A Method of Cardiovascular Disease Prediction using Machine Learning

A.Geetha Devi
ECE
Prasad V.Potluri Siddhartha Institute of Technology,
Vijayawada, A.P., India

Surya Prasada Rao Borra
ECE
Prasad V.Potluri Siddhartha Institute of Technology,
Vijayawada, A.P., India

K. Vidya Sagar
EIE
VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India,

***Abstract*— In the recent times, the major reason for increasing death rate is cardiac diseases. It is illogical for a typical man to experience exorbitant tests like the ECG as often as possible. Hence, there is an immediate need to forestall the death rate by setting up a framework for identifying the cardiovascular diseases in the initial stage, as there is a wild increase in the rate of cardiac arrests at adolescent age. For this purpose there are various classification algorithms of machine learning which can forestall the weakness of heart from the given essential indications corresponding to age, sex, cholesterol, glucose levels, heartbeat rate and so on. In this paper a popular classification methodology K Nearest Neighbor (KNN) algorithm has been utilized for detecting the heart diseases at the early stage. The UCI dataset has been utilized for classification which contains the medical records of 303 patients. An accuracy of 87% has been obtained by the KNN algorithm.**

***Keywords*— *Cardiovascular disease (CVD), classification algorithms, K Nearest Neighbour (KNN) algorithm, Machine learning*.**

I. INTRODUCTION

Heart pumps blood to various organs of human body and is the vital organ of circulatory system. Heart disease is the range of disorders or conditions that affect the heart functioning. Any heart problem may lead to critical health issues and even premature death. Heart diseases takes the first position in the cause of deaths. As per World Health Organization (WHO), each year there are around 17.9 million deaths are due to heart diseases i.e., 31% of deaths are due to cardio vascular diseases (CVD’s) and in this 85% of are due to heart stroke. Hence, it is very necessary to predict the heart diseases at the earliest possible. There is no shortage of records with respect to healing indications of patients persistent heart attacks. Anyway the latent they need to assist us with prognosticating comparable potential outcomes in deceptively solid grown-ups are going unobserved. For example: based on the records of the Indian Heart Association, 50% of heart strokes are under the age of 50 years and 25% are under the age of 40 years in Indians. The populace from urban areas is thrice as helpless against coronary episodes as country population.

Heart diseases are mainly caused due to diabetes, high blood pressure, increased stress, cholesterol, obesity, age and family history. The syndromes can be found out using medical science and data mining concepts. There are several ways in prevision of heart diseases in which data mining plays a significant role. The extremity of the cardiovascular disease can be estimated by the use of classification methods such as K-Nearest Neighbour Algorithm (KNN), Random forest algorithm, Support Vector Machine (SVM) algorithm and decision tree algorithm. In our proposed algorithm some attributes for heart disease prediction are utilised. The results obtained demonstrate a better level of accuracy compared to the other methods available in literature. Machine Learning Algorithms are utilized to obtain the results, which has high performance in the estimation of heart disease. This model, has an accuracy up to 87%.

Machine learning is an emerging technology which provides the systems the ability to think and learn from the experience. The main goal of machine learning is to deploy computer programs to access the data and use it for making prediction of the newer data.

Supervised learning and unsupervised learning are the two types of Machine learning(ML). The training data contains the outputs in supervised learning algorithms, On the other hand the unsupervised learning does not include the outputs. The applications of Machine learning algorithms are found in both classification and regression. In classification, the function that is being learned will be discrete that is the output will be either 1 or 0 whereas in regression we get continuous outputs.

In this paper KNN Algorithm has been utilized to predict the heart diseases. The main objective is to increase the efficiency in the prediction rate of cardiovascular disease. This experimental results have high capacity compared to other models.

II. RELATED WORK

In the year 2000, ShusakuTsumoto [1]conducted a research and found out that humans cannot arrange huge data in an order or in patterns. Hence, data mining concepts can be utilized for finding various patterns from the available huge database and perform various operations on it.

The heart disease prediction can be carried out using various algorithms such as Support Vector Machine (SVM) classifier, decision tree and random forest algorithm[2-3]. But these are supervised ML algorithms.

Each one of them has their own advantages and disadvantages.

