Analysis and Forecast of Heart Syndrome by Intelligent Retrieval Approach

Noor Basha, K. Manjunath, Mohan Kumar Naik, P. S. Ashok Kumar, P. Venkatesh, and M. Kempanna

Abstract At present scenarios in the world, heart disease analysis and prediction are two demanding factors to be faced by the doctors that are very ridiculous, and in this regard, health industries will generate enormous amount of data. To reduce huge range of deaths from diseases like heart disease, cancer, tumour and Alzheimer's disease, doctors must find the rapid and effectual analysis and detection techniques to be used, where various algorithms are used to learning the machines and create very important responsibilities in study and prediction of various diseases in humans. The key intension of this article is characterized on forecasting and analysis of various heart-related syndromes in patients with wide range of age by means of machine learning algorithms and techniques. In this case study, many parameters are considered to do analysis and predict heart disease of patients, where KNN, logistic regression and decision tree algorithm are used to calculate accuracy and performance.

Keywords Heart syndrome · K-nearest neighbour · Logistic regression · Decision tree

N. Basha
Department of CSE, VIT, Bengaluru, India

K. Manjunath
Department of CSE, Govt. Polytechnic, Chennasandra, Bengaluru, India

M. K. Naik
Department of ECE, NHCE, Bengaluru, India

P. S. A. Kumar (✉)
Department of CSE, DBIT, Bengaluru, India

P. Venkatesh
Department of TCE, DBIT, Bengaluru, India

M. Kempanna
Department of CSE, BIT, Bengaluru, India

S.-L. Peng et al. (eds.), *Intelligent Computing and Innovation on Data Science*,
Lecture Notes in Networks and Systems 118,
https://doi.org/10.1007/978-981-15-3284-9_57

1 Introduction

Due to busy schedule as well as routine assignments, peoples are facing severe stress and anxiety. Moreover, some other peoples are addicted with chronic habitual behaviour, like consumption of cigars and gutka, and those peoples are suffering from chronic diseases like heart diseases, cancer, liver problems, kidney failures, etc. To cure such patients with chronic diseases, is a very big hurdle to creative medical practitioners and medical researchers to solve the current world issue and objectives. Regarding this new challenge, IT professionals are provided hand-to-hand support to predict such disease early and cure as well as recover the patients from the chronic disease.

1.1 Heart Syndrome—Case Study

In this world, each person is unique in his attributes and his behaviour, out of which each person may have dissimilar readings of pulse rate and blood pressure. In general, a healthy human pulse rate must be in the range of 60–100 bpm and BP with a range 120/80–140/90 (mm Hg), and these benchmarks are proved.

Nowadays in throughout the world, for accidental or abrupt death, heart syndrome is one key basis, i.e., many peoples are affected by heart disease which is regardless of age in both men as well as women. This is because of improper dieting and consumption of alcoholic contents, cigars, etc., on irrespective of attributes like gender, diabetes, age, BMI, etc., which also added up this disease to humans rapidly. In this paper, we tried to do analysis and predict the heart disease in view of various factors like age, gender, blood pressure, heart rate, diabetes, etc., even though prediction of heart disease is one of the tricky jobs to researchers as well as doctors.

At present, various tools and techniques are available in the market to predicting the diseases, but still we expected some flaws in the analysis and predicting algorithms. Nowadays, based on big data approach, machine learning algorithm plays very important responsibility to explore and develop concealed knowledge and information about the chronic diseases.

2 About Literature

Coronary heart disease narrows down the coronary arteries. Basically, coronary arteries will supply both oxygen as well as blood to heart, and if the heart functioning is not proper, then it causes to malfunctioning of heart which leads to ill or death to a person.

Dewan et al. [1] discussed in detail the cardiovascular disease and different symptoms of heart attack. The different types of classification and clustering algorithms and tools were used.

Thomas and Theresa Princy [2] made use of many classification algorithms to predict the severe heart syndromes based on risk rate, where the author specifically used data mining approach.

Amin et al. [3] made use of an artificial neural network and genetic algorithm to predict health diseases. In this reference, author collaborates the data mining approach with association rules and classification techniques. In this regard, the model developed by the author is so efficient on predicting the heart syndrome. Shilaskar and Ghatol [4], evolution based feature selection is one of the effective method to select critical feature in a data set. Purusothaman and Krishnakumari [5], with critical factors and effective model an experienced medical practitioner can predict heart disease. Miao et al. [6] made use of various classification algorithms to create effective data analysis model on prediction of severe heart syndromes. Usually, dataset may contain noise features, and it abruptly corrupts the valid data, so they tried to reduce the noise by cleaning and pre-processing the dataset and also tried to reduce the dimensionality of the dataset. They found that good accuracy can be achieved with neural networks.

