REPORT

Name of the Student

SHAIK AHMED ALI (Roll No. AP21110010916)

Name of the Course

INFORMATION RETRIEVAL

Supervisor: DR.Bala Venkateswarlu

Department of Computer Science and Engineering

SRM University AP

Table of Contents

Abstract

Heart disease remains one of the most critical and widespread health problems globally, accounting for a significant percentage of mortality every year. Early diagnosis and prediction play a vital role in preventing severe outcomes and reducing the risk of life-threatening complications. Traditional diagnostic procedures often rely heavily on clinical expertise, manual interpretation of medical data, and invasive tests, which can be time-consuming, costly, and prone to human error. With the advancement of technology, Machine Learning (ML) has emerged as a powerful tool that can support early medical decision-making by identifying complex patterns in patient data that may not be easily visible through conventional methods.

This project focuses on developing an automated heart disease prediction model using multiple machine learning algorithms and comparing their performance to identify the most accurate model. The dataset used in this study consists of key clinical features such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise-induced angina, oldpeak, slope, number of major vessels and thalassemia type. These features serve as important indicators in assessing the likelihood of heart-related complications. The data undergoes preprocessing steps including normalization, conversion of categorical values, and splitting into training and testing sets to ensure reliability and performance consistency.

Eight machine learning models were implemented and evaluated—Logistic Regression, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, XGBoost, and Neural Network (Multi-Layer Perceptron). The performance of each model was assessed using accuracy as the primary evaluation metric. Experimental results demonstrate that the **Random Forest classifier achieved the highest accuracy of 90.16%**, outperforming all other models. This is mainly because Random Forest reduces overfitting through ensemble learning and performs effectively on smaller structured datasets. Neural Network achieved a moderate accuracy of 81.97%, as deep learning generally requires larger datasets and extensive hyperparameter tuning for optimal performance.

The results clearly show that Machine Learning can significantly enhance the ability to detect heart disease early, enabling preventive treatment and improving patient outcomes. The finalized model can be implemented as a predictive tool in hospitals, integrated into mobile applications, or combined with wearable health monitoring devices for real-time risk assessment. This research highlights the potential of data-driven decision-support systems in healthcare and demonstrates how machine learning models can contribute to advanced diagnostic solutions.

Introduction

Heart disease is one of the most serious and rapidly growing health concerns worldwide, contributing to a large portion of global mortality each year. According to the World Health Organization (WHO), cardiovascular diseases account for nearly 32% of all global deaths annually, demonstrating the urgent need for early detection and proactive medical intervention. With lifestyle changes, increasing stress levels, and genetic predispositions, the prevalence of heart disease continues to rise, particularly among younger populations. In many cases, patients may not exhibit noticeable symptoms until the condition becomes critical, making early prediction essential for preventive treatment and long-term risk reduction.

Traditional diagnostic processes for heart disease rely on clinical tests, physician expertise, and manual interpretation of medical data. These methods can be time-consuming, cost-intensive, and subject to potential human error. Additionally, with the large amount of medical data being generated regularly, manual evaluation becomes challenging and inefficient. In this scenario, the role of intelligent computational systems becomes vital in assisting doctors by analyzing complex patterns in patient records and enabling faster, data-driven decision-making.

Machine Learning (ML) has emerged as a powerful approach for medical diagnosis due to its ability to learn from historical data, identify hidden correlations, and classify patient conditions with high accuracy. By utilizing machine learning algorithms, it is possible to predict the likelihood of heart disease based on multiple clinical parameters such as age, sex, chest pain type, blood pressure, cholesterol levels, ECG results, maximum heart rate, and other physiological indicators. These models can significantly improve prediction effectiveness, reduce diagnostic delays, and support healthcare professionals in delivering early diagnosis and treatment.

This project aims to develop an intelligent system that predicts the presence of heart disease using multiple machine learning techniques and compares their performance to identify the most suitable model. The research involves implementing and evaluating eight machine learning algorithms—Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, XGBoost, and a Neural Network model. The goal is to identify the model that provides the highest accuracy and can be used in real-world medical applications.

