Heart Disease Prediction

17.07.2024

Group 7

Ganesh Krishna - AM.EN.U4AIE22116

Abhai A - AM.EN.U4AIE22155

Vyvidh S Krishna - AM.EN.U4AIE22156

Aravind Krishnan - AM.EN.U4AIE22161

Table of Contents

1. Abstract

Heart disease, also known as cardiovascular disease (CVD), is a term that encompasses a range of heart-related conditions and has become the leading cause of annual deaths globally, driven largely by changes in lifestyle. Early identification of high-risk patients is crucial in mitigating the impact of heart disease and reducing its prevalence. This project aims to develop robust machine learning models to accurately classify the presence or absence of heart disease in patients using three distinct datasets.

Classification, a supervised machine learning technique, will be the primary method employed to predict heart disease. By identifying patterns and relationships between various patient attributes—such as age, gender, blood pressure, cholesterol levels, and other relevant health metrics—and the likelihood of heart disease, the models will provide valuable insights into patient risk profiles.

In this project, we will explore and compare multiple classification algorithms, including K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Support Vector Machine (SVM), to determine which model offers the highest predictive accuracy. The evaluation of these models will be based on metrics such as Precision-Recall Area Under Curve (PR AUC) scores to ensure the most effective and reliable prediction outcomes.

By leveraging advanced machine learning techniques and diverse datasets, this project not only aims to enhance the predictive accuracy of heart disease classification but also seeks to contribute to the broader field of healthcare by enabling earlier interventions. Ultimately, the goal is to provide a tool that can assist healthcare professionals in identifying and managing high-risk patients, thereby potentially saving many lives and alleviating the burden of heart disease on a global scale.

2. Introduction

Background Information on Heart Disease and Its Global Impact

Heart disease, also known as cardiovascular disease (CVD), refers to a group of disorders affecting the heart and blood vessels. These conditions include coronary artery disease, arrhythmias, heart failure, and congenital heart defects. Heart disease is the leading cause of death globally, responsible for approximately 17.9 million deaths each year, which accounts for 31% of all global deaths. The primary risk factors for heart disease include unhealthy diet, lack of physical activity, tobacco use, and excessive alcohol consumption. As lifestyles have evolved, the prevalence of these risk factors has increased, exacerbating the incidence of heart disease.

The burden of heart disease extends beyond mortality rates. It also imposes a significant economic strain on healthcare systems and affects the quality of life for millions of individuals. Early detection and management of heart disease are crucial in mitigating its impact, as timely intervention can prevent severe complications and improve patient outcomes.

Motivation for the Project and Its Significance

Given the substantial global impact of heart disease, there is a pressing need for effective methods to identify individuals at high risk of developing this condition. Traditional diagnostic methods, while effective, can be time-consuming and expensive. With advancements in technology, particularly in the field of machine learning, there is an opportunity to develop more efficient and accurate predictive models that can aid in the early detection of heart disease.

The motivation for this project stems from the potential of machine learning to revolutionize healthcare by providing tools that can predict heart disease with high accuracy. By leveraging machine learning algorithms, we can analyze large datasets of patient information to uncover patterns and relationships that are not immediately apparent to human clinicians. This project aims to contribute to the growing body of research in this area and provide a practical solution that can be

implemented in clinical settings to improve patient outcomes and reduce the burden of heart disease.

Objectives and Goals of the Project

The primary objective of this project is to develop robust machine learning models that can accurately classify the presence or absence of heart disease in patients. To achieve this, we will utilize three different datasets, each containing a variety of patient attributes related to heart health.

By achieving these goals, this project aims to provide a valuable tool for healthcare professionals, aiding in the early detection and management of heart disease, ultimately saving lives and reducing healthcare costs.

3. Literature Review

Overview of Existing Research and Studies Related to Heart Disease Prediction Using Machine Learning

Heart disease prediction has been a prominent area of research within the field of machine learning. Numerous studies have been conducted to develop models that can accurately predict the likelihood of heart disease based on various patient attributes. These studies typically utilize datasets containing medical records and health indicators such as age, gender, blood pressure, cholesterol levels, and lifestyle factors. The primary goal is to identify patterns and correlations that can be used to predict heart disease.

