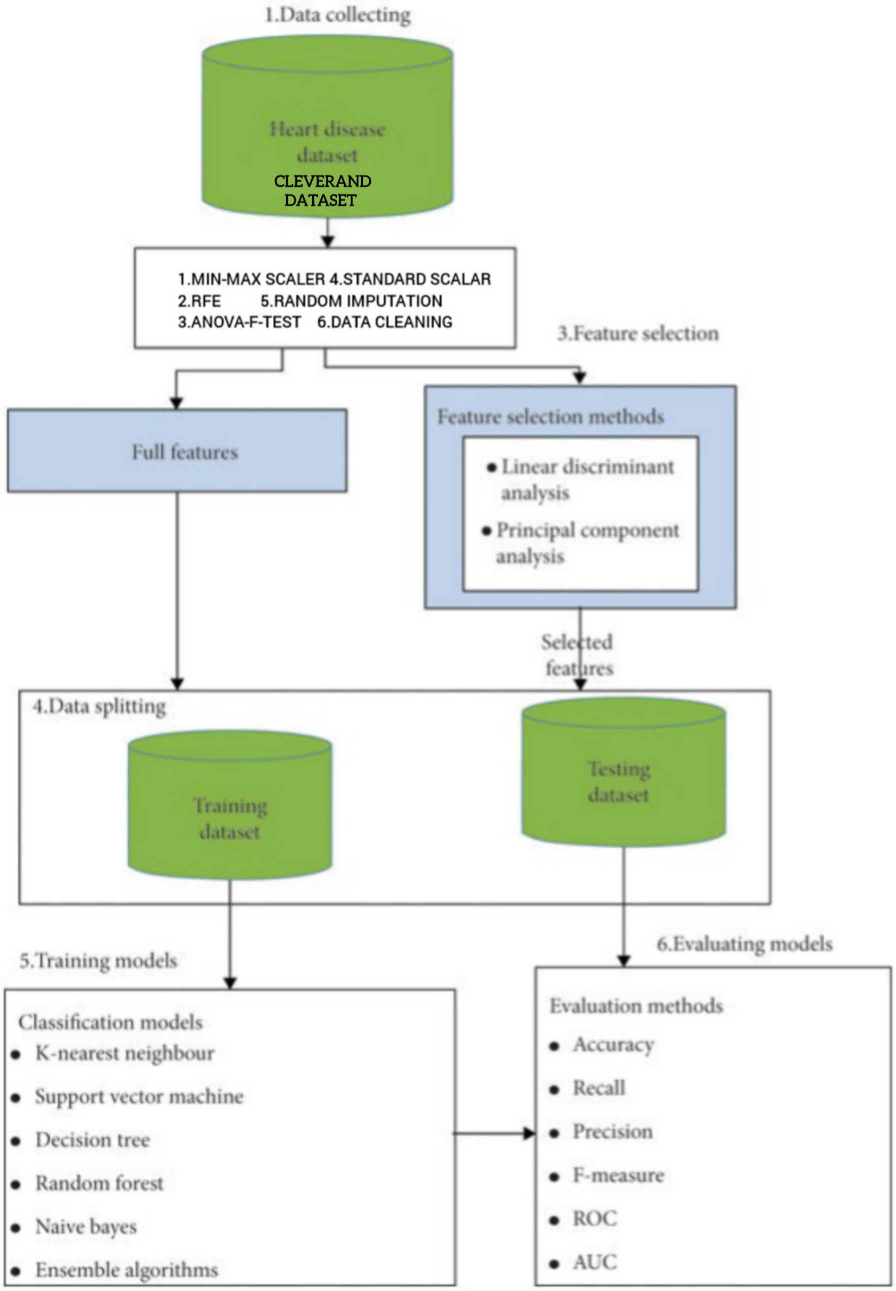
Supports healthcare professionals in making accurate, data-driven diagnoses.

## Advantages and Uses of ML Algorithms

- Improved accuracy and speed in diagnosing conditions.
- Ability to analyze large datasets for pattern recognition.
- Enhanced predictive capabilities leading to timely interventions.

