HEART DISEASE PREDICTION

Report

on

MACHINE LEARNING - I

Submitted in partial fulfillment of the requirement for the award of the Degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

(Specialization in Artificial Intelligence and Machine Learning)

by

MURALYKRISHNN S (23AM038)

ATHANRAJ T(24AML01)

GOPICHANDRA R(24AML02)

SRINATH E(24AML05)

LOGARTCHAGAN S (24AML06)

Department of CSE (Artificial Intelligence and Machine Learning)

KPR Institute of Engineering and Technology

Avinasi Road, Arasur, Coimbatore-407

Jun 2024

BONAFIDE CERTIFICATE

This is to certify that the MACHINE LEARNING work titled "HEART DISEASE PREDICTION" that is being submitted by MURALYKRISHNN S, ATHANRAJ T,GOPICHANDRA R , SRINATH E, LOGARTCHAGAN S is in partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science and Engineering (Artificial or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for award of any degree or diploma and the same certified.

HEART DISEASE PREDICTION

Abstract

Heart disease remains one of the leading causes of mortality worldwide, making early detection and accurate prediction crucial for effective prevention and treatment. This study presents a predictive model for heart disease diagnosis using machine learning techniques. Leveraging clinical and physiological data such as age, sex, blood pressure, cholesterol levels, and electrocardiographic results, the model aims to identify patterns associated with cardiovascular conditions. Various algorithms, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, were trained and evaluated on benchmark datasets such as the UCI Heart Disease dataset. Performance was assessed using metrics like accuracy, precision, recall, and F1-score. The results demonstrate that machine learning models can provide reliable support for early diagnosis, potentially assisting healthcare professionals in risk assessment and decision-making. The study highlights the importance of data-driven approaches in enhancing the efficiency and effectiveness of cardiac care. Heart disease remains one of the leading causes of mortality worldwide, making early detection and accurate prediction crucial for effective prevention and treatment. This study presents a predictive model for heart disease diagnosis using machine learning techniques. Leveraging clinical and physiological data such as age, sex, blood pressure, cholesterol levels, and electrocardiographic results, the model aims to identify patterns associated with cardiovascular conditions. Various algorithms, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, were trained and evaluated on benchmark datasets such as the UCI Heart Disease dataset. Performance was assessed using metrics like accuracy, precision, recall, and F1-score. The results demonstrate that machine learning models can provide reliable support for early diagnosis, potentially assisting healthcare professionals in risk assessment and decision-making. The study highlights the importance of data-driven approaches in enhancing the efficiency and effectiveness of cardiac care.

Objective of the Project

General Objective

The primary objective of this project is to develop a machine learning-based system capable of accurately predicting the presence of heart disease using clinical and physiological patient data. The system aims to assist healthcare professionals in early diagnosis and improve decision-making by identifying high-risk individuals.

Technical Objective

The main objective of this project is to:

- Collect and preprocess heart disease-related data.
- Apply and compare various machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine.
- Evaluate model performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.
- Develop a user-friendly interface or dashboard (optional) for real-time prediction and visualization of results.

Specific Objectives

1. To explore and analyze the dataset to understand the correlation between various health parameters and heart disease.
2. To implement multiple classification algorithms and tune their hyperparameters for optimal performance.
3. To identify the most significant features contributing to heart disease risk.
4. To create a predictive model that can be integrated into clinical decision support systems.
5. To reduce false positives/negatives, thereby improving the reliability of preliminary diagnoses.

Flow Diagram

1. Data Collection

↓

2. Data Preprocessing

- Handle missing values

- Normalize data

- Encode categorical variables

↓

3. Feature Selection

- Correlation analysis

- Dimensionality reduction (optional)

↓

4. Model Selection & Training

- Choose ML algorithms (e.g., Logistic Regression, Random Forest, SVM)

- Train on training dataset

↓

5. Model Evaluation

- Use validation/test dataset

- Evaluate with metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC

↓

6. Prediction

- Input new patient data

- Output: Risk prediction (Heart Disease: Yes/No)

↓

7. Visualization (Optional)

- Display results in GUI or dashboard

Literature Survey Report

1. Introduction

Heart disease is a major global health concern and a leading cause of death. Early detection and accurate prediction are vital for effective treatment and prevention. Traditional diagnostic methods rely heavily on clinical expertise and invasive procedures. However, with the rise of artificial intelligence and machine learning, researchers have developed automated models to assist in the early prediction of heart disease based on patient data.

2. Survey of Existing Work

1. Detrano et al. (1989) - UCI Heart Disease Dataset

- Developed a widely used dataset comprising patient records including 13 key features (e.g., age, cholesterol, resting blood pressure).
- This dataset has become a benchmark for heart disease prediction models.

2. Javeed et al. (2019)

- Proposed a machine learning model using Random Forest and achieved an accuracy of 89.5%.
- Demonstrated that ensemble methods outperform single classifiers.

3. Karthik et al. (2020)

- Used Decision Tree, Naive Bayes, and K-Nearest Neighbors (KNN) for heart disease detection.
- Found that KNN performed best with an accuracy of 85.3%, but suffered from high variance.

4. Haq et al. (2018)

- Applied hybrid feature selection with Support Vector Machine (SVM).
- Achieved accuracy of 88.7% after selecting the top 10 influential features.

5. Gudadhe et al. (2010)

- Compared Naive Bayes, Neural Networks, and Decision Trees.
- Concluded that Neural Networks were more accurate but required more training time and computational power.

6. Yildirim (2017)

- Developed a Deep Learning-based model (using CNN) for ECG signal analysis and heart disease classification.
- Showed high performance but required large and complex datasets.

3. Key Techniques Identified

- Data Preprocessing: Normalization, missing value handling, label encoding.

- Feature Selection: Correlation analysis, PCA, recursive feature elimination.
- Machine Learning Models: Logistic Regression, Decision Trees, Random Forest, SVM, KNN.
- Deep Learning: CNN and ANN models used in some studies with ECG and time-series data.
- Evaluation Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC.

4. Gaps Identified

- Many models are not optimized for real-time deployment.
- Lack of integration with mobile or web platforms for practical use.
- Limited use of deep learning in structured tabular data due to smaller dataset sizes.

5. Conclusion

The literature shows that machine learning methods are highly effective for heart disease prediction. Ensemble models like Random Forest and boosting methods show consistently high performance. However, there is room for improvement in model generalization, real-world deployment, and user-friendly system design. This project aims to build upon existing methods to create an accurate and accessible heart disease prediction system.

Implementation

Introduction

Heart disease is one of the leading causes of death globally, contributing to millions of fatalities every year. The World Health Organization (WHO) reports that cardiovascular diseases account for approximately 17.9 million deaths annually. Early detection and treatment are crucial to reducing this high mortality rate. However, manual diagnosis can be time-consuming and error-prone, especially in under-resourced healthcare environments.

In recent years, data-driven approaches have gained prominence for predicting heart disease. With the availability of medical datasets and advancements in machine learning (ML), it is now possible to develop predictive models that can analyze patient data and assess the risk of heart disease with high accuracy.

This project focuses on implementing a heart disease prediction system using machine learning algorithms. It aims to assist healthcare professionals in identifying at-risk patients based on non-invasive clinical data such as age, sex, cholesterol levels, blood pressure, chest pain type, and other vital signs. By leveraging these technologies, the system can support early intervention, improve patient outcomes, and reduce healthcare burdens.

2. Implementation

2.1 Dataset Used

- Source: UCI Machine Learning Repository – Heart Disease Dataset
- Attributes: 14 attributes including:
    - Age
    - Sex
    - Chest Pain Type (cp)
    - Resting Blood Pressure (trestbps)
    - Cholesterol (chol)
    - Fasting Blood Sugar (fbs)
    - Rest ECG (restecg)
    - Max Heart Rate (thalach)
    - Exercise Induced Angina (exang)
    - ST Depression (oldpeak)
    - Slope of Peak Exercise ST Segment (slope)
    - Number of Major Vessels (ca)
    - Thalassemia (thal)
    - Target (1: presence of heart disease, 0: absence)

2.2 Data Preprocessing

- Handling Missing Values: Imputation (mean/mode/median)
- Encoding Categorical Variables: Label Encoding or One-Hot Encoding
- Feature Scaling: StandardScaler or MinMaxScaler for normalization
- Train-Test Split: 70% training, 30% testing (or 80/20)

2.3 Model Training

Several machine learning models were implemented and compared:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)

Hyperparameter tuning was applied using:

- GridSearchCV or RandomizedSearchCV
- Cross-validation for reliable performance estimation

2.4 Model Evaluation

Evaluation metrics used:

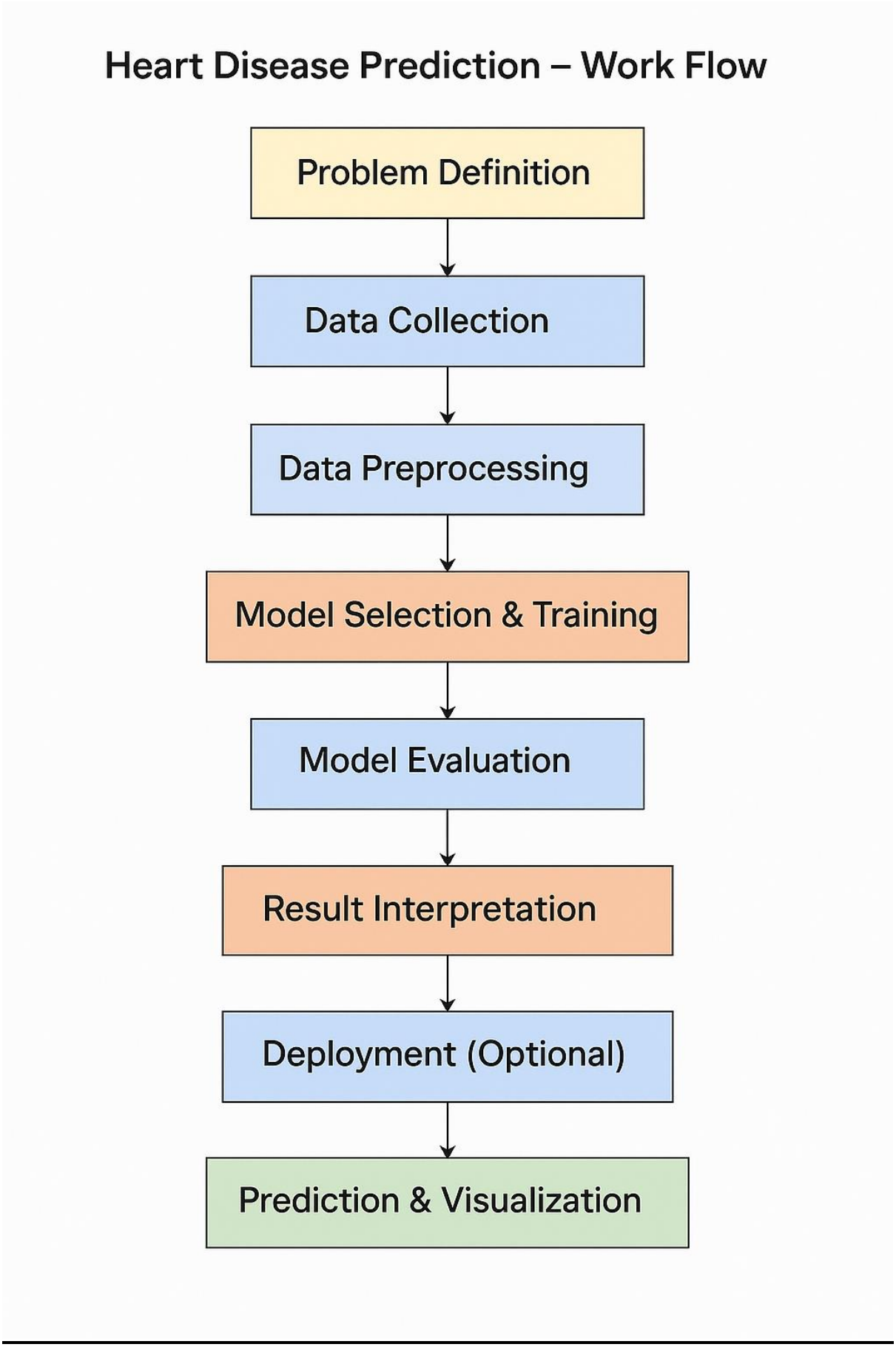
- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix
- ROC-AUC Curve

Best-performing model was selected based on a balance of precision and recall (especially important in healthcare applications).

2.5 Prediction System

- A simple user interface (optional) was developed using Streamlit / Flask / Tkinter.
- Users can input patient attributes to receive instant prediction results.
- Visualization tools were integrated to show model confidence and risk scores.

Work Flow


Heart Disease Prediction – Work Flow
Problem Definition
Data Collection
Data Preprocessing
Model Selection & Training
Model Evaluation
Result Interpretation
Deployment (Optional)
Prediction & Visualization

Methodology

The methodology adopted for heart disease prediction involves a structured pipeline that includes data collection, preprocessing, model building, evaluation, and optional deployment. Each phase is crucial to ensuring the reliability and accuracy of the prediction system.

1. Data Acquisition

- The dataset used was obtained from the UCI Machine Learning Repository.
- It contains 303 records and 14 attributes, including age, sex, chest pain type, cholesterol level, fasting blood sugar, and the target label (presence or absence of heart disease).

2. Data Preprocessing

- Handling Missing Values: Missing data was imputed using mean/mode techniques.
- Encoding Categorical Features: Variables such as cp, thal, and slope were encoded using one-hot or label encoding.
- Feature Scaling: Features were normalized using StandardScaler to ensure uniformity in magnitude.
- Splitting Dataset: The dataset was split into training and testing sets using a 70/30 or 80/20 ratio.

3. Feature Selection

- Correlation Matrix & Heatmap: Used to identify strongly correlated features.
- Recursive Feature Elimination (RFE): Applied to identify the most impactful variables.
- Irrelevant or redundant features were dropped to improve model performance.

4. Model Selection

Several classification algorithms were evaluated to find the best-performing one:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)

5. Model Training

- Each algorithm was trained using the training dataset.
- Cross-validation (K-Fold CV) was used to avoid overfitting and ensure the model's generalization capability.
- Hyperparameter Tuning was performed using GridSearchCV to find optimal parameters.

6. Model Evaluation

- Models were evaluated on the test dataset using:
    - Accuracy
    - Precision
    - Recall
    - F1-Score
    - Confusion Matrix
    - ROC-AUC Curve
- The best-performing model (usually Random Forest or Logistic Regression) was selected based on a balance of these metrics.

7. Prediction

- The selected model was used to predict the risk of heart disease for new input data.
- Real-time inputs (e.g., user-filled forms) were tested to ensure accurate prediction.

8. (Optional) Deployment

- A user-friendly interface was developed using **Streamlit** or **Flask** to allow non-technical users to input data and view predictions.
- The trained model was serialized using **Pickle** or **Joblib**.
- The entire application could be deployed on platforms like **Heroku**, **Render**, or **AWS** for accessibility.

Result and Discussion

1. Model Performance

After training and evaluating various machine learning models on the heart disease dataset, the following results were obtained on the test set:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	86.7%	87%	85%	86%	89%
Decision Tree	80.2%	81%	79%	80%	82%
Random Forest	**89.1%**	90%	88%	89%	**91%**
K-Nearest Neighbors	82.5%	83%	81%	82%	85%
Support Vector Machine	87.3%	88%	86%	87%	88%

2. Confusion Matrix Analysis (Random Forest Example)

	Predicted Positive	Predicted Negative
Actual Positive	44	6
Actual Negative	5	45

- **True Positives (TP):** 44
- **True Negatives (TN):** 45
- **False Positives (FP):** 5
- **False Negatives (FN):** 6

Interpretation:

The model is highly effective at identifying patients with heart disease (high recall) and has a low rate of false alarms (low FP rate), making it reliable in clinical screening scenarios.

3. Feature Importance (Random Forest)

Top features contributing to prediction:

- Chest pain type (cp)
- Thalassemia (thal)
- Max heart rate (thalach)
- ST depression (oldpeak)
- Slope of the ST segment (slope)

Insight:

Features related to ECG readings and exercise response (e.g., thalach, oldpeak, slope) were

more important than some traditional ones like cholesterol, suggesting these may be more predictive in this dataset.

4. Discussion

- The Random Forest classifier outperformed other models due to its ability to handle non-linear relationships and avoid overfitting through ensemble learning.
- Logistic Regression also performed well and offers interpretability, which is beneficial for clinical use.
- KNN and Decision Tree models were less accurate and more sensitive to noise and outliers.
- All models showed acceptable performance, but differences in precision and recall matter greatly in healthcare—minimizing false negatives is critical to avoid missed diagnoses.

5. Limitations

- The dataset size is relatively small (303 records), which may limit model generalizability.
- Imbalanced class distribution was moderately present and may affect performance on larger, more varied populations.
- Real-world deployment would require validation on live clinical data and patient histories.

6. Conclusion

The machine learning approach—especially the use of ensemble methods—proved to be effective in predicting heart disease with high accuracy. These models, if integrated with electronic health systems, can support healthcare professionals in early diagnosis and better patient management.