Discussion of Methods and Findings from Previous Work

1. **Logistic Regression and Decision Trees**: One of the foundational studies in this domain employed logistic regression and decision tree algorithms to predict heart disease. The findings indicated that logistic regression, a statistical method for binary classification, provided a straightforward approach with reasonably good accuracy. Decision trees, on the other hand, offered better interpretability, allowing healthcare professionals to understand the decision-making process of the model (Dua & Graff, 2019).
2. **Random Forest and Ensemble Methods**: More recent studies have explored the use of ensemble methods, such as Random Forest, to improve predictive performance. Random Forest combines multiple decision trees to enhance accuracy and robustness. Research has shown that Random Forest models outperform single decision tree models by reducing overfitting and capturing more complex interactions between features (Caruana & Niculescu-Mizil, 2006).
3. **Support Vector Machines (SVM)**: Support Vector Machines have also been widely used in heart disease prediction. SVMs are effective in high-dimensional spaces and have been shown to perform well with both linear and non-linear relationships. Studies have demonstrated that SVMs can achieve high accuracy rates, particularly when combined with techniques like kernel functions to handle non-linearity (Akay, 2009).

4. **Neural Networks and Deep Learning**: With the advent of deep learning, neural networks have gained traction in medical predictions. Neural networks, particularly deep neural networks (DNNs), can model complex patterns in data. Research has shown that DNNs can achieve high predictive accuracy, though they require large amounts of data and computational resources. Studies leveraging neural networks have reported improved performance over traditional machine learning models, especially when dealing with large and complex datasets (Ravi et al., 2017).

Identification of Gaps in Current Research

Despite the significant advancements in heart disease prediction using machine learning, several gaps remain:

1. **Data Quality and Availability**: Many studies rely on publicly available datasets that may not fully represent diverse populations or contain comprehensive patient information. There is a need for more extensive and varied datasets to improve model generalizability and applicability across different demographic groups.
2. **Model Interpretability**: While complex models like deep neural networks offer high accuracy, they often lack interpretability. Healthcare professionals require models that not only provide accurate predictions but also offer insights into the decision-making process to ensure trust and accountability.
3. **Integration into Clinical Practice**: Many studies focus on developing predictive models but do not address the practical challenges of integrating these models into clinical workflows. There is a need for research on the implementation of machine learning models in real-world healthcare settings, including considerations for user interface design, data security, and regulatory compliance.

Addressing Research Gaps

This project aims to address these gaps by:

1. **Utilizing Multiple Datasets**: We will use three different datasets to enhance the diversity and representativeness of the data, improving the generalizability of the models.
2. **Balancing Accuracy and Interpretability**: We will compare multiple machine learning models, including both traditional and advanced algorithms, to balance predictive accuracy with interpretability.

By addressing these gaps, this project aims to contribute heart disease prediction by comparing various models such as KNN, Logistic Regression, Decision Trees, Random Forests and SVM

4. Datasets Used

Description of the Datasets Used

For this project, we have utilized three distinct datasets, each offering unique features and patient records to ensure a comprehensive analysis. Below is a detailed description of each dataset, including their sources, number of records, and types of features:

1. Heart Disease Dataset (heart.csv)

This dataset was created by combining five different heart disease datasets, each containing similar features. The combined dataset provides a robust collection of patient records, making it the largest heart disease dataset available for research purposes.

- **Source**: The datasets were collected from the UCI Machine Learning Repository, specifically from the heart disease datasets.
- **Original Datasets**:
    - **Cleveland**: 303 observations
    - **Hungarian**: 294 observations
    - **Switzerland**: 123 observations
    - **Long Beach VA**: 200 observations
    - **Stalog (Heart) Data Set**: 270 observations
    - **Total**: 1190 observations
    - **Duplicated**: 272 observations
    - **Final Dataset**: 918 observations

Features:

- Age: Age of the patient (years)
- Sex: Sex of the patient (M: Male, F: Female)
- ChestPainType: Chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)
- RestingBP: Resting blood pressure (mm Hg)
- Cholesterol: Serum cholesterol (mg/dl)

- FastingBS: Fasting blood sugar (1 if FastingBS > 120 mg/dl, 0 otherwise)
- RestingECG: Resting electrocardiogram results (Normal, ST, LVH)
- MaxHR: Maximum heart rate achieved
- ExerciseAngina: Exercise-induced angina (Y: Yes, N: No)
- Oldpeak: ST depression induced by exercise relative to rest
- ST_Slope: The slope of the peak exercise ST segment (Up: upsloping, Flat: flat, Down: downsloping)
- HeartDisease: Presence of heart disease (1: Yes, 0: No)

Link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

2. Cardiovascular_Disease_Dataset

This dataset was acquired from a multispecialty hospital in India. It includes comprehensive patient records with features relevant to heart disease prediction.

