Heart Disease Prediction System Using Neighboring Distance Based Outlier Detection Approach

D.Priyanth
PG Student, Department of CSE, Jawaharlal Nehru Technological University College of Engineering (Autonomous) Anantapuramu

***Abstract*—The clinical or hospital information is stored in large volume in medical database need intelligent based discovery. Medical data is huge and require more possibilities analysis. Data mining the procedure to analyze a huge volume of available medical data from various futuristic potential and provides knowledge information to the physician to predict the patient disease accurately. Existing system used a heart disease prediction model (HDPM) approach for Clinical Decision Support System (CDSS). This was a web-based application. It will collect the data of patients and the information sent to Heart Disease Clinical Decision Support System. This was designed for medical practitioners. The disadvantage of system is it works on limited data set and accuracy is also low. The propose uses neighboring distance-based outlier detection approach. It can give better results in terms of accuracy. This can help patients in getting a quick diagnosis with a lot less cost.**

***Index Terms*—Heart disease, outlier data, XGBoost, Accuracy, machine learning.**

I. INTRODUCTION

Present time, data has been scattered as Statistics, Reports and Forms and so forth. It is a vast benefit which allows the making of outcome in genuine time conditions. In spite of that, a group of study has been conducted in various areas, health care has a wide extension to utilize officially accessible information and determine results which will be available to the world. Cardiovascular illnesses consist of Heart and blood vessel sicknesses that comprise of many problems, lot of which are linked to an operation termed atherosclerosis. When a material termed plaque accumulates in the walls of the arteries and evolves that case is termed Atherosclerosis. This accumulates and tightens the arteries making them harder for blood to flow out of the arteries. The term Myocardial Infarction or stroke is when the blood becomes clot which can also cause a heart attack. Respiratory failure infections are the significant explanation of death at a normal level around the world. In 2015, 17.7 million passings which are caused from cardiovascular infection are assessed to be roughly 31% around the world, according to the World Health Organization. According to this report, 82% of them are in low and middle-income countries, 17 million are under 70 years of age which are prone to non-communicable diseases, 6.7 million due to stroke and 7.4 million were due to coronary heart disease When a heart attack happens, we have to quicken medical attention to prevent heart damage and to maintain the life of a patient with a heart attack. These days, the utilization of computer technology for medicine is very high. In order to realize our goals in this complex phase, active hybrid fuzzy expert systems that the doctor may need and that can prophecy the probability of a patient getting a heart illness problem and being able to assist in embodying the illness. The purpose of classification is to look for a pattern to predict the category of objects whose classification is unknown and depicts them in distinguishing data categories or concepts.

II. RELATED WORK

According to Ordonez [1] the heart disease can be predicted with some basic attributes taken from the patient and in their work have introduced a system that includes the characteristics of individual human being based on totally 13 basic attributes like sex, blood pressure, cholesterol and others to predict the likelihood of a patient getting affected by heart disease. They have added two more attributes i.e., fat and smoking behavior and extended the research dataset. The data mining classification algorithms such as Decision Tree, Naive Bayes, and Neural Network are utilized to make predictions and the results are analyzed on heart disease database. Yilmaz [2] have proposed a method that uses least