The results of this research demonstrate that the **Random Forest model achieved the highest accuracy of 90.16%**, outperforming all other algorithms, including deep learning-based neural networks. This model can be integrated into a real-time medical decision support system, web-based application, or mobile healthcare platform to assist doctors and patients in early diagnosis and preventive care.

Dataset Description

The dataset used in this project plays a crucial role in building an accurate and reliable machine learning model for heart disease prediction. It contains various clinical and physiological attributes collected from patient medical records, which are commonly used by cardiologists to assess cardiovascular health. The dataset represents real-world medical patterns and serves as the foundation for training and testing different machine learning algorithms.

The dataset includes **14 input features** and **one target output attribute**. Each feature corresponds to an important medical indicator that may be associated with heart disease risk. These features help the machine learning model understand the health condition of a patient and classify whether they are likely to have heart disease or not. The **target variable**—also known as the dependent variable—is binary, where:

- **1 indicates the presence of heart disease**, and
- **0 indicates the absence of heart disease**.

Features Included in the Dataset

Feature	**Description**
Age	Age of the patient in years
Sex	Gender of the patient (1 = Male, 0 = Female)
Chest Pain Type (cp)	Type of chest pain experienced (4 categories)
Resting Blood Pressure (trestbps)	Blood pressure level at rest (mm Hg)
Cholesterol (chol)	Serum cholesterol level (mg/dL)
Fasting Blood Sugar (fbs)	Whether fasting blood sugar > 120 mg/dL (1 = True, 0 = False)
Resting ECG (restecg)	Resting electrocardiographic results
Maximum Heart Rate (thalach)	Maximum heart rate achieved
Exercise-Induced Angina (exang)	Pain due to exercise (1 = Yes, 0 = No)
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment

Feature	Description
Ca	Number of major vessels colored by fluoroscopy
Thal	Thalassemia test result (3 = normal, 6 = fixed defect, 7 = reversible defect)

Dataset Characteristics

- Contains numerical as well as categorical attributes
- Consists of a mix of linear and non-linear relationships
- Balanced dataset suitable for classification
- Requires scaling for algorithms like SVM, KNN, and Neural Network

Data Source and Format

The dataset is widely used in academic research and benchmarking studies for cardiovascular disease prediction. It is originally derived from a medical research study and is commonly available through public repositories like UCI Machine Learning Repository and Kaggle. The dataset was loaded into the development environment using Python's Pandas library in **CSV format**, which allowed easy processing and analysis.

Importance of Dataset for ML Model

Accurate prediction highly depends on well-structured data. Therefore, preprocessing steps such as standardization, encoding of categorical values, handling missing data, and splitting into training and testing subsets were applied. These steps ensure that the machine learning models learn effectively, avoid bias, and can generalize well to new unseen data.

In summary, the dataset is comprehensive, medically relevant, and appropriate for evaluating multiple machine learning algorithms. The diverse feature set enables the study to analyze various risk factors, making the model useful for early screening and preventive cardiovascular healthcare.

Problem Statement & Objectives

Problem Statement

Heart disease is a major health concern worldwide, affecting millions of people and contributing significantly to global mortality rates. Timely detection and diagnosis are essential for successful treatment and prevention of severe complications such as heart attacks and strokes. However, traditional diagnostic methods rely heavily on clinical tests, cardiologist expertise, and manual interpretation of patient records. These processes can be time-consuming, costly, and prone to human error. In many cases, patients are diagnosed only after reaching a critical stage, reducing the chances of successful intervention and increasing the cost of treatment.

In recent years, large amounts of patient medical data have become available through hospital records and medical research. However, manually analyzing such large and complex datasets is extremely challenging and inefficient. Conventional diagnostic methods are not capable of automatically identifying hidden patterns in medical data that may indicate early risks of heart disease. Therefore, there is a strong need for an intelligent system that can analyze clinical data accurately and support doctors in predicting the possibility of heart disease at an early stage.