- **Source**: Kaggle
- **Number of Records**: 1000 subjects
- **Number of Features**: 14 features

Features:

- Age: Age of the patient (years)
- Sex: Sex of the patient (M: Male, F: Female)
- ChestPainType: Chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)
- RestingBP: Resting blood pressure (mm Hg)
- Cholesterol: Serum cholesterol (mg/dl)
- FastingBS: Fasting blood sugar (1 if FastingBS > 120 mg/dl, 0 otherwise)
- RestingECG: Resting electrocardiogram results (Normal, ST, LVH)
- MaxHR: Maximum heart rate achieved (numeric value between 60 and 202)
- ExerciseAngina: Exercise-induced angina (Y: Yes, N: No)
- Oldpeak: ST depression induced by exercise relative to rest (numeric value)
- ST_Slope: The slope of the peak exercise ST segment (Up: upsloping, Flat: flat, Down: downsloping)
- HeartDisease: Presence of heart disease (1: Yes, 0: No)

Link: https://data.mendeley.com/datasets/dzz48mvjht/1

3. Framingham Heart Study Dataset (framingham.csv)

This dataset originates from the Framingham Heart Study, an ongoing cardiovascular study involving the residents of Framingham, Massachusetts. It includes demographic, behavioral, and medical history data to predict the 10-year risk of coronary heart disease.

- **Source**: Kaggle
- **Number of Records**: Data from multiple cohorts, spanning several decades

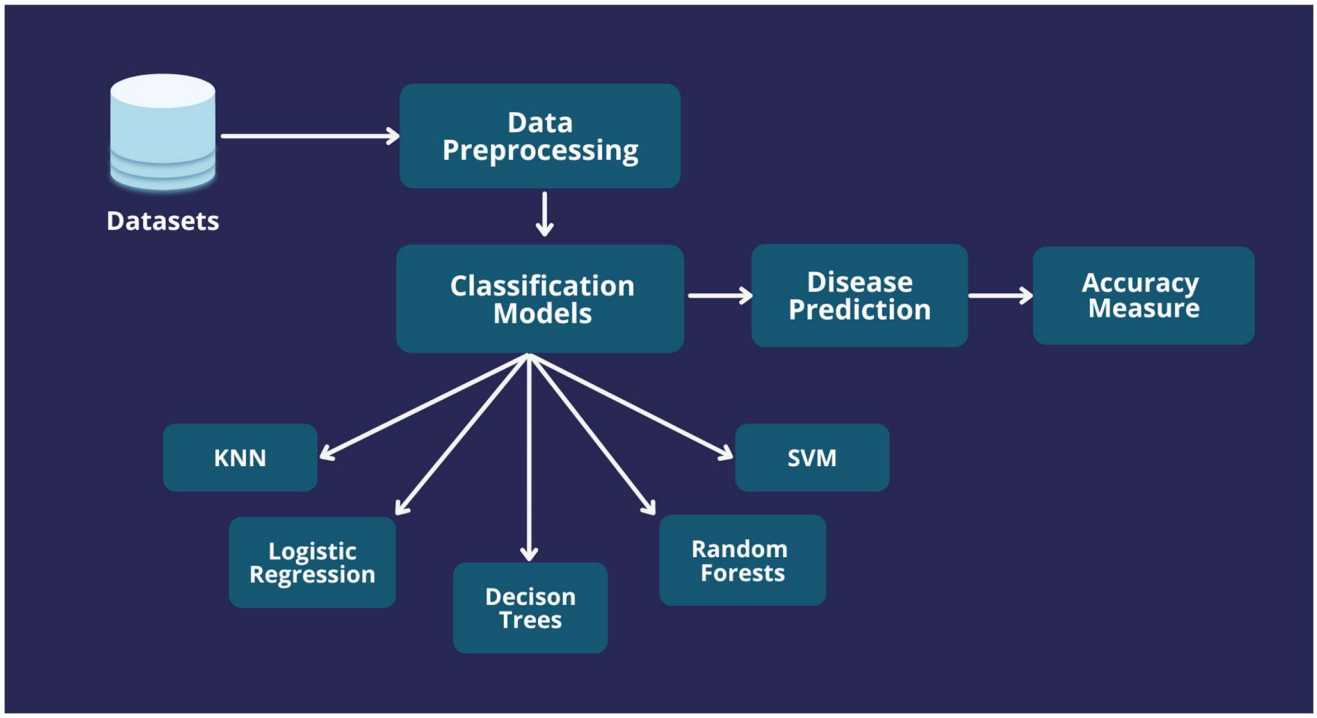
Features:

- Sex: Male or Female (Nominal)
- Age: Age of the patient (Continuous)
- Current Smoker: Whether the patient is a current smoker (Nominal)
- Cigs Per Day: Number of cigarettes smoked per day (Continuous)
- BP Meds: Whether the patient is on blood pressure medication (Nominal)
- Prevalent Stroke: History of stroke (Nominal)
- Prevalent Hyp: History of hypertension (Nominal)
- Diabetes: Whether the patient has diabetes (Nominal)
- Tot Chol: Total cholesterol level (Continuous)
- Sys BP: Systolic blood pressure (Continuous)
- Dia BP: Diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- Heart Rate: Heart rate (Continuous)
- Glucose: Glucose level (Continuous)
- 10 Year CHD: 10-year risk of coronary heart disease (Binary: 1 = Yes, 0 = No)

Link:http://www.who.int/mediacentre/factsheets/fs317/en/

These datasets will be crucial for developing and validating our machine learning models, allowing us to build a comprehensive tool for heart disease prediction.

5. Methodology



Data Preprocessing Steps

Data preprocessing is a crucial step to ensure the datasets are clean and suitable for analysis. Below are the steps taken for preprocessing each dataset.

Cardiovascular Disease Dataset

1. **Removed duplicate rows:**
    - The dataset initially contained 1190 observations, with 272 duplicates. Removing these duplicates resulted in a final dataset of 918 observations.
2. **Handled missing values**:
    - Identified 53 occurrences of 0 values and 5 instances of values exceeding 564 in the 'Serum Cholesterol' feature, which was inconsistent with the expected range of 126-564 mg/dL.
    - Replaced 0 values with NaN to address this anomaly.
3. **Dealing with Noise**:

    - Applied KNN imputation to handle missing values and reduce noise. This improved the correlation between the target variable and serum cholesterol from 0.19 to 0.3777.
4. **Feature Engineering**:
    - Verified that the 'patientid' column contained unique values for each patient and subsequently dropped it as it was no longer necessary for analysis.

Framingham Heart Study Dataset (framingham.csv)

1. **Handled missing values**:
    - For categorical features such as 'BPmeds' and 'education', the missing values were filled using the mode.
    - For numerical features like 'BMI' and 'heart rate', the mean was used.
    - Applied KNN imputation for the remaining features.
2. **Handling Imbalanced Data**:
    - Identified highly imbalanced features such as 'BPMeds', 'prevalentStroke', 'diabetes', and 'TenYearCHD'.
3. **Outlier Detection and Handling**:
    - Detected outliers in features like 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', and 'glucose'.
    - Two approaches used for handling outliers:
        - Replacing outliers with values of the lower or upper whisker.
        - Replacing outliers with the mean values of the respective feature.

Heart Disease Dataset (heart.csv)

1. **Encoding Categorical Features**:
    - Label Encoding for binary features ('Sex', 'ExerciseAngina').
    - Applied different encoding methods for non-binary categorical features:
        - Label Encoding in DataFrame df1.
        - One-Hot Encoding in DataFrame df2.
        - Ordinal Encoding in DataFrame df3.

2. **Handling Outliers**:
    - Detected outliers in 'RestingBP', 'Cholesterol', 'MaxHR', and 'Oldpeak'.
    - Replaced outliers using:
        - Values below the lower whiskers with the lower whiskers.
        - Values above the higher whiskers with the higher whiskers.
        - Alternatively, replaced outliers with the mean value of the respective feature.

