IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DIAGNOSIS OF HEART DISEASES USING MACHINE LEARNING

Ayushi Bharti*1, Lokesh Malik2

1,U.G.Scholar, 2P.G.Scholar,1,Electronics and communication engineering,2Computer Science and Engineering, 1MSIT,2 DU, New Delhi,India

ABSTRACT

The diseases of the heart, scientifically termed as Cardiovascular Disease, describes a huge range of conditions from which our heart gets affected. Some examples are Heart Valve Disease, Heart Infection etc. Although many of these can be stopped pre-handedly if we choose a healthy lifestyle, still most of the people get diagnosed with it and we can see it's proof in the worldwide record that maximum death occurs due to cardiovascular diseases. Because of the involvement of various risks, we need the most accurate model which can predict the occurrence of heart disease beforehand, so that we can take precautionary steps to avoid the disease. Due to such a large amount of data in the healthcare department, many researchers have applied different types of techniques to analyze the data and help our medical department in successfully predicting cardiovascular disease. In this paper, we used a dataset from an existing database of Cleveland UCI repository of patients of the heart. This data contains a total of 76 attributes and 303 instances. Now, to do the testing, we reserved 14 attributes out of the 76 attributes. On the basis of the results acquired from the test set, we will find out the performance of each algorithm. We will be taking help of supervised learning, in which we will have the data tested using Decision Tree algorithm, Naïve Bayes algorithm, Random Forest algorithm and K-Nearest Neighbor algorithm. This paper shows that the model with maximum accuracy has been made when we used the K nearest neighbor algorithm.

INTRODUCTION

According to WHO (World Health Organization), Cardio-Vascular Diseases, also termed as CVDs, are the major global cause of death. In 2019, 32% of global deaths were from CVDs and it is expected that by the time we reach the year 2030, the global deaths from CVDs will amount to 23.6 million a year. These deaths mainly occur in countries with low/middle income. Though there are many factors which bring the risk of getting diagnosed with CVDs, some of them include: too much intake of alcohol, smoking, lack of physical activities, obesity, high cholesterol level and also genetic predisposition. Therefore, having an early diagnosis of CVDs is very important as it can save a life.

From all the fields of Artificial Intelligence, Machine Learning is emerging a lot and also evolving at a very fast rate as most researchers use data mining and machine learning techniques and algorithms in their research. The healthcare field contains huge data and these algorithms help us to analyze this huge data. Now using machine learning and data mining, predictive modeling will be done which will tell the most likely outcomes with the help of previous and current data. So data mining means extracting important data and information from a big amount of databases which will help in decision making.

Professionals of the medical domain do analysis of this data to make effective decisions. Classification algorithms are used to deliver clinical aid via analysis. This algorithm is tested to predict CVDs in patients. Now, various techniques of classification like Decision Tree, Naïve Bayes, Random forest and K-Nearest Neighbor and data mining techniques like Clustering, Regression and Association Rule are used to classify the disease's attributes for prediction. In this research, we took the UCI repository data. For prediction, a model is made using classification algorithms which are mentioned above. Also, various other algorithms are also told about in this paper which are used for CVDs prediction. It also tells future areas where development and research is needed.

CONTEXT

CVDs affect a lot of people around the world and are a major reason for death all over the world. To reduce the cost of diagnostic tests, the diagnosis should be aided with computerized techniques, it should be reliable and proficient too. Data mining is a technology which helps the system to classify the various attributes. This paper uses classification for prediction. This section tells about machine learning and methods used in it with a little of their description, pre-processing of data, measurement evaluation and details about the dataset that we used.

Machine Learning

It is one of all emerging subdomains of Artificial Intelligence. Its main purpose is designing systems and making them learn and predict on the basis of its experience and the experience is acquired when it is trained using a training dataset. Then there is testing data which is treated as the input data and the model made through training is then tested on this input data and prediction is made. Hidden patterns are also detected by machine learning in the dataset to make a model. The missing values of the dataset are filled. Then the model uses the testing data for prediction of CVDs and then its accuracy is measured.

Various techniques of machine learning are discussed below:

- **Supervised Learning**
  - A labeled dataset is used in this technique to train the data. In this, it has some input data and its outcome is already given (In our case, the outcome can be “no risk, mild risk, high risk and very high risk”). Now, this data is divided into 2 categories, one for training and one for testing. Training dataset will be used to make the model learn and train; and testing dataset will be used to measure its accuracy.
- **Unsupervised Learning**
  - In this technique, the dataset is not labeled. The main focus in this is to find the hidden patterns in data. Thus, developing the patterns is what the model is trained for. It will easily predict the hidden patterns for any testing dataset, but it will describe the patterns after exploring the data. No response is seen in the dataset in this technique. An example of unsupervised learning is the clustering method.
- **Reinforcement Learning**
  - In this, the results or outcomes in the training set are not available, thus the model learns by experiencing each time. It keeps on improving itself by trying every possible solution and then takes the best solution and gets experience through it.