Decision tree [2] algorithm is a flowchart like structure in which each internal node corresponds to a test on an attribute in the dataset and each branch corresponds to the output of the test. This classifier got an accuracy of 79% for the heart disease prediction.

Random forest algorithm [3] merges many decision trees for finding the final output rather than depending on the single decision tree. It has lesser variance than a single decision tree. This algorithm got an accuracy of 79% for the heart disease prediction.

In SVM [1] classifier each member of the data is represented as a point in the space defined by a separating hyper plane. This classifier got an accuracy of 83% for the heart disease prediction.

III. METHODOLOGY TO PREDICT HEART DISEASE

In this paper, we have used the Cleveland UCI dataset and Google Co labs for the execution of the K Nearest Neighbor (KNN) algorithm. The following block diagram gives the organization of the data.

The Fig.1 illustrates different stages in machine learning. They are data pre-processing, Feature selection and reduction, splitting.

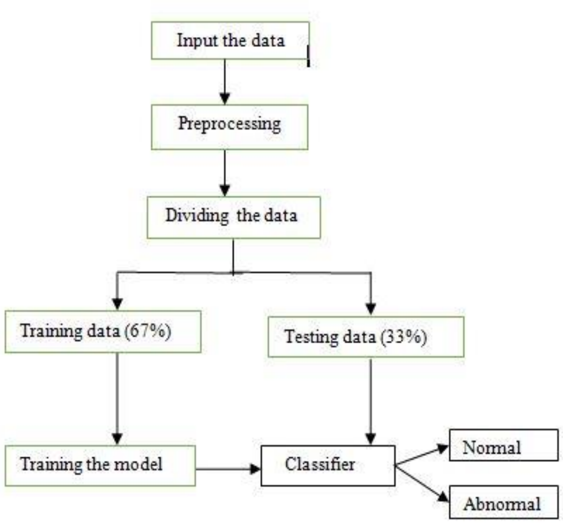

Fig.1. Flow Chart of Machine Learning Algorithm

A. Data Preprocessing

After assortment of various records, heart Disease information is pre-processed. There are a sum of 303 patient records were loaded in the dataset, but 6 records were found with some missing qualities hence they expelled from the dataset. The preprocessing has been performed on the remaining 297 patient records. The characteristics of the given dataset presented are the multiclass variable and double classification. The nearness or nonattendance of heart Disease is checked by multi-class variable.

B. Feature Selection and Reduction

In order to recognize the individual data of the patient, two parameters relating to age and sex are utilized among the 13 qualities of the informational collection, The remaining 11 properties are considered important as they contain significant clinical records. These Clinical Records are used for the diagnosis and for learning the severity of heart disease .Several Machine Learning Techniques are used like Decision Tree ,Support Vector, Random Forest. All these Machine Learning techniques use all 13 attributes. In this experiment we use K-Nearest Neighbour Algorithm. Fig. 2. describes the bar chart of prediction of heart disease using K-Nearest Neighbour Algorithm.

C. *Splitting*

The entire dataset is divides into training data and testing data. The percentage of training data is 67% and the remaining is used for testing i., 33% of data.

The outcome of the algorithm will be either 0 (indicating the absence of heart disease) or 1(indicating the presence of heart disease). The prediction for the new instance (datum) in KNN algorithm, is made by searching through the entire training dataset for the k (a numerical) most similar instances. This algorithm classifies new data point based on similarity measures (distance function) on data. In KNN algorithm, new data point will be allotted a value based on how closely it matches the points in training dataset i.e., it uses feature similarity.

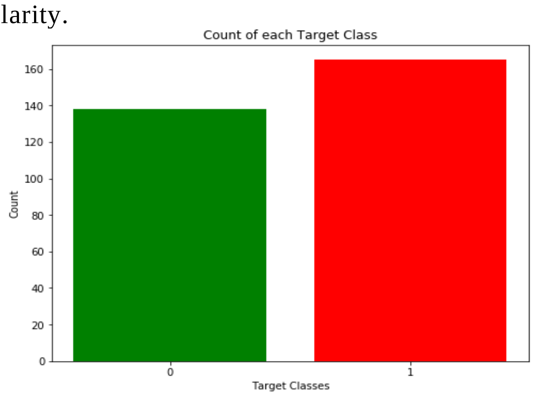

Fig.2: Bar chart of prediction of Heart disease of the target class by KNN algorithm

IV. RESULTS AND DISCUSSION

The heart disease prediction carried out by KNN algorithm as follows

A. *Loading the data*

In this paper the UCI dataset, that contains medical records of 303 patients has been utilized. In this, 67% of data is used as training data and 33% of data for testing the model. The dataset should be checked whether it is balanced or not before training the model. The dataset which we have utilized is a balanced dataset.

B. *Initialization of K to the chosen number of neighbors*

The value of K means the number of neighbors that has been choosen. The accuracy of the model changes if the value of K is varied. The value of K is varied from 1 to 21 to get the required accuracy measure for different number of neighbors.

TABLE I. PARAMETERS TO BE CONSIDERED FOR CLASSIFICATION

Parameter	Description	Type
Age	*Patient Age*	*Numeric*
Sex	*Patient Gender*	*Nominal*
Cp	*Chest Pain is of 4 values* *1.Typical Agina* *2.Atypical Agina* *3.Non-anginal pain* *4.Asymptomatic*	*Nominal*
Test bps	*Value of Blood Pressure at resting mode*	*Numeric*
Chol	*Serum Cholesterol*	*Numeric*
FBS	*Blood Sugar Levels on Fasting*	*Nominal*
Resting	*Results of ectrocardiogram while at rest*	*Nominal*
Thali	*The accomplishment of maximum rate of heart*	*Numeric*
Exang	*Agina induced by exercise*	*Nominal*
Oldpeak	*Exercise induced ST depression in comparison with state of rest*	*Numeric*
Slope	*ST segment measurement*	*Nominal*
Ca	*Fluoroscopy coloured major vessels*	*Numeric*
Thal	*Status of heart through 3 values*	*Nominal*

ALGORITHM

C. For every data point the following steps have been performed

The distance between the query example and the current example from the data has been calculated: The KNN algorithm assumes that indistinguishable things exist in close proximity. Hence, the distance between the current point and all the other points in the dataset has to be calculated. Any distance measure like Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance can be used. In this paper, Euclidean distance is utilized and calculated using the following formula described in eq.(1)

$$\textit{Euclidean Distance} = \sqrt{\sum \left(x_i - x_j\right)^2} \quad \dots \quad (1)$$

The distance and the index of the example to an ordered collection are stored: The result of the Euclidean distance is stored into a table.

D. Arrange the ordered collection of distances and indices in ascending order by the distances

After calculating the distances and storing them in the table, now we should arrange them in the ascending order i.e., smallest values to largest values. So the point which has smallest distance will be on the top of the table and the point which has largest value will at the bottom of the table.

E. Take the first K entries from the sorted collection

Pick the first K entries from the table which is sorted. For example if K=5, then first values present in the table should be considered.

F. Get the labels of the selected K entries

The results of the first K entries are retrieved for predicting the outcome of the present input.

G. Return the mode of the K labels

Finally calculate the average of the selected K entries. This is the final outcome of the model for the given input values.

TABLE II. RANGE OF VARIOUS PARAMETERS CONSIDERED FOR CLASSIFICATION

Age	*Numeric [29 to 71; unique=41; mean=54.4; median=56]*
Sex	*Nominal [0 to 1; unique=2; mean=0.68; median=1]*
CP	*Numeric [1 to 4; unique=4; mean=3.16; median=3]*
TESTBPS	*Numeric [94 to 200; unique=50; mean=131.69; median=130]*
CHOL	*Numeric [126 to 554; unique=152; mean=246.69; median=241]*
FBS	*Nominal [0 to 1; unique=2; mean=0.15; median=0]*
RESTECG	*Numeric [0 to 2; unique=3; mean=0.99; median=1]*
THALACH	*Numeric [71 to 202; unique=91; mean=149.61; median=153]*
EXANG	*Numeric [0 to 1; unique=2; mean=0.33; median=0.00]*
OLPEAK	*Numeric [0to6.20; unique=40; mean=1.04; median=0.80]*
SLOPE	*Numeric [1 to 3; unique=3; mean=1.60; median=2]*
CA	*Categorical[5 levels]*
THAL	*Categorical[4 levels]*
Target	*Numeric[0 to 4; unique=5; mean=0.94; median=0.00]*

$$P\ (class=0) = count\ (class=0)\ /(count\ (class=0) + count\ (class=1))$$

There are four parameters in a confusion matrix. They are true positive, false negative, true negative and false positive.

Positive: Original value is positive.

Negative: Original value is not positive.

True Positive (T_P): Original value is positive and the classifier also predicted as positive.

False Negative (F_N): Original value is positive but the classifier predicted as negative.

True Negative (T_N): Original value is negative positive and also the classifier predicted as negative.

False Positive (F_P): Original value is negative but the classifier predicted as positive.

Accuracy= *(T_P+T_N)/(T_P+T_N+F_P+F_N)*

Recall: Recall is the ratio of total number of correctly categorized positive examples to the total number of positive examples

Recall= T_P/(T_P+F_P)

Precision: Precision is the ratio of total number of correctly categorized positive examples to the total number of predicted positive examples

Precision= *T_P/(T_P+F_P)*

F-measure: F-measure calculates the harmonic mean using the precision and recall parameters.

F-measure=(2 x recall x precision) / (recall + precision)

By using the KNN algorithm we got an highest accuracy of 87% for 8 neighbors and the parameters of the confusion matrix for 8 neighbours are T_P=42,F_N=6,F_P=7and F_N=45. The accuracy is 87% , precision is 0.8571,recall is 0.4827 and F-measure is 0.617. From the above figure we can say that the K-Nearest Classifier has highest accuracy of 87% for eight neighbours.

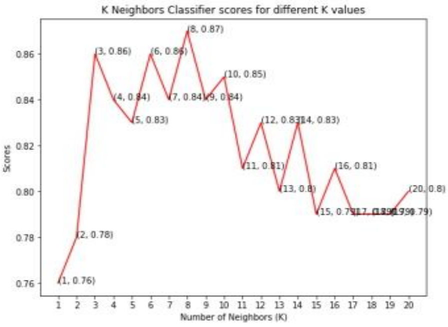

Fig.3. Output of the KNN Classis

V. CONCLUSIONS

Prediction of Heart disease is a challenging and very necessary in the medical field. The recognition of heart diseases through the processing of raw health care information will help in the long term saving of human lives. The mortality rate can be controlled if the disorder is detected at early stages and preventative measures are adopted as soon as possible It is helpful in the early detection of abnormalities in heart. In this paper, KNN algorithm is utilised to perate information and furnish a method towards heart disease by which an accuracy of 87% for K=8 neighbors has been achieved. However, a. further extension of the work is highly desirable to direct the investigations towards real world data instead of theoretical methods and simulations.

REFERENCES

[1] C.Sowmiya and Dr.P.Sumitra, "Analytical Study of Heart Disease Diagnosis Using Classification Techniques" https://doi.org/10.1109/ITCOSP.2017.8303115.

[2] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", J. King Saud Univ.-Comput. Inf.Sci., vol. 24, no. 1, pp.27–40, Jan.2012.Doi:10.1016/j.jksuci.2011.09.002.

[3] A. S. Abdullah and R. R. Rajalaxmi, "A data mining model for predicting the coronary heart disease using random forest classifier", in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25

[4] H. A. Esfahani and M. Ghazanfari,"Cardiovascular disease detection using a new ensemble classifier", in Proc. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.

[5] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009. doi: 10.1016/j.eswa.2008.09.013.

[6] M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining", in Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl.Manage. (ABLAZE), Feb. 2015, pp. 520–525.

[7] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning", in Proc. 2nd Int. Conf. Electron.,Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278.

[8] B. S. S. Rathnayakc and G. U. Ganegoda, "Heart diseases prediction with data mining and neural network techniques", in Proc. 3rd Int. Conf.Converg. Technol. (I2CT), Apr. 2018, pp. 1–6.

[9] Hai Wang et.al., "Medical Knowledge Acquisition through Data Mining", Proceedings of 2008 IEEEInternational Symposium on IT in Medicine and Education 978-1-4244- 2511-2/08©2008 Crown.

[10] VikasChaurasia, Saurabh Pal, "Early Prediction of Heart disease using Data mining Techniques", Caribbean journal of Science and Technology,2013

[11] K.Sudhakar, Dr. M. Manimekalai, "Study of Heart Disease Prediction using Data Mining", IJARCSSE 2016. Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Technique