The problem addressed in this project is the **development of an automated machine learning-based prediction model that can accurately classify whether a patient is likely to have heart disease, based on multiple clinical parameters**. Additionally, the project aims to compare different machine learning techniques and determine which algorithm performs best in terms of prediction accuracy, efficiency, and reliability.

Objectives of the Project

The main objectives of this research work are:

- **To apply machine learning techniques to classify heart disease cases** and assist in early diagnosis using medical parameters.
- **To implement multiple ML algorithms** such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, XGBoost, and Neural Network, and evaluate their performance.
- **To analyze and compare model accuracy** and identify the most suitable machine learning model for predicting heart disease.
- **To preprocess and prepare the dataset** using techniques such as normalization, encoding, and train–test splitting, ensuring better training and generalization.

- **To examine the effect of different algorithms on clinical data**, understanding how linear vs. non-linear models behave.
- **To develop a system that can assist medical practitioners** in decision-making through reliable predictions.
- **To lay a foundation for future deployment** as a web or mobile application, enabling real-time medical support and integration with wearable health monitoring devices.

Expected Outcome

The outcome of this project is to identify the model with the highest prediction accuracy and evaluate its potential use for real-world medical decision support. Based on the experimental results, **Random Forest achieved the highest accuracy of 90.16%**, proving to be the best-performing model for predicting heart disease in this study.

Methodology

The methodology adopted in this project consists of a series of systematic steps that enable efficient processing, analysis, and prediction of heart disease using machine learning techniques. The goal of the methodology is to transform raw medical data into a reliable prediction model by applying appropriate preprocessing, selecting suitable machine learning algorithms, training the models, evaluating their performance, and selecting the best-performing approach. The following methodological steps were implemented to successfully complete the project:

1. Understanding and Analyzing the Dataset

The first step involved studying the dataset structure, identifying the available features, and analyzing their relevance to the prediction of heart disease. Descriptive statistics and visual inspection techniques were used to examine data distribution, correlation between features, and initial patterns. Checking for missing values and class distribution helped determine necessary preprocessing needs.

2. Data Preprocessing

Data preprocessing is essential to improve the quality of input data and enable machine learning algorithms to perform efficiently. This step included:

- **Cleaning and handling missing values**
- **Encoding categorical features** such as chest pain type and thalassemia values into numerical format
- **Feature scaling using StandardScaler**, ensuring all numerical attributes lie on a standardized scale (mean = 0, variance = 1), which is especially important for distance-based algorithms like KNN and SVM
- **Splitting the dataset into training and testing subsets** using an 80:20 ratio to ensure unbiased evaluation

3. Model Selection and Training

To determine the most effective algorithm for heart disease prediction, eight different machine learning models were selected and trained. These include Logistic Regression, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, XGBoost, and a Neural Network model using the Multi-Layer Perceptron (MLP). Each model was trained on the same training dataset using Scikit-learn and Keras frameworks to ensure fair comparison.

4. Model Evaluation

After training, each model was evaluated using the testing data. **Accuracy score** was used as the primary performance metric to measure how well each model predicted heart disease outcomes. The results were compared to determine which algorithm best generalized to unseen data. The evaluation was printed and tabulated for analysis.

5. Selection of Best Model

Based on the evaluation, the **Random Forest algorithm achieved the highest accuracy of 90.16%**, outperforming all other models. This is due to its ensemble learning approach, reduction of overfitting, and robustness on small datasets. The neural network achieved 81.97% accuracy, demonstrating that deep learning requires larger datasets and more tuning to reach optimal performance.

6. Implementation and Output Interpretation

The complete implementation was executed using **Python on Google Colab**, leveraging libraries such as Pandas, NumPy, Scikit-learn, Keras, XGBoost, and Matplotlib for visualization. The final output displays accuracy values for all models and highlights Random Forest as the superior model for deployment.

Data Preprocessing

Data preprocessing is a crucial phase in any machine learning project, especially in healthcare applications where the quality and correctness of data directly impact the model's predictive performance and clinical reliability. Raw medical datasets often contain inconsistencies, missing values, noise, and variations in scale, which can degrade model accuracy if not treated properly. Therefore, a systematic preprocessing pipeline was applied to prepare the heart disease dataset for effective machine learning analysis.

The preprocessing steps involved several key operations, including dataset exploration, cleaning, encoding, scaling, and splitting. Initially, the dataset was examined to understand its structure, attribute characteristics, and statistical properties using descriptive analysis techniques such as .info(), .describe(), .shape(), and .isnull().sum(). This analysis verified that the dataset contained no significant missing values, making it suitable for direct modeling without imputation. The dataset consisted of 14 input features and a target output variable indicating whether heart disease was present (1) or absent (0).

The dataset included both numerical and categorical attributes. Since machine learning models require numerical inputs, **categorical values were encoded into numerical format**. For example, attributes like chest pain type (cp), thalassemia (thal), and slope (slope) were converted into numerical classes using label encoding techniques. This step enabled the algorithms to understand categorical patterns effectively without misinterpreting them as text.

One of the most important preprocessing steps was **feature scaling**. Many machine learning algorithms, such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Neural Networks, rely on distance calculations or gradient-based optimization. If features have different ranges (for example, cholesterol values in hundreds and fasting blood sugar in tens), larger values dominate model prediction, causing bias. To resolve this, **StandardScaler was used**, which transforms the data such that each feature has a mean of 0 and variance of 1. This normalization helps in faster convergence and improves accuracy for models sensitive to scale.

After scaling, the dataset was divided into **training and testing subsets** using an 80:20 split ratio. The training set was used to teach the machine learning models, while the testing set evaluated their performance on unseen samples to prevent overfitting and ensure generalization capability.

Overall, preprocessing established a clean and standardized dataset, enabling fair comparison between different machine learning models. Proper preprocessing significantly contributed to the success of the project, resulting in accurate and stable performance, specifically enabling the Random Forest model to achieve the highest prediction accuracy.

Models Used

To identify the most effective algorithm for predicting heart disease, eight different machine learning models were implemented and evaluated in this project. Each algorithm applies a different learning approach, mathematical structure, and prediction strategy. The purpose of using multiple models was to analyze performance differences, understand how each technique handles medical data, and determine the most reliable model for real-world healthcare applications. Below is a detailed explanation of each model used in the study:

1. Logistic Regression

Logistic Regression is a supervised learning algorithm used for binary classification problems. It models the probability of a class using a sigmoid function and predicts outputs between 0 and 1. Due to its linear decision boundary, logistic regression serves as an effective baseline model in medical diagnosis. In this project, it achieved 85.25% accuracy, showing that the dataset contains partially linear relationships.

2. Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that features are conditionally independent, which makes it extremely fast and efficient on small datasets. Although real clinical data features are not fully independent, Naive Bayes still performed well with an accuracy of 85.25%, close to logistic regression.

3. Support Vector Machine (SVM)

SVM is a powerful classification algorithm that finds the optimal hyperplane separating classes by maximizing the margin. It works well for high-dimensional datasets. In this project, SVM achieved 81.97% accuracy. However, performance was affected due to overlapping feature distributions and limited data size.

4. K-Nearest Neighbors (KNN)

KNN is an instance-based classifier that makes predictions by finding the closest neighbors using distance metrics such as Euclidean distance. It does not learn from data but memorizes training samples. Due to noise sensitivity and scaling issues in medical features, KNN achieved the lowest accuracy of 67.21%.

5. Decision Tree

Decision Tree creates a hierarchical tree based on splitting rules using metrics like information gain and Gini index. It captures non-linear relationships effectively and is easy to interpret. However, due to high variance and overfitting risk, the accuracy remained at 81.97%.

6. Random Forest

Random Forest is an ensemble learning technique that combines multiple decision trees using bootstrap sampling and random feature selection. It reduces overfitting and improves accuracy and stability. In this study, Random Forest achieved the highest accuracy of **90.16%**, making it the best model for deployment.

7. XGBoost

XGBoost is an advanced gradient boosting technique that improves model accuracy by sequentially correcting errors from previous models. It is highly effective on large datasets and structured tabular data. Here, XGBoost achieved 83.61% accuracy, slightly lower than Random Forest due to limited data and minimal tuning.