Classification Techniques

Classification is done for predicting further cases based on previous information. Data mining techniques like Neural Network, Naïve Bayes and Decision Tree are applied by many researchers for having precision diagnosis of cardiovascular disease i.e., heart disease. The accuracy of models using different techniques depend upon how many attributes are present. This research paper will give the score of accuracy for better health results. In the pre-processing step of our data, we used WEKA tool which is in attribute-relation file i.e., ARF format. From a total of 76 attributes, we have reserved 14 attributes only for testing and analysis purposes. By analyzing and comparing using the approach of different algorithms along with the WEKA tool, prediction of CVDs can be done and people can be treated early if required.

APPROACH

The objective of this paper is to tell before-hand the odds of getting cardio-vascular disease, which will be helpful for patients and medical professionals. To fulfill this objective, we have already discussed above many machine learning techniques to be applied on dataset and also the analysis of dataset is mentioned in this paper along-with the results of other authors who tried this. This paper also tells which attributes play a major role in anticipating higher precision. Due to this, a patient will not have to be tested on all attributes, but only for the major attributes and this may result in less expense, as not all the attributes contribute a big amount to the outcome.

Data Source

We have used a dataset from an existing database of UCI Machine Learning repository (Cleveland database). Contains around 300 instances of data and 14 attributes which include 13 predictors and 1 class. The attributes are like ECG results, blood pressure, type of chest pain etc. You can see all the attributes in the following table:

Table 1: Attribute table

S.NO.	**Attribute**	**Symbol used**	**Remarks**
1	Age	Age	Age of patients in years
2	Sex	Sex	Gender of patients; Female(0) and Male(1)
3	Resting ECG	RestECG	Normal(0), Having ST-T wave abnormality(1), Left Ventricular Hypertrophy(2)
4	Resting BP	RestBP	The blood pressure of a patient at the time of admission in mmHG.
5	Chest pain	CP	Typical Angina(1), Atypical Angina(2), Non-Anginal Pain(3), Asymptomatic(4)
6	ST depression	Depr	ST depression induced by exercise in relation with rest
7	Slope	Slope	Slope of peak exercise; Up-sloping(1), Flat(2), Down Sloping(3)
8	Max Heart Rate	MHR	Maximum Heart Rate achieved
9	Fasting Blood Sugar	FBS	Fasting blood sugar > 120 mg/dl; False(0) and True(1)
10	No. of Vessels	Ves	Number of Major Vessels (0-3) colored by fluoroscopy
11	Thalassemia	Thal	Types of Defect; Normal(3), Fixed Defect(6), Reversible Defect(7)
12	Serum Cholesterol	SeChol	Serum Cholesterol in mg/dl
13	Exercise-Induced Angina	EIA	Yes(1), No(0)
14	**Num(Class Attribute)**	Class	Status of diagnosis of CVD; No risk(0), Mild Risk(1), High Risk(2), Very High Risk(3)

Total 4 algorithms are used in this research to get reasons for getting CVDs and then a model is created with maximum possible accuracy.

Data Pre-Processing

The real-life information contains a lot of noisy data and missing values data. So, to make good predictions, we first pre-process the data to overcome the issues which might come from missing values and noisy data. We also have made a flow-graph of our proposed model, which can be seen below:

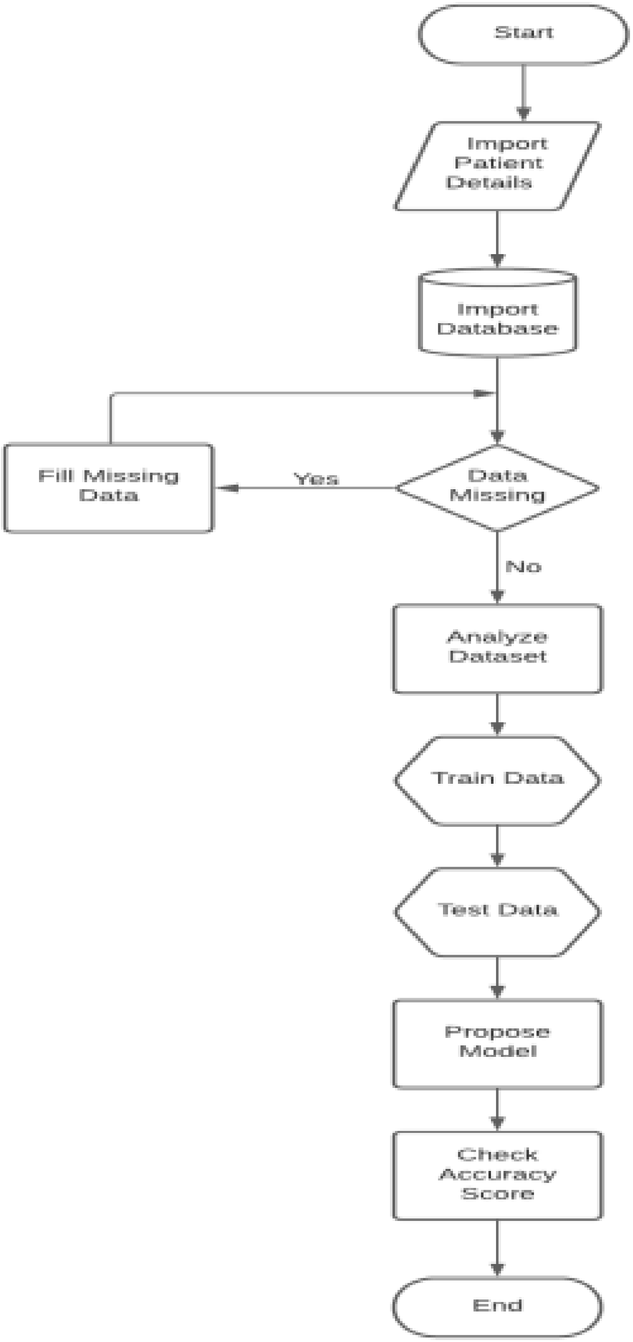

Figure 1: Model Flowchart

Step 1 Clean: To get accurate results, the data must be cleaned and missing values must be filled.

Step 2 Transform: In this, we use smoothing, aggregation and normalization tasks to make data more comprehensible by changing the format of the data.

Step 3 Integration: Before processing, we need to integrate the data as it might be acquired from various sources, not necessarily from a single one.

Step 4 Reduction: The acquired data is very complex and needs to be formatted for achieving desired results.

The data is then classified and divided into test and training data which are then tried on different algorithms to get accuracy score results.

Algorithms Used

Decision Tree

It is a classification algorithm which works on numerical and categorical data. Tree-like structures are made using this algorithm. It is very simple and often used in healthcare datasets.

There are 3 nodes on which the model makes analysis. They are:

- Root Node: This is the main node and all other nodes function based on this node.
- Interior Node: It handles various attributes.
- Leaf Node: It represents each test's results.

It splits the data into analogous sets (two or more) on the basis of the most important indicator. We calculate each attribute's entropy (E) and then we divide the data on the basis of minimum entropy or maximum information gain.

$$\textbf{Entropy(S)} = \Sigma_{□=1}^{□} -□_i \mathbf{log_2 P_i}$$

$$\textbf{Information Gain(S,A) = Entropy(S) -} \sum_{□\ □\ □□□□□□(□)} \frac{|□□|}{□} \textbf{Entropy(Sv)}$$

Reading and interpreting the obtained results is easier.

Because the Decision Tree Algorithm analyzes the data in tree-like structure, that is why, in comparison with other algorithms, the decision tree algorithm has high accuracy. But the data might be over classified and at a time only one attribute gets tested for decision making.

With this algorithm, the accuracy achieved is 71.43%.

Naïve Bayes Classifier

It's a supervised learning algorithm. It uses Bayes theorem whose concept is probability. The predictors do not relate to each other and also do not have correlation. Each attribute contributes independently to maximize the probability. Bayesian methods are not used by this classifier but it works with the Naïve Bayes Model. Many real-world complex situations use this classifier:

$$□(□/□) = \frac{□(□/□) * □(□)}{\mathbf{P(Y)}}$$

Here, P(X/Y) denotes posterior probability.

P(X) denotes class prior probability.

P(Y) denotes predictor prior probability.

P(Y/X) denotes probability of a predictor which is called likelihood probability.

Like the Decision tree algorithm, it is also simple and easy to implement and it can handle non-linear and complicated data very well. But, there is a little accuracy loss because it's based on assumptions and conditional class independence.

The accuracy using this algorithm was seen to be 84.1584% and only 10 attributes were kept which were the most important ones. And when we used all 13 attributes, our accuracy decreased to 83.49%.

Random Forest Algorithm

Like Naïve Bayes, it is also a Supervised Learning Algorithm technique. This algorithm creates a forest from several trees. Each tree in the forest will give its output class and the class with highest votes will be the final output.

More the number of trees, the more will be its accuracy. Common approaches include the following:

- Forest RI (Random Input Choice)
- Forest RC (Random Blend)
- Combination of Forest RC and Forest RI.

Random Forest Algorithm can be used for both regression and classification but does very well in classification tasks and takes care of missing values. But results are not accountable for, as it is slow in obtaining prediction because of the large dataset and more trees requirement.

With the Cleveland dataset, Random Forest Algorithm obtained accuracy of 91.6%.

K-Nearest Neighbor (K-NN)

It is also one of supervised learning algorithms. It classifies the objects which are dependent on the nearest neighbor. It is an example of instance-based learning. Euclidean Distance formula is used to measure the distance between an attribute and its neighbors. The group containing named points is used to mark other points. Based on similarity among groups, the data is clustered. It is very much

possible with KNN algorithm to fill the missing values in the dataset. After the missing values get filled, many prediction techniques are applied to the dataset. By combining various algorithms, it's possible to obtain better accuracy. K-NN is simple to execute without the need of creating a model or having to make any other assumptions. This algorithm is very flexible and used for regression, search and classification. Although it is the simplest of all, but still, the accuracy gets affected by irrelevant and noisy features. In the study of Pouriyeh et alHe took the value K as 9 and achieved an accuracy of 83.16%.

RESULTS & ANALYSIS

The objective of this research paper was to predict whether a person is prone to getting a heart disease or not. On the UCI repository data of Cleveland, we used various supervised learning techniques using Decision Tree, Naïve Bayes, Random Forest and K-Nearest Neighbor. The WEKA tool was used with different classifying algorithms in various experiments. For this experiment, we recommend that one must use a system with at-least 16GB RAM and generation of Intel Processor shall be at-least 9th Generation or above. Dataset was split into 2 sets, one set for training and one for testing purposes. After doing pre-processing, supervised classification techniques were applied like Decision Tree, Naïve Bayes, Random Forest and K-Nearest Neighbor for obtaining the accuracy score. We used python programming for noting the results of applied techniques on test and training dataset.

The respective accuracy score percentage is shown below for various algorithms through a table and bar graph.

	Decision Tree	**Naïve Bayes**	**Random Forest**	**K-Nearest Neighbor**
Accuracy				
Training Data	80.2631 (gini)	81.0572	82.8947 (gini)	90.7894 (K=7)
Testing Data	73.6842 (entropy)	88.1578	84.2105 (entropy)	78.9473 (K=2)

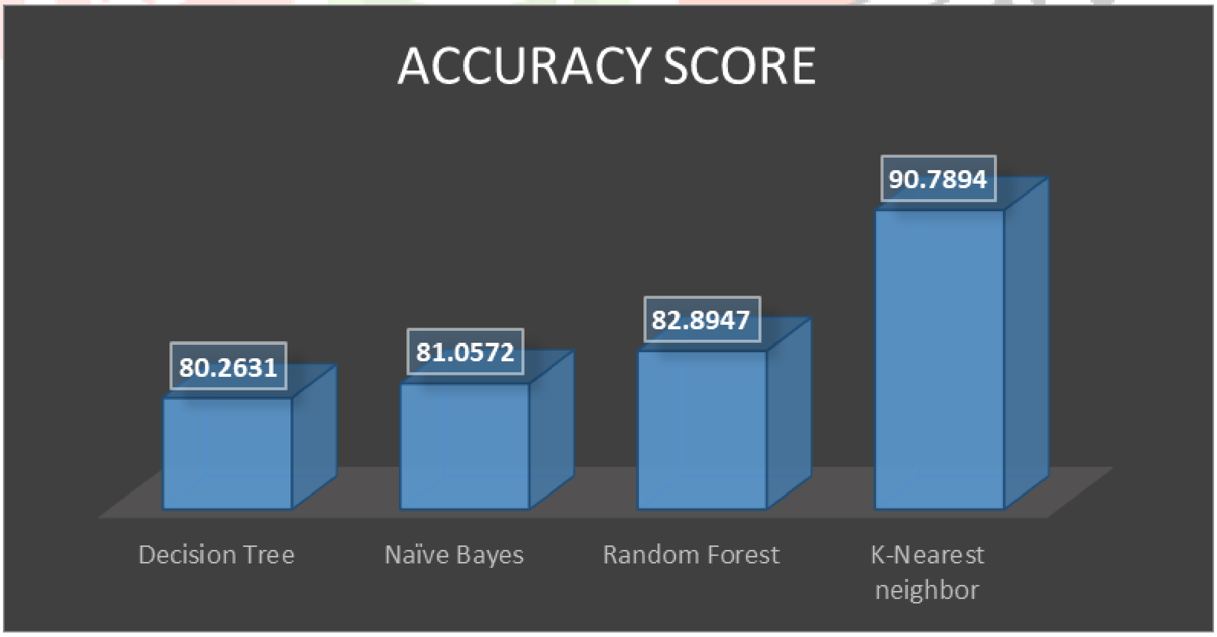

Figure 2: Accuracy Score Comparison of Algorithms Used

As we can see that while training, the maximum accuracy is achieved by K-NN algorithm i.e., 90%.

Also we have shown the comparison of the accuracy score obtained by other researchers and authors who also proposed a model for prediction of heart disease (given in table below).

Table 2: Accuracy obtained by different researchers using different algorithms

Author	**Technique**	**Accuracy Score**
Kumar Dwivedi	Naïve Bayes	83%
	Classification Tree	77%
	K-NN	80%
	Logistic Regression	85%
	SVM	82%
	ANN	84%
Seema et al.	Naïve Bayes	93.85%
	Decision Tree	92.59%
	SVM	95.2%
	Artificial Neural Networks	94.27%
Chaurasia et al.	J48	84.35%
	Bagging	85.03%
	SVM	94.60%
Parthiban et al.	Naïve Bayes	74%
Vembandasamy et al.	Naïve Bayes	86.419%
Otoom et al.	Naïve Bayes	84.5%
	SVM	84.5%
	Functional Trees	84.5%
Model proposed in this paper	Decision Tree	80.26%
	Naïve Bayes	81.05%
	Random Forest	82.89%
	K-Nearest Neighbor	90.78%

Highest accuracy is 95.2%, which was achieved using the SVM algorithm by Seema et al.

CONCLUSION

The overall objective was to describe different techniques of data mining which can prove to be useful in predicting effective heart disease. Our goal was to get as high accuracy with a lower number of attributes as possible.

We took only 14 essential attributes and used 4 classification techniques namely Decision Tree, Naïve Bayes, Random Forest and K-NN. The data was used in the model after pre-processing it. The best score was obtained by Naïve Bayes, Random Forest and K-Nearest Neighbor algorithm and highest accuracy was by K-NN algorithm with K=7. In the future, we can apply more techniques to get even greater accuracy scores for early prediction of CVDs.

REFERENCES

[1]. Chaurasia V, pal S. Data Mining Approach to detect heart diseases. Int J Adv Computer Science Information Technology, 2014

[2]. Deepika K, Seema S. Predictive analytics to prevent and control chronic diseases. In: 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT). IEEE. p. 381–86.

[3]. Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.

[4]. Otoom AF, Abdallah EE, Kilani Y, Kefaye A, Ashour M. Effective diagnosis and monitoring of heart disease. Int J Softw Eng Appl. 2015;9(1):143–56.

[5]. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9:1–16. https://doi.org/10.4236/jilsa.2017.91001.

[6]. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low-and middle-income countries. Curr Probl Cardiol. 2010;35(2):72–115.

[7]. Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.

[8]. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.

[9]. Vembandasamy K, Sasipriya R, Deepa E. Heart diseases detection using Naive Bayes algorithm. Int J Innov Sci Eng Technol. 2015;2(9):441–4.

[10] Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29(10):685–693.

[11].Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. Int J Appl Inf Syst (IJAIS). 2012;3(7):25–30.

[12]. Xu S, Zhang Z, Wang D, Hu J, Duan X, Zhu T. Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. In: 2017 IEEE 2nd international conference on big data analysis (ICBDA). IEEE. p. 228–32.

[13]. Pahwa K, Kumar R. Prediction of heart disease using hybrid technique for selecting features. In: 2017 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics (UPCON). IEEE. p. 500–504.

[14]. Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p. 482–86.

[15]. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p. 204– 207.

[16]. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[17]. Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Dis. 2015;7(1):129–37.

[18]. https://link.springer.com/article/10.1007/s42979-020-00365-y

[19]. https://www.heart.org/idc/groups/ahamah
public/@wcm/@sop/@smd/documents/downloadable/ucm_470704.pdf