A Novel Approach for Classification of Heart Disease Using Machine Learning

P. V. Kavitha¹, C. Naveen Kumar², M. Ajith³, K. Dayasudan⁴
¹Assistant Professor, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India
²,³,⁴Student, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India

*Corresponding author: naveen300998@gmail.com

Abstract: Heart disease is one of the hugest reasons for mortality on the planet today. Expectation of cardiovascular disease is a basic test in the region of clinical data investigation. Health care field has an enormous proportion of data, for setting up those data certain methods are used. Data mining is one of the strategies routinely used. Heart disease is the Leading purpose behind death around the globe. This System predicts the developing possibilities of Heart Disease. The results of this System give the odds of happening heart disease as far as percentage. The datasets used are classified with respect to clinical parameters. This System assesses those parameters using data mining classification strategy. In this paper, we propose a novel strategy that targets finding noteworthy highlights by applying AI strategies bringing about improving the precision in the forecast of cardiovascular disease. The expectation model is presented with various mixes of highlights and a few known classification methods. The datasets are dealt with in python programming using two major Machine Learning Algorithm to be explicit Random Forest Algorithm and Gradient Boosting Tree Algorithm which shows the best algorithm among these two in regards to precision level of heart disease.

Keywords: Cardiovascular, Data mining, Machine learning techniques, Random Forest, Gradient boosting Tree algorithm.

1. Introduction

Heart diseases have developed as one of the most conspicuous reason for death all around the globe. Heart diseases have executed 1.7 million Indians in 2016, as indicated by the 2016 Global Burden of Disease Report. World Health Organization (WHO), recommend that India have lost up to $237 billion, from 2005-2015, because of heart related or Cardiovascular diseases. In this manner, practical and exact expectation of heart related diseases is significant. In this day and age hard to recognize heart disease on account of a few contributory hazard factors, for example, diabetes, hypertension, elevated cholesterol, anomalous heartbeat rate and numerous different components. Heart disease is a term covering any confusion of the heart. In contrast to cardiovascular disease, which depicts issues with the veins and circulatory framework just as the heart, heart disease alludes to issues and distortions in the heart itself. Different procedures in data mining and neural systems have been utilized to discover the seriousness of heart disease among people. The seriousness of the disease is classified dependent on different techniques like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB). The idea of heart disease is mind boggling and consequently, the disease must be taken care of cautiously. Not doing so may influence the heart or cause sudden passing. The point of view of clinical science and data mining are utilized for finding different sorts of metabolic conditions. Data mining with classification assumes a noteworthy job in the expectation of heart disease and data examination. Different techniques have been utilized for information deliberation by utilizing known strategies for data mining for forecast of heart disease. We can likewise observe the different AI model for anticipating heart disease. However, there is exactness not meet all the case. They are utilized different component for foreseeing heart disease. At the point when we are utilizing an electrocardiogram (ECG) flags that is a finished portrayal of the electrical movement of the heart on the outside of the human body, and it is broadly applied in the clinical conclusion of heart diseases. ECG signals have been generally utilized for recognizing heart diseases because of its straightforwardness and non-obtrusive nature. Highlights of ECG signs can be figured from ECG tests and separated utilizing some product (ex: MATLAB). For example, a large number of individuals experience the ill effects of sporadic heartbeats which can be deadly now and again. Numerous investigations have created arrhythmia classification moves toward that utilization programmed examination and conclusion frameworks dependent on ECG signals. The most significant components for the investigation and finding of cardiovascular diseases are highlights extraction and beats classification. Various strategies for characterizing ECG signals were proposed as of late and great outcomes accomplished. Enormous data systematic assumes a crucial job in dealing with the gigantic measure of health-care data and improving the nature of health-care administrations offered to patients. In this specific situation, one of the difficulties lies in the classification of data, which depends on adequately circulated preparing stages, propelled data mining and AI methods. In this manner, a Big data strategy is acquainted in this work with address the difficulties looked by group the ECG beats.
2. Motivation

This section provides an introduction of the existing methods in predicting the diseases related to heart and their accuracies.


Heart disease is one of the most critical reasons for mortality on the planet today. Expectation of cardiovascular disease is a basic test in the region of clinical data examination. AI (ML) has been demonstrated to be powerful in helping with settling on choices and expectations from the enormous amount of data delivered by the healthcare business. We have likewise observed ML strategies being utilized in ongoing improvements in various territories of the Internet of Things (IoT). Different investigations give just a brief look into foreseeing heart disease with ML procedures. In this paper, we propose a novel strategy that targets finding huge highlights by applying AI strategies bringing about improving the exactness in the expectation of cardiovascular disease. The forecast model is presented with various mixes of highlights and a few known classification strategies. We produce an upgraded presentation level with a precision of 88.7% through the forecast model for heart disease with the mixture arbitrary woods with a straight model (HRFLM).