In some other literature, we referred an analysis using data mining. The analysis showed that using different techniques and taking different number of attributes gives different accuracies for predicting heart diseases. Even though the heart disease data set may contains many features and syntactically related duplicate information, in this regard we must refine the data set efficiently. This has to be pre-processed. Also, they say that feature selection has to be done on the dataset for achieving better results.

3 Methodology and Data Analysis

Generally in diabetic patients, high glucose content in the blood may cause damage in blood vessels as well as nerves in the body. If a person is suffering from diabetes on long period, then in future that person has higher probability to get heart disease, i.e., imagine if the person is diabetic, and addicted with alcoholic contents and chain smokers will naturally raise and develop the riskier heart disease.

Figure 1 symbolizes the graph of heart syndrome on people with age and count, if a person is directed by stressful life, he can easily damage his arteries and is accepting very big chance of coronary heart disease.

With the symptoms of high blood pressure, it formulates the person's heart to work very harder to pump the blood, it causes to strain the heart, and moreover, it relates to damage many blood vessels. Abnormal cholesterol levels in the body promote heart diseases and corpulence (obesity). Along with this, improper dieting as well as family history also causes chronic heart disease to the individual. In general, senior citizens and age-old peoples will easily hit by the heart diseases, due to many factors like age, gender, abnormal or unhealthy diet and stress, etc. In this regard, men are big victims or big risk of heart disease prone.

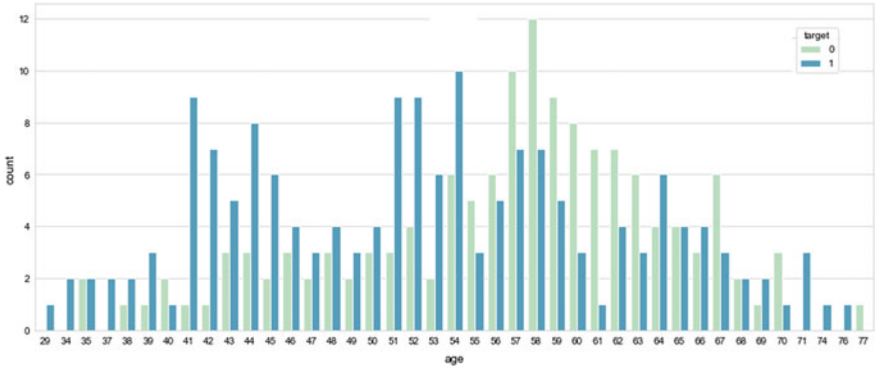


Fig. 1 Number of people who have heart disease based on age

At present, huge amount of research work is related towards heart diseases analysis and prediction system, where many researchers used various techniques and algorithms of machine learning and deep learning. The aim of ML and DL techniques is to achieve better accuracy and efficiency, so that doctors and patients can easily analyze and predict the heart attack chances.

3.1 Data Sources

The dataset used in this article is from Kaggle Web. Basically, Kaggle supports various dataset publication contents, where datasets are open source, very easily accessible data formats and supported to all platforms and work with any tools.

Table 1 specifies the features and represents the various conceptions used to create an effective system model to classify and validate the severity of heart syndrome in critical patients.

Table 1 Various characteristics used in system model

Sl. no.	Features and conceptions
1	Age (referred in months and years)
2	Sex (male = 1, female = 0)
3	CP (chest pain) in patients, with category, like (normal angina = 1, atypical angina = 2, non-angina = 3, asymptomatic = 4)
4	Chol (checked cholesterol content in serum)
5	FBS (fastening blood sugar)
6	Thalach (it referred with respect to maximum heart rate)
7	Exang (exercise-induced angina)