8. Neural Network (MLP Classifier)

A Multi-Layer Perceptron (MLP) neural network was implemented with fully connected layers using ReLU activation and a sigmoid output. Trained for 300 epochs using the Adam optimizer and binary cross-entropy loss, the model achieved 81.97% accuracy. Neural networks typically require large datasets and tuning, which explains the moderate performance in this case.

Implementation

The implementation of the Heart Disease Prediction system was carried out using the Python programming language on the **Google Colab** platform. Google Colab was chosen due to its easy access to GPU support, integrated environment, and ability to handle large datasets efficiently without requiring local system resources. The implementation process included importing necessary libraries, preprocessing the dataset, training multiple machine learning models, evaluating the results, and selecting the best-performing model based on accuracy.

Development Environment

- **Platform:** Google Colab (Jupyter notebook environment)
- **Programming Language:** Python 3
- **Libraries Used:**
    - Pandas and NumPy for data manipulation and analysis
    - Matplotlib and Seaborn for visualizations
    - Scikit-learn for classical machine learning algorithms
    - Keras and TensorFlow for deep learning implementation
    - XGBoost package for gradient boosting

Implementation Steps

The implementation followed a structured approach as described below:

1. Importing Libraries

The project started by importing required Python libraries for reading the dataset, preprocessing, visualization, model training, and performance evaluation.

2. Loading and Exploring the Dataset

The dataset was loaded using the read_csv() function from Pandas. The .info(), .head(), .shape(), and .describe() commands were used to inspect the dataset and understand its characteristics. The analysis confirmed that the dataset was clean and suitable for modeling.

3. Data Preprocessing and Feature Preparation

Preprocessing steps such as encoding categorical values and applying StandardScaler() for feature scaling were implemented. The data was then split into training and testing sets using the train_test_split() function (80% training, 20% testing).

4. Model Training

Eight machine learning models were trained using Scikit-learn's built-in classifiers. Each model was trained using the training dataset and then evaluated using the testing dataset. The models implemented include:

- Logistic Regression
- Naive Bayes
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- XGBoost
- Neural Network (using Keras Sequential API)

Neural Network Implementation Details

A Multi-Layer Perceptron (MLP) model was built using the Keras Sequential model. It consisted of:

- Input layer with 14 neurons (features)
- Hidden layers with **ReLU** activation for non-linearity
- Output layer with **sigmoid** activation for binary classification (0 or 1)
- Optimizer: **Adam**
- Loss Function: **Binary Cross Entropy**
- Epochs: **300**
  The model was trained through forward propagation and backpropagation, updating weights to minimize error.

5. Model Evaluation

All models were evaluated using the accuracy_score metric. A comparison table was created to identify the best-performing model. Random Forest delivered the highest accuracy of **90.16%**, outperforming other models, including the neural network.

6. Final Output and Interpretation

The results were displayed in a structured output showing accuracy values for each model. Visual comparison charts were generated to illustrate model performance differences. Based on evaluation random forest has the highest accuracy.

Results

The primary objective of this project was to implement and compare multiple machine learning algorithms to determine the most accurate predictive model for heart disease detection. After preprocessing the dataset and training eight different machine learning models, the performance of each model was evaluated using the **accuracy score** as the key performance metric. Accuracy was selected because the dataset is relatively balanced, making accuracy a reliable evaluation measure in this context.

Model Accuracy Comparison

The accuracy results for all eight models are displayed in the table below:

Model	**Accuracy (%)**
Logistic Regression	85.25%
Naive Bayes	85.25%
Support Vector Machine (SVM)	81.97%
K-Nearest Neighbors (KNN)	67.21%
Decision Tree	81.97%
Random Forest	**90.16%**
XGBoost	83.61%
Neural Network (MLP)	81.97%

Among all the models tested, **Random Forest achieved the highest accuracy of 90.16%**, making it the best-performing model for predicting heart disease in this research. Random Forest combines multiple decision trees through bagging and bootstrap sampling, which significantly reduces overfitting and variance, allowing it to generalize effectively on the test dataset. Its ability to handle both linear and non-linear relationships within the features contributed to its superior performance.