This examination proposed an ECG (Electrocardiogram) classification approach utilizing AI dependent on a few ECG highlights. An electrocardiogram (ECG) is a sign that gauges the electric action of the heart. The proposed approach is actualized utilizing ML/libs and Scala language on Apache Spark system; MLlib is Apache Spark's versatile AI library. The key test in ECG classification is to deal with the inconsistencies in the ECG signals which is essential to distinguish the patient status. Hence, we have proposed a proficient way to deal with arrange ECG signals with high exactness Each heartbeat is a mix of activity drive waveforms delivered by various particular cardiovascular heart tissues. Heartbeats classification faces a few challenges in light of the fact that these waveforms contrast from individual to another, they are depicted by certain highlights. These highlights are the contributions of AI algorithm. All in all, utilizing Spark–Scala apparatuses rearranges the use of numerous algorithms, for example, AI (ML) algorithms. On other hand, Spark–Scala is liked to be utilized more than different apparatuses when size of preparing data is excessively enormous. For our situation, we have utilized a dataset with 205,146 records to assess the presentation of our methodology. AI libraries in Spark–Scala give simple approaches to actualize numerous classification algorithms (Decision Tree, Random Forests, Gradient-Boosted Trees (GDB), and so forth.). The proposed technique is assessed and approved on pattern MIT-BIH Arrhythmia and MIT-BIH Supraventricular Arrhythmia database. The outcomes show that our methodology accomplished a general exactness of 96.75% utilizing GDB Tree algorithm. For multi class classification, it accomplished to 98.03% precision utilizing Gradient Boosting tree bolsters just twofold classification.


This has high mortality brought about via cardiovascular diseases. To limit the mortality, we manufacture a tele-ecg framework for heart diseases early recognition and observing. In this examination, the tele-ecg framework was improved utilizing Hadoop structure, so as to manage huge data preparing. The framework was based on bunch PC with 4 hubs. The server can deal with 60 demands simultaneously. The framework can arrange the ECG data utilizing choice tree and irregular woods. The precision is 97.14% and 98.92% for choice tree and arbitrary woods separately. Preparing process in irregular backwoods is quicker than in choice tree, while testing process in choice tree is quicker than in arbitrary woodland.

3. Proposed System

The proposed system uses Regression algorithm for improving the accuracy of classifying heart disease. It compares the proposed approach with Random Forest and Gradient-Boosted Trees (GBT).

A. Overview

In this paper, multiples classifiers are proposed for Heart disease classification, these classifiers are used mostly in Machine Learning fields. Each classifier influences the final decision according to its performance on the training data. Parameters of each classifier are adjusted on the basis of an individual classifier’s performance on the training data by applying the ML techniques. The classification performance was validated on a set of 70000 records with different features. So, our work is distinguished by:

- To characterize the dataset as indicated by the heart diseases
- Train the model using the classified dataset
• Using Machine learning algorithms for classification. Increase the efficiency of the model and compare it with existing model.

B. Problem definition

To build up the Machine Learning model for classification of heart disease by utilizing various kinds of highlights who have just influenced heart disease. To acquire high precision even within the sight of greatest number of missing highlights with less manual work by utilizing Random Forest model in AI.

C. Dataset collection

The development of this work requires a database with digital ECG records for computational analysis of many different patients with different pathologies. The dataset was taken from “https://www.kaggle.com/sulianova/cardiovasculardisease-dataset/data” is contains 70000 records and 13 features. And another one dataset was taken from https://archive.ics.uci.edu/ml/datasets/Heart+Disease is contains 300 record and 12 features.

D. Feature acquisition and storing

Features were obtained using Jupyter notebook software and stored in csv file with known columns types; some columns are integer type and others are double types.

E. Preprocessing

Processing was implemented using Jupyter notebook software for classification. Jupyter notebook software is very helpful tool in classification problems on other hand, machine learning algorithms are not implemented easily in MATLAB. This case can be summarized as follows:

• Dataset size is too large.
• Need to implement algorithm like: Decision Tree, Random Forests, Gradient Boosted Trees.
• Processing speed

So, using Jupyter notebook would be helpful in this case. Every generic model of machine learning consists of some components independent of the algorithm adopted. In our case, they are:

• Pre-processing
• Features Selection
• Using algorithm Random Forest, GBT

F. Approaches

1) Random Forest model (RF model)

Random forests are a troupe learning technique for characterization, relapse and various endeavors that works by building countless choice trees at getting ready time and yielding the class that is the strategy for the classes (order) or mean desire (relapse) of the individual trees. The made model depended on a very basic level upon two parameters of random woods; these two parameters are most noteworthy parameters of RF. RF arranged model was worked by various estimations of Number of Trees and Max Depth.