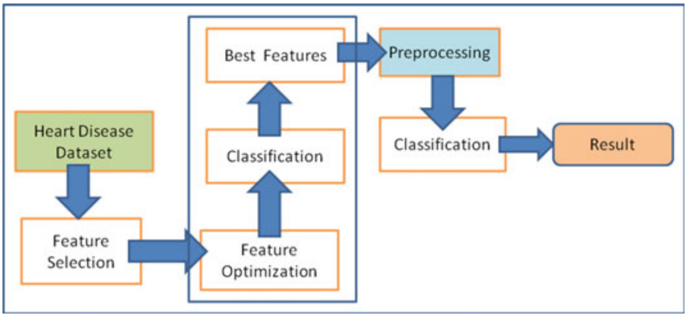


Fig. 2 Proposed architecture of the system

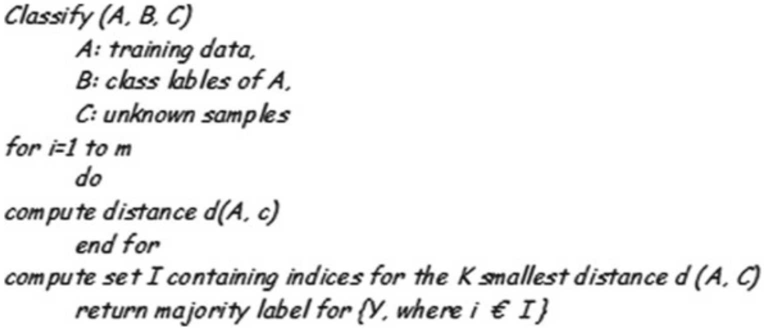
Some of the general symptoms of heart syndrome in severe and critical patients are like severe chest pain, shortness of breath, indigestion, burning sensation in chest, severe pain in stomach, sweating and fatigue, vomiting sensation, dizziness with anxiety and variations in heartbeat.

Figure 2 represents the proposed architecture of the system used to apply various machine learning algorithms to examine and forecast the syndrome of severe and critical coronary heart patients. In this model, heart disease data is considered to be an input data, and then, data is pre-processed by replacing non-available values with column means.

3.2 K-Nearest Neighbours (KNN)

In KNN algorithm, data is classified and regressed, where algorithm learns to evaluate the outcome from specific dataset. It performs well even if the training data is large and contains noisy values. In KNN algorithm, data is divided into two sets, i.e., training data and test sets, where experimental result set is mint for model building and training, where *k*-value is decided. Now, test data to be predicted on the model is built. There are different distance measures.

Pseudocode of KNN algorithm



```
Classify (A, B, C)
        A: training data,
        B: class lables of A,
        C: unknown samples
for i=1 to m
        do
compute distance d(A, c)
        end for
compute set I containing indices for the K smallest distance d (A, C)
        return majority label for {Y, where i € I}
```

3.3 Logistic Regression

The importance of logistic regression algorithm is used to classification tasks, where many classification tasks are done at routine pattern, i.e., logistic regression is used for multiclass classification.

For example, in e-mail classification task, to check whether those received e-mails are spam e-mails or not, or while doing online transaction, the user must be aware about whether the classified Website is fraudulent or not, etc.

Logistic regression is one of the statistical models utilized for binary classification, where it predicts the type (this *or that*, *yes or no*, *A or B*, etc.).

Logistic regression is a classification algorithm 1 that works by trying to learn a function that approximates $P(Y \mid \mathbf{X})$. It makes the central assumption that $P(Y \mid \mathbf{X})$ can be approximated as a sigmoid function applied to a linear combination of input features. Logistic regression is the building block for artificial neural networks.

Mathematically, for a single training data point (x, y) as,

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \sigma(z)\ \text{where}\ z = \theta_0 + \sum_{i=1}^{m} \theta_i x_i$$

Equivalent forms of above equation can be written as,

$$\begin{aligned} P(Y = 1|\mathbf{X} = \mathbf{x}) &= \sigma\left(\theta^{\mathrm{T}}\mathbf{x}\right) && \text{where we always set } x_0 \text{ to be 1} \\ P(Y = 0|\mathbf{X} = \mathbf{x}) &= 1 - \sigma\left(\theta^{\mathrm{T}}\mathbf{x}\right) && \text{by total law of probability} \end{aligned}$$

In the probability of data, probability of $Y \mid \mathbf{X}$ algorithm is used to create and select the maximized theta value. State log probability function and partial derivatives with respect to theta can be written as,

(a) An algorithm that can choose optimal values of theta.
(b) How the equations is derived.

Finally, logistic regression algorithm totally depends on its θ value.

3.4 Decision Tree

Decision tree is one of the intelligent retrieval techniques for regression and categorization of datasets, where algorithm perform very effectively on continuous and categorical attributes, i.e., decision tree algorithm divides the population into two or more similar sets based on the most significant predictors. In this algorithm, entropy will calculate each and every attribute, and then, it split the dataset with the help of other predictors with maximum information gain or minimum entropy.

Decision Tree Algorithm Pseudocode

```
Step 1: select the root node in tree
Step 2: find the best attribute in a set
Step 3: split the set into subsets
Step 4: generate the subset with unique value
Step 5: redo from step 1
Step 6: generate subset
Step 7: check leaf node in tree.
```

4 Result Analysis

The above-mentioned machine learning algorithms are used in this dataset implementations, where logistic regression algorithm has very high accuracy compared to other two algorithms, given in Table 2.

Below result set represents the various key performance indices of the patient's dataset like precision, recall, f1-score and support of all three individual algorithms KNN, decision tree and logistic regression algorithm to determine the accuracy score.

Accuracy score of KNN algorithm is: 72.527%

Accuracy	0.73	0.73		91
Macro_avg	0.73	0.73	0.73	91
Weighted_avg	0.74	0.74	0.73	91

Accuracy score of decision tree algorithm is: 73.27%

Accuracy	0.73			91
Macro_avg	0.73	0.73	0.73	91
Weighted_avg	0.74	0.73	0.73	91

Accuracy score of logistic regression is: 82.52%

Accuracy	0.82			91
Macro_avg	0.82	0.82	0.82	91
Weighted_avg	0.82	0.82	0.82	91

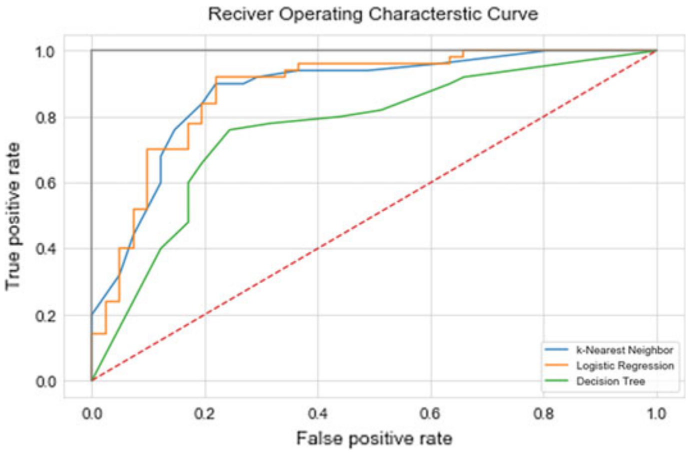


Fig. 3 ROC of KNN, logistic regression and decision tree

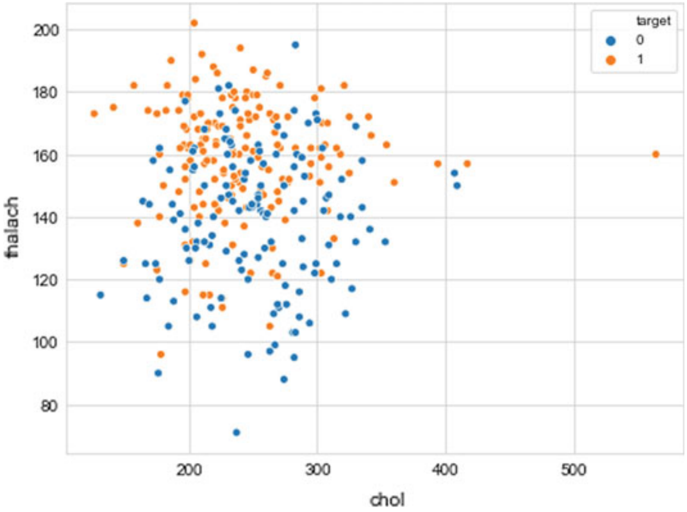


Fig. 4 Scatter plot for thalach versus chol

Table 2 Accuracy of algorithm

Approach	Accuracy
KNN	72.52
Decision tree	73.27
Logistic regression	82.52

The machine learning models are evaluated using the ROC metric. This can be used to understand the model performance, and it is shown in Fig. 3.

Figure 4 represents the scattered plot of heart syndrome with respect to thalach versus chol.

5 Conclusion and Future Work

Parental history or hereditary symptoms will lead to many chronic diseases to peoples, out of which heart disease is one among. If we identify the chronic diseases in early stage, it can be cured, so medical or hospital dataset is collected from Kaggle Web to analysis with different algorithm to check the accuracy score on key attribute with heart disuse patients. While implemented on this system model for heart disease patients by KNN, decision tree and logistic regression algorithm, with verification of key attributes, we found that logistic regression algorithm performs very effective and efficient performance on accuracy score for heart disease prediction. With inference of this customized model, machine learning algorithm provides very valuable knowledge on analysis and prediction of many chronic diseases. In this regard, researchers are helpful to the needy persons, doctors and society.

References

1. Dewan A, Sharma M (2015) Prediction of heart disease using a hybrid technique in data mining classification. 978-9-3805-441 6-8/15/$31.00 c 2015 IEEE
2. Thomas J, Theresa Princy R (2016) Human heart disease prediction system using data mining techniques. In: ICCPCT
3. Amin SU, Agarwal K, Beg R (2013) Genetic neural network based data mining in prediction of heart disease using risk factors. In: IEEE conference in ICT, pp 1227–1231
4. Shilaskar S, Ghatol A (2013) Feature selection for medical diagnosis: evaluation for cardiovascular diseases. Expert Syst Appl 40(10):4146–4153
5. Purusothaman G, Krishnakumari P (2015) A survey of data mining techniques on risk prediction: heart disease. Indian J Sci Technol 8(12):1
6. Miao KH, Miao JH, Miao GJ (2016) Diagnosing coronary heart disease using ensemble machine learning. IJACSA 7(10):30–39