The second best-performing models were Logistic Regression and Naive Bayes, both achieving an accuracy of 85.25%. These models assume linear or probabilistic patterns among features, which explains why they performed reasonably well but were unable to capture deeper non-linear relationships.

The Neural Network model achieved an accuracy of **81.97%**, which is lower than the Random Forest model. Although neural networks typically outperform classical algorithms,

deep learning requires a large amount of training data and careful hyperparameter tuning. In this project, due to a smaller dataset and limited epochs, the neural network struggled to converge and showed fluctuating accuracy during training. This demonstrates the importance of dataset size and network optimization for achieving optimal deep learning performance.

K-Nearest Neighbors produced the lowest accuracy (67.21%) due to its sensitivity to noise and scaling issues. Since the algorithm relies on distance calculations, feature discrepancies negatively influenced its predictions, making it unsuitable for this dataset.

Conclusion

This project successfully demonstrates the application of machine learning techniques for predicting heart disease based on clinical patient data. By utilizing a structured methodology that included data preprocessing, training of multiple machine learning algorithms, and comparative evaluation, an efficient model was developed to assist in early diagnosis. The project aimed to determine the most effective algorithm for heart disease prediction by analyzing the performance of eight different machine learning models: Logistic Regression, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, XGBoost, and a Neural Network (MLP Classifier).

The results of the study indicate that **Random Forest achieved the highest accuracy of 90.16%**, outperforming all other models tested. The superior performance of Random Forest can be attributed to its ensemble learning mechanism, which combines multiple decision trees to reduce overfitting and improve prediction stability. Its ability to capture both linear and non-linear patterns in structured data made it particularly suitable for this healthcare application. As a result, Random Forest was identified as the most reliable model for heart disease prediction within the scope of this dataset.

The Neural Network model, although powerful in theory, achieved an accuracy of 81.97%, which was lower than the Random Forest model. This performance gap highlights an important observation: deep learning models require significantly larger datasets and intensive hyperparameter tuning to deliver optimal results. In the case of this dataset, which contained a limited number of samples, classical ensemble techniques proved more advantageous than deep neural structures.

Furthermore, K-Nearest Neighbors produced the lowest accuracy at 67.21%, primarily due to its sensitivity to feature scaling, noise, and the curse of dimensionality. The performance comparison clearly emphasizes that model suitability depends on the characteristics of the dataset and the complexity of the relationships within the data.

Overall, the project achieves its objective by identifying a highly accurate heart disease prediction model, demonstrating the potential of machine learning in medical decision support. By automating risk prediction, such models can assist healthcare professionals in early diagnosis, improve treatment planning, and enhance patient care outcomes. This research contributes to advancing intelligent healthcare applications and establishes a foundation for integrating predictive analytics into clinical systems.

The successful implementation and findings of this project confirm that machine learning can be a powerful asset in healthcare analytics and can significantly improve the efficiency and reliability of disease prediction systems.

References

1. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Heart Disease Dataset*. University of California, Irvine, School of Information and Computer Sciences. Retrieved from https://archive.ics.uci.edu

2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

3. Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

4. Chollet, F. (2015). *Keras: Deep Learning Library for Theano and TensorFlow*. GitHub repository. Retrieved from https://keras.io

5. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., & Chen, Z. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. Google Research. https://www.tensorflow.org

6. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

7. Jović, A., Brkić, K., & Bogunović, N. (2015). *Review of Feature Selection Methods in Medical Data Processing*. IEEE 38th International Convention on Information and Communication Technology.

8. World Health Organization. (2021). *Cardiovascular Diseases (CVDs)*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases

9. Palaniappan, S., & Awang, R. (2008). *Intelligent Heart Disease Prediction System Using Data Mining Techniques*. IEEE Computer and Communication Engineering, 108–115.

10. Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-measure to ROC, Informedness and Correlation*. Journal of Machine Learning Technologies, 2(1), 37–63.