Precondition: A training set $S := (x_1, y_1), \ldots , (x_n, y_n)$, features F, and number of trees in forest $B$.

```
1. function RandomForest (S, F)
2.     H ← ∅
3.     for i ∈ 1, . . . , B do
4.         S (i) ← A bootstrap sample from S
5.         hi ← RandomizedTreeLearn (S (i), F)
6.         H ← H ∪ {hi}
7.     end for
8.     return H
9. function RandomizedTreeLearn (S, F)
10.     At each node:
11.         f ← very small subset of F
12.         Split on best feature in f
13.         return the learned tree
14. end function
```

2) KNN

Naive Bayes classifier is a supervised algorithm which characterizes the dataset based on Bayes theorem. The Bayes theorem is a standard or the numerical idea that is utilized to get the likelihood is called Bayes theorem. Bayes theorem requires some autonomous supposition and it requires free factors which is the essential presumption of Bayes theorem. Bayes theorem on Mathematical Representation:

$$P(A | B) = \frac{P(B | A) * P(A)}{P(B)}$$

Here,

$P(A) \Rightarrow$ independent probability of A (prior probability)
$P(B) \Rightarrow$ independent probability of B
$P(B|A) \Rightarrow$ conditional probability of B given A
$P(A|B) \Rightarrow$ conditional probability of A given B (posterior
Naive Bayes is a straightforward and amazing algorithm for predictive modeling. This model is the best and proficient classification algorithm which can deal with huge, entangled, non-straight, subordinate data. Naive involves two section to be specific naive and Bayes where naive classifier expects that the nearness of the specific element in a class is disconnected to the nearness of some other component.

3) Decision Tree Algorithm

The decision tree is a supervised machine learning algorithm. It handles both the categorical data and numerical data. Based on certain conditions it gives a categorical solution such Yes/No, True or false, 1 or 0. For handling medical dataset the Decision tree Classification algorithm is widely used. The result of this model differing from the other models like the knn model, SVM model. The output consists of horizontal and vertical line splits based on the condition depends on the dependent variables. The accuracy level of this algorithm is quite higher than the other algorithms. The reason for the higher accuracy of this algorithm is this model analysis the dataset in the tree shape format. Thus, each and every attribute of the dataset is been analyzed. Thus, the accuracy rate of this model is higher. This model analyzes the data in the tree-shaped structure. Tree shaped diagram determines the course of actions. The decision tree model analyzes the data on the basis of three nodes namely

- Root node - Basis of this node all other perform it Function
- Interior node - the condition of dependent variables
- Leaf node - the final result is leaf node.

Formula for finding root node (Information Gain)

\[
\text{Information Gain} = \text{Class Entropy} - \text{Entropy Attributes}
\]

To find Class Entropy:

\[
(P_i + N_i) = - \frac{P}{P + N} \log_2 \frac{P}{P + N}
\]

Here => P, Possibilities of Yes.
=> N, Possibilities of No.
To find Entropy Attributes:
Entropy attribute = \( \sum \frac{P_i + N_i}{P + N} \)

4) Gradient Boosting Tree

Gradient boosting is an AI strategy for regression and classification issues, which delivers an expectation model as an outfit of feeble forecast models, regularly decision trees. The target of any supervised learning algorithm is to characterize a misfortune work and limit it.

4. Result and Discussion

This section shows the results and discussion of the proposed method.

A. Accuracy: It describes the closeness of a measurement to the true value.

\[
\text{accuracy} = \frac{(tn+tp)}{(tn+tp+fn+fp)}
\]

B. Precision: It describes the closeness of agreement among a set of results.

\[
\text{precision} = \frac{tp}{(tp+fp)}
\]

C. Sensitivity: Test’s ability to identify a condition correctly.

\[
\text{sensitivity} = \frac{tn}{(tn+fp)}
\]

D. Specificity: Test’s ability to exclude a condition correctly.

\[
\text{specificity} = \frac{tp}{(tp+fn)}
\]

E. Recall: Test’s ability to retrieve a relevant item.

\[
\text{Recall} = \frac{tp}{(tp+fn)}
\]

tp - true positive (sick people correctly diagnosed as sick)

tn - true negative (healthy people correctly identified as healthy)

fp - false positive (healthy people incorrectly identified as sick)

fn - false negative (sick people incorrectly identified as healthy)

5. Conclusion and Future Work

The aim of this project is to know whether the patient has heart disease or not. The records in the datasets are divided into training set and test sets. After preprocessing the data, data mining classification technique namely Random Forest and
GBT were applied. This section shows the results of those classification model done using Python Programming. The results are generated for both training datasets and test data sets. After these models compare to other models are follow Decision Tree, KNN and GBT algorithms.

References