squares support vector machine (LS-SVM) utilizing a binary decision tree for classification of cardiotocogram to find out the patient condition. Duff, et al. [3] have done a research work involving five hundred and thirty-three patients who had suffered from cardiac arrest and they were integrated in the analysis of heart disease probabilities. They performed classical statistical analysis and data mining analysis using mostly Bayesian networks. Frawley, et al. [4] have performed a work on prediction of survival of coronary heart disease (CHD) which is a challenging research problem for medical society. They also used 10-fold cross-validation methods to determine the impartial estimate of the three prediction models for performance comparison purposes. Lee et al. [5] proposed a novel methodology to expand and study the multi-parametric feature along with linear and nonlinear features of Heart Rate variability diagnosing cardiovascular disease. They have carried out various experiments on linear and non-linear features to estimate several classifiers, e.g., Bayesian classifiers, CMAR, C4.5 and SVM. Based on their experiments, SVM outperformed the other classifiers. Noh et al. [6] suggested a classification method which is an associative classifier that is constructed based on the efficient FP-growth method. Because the volume of patterns can be diverse and huge, they offered a rule to measure the cohesion and in turn allow a tough choice of pruning patterns in the pattern-generating process. Parthiban, et al. [7] have proposed a new work in which the heart disease is identified and predicted using the proposed Coactive Neuro-Fuzzy Inference System (CANFIS). Their model works based on the collective nature of neural network adaptive capabilities and based on the genetic algorithm along with fuzzy logic in order to diagnose the occurrence of the disease. The performance of the proposed CANFIS model was evaluated in terms of training performances and classification accuracies. Finally, their results show that the proposed CANFIS model has great prospective in predicting the heart disease. Singh, et al. [8] have done a work using, one partition clustering algorithm (k-Means) and one hierarchical clustering algorithm (agglomerative). k-means algorithm has higher effectiveness and scalability and converges fast when production with large data sets. Hierarchical clustering constructs a hierarchy of clusters by either frequently merging two smaller clusters into a larger one or splitting a larger cluster into smaller ones. Using WEKA data mining tool, they have calculated the performance of k-means and hierarchical clustering algorithm on the basis of accuracy and running time. Guru, et al. [9] have proposed the computational model based on a multilayer perceptron with three layers is employed to enlarge a decision support system for the finding of five major heart diseases. The proposed decision support system is trained using a back propagation algorithm amplified with the momentum term, the adaptive learning rate and the forgetting mechanics.

Palaniappan, et al. [10] have carried out a research work and have built a model known as Intelligent Heart Disease Prediction System (IHDPS) by using several data mining techniques such as Decision Trees, Naïve Bayes and Neural Network. Shanta Umar, et al. [11] have done a research work in which the intelligent and effective heart attack prediction system is developed using Multi-Layer Perceptron with Back Propagation. Accordingly, the frequency patterns of the heart disease are mined with the MAFIA algorithm based on the data extracted. Yanwei, et.al [12] have built a classification method based on the origin of multi parametric features by assessing HRV (Heart Rate Variability) from ECG and the data is pre-processed and heart disease prediction model is built that classifies the heart disease of a patient. Data mining plays an important role in the field of heart disease prediction. [13] Medical Data mining has great potential like exploring the hidden patterns which can be utilized for clinical diagnosis of any disease dataset [14]. Several data mining techniques are used in the diagnosis of heart disease such as Naive Bayes, Decision Tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies. Naive Bayes is one of the successful classification techniques used in the diagnosis of heart disease patients. Peter et al. [15] talked about a new feature selection method algorithm which is the hybrid method which combined CFS and Bayes theorem (CFS+FilterSubset Eval) and evaluated accuracy 85.5%. Shouman [16] presented work by integrating k-means clustering with Naive Bayes using different initial centroid selection to improve the Naive Bayes

accuracy for diagnosing heart disease patients and accuracy was 84.5%.

III. PROPOSED WORK

The proposed approach in this paper uses neighbouring based outlier detection mechanism to attain high prediction rate of heart disease from the used dataset. The proposed framework is as shown in fig1.

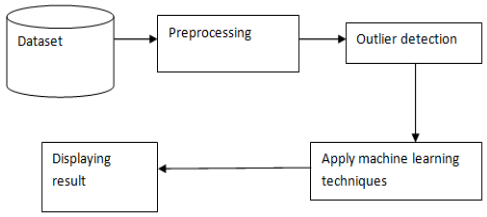

Fig1:proposedsystem

a) Data set: The dataset selected for the prediction of cardiovascular diseases (heart diseases) is collected from Statlog and Cleveland. Data set description is as given in Table1.

b) Preprocessing: Coronary illness data is pre-processed after the collection of different records. The dataset covers an aggregate of 303 patient histories, where 6 records are for certain missing qualities. Those 6 records have been taken out from the dataset and the leftover 297 patient records are utilized in pre-preparing. In case of the patient having a coronary illness, the worth is set to 1, else the worth is set to 0 signifying the shortfall of heart illness in the patient. The pre-preparing of information is done by changing over clinical records into determination esteems. The consequences of information pre-handling for 297 patient records demonstrate that 137 records show the estimation of 1 setting up the presence of heart illness while the excess 160 mirrored the estimation of 0 showing the shortfall of coronary illness.

c) Outlier detection:The definition suggests to us that an outlier is something which is an odd-one-out or the one that is different from the crowd. Some statisticians define outliers as 'having a different underlying behavior than the rest of the data'. Alternatively, an outlier is a data point that is distant from other points. In this paper Squeezer algorithm is used for cluster formation.