Table 1: Excel Data from Research Papers

| **PAPER NO** | **DATA SET NAME** | **ATTRIBUTES** | **EVALUATION METRICS** | **ALGORITHM USED** | **DESCRIPTIO** |
|---|---|---|---|---|---|
| main paper | Cleveland Heart Disease Dataset 2016 | Blood pressure, Cholesterol level, Blood sugar | Accuracy (90%), Precision | ML: KNN, Pre: Standard Scaler, Min-Max Scaler | Heart disease d study achieving curacy. |
| paper.1 | COVID-19 Blood Biomarker Dataset | Age, Sex, BMI, Cholesterol, Blood Sugar | 99% Accuracy | ML: ANN, Pre: MinMaxScaler | Prediction of C 19 based on bior |
| paper.6 | Scikit-Learn | Age, Gender, Chest Pain, BP | FACO outperforms by 15% | ML: KNN, Pre: Feature Selection | Hybrid feature s method for he ease. |
| paper.8 | Framingham CHD Dataset | Age, Smoker, BP, Cholesterol | Accuracy: 84.7% → 86.3% | ML: Logistic Regression, Pre: Imputation | Optimization o tic Regression n |
| paper.10 | BRFSS Dataset | Age, Sex, Chest Pain, BP | Accuracy: 90.67% | ML: Logistic Regression, KNN, Decision Tree | Use of ML for risk prediction. |
| paper.12 | CDC Heart Patients Dataset | Hypertension, BMI, Stress | Accuracy: 97%, Precision >90% | ML: PAC, SGD, Gradient Boosting | Stacking model ter prediction. |
| paper.14 | Diabetes Dataset (Kaggle) | High BP, Cholesterol, BMI | Accuracy: 86.51% | ML: Logistic Regression, Decision Tree | Feature selecti proves classifica curacy. |
| paper.16 | UCI Cleveland Dataset | Age, Sex, Cholesterol, BP | Accuracy: 88% | ML: Random Forest, XGBoost | Incorporating history in heart detection. |
| paper.17 | Kaggle Datasets (UCI) | Age, BMI, Diabetes | Accuracy: 100% | ML: EXBNet (Boosted NN) | EXBNet outp standard ML m |


1.Data collecting
Heart disease dataset
CLEVERAND DATASET
1.MIN-MAX SCALER 4.STANDARD SCALAR
2.RFE 5.RANDOM IMPUTATION
3.ANOVA-F-TEST 6.DATA CLEANING
3.Feature selection
Full features
Feature selection methods
• Linear discriminant analysis
• Principal component analysis
Selected features
4.Data splitting
Training dataset
Testing dataset
5.Training models
Classification models
• K-nearest neighbour
• Support vector machine
• Decision tree
• Random forest
• Naive bayes
• Ensemble algorithms
6.Evaluating models
Evaluation methods
• Accuracy
• Recall
• Precision
• F-measure
• ROC
• AUC

# 2 Architecture Diagram

# 3 Table I: Confusion Matrix

| Model | True Positive | False Positive | Accuracy (%) |
|---|---|---|---|
| KNN | 135 | 26 | 82.0 |
| SVM | 111 | 56 | 68.0 |
| Random Forest | 140 | 25 | 85.0 |
| Naive Bayes | 82 | 82 | 50.0 |

Table 2: Classification Accuracy for Different Algorithms

# 4 Data Preprocessing

## Objective

Clean and prepare data for analysis.

## Steps Involved

- **Handling Missing Values:** Imputation strategies (mean, median, mode), removal of incomplete records.
- **Data Normalization:** Scaling data to a uniform range (e.g., Min-Max scaling).
- **Data Encoding:** Converting categorical variables into numerical format using one-hot or label encoding.

# 5 Data Sheet Description with Visualization

- **Types of Data Sheets Used:** Heart disease dataset including clinical parameters and patient history.
- **Number of Variables:** Total variables analyzed (e.g., age, gender, cholesterol levels, etc.).

# 6 Algorithms

- **Mathematical Aspects:** Overview of statistical methods and formulas used in ML algorithms.
- **Optimization Techniques:** Discussion on loss functions and training improvements.

# 7 Feature Selection

- **LDA:** Reduces dimensionality while maintaining class separability.
- **PCA:** Transforms features into uncorrelated principal components.
- **Recursive Feature Elimination:** Iteratively removes features to find the best subset.

# 8 Data Splitting

- **Train-Test Split:** Typically 70-30 or 80-20 ratio.
- **Cross-Validation:** Uses k-fold strategy for robust evaluation.

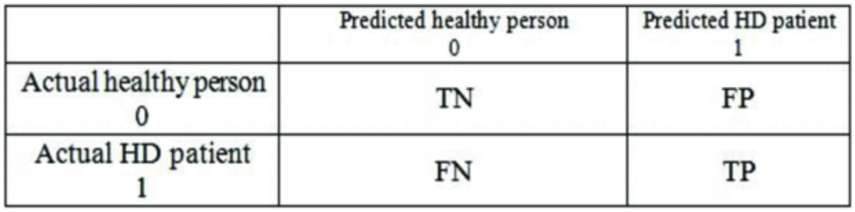
# 9 Training Models

- **KNN:** Classifies based on proximity to other data points.
- **SVM:** Finds optimal hyperplane maximizing margin.
- **Decision Tree:** Flowchart-based classification.
- **Random Forest:** Ensemble of decision trees.
- **Naive Bayes:** Probabilistic classifier using Bayes' theorem.
- **Ensemble Algorithms:** Boosting and bagging to improve performance.