Detailed Explanation of Machine Learning Models and Algorithms Used

1. **K-Nearest Neighbors (KNN)**:
    - A simple, non-parametric algorithm used for classification and regression.
    - Classifies a data point based on the majority class among its k-nearest neighbors.
2. **Logistic Regression**:
    - A statistical model that uses a logistic function to model a binary dependent variable.
    - Effective for binary classification problems like heart disease prediction.
3. **Decision Trees**:
    - A tree-like model of decisions and their possible consequences.
    - Splits data into subsets based on the value of input features.
4. **Random Forest**:
    - An ensemble method that creates multiple decision trees and merges them to get a more accurate and stable prediction.
    - Reduces overfitting and improves accuracy.
5. **Support Vector Machine (SVM)**:
    - A supervised learning model used for classification and regression analysis.
    - Finds the hyperplane that best divides a dataset into classes.

Feature Selection and Engineering Processes

1. **Cardiovascular Disease Dataset**:
    - Dropped the 'patientid' column.
    - Encoded categorical variables using appropriate methods.
2. **Framingham Heart Study Dataset**:
    - Filled missing values using mode and mean, and applied KNN imputation for other features.
    - Handled outliers using two different approaches for robustness.
3. **Heart Disease Dataset**:
    - Encoded categorical features using label encoding, one-hot encoding, and ordinal encoding.
    - Handled outliers by replacing them with whisker values or mean values.

Model Training and Validation Process

1. **Data Splitting**:
    - Split each dataset into training and testing sets to evaluate model performance.
2. **Model Training**:
    - Trained each model using the training set.
    - Applied cross-validation to ensure the model's generalization capability.
3. **Model Validation**:
    - Validated the models using the testing set.
    - Evaluated performance using appropriate metrics.

Techniques Used for Model Evaluation

Precision-Recall Area Under the Curve (PR AUC):

- Used to measure the trade-off between precision and recall for different thresholds, especially for imbalance datasets
- This is because when we are evaluating we need to ensure that those with heart disease(positive class) are not incorrectly predicted as negative. The vice versa is manageable in the further detection process.

6. Results

Cardiovascular_Disease_Dataset

Presentation of Model Performance Metrics

Here are the PR AUC scores for each model across the three datasets:

Cardiovascular Disease Dataset:

Model	DataFrame	PR AUC
KNN	df1	0.99
Logistic Regression	df1	1.00
Decision Tree	df1	0.97
Random Forests	df1	1.00
SVM	df1	1.00

df : with noise
df1 : without noise

Framingham Dataset (framingham.csv)

Model	DataFrame	PR AUC
KNN	df2	0.25
Logistic Regression	df1	0.21
Decision Tree	df1	0.23
Random Forests	df1	0.25
SVM	df1 & df2	0.21

df : original
df1 : outliers were replaced with the values of the lower or upper whisker
df2: outliers were replaced with the mean values

Heart Disease Dataset (heart.csv)

Model	DataFrame	PR AUC
KNN	df3_1	0.96
Logistic Regression	df2_2	0.94
Decision Tree	df1_1	0.95
Random Forests	df1_1 & df3_1	0.96
SVM	df1_1	0.97

df1 : label encoding
df2 : one hot encoding
df3 : ordinal encoding

df*_1: replacing outliers with lower and higher whiskers
df*_2: replacing outliers with mean value of the respective feature
*** = [1,2,3]**

Comparison of Model Performance

The models performed differently across the datasets:

- **Cardiovascular Disease Dataset**: Logistic Regression, Random Forests, and SVM achieved perfect PR AUC scores (1.00), indicating excellent performance. KNN also performed well with a score of 0.99, while the Decision Tree had a slightly lower score of 0.97.
- **Framingham Dataset**: All models struggled with lower PR AUC scores, with KNN and Random Forests achieving the highest score of 0.25. Logistic Regression, Decision Tree, and SVM had scores around 0.21-0.23, indicating poor performance, likely due to the imbalanced nature of the dataset.
- **Heart Disease Dataset**: SVM achieved the highest PR AUC score of 0.97, followed by KNN and Random Forests with 0.96, Decision Tree with 0.95, and Logistic Regression with 0.94. All models performed well, but SVM had a slight edge.

7. Discussion

Interpretation of the Results

- **Cardiovascular Disease Dataset**: The models performed exceptionally well, with Logistic Regression, Random Forests, and SVM achieving perfect PR AUC scores. This suggests that the dataset is well-suited for these algorithms and the preprocessing steps were effective.
- **Framingham Dataset**: The low PR AUC scores indicate that the models struggled with this dataset, likely due to its imbalanced nature. Despite preprocessing efforts, the imbalance in the target variable may have led to poor model performance.
- **Heart Disease Dataset**: All models performed well, with SVM achieving the highest PR AUC score. This indicates that the dataset is well-suited for these algorithms, and the encoding and outlier handling techniques were effective.

Discussion of Strengths and Limitations of Each Model

- **KNN**:
    - **Strengths**: Simple and intuitive, performs well with a sufficient amount of data.
    - **Limitations**: Sensitive to the choice of k and distance metric, computationally expensive for large datasets.
- **Logistic Regression**:
    - **Strengths**: Easy to implement, interpretable, performs well on linearly separable data.
    - **Limitations**: Assumes linear relationship between features and target, may struggle with complex relationships.
- **Decision Tree**:
    - **Strengths**: Easy to interpret, handles both numerical and categorical data, captures non-linear relationships.
    - **Limitations**: Prone to overfitting, sensitive to small changes in data.
- **Random Forests**:

    - **Strengths**: Reduces overfitting, handles high-dimensional data well, robust to noise.
    - **Limitations**: Less interpretable than individual decision trees, can be computationally expensive.
- **SVM**:
    - **Strengths**: Effective in high-dimensional spaces, works well with clear margin of separation.
    - **Limitations**: Computationally intensive, sensitive to the choice of kernel and parameters.

Insights Gained from the Project

- Preprocessing steps like handling missing values, encoding categorical features, and dealing with outliers significantly impact model performance.
- Imbalanced datasets, like the Framingham dataset, require special techniques like undersampling and stratified k-fold cross-validation to improve model performance.
- Cross-validation is essential to ensure models are not overfitting or underfitting.

8. Conclusion

Summary of Key Findings and Their Implications

- **Cardiovascular Disease Dataset**: Logistic Regression, Random Forests, and SVM are highly effective, achieving perfect PR AUC scores. This indicates these models are well-suited for heart disease prediction on this dataset.
- **Framingham Dataset**: The models struggled with low PR AUC scores due to dataset imbalance. Techniques like undersampling and stratified k-fold cross-validation are necessary to address this issue.
- **Heart Disease Dataset**: SVM performed the best, followed closely by KNN and Random Forests, indicating these models are well-suited for heart disease prediction on this dataset.

Potential Impact of the Project on Heart Disease Diagnosis and Management

- This project demonstrates the potential of machine learning models to accurately predict heart disease, which can lead to early diagnosis and better management of the condition.
- Implementing these models in clinical settings can help healthcare professionals make informed decisions, ultimately improving patient outcomes.

Recommendations for Future Research and Improvements

- **Address Imbalanced Data**: Apply techniques like undersampling, oversampling, and stratified k-fold cross-validation to improve model performance on imbalanced datasets.
- **Explore Additional Models**: Investigate other machine learning models like Gradient Boosting, XGBoost, and Neural Networks for potentially better performance.
- **Model Interpretability**: Focus on model interpretability to ensure they provide insights that can be understood by healthcare profession

9. References

1. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K. (2003). KNN Model-Based Approach in Classification. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg.

2. L. Li et al., "Classification of heart sound signals with BP neural network and logistic regression," 2017 Chinese Automation Congress (CAC), Jinan, China, 2017, pp. 7380-7383, doi: 10.1109/CAC.2017.8244111.

3. M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi and C. S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 3, pp. 559-566, May 2010, doi: 10.1109/TITB.2009.2038906.

4. A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," in IEEE Access, vol. 7, pp. 180235-180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

5. Ch. Usha Kumari, A. Sampath Dakshina Murthy, B. Lakshmi Prasanna, M. Pala Prasad Reddy, Asisa Kumar Panigrahy,An automated detection of heart arrhythmias using machine learning technique: SVM,Materials Today: Proceedings,Volume 45, Part 2,2021,Pages 1393-1398,ISSN 2214-7853,