Table1.Descriptionofthe AttributesinDataset

Variable	Description	Type
Age	Age in years	Integer
Sex	1=Male, 0=Female	Binary
Cp	cp: chest pain type 0: asymptomatic 1: atypical angina 2: non-anginal pain 3: typical angina	Categorical
Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Continuous
Chol	Serum cholesterol in mg/dl	Continuous
Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Binary
Restecg	Resting electrocardiographic results 0: showing probable or definite left ventricular hypertrophy by Estes' criteria; 1: normal; 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)	Categorical
Thalach	Maximum heart rate achieved	Continuous
Exang	Exercise induced angina (1 = yes; 0 = no)	Binary
Oldpeak	ST depression induced by exercise relative to rest	Continuous
Slope	The slope of the peak exercise ST segment 0: downsloping; 1: flat; 2: upsloping	Categorical
Ca	number of major vessels (0-3) colored by fluoroscopy *(4 missing values)	Integer
Thal	Thallium stress test result 1 = fixed defect; 2 = normal; 3 = reversable defect) *(2 missing values)	Categorical
Target	Presence of heart disease: 0 = disease, 1 = no disease	Binary

Algorithm:

Step1: Read in an object from D. If it is the first object, establish a new clustering structure. Otherwise, goto step 2;

Step2: For each existed cluster C, calculate the similarity between C and each object, according to Eq. (1);

$$Sim\ (C, tid) == \sum_{i=1}^{m} \left(\frac{Sup\ (ai)}{\sum_{ai \in VAL\ i(C)} Sup\ (ai)}\right) \ldots\ldots eqn\ (1)$$

object t with tid∈TID (TID is the set of unique identifiers of every object),

where Sup(ai) is the support of ai in cluster C and VALi(C) is the set of different attribute values respect to C.

Step3: Get the maximum of similarity sim_max and corresponding cluster index from the results computed in step 2;

Step4: If sim_max ≥s, assign the object to the corresponding cluster. Otherwise, goto step 5;

Step5: Construct a new clustering structure;

Step 6: If all the objects in D are processed, output the clustering results. Otherwise, goto step 1.

In the next phase calculate the outlier score of an instance is calculated as shown in equation 2. Based

on the distance between the instance and the center of its nearest cluster available from Squeezer algorithm.

$$Score=\frac{distance\ (o.Co)}{L} \ldots\ldots\ldots eqn\ (2)$$

In this formula, distance (o, Co) represents the distance between instance o and cluster center Co, whereas L indicates the mean distance of that cluster. So, Anomaly Score in Eqn.2 measures the ratio of the distance of each instance from the cluster center to the mean distance of that cluster. The further away an instance from the center of its cluster, the more likely that o is an anomaly instance. Next, we calculate the minimum anomaly threshold score and maximum anomaly threshold score for each cluster using eqn. 3 and eqn. 4 respectively, where Q1 represents 25th percentile of the data and Q3 represents 75th percentile of the data. The Interquartile range (IQR) is the difference between Q3 and Q1 as shown in eqn. 5. Finally, all the instances having an anomaly score greater than Max Threshold (MaxT) or less than Min Threshold (MinT) will be detected as an anomaly.

*Min Threshold (MinT)=Q1-1.5*IQR ...eqn (3)*
*Max Threshold (MaxT)=Q3+1.5*IQR ...eqn (4)*
Interquartile range (IQR)=Q3-Q1 ...eqn (5)

d) Nearest Neighbor Algorithm

The K-Nearest Neighbor (K-NN) classification algorithm is one of the many supervised classification algorithms. K-NN classifies object based on the similarity measures, which could be the distance functions. The K-NN algorithm is a non-parametric algorithm; no assumptions are required on the underlying data. K-NN classification works by calculating the distance between a new input with all the observations in the dataset, and then the algorithm classifies the new input with its nearest neighbor. There are multiple calculation methods to calculate the distance between points in a graph.

IV. EXPERIMENTAL RESULTS

Proposed system is experimented with different values. As shown in fig1 the web application home page is displayed. Entering patient details is as shown in fig2. The final result will be displayed based on input values as shown in fig3.