# 10 Evaluating Models

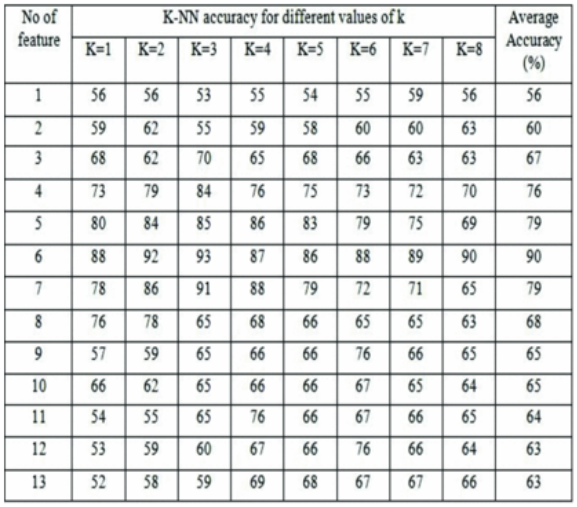
- **Accuracy:** Correct predictions / total predictions.
- **Recall:** Ability to identify true positives.
- **Precision:** Ratio of true positives to predicted positives.
- **F-measure:** Harmonic mean of precision and recall.
- **ROC:** Visual trade-off between sensitivity and specificity.
- **AUC:** Area under ROC curve; 1 indicates perfect model.

Table I. Confusion metrics[25]–[26]–[27]



| | Predicted healthy person<br>0 | Predicted HD patient<br>1 |
|---|---|---|
| Actual healthy person<br>0 | TN | FP |
| Actual HD patient<br>1 | FN | TP |

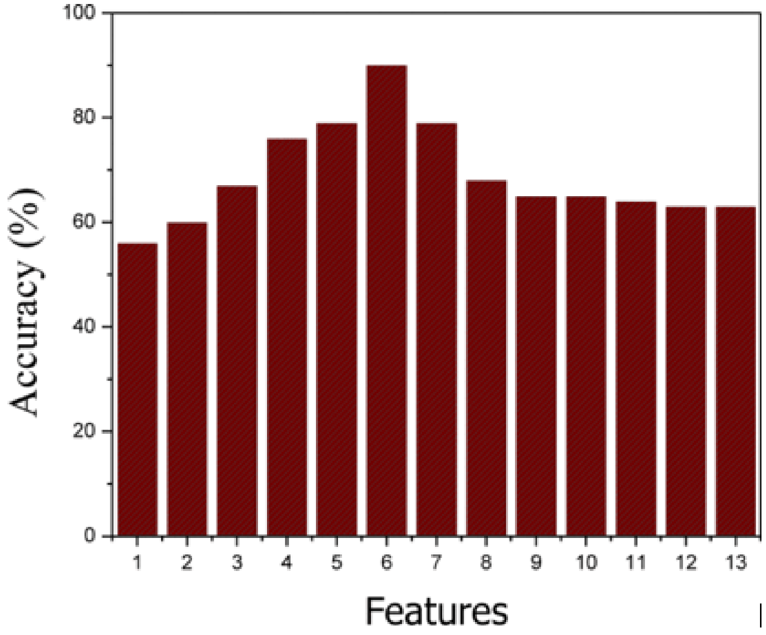
**Table II.** Classification accuracy of K-NN with different number of features



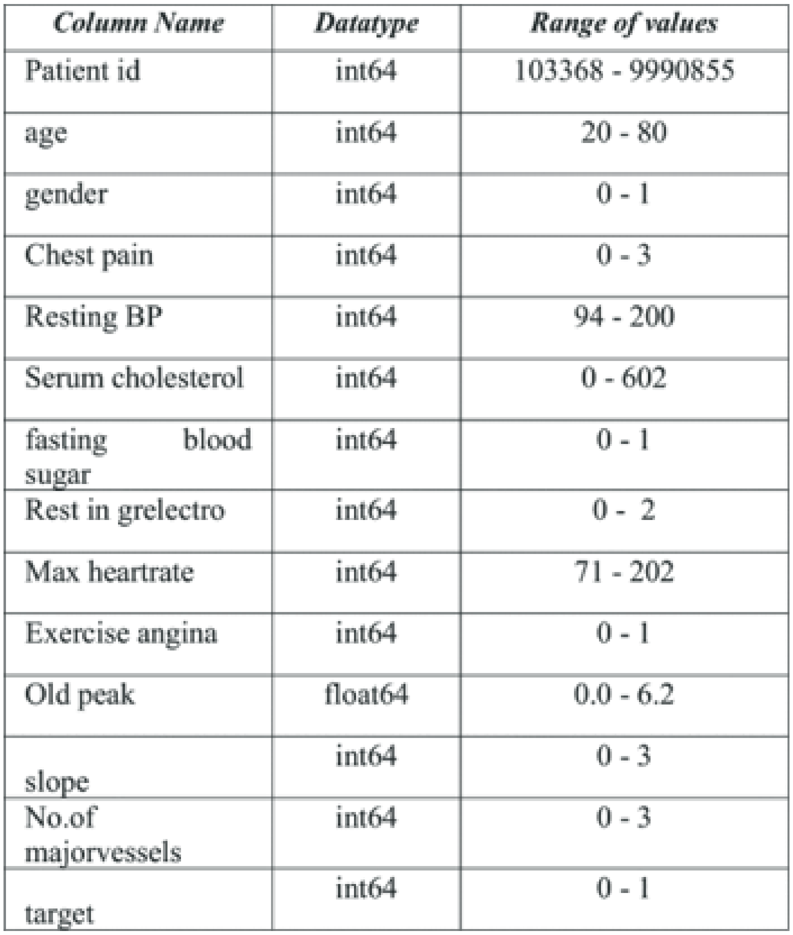
| No of feature | K-NN accuracy for different values of k | | | | | | | | Average Accuracy (%) |
|---|---|---|---|---|---|---|---|---|---|
| | K=1 | K=2 | K=3 | K=4 | K=5 | K=6 | K=7 | K=8 | |
| 1 | 56 | 56 | 53 | 55 | 54 | 55 | 59 | 56 | 56 |
| 2 | 59 | 62 | 55 | 59 | 58 | 60 | 60 | 63 | 60 |
| 3 | 68 | 62 | 70 | 65 | 68 | 66 | 63 | 63 | 67 |
| 4 | 73 | 79 | 84 | 76 | 75 | 73 | 72 | 70 | 76 |
| 5 | 80 | 84 | 85 | 86 | 83 | 79 | 75 | 69 | 79 |
| 6 | 88 | 92 | 93 | 87 | 86 | 88 | 89 | 90 | 90 |
| 7 | 78 | 86 | 91 | 88 | 79 | 72 | 71 | 65 | 79 |
| 8 | 76 | 78 | 65 | 68 | 66 | 65 | 65 | 63 | 68 |
| 9 | 57 | 59 | 65 | 66 | 66 | 76 | 66 | 65 | 65 |
| 10 | 66 | 62 | 65 | 66 | 66 | 67 | 65 | 64 | 65 |
| 11 | 54 | 55 | 65 | 76 | 66 | 67 | 66 | 65 | 64 |
| 12 | 53 | 59 | 60 | 67 | 66 | 76 | 66 | 64 | 63 |
| 13 | 52 | 58 | 59 | 69 | 68 | 67 | 67 | 66 | 63 |

**Fig 1.**

K-NN performance on reduced number of features sets



## 2. DATASET DETAILS



| *Column Name* | *Datatype* | *Range of values* |
|---|---|---|
| Patient id | int64 | 103368 - 9990855 |
| age | int64 | 20 - 80 |
| gender | int64 | 0 - 1 |
| Chest pain | int64 | 0 - 3 |
| Resting BP | int64 | 94 - 200 |
| Serum cholesterol | int64 | 0 - 602 |
| fasting blood sugar | int64 | 0 - 1 |
| Rest in grelectro | int64 | 0 - 2 |
| Max heartrate | int64 | 71 - 202 |
| Exercise angina | int64 | 0 - 1 |
| Old peak | float64 | 0.0 - 6.2 |
| slope | int64 | 0 - 3 |
| No.of majorvessels | int64 | 0 - 3 |
| target | int64 | 0 - 1 |

**Fig. 1.**

Risk categories in the dataset (down)

**Fig. 1.**

Risk categories in the dataset (down)

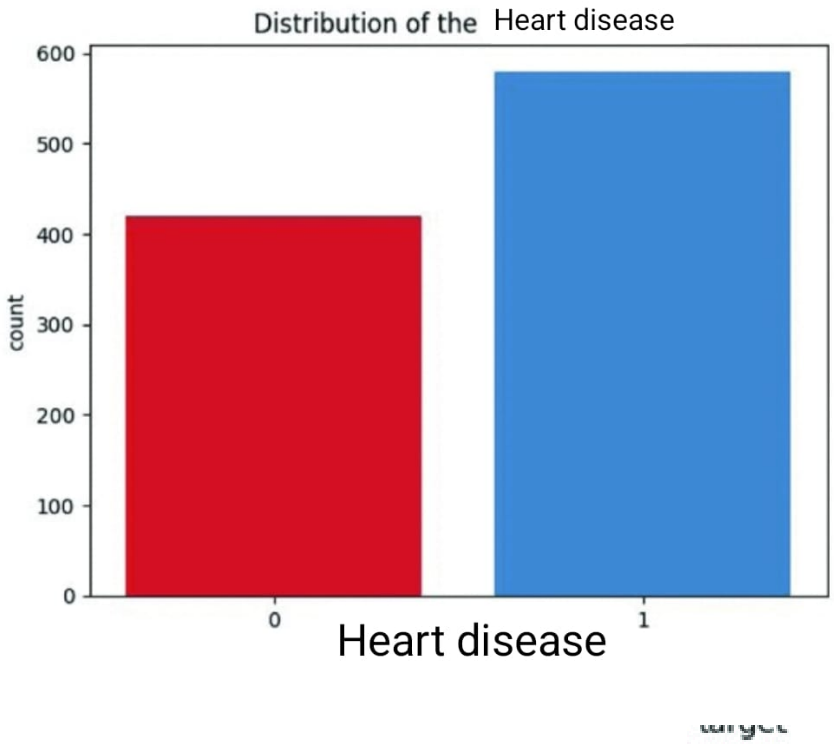
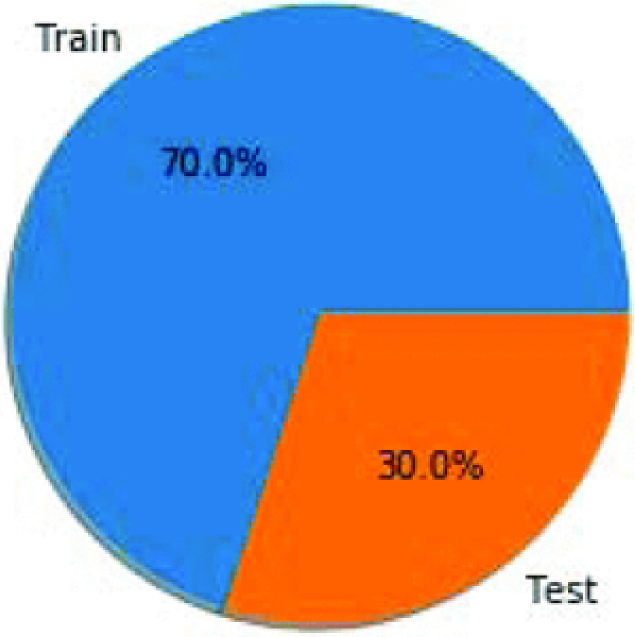


**Fig. 2.**

Dataset division for model



### 3. Architecture of ANN

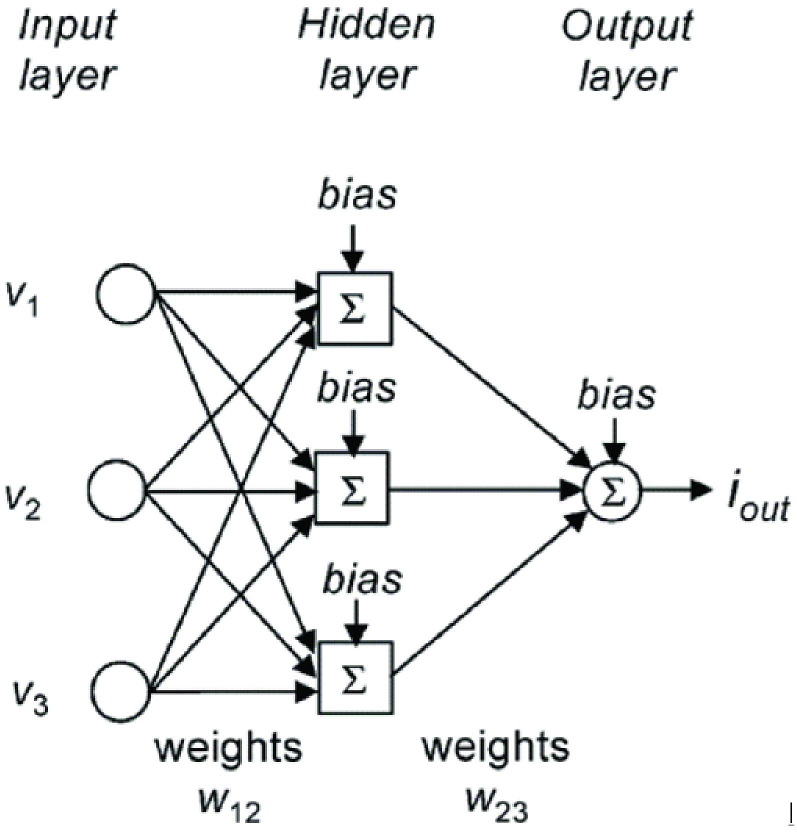
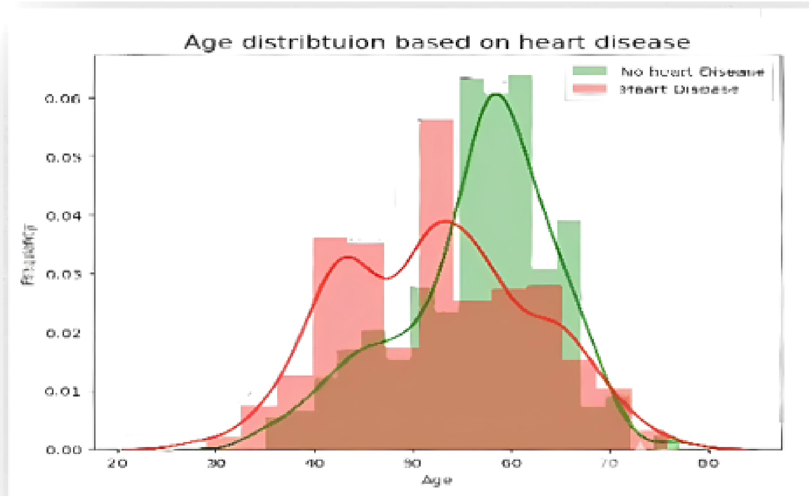
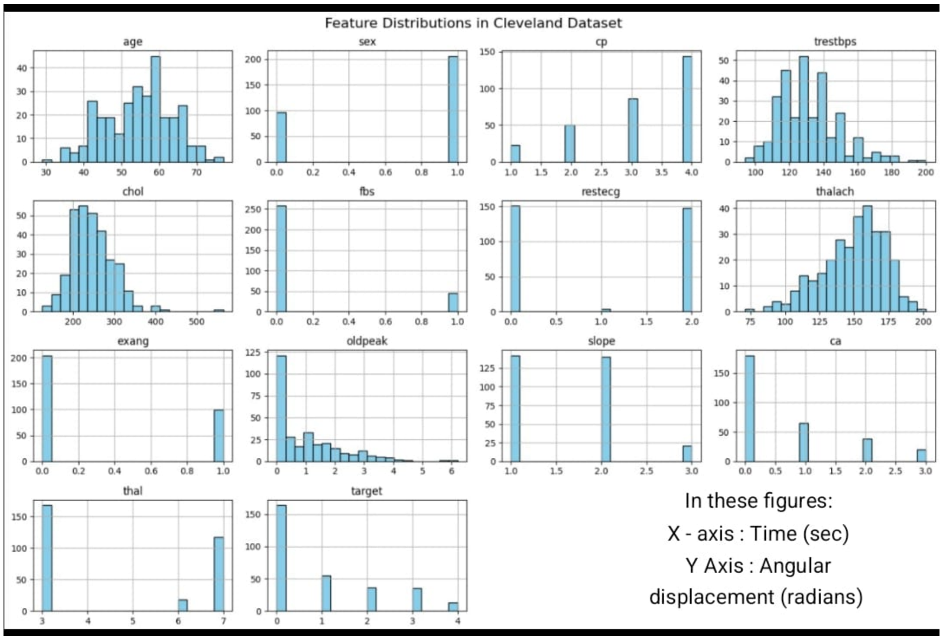
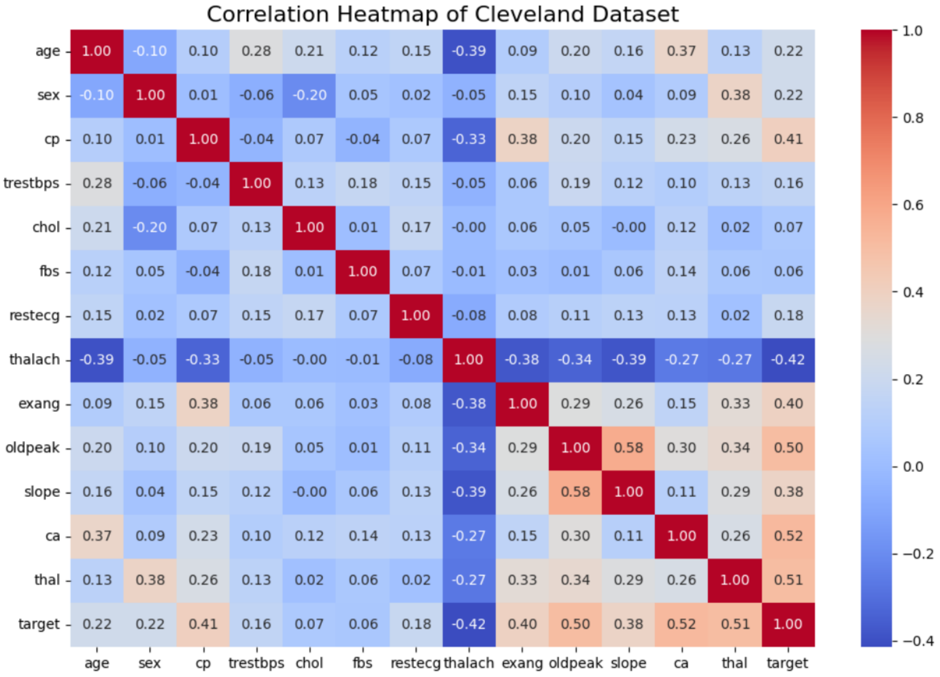


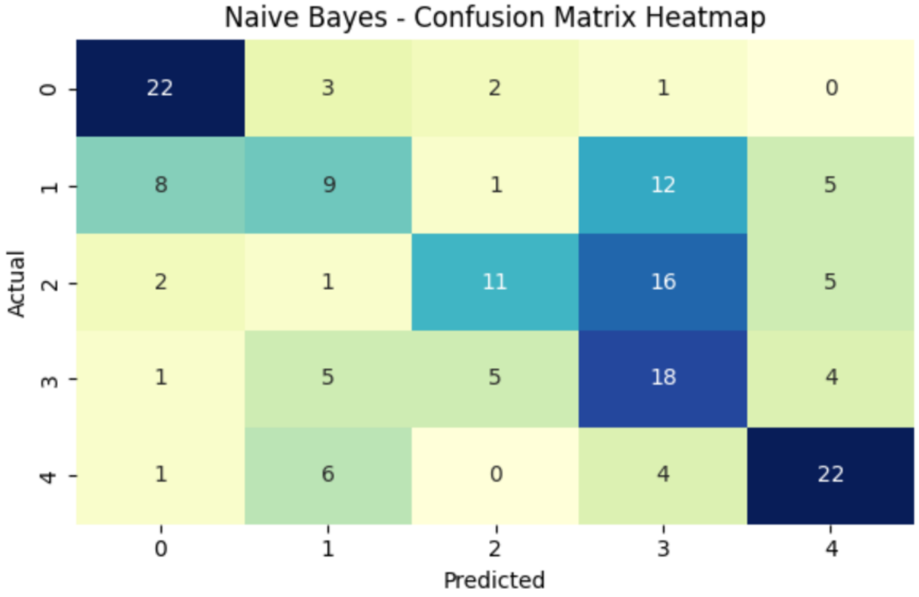
**FIG :**



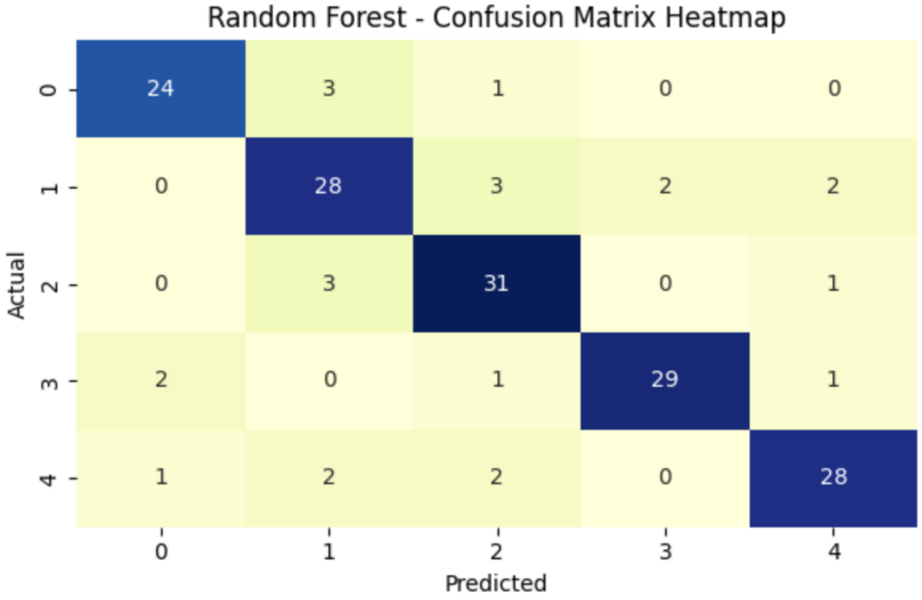


In these figures:
X - axis : Time (sec)
Y Axis : Angular displacement (radians)

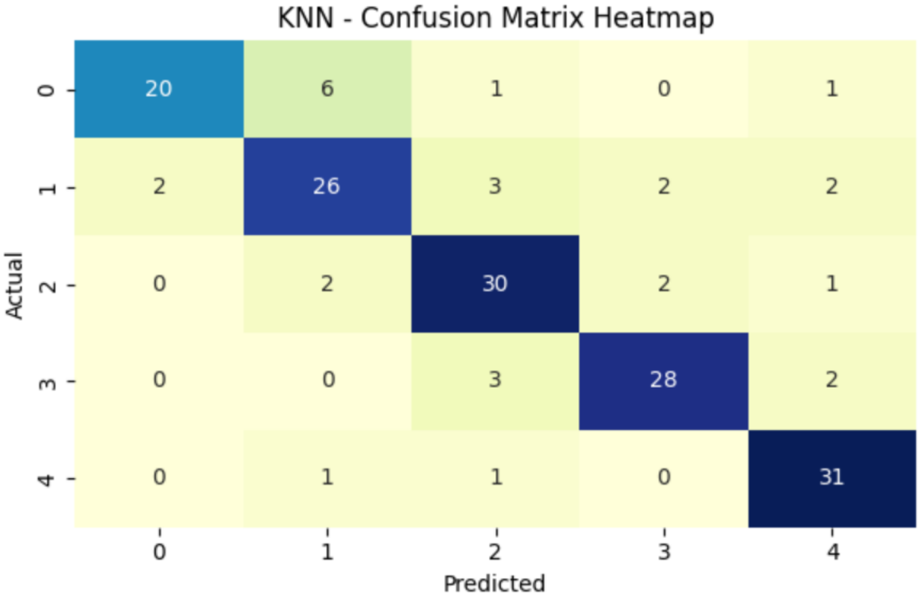




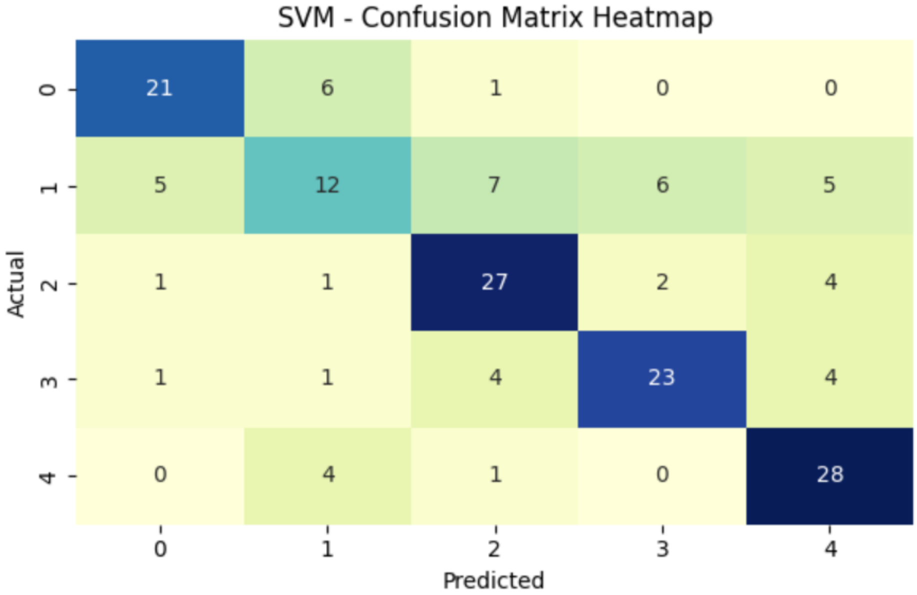
Naive Bayes - Confusion Matrix Heatmap



Random Forest - Confusion Matrix Heatmap


KNN - Confusion Matrix Heatmap
Actual
Predicted
0 1 2 3 4
0: 20 6 1 0 1
1: 2 26 3 2 2
2: 0 2 30 2 1
3: 0 0 3 28 2
4: 0 1 1 0 31

H


SVM - Confusion Matrix Heatmap
Actual
Predicted
0 1 2 3 4
0: 21 6 1 0 0
1: 5 12 7 6 5
2: 1 1 27 2 4
3: 1 1 4 23 4
4: 0 4 1 0 28

# 11 Conclusion

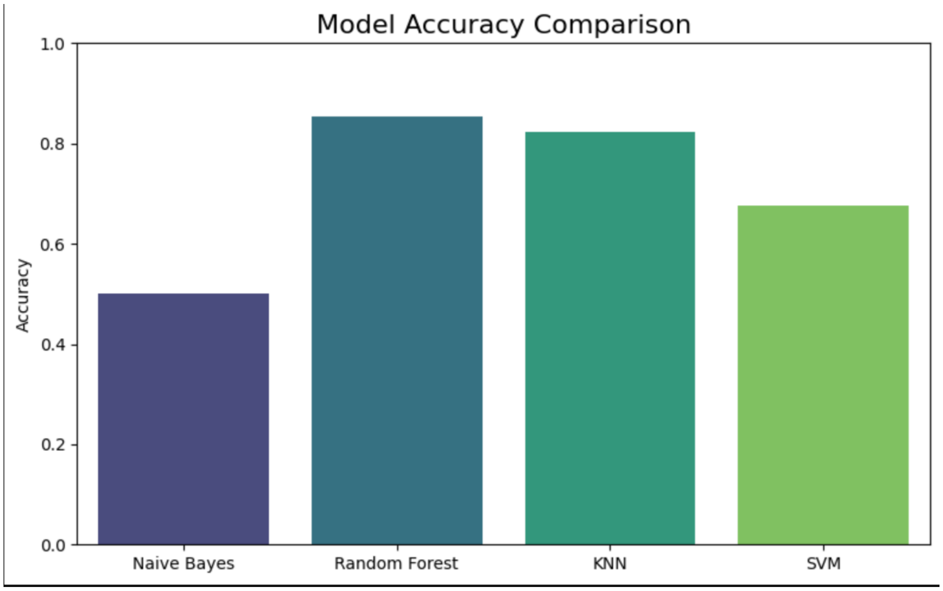
In this research, we investigated the application of various supervised machine learning algorithms—Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest—for the prediction of heart disease using the Cleveland dataset. Each algorithm was trained and evaluated based on its predictive accuracy and corresponding confusion matrix.

The comparative analysis revealed that the Random Forest algorithm achieved the highest accuracy of 85%, significantly outperforming Naive Bayes (50%), SVM (68%), and KNN (82%). This indicates that ensemble methods like Random Forest, which leverage multiple decision trees and majority voting mechanisms, are more effective in handling complex, non-linear patterns present in medical data.

Furthermore, the confusion matrices provided deeper insight into the classification performance, highlighting the trade-offs between sensitivity and specificity for each model. Based on these findings, Random Forest not only demonstrated the best generalization capability but also showed promise as a robust predictive tool in the domain of heart disease diagnosis.

This study emphasizes the importance of algorithm selection in medical data analysis and sets a strong foundation for future work, where techniques such as hyperparameter tuning, feature engineering, and deep learning could be explored to further improve diagnostic accuracy and reliability.

- The darker the shade (more blue), the higher the count of samples.
- The lighter the shade (more yellow/white), the lower the count.
- **X-axis (Predicted):** This shows what the model predicted each instance to be.
- **Y-axis (Actual):** This shows the actual or true labels from the dataset.


Model Accuracy Comparison
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Naive Bayes
Random Forest
KNN
SVM