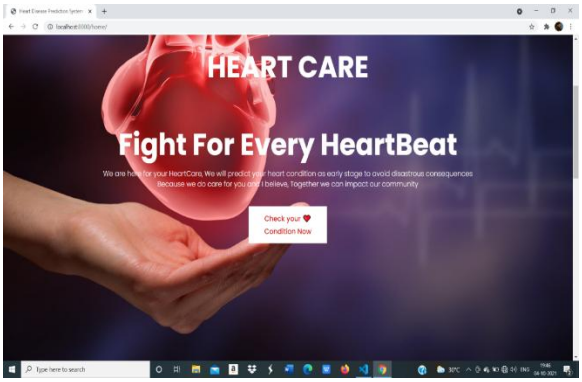

Fig.1:Homepage

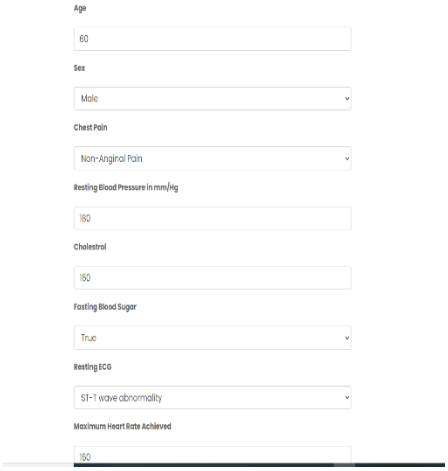

Fig2: Entering details page

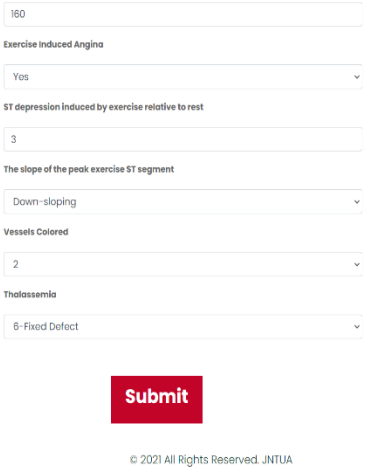

Fig3: Entering details page

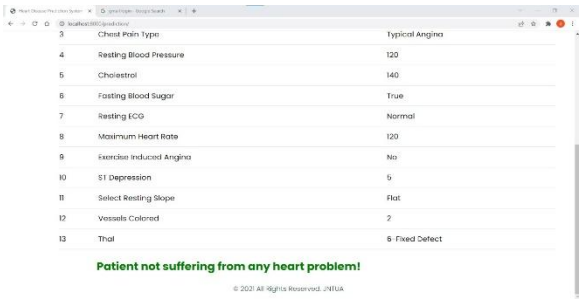

Fig4:Resultpage

Fig5:Comparison of Accuracy with existing system

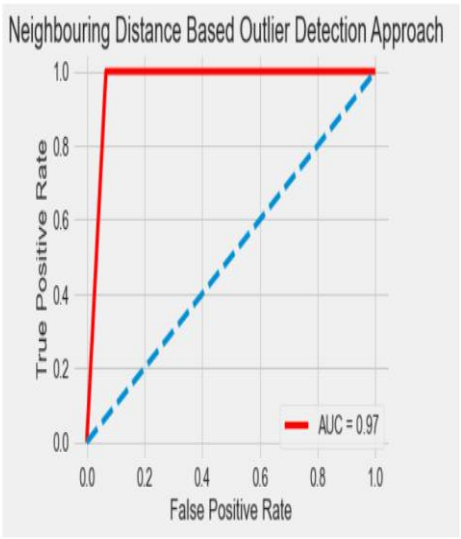

Fig6:AUCofproposedsystem

Accuracy of the proposed system is given in fig6 and it is shown that the proposed system gives better results than existing system XGboost. areaundercurve(AUC)oftheproposedsystemisasshown infig6.

V. CONCLUSIONS

Machine learning is utilized across numerous circles all throughout the planet. The medical care industry is no exemption. AI can assume a fundamental part in anticipating presence/nonattendance of Locomotors problems. Heart illnesses and that's just the beginning. Such data, whenever anticipated well ahead of time, can give significant bits of knowledge to specialists who would then be able to adjust their analysis and treatment per patient premise. The current framework utilizes Dbscan based strategies for foreseeing coronary illness yet this procedure isn't suit or huge datasets. To avoid this problem this paper proposed a new outlier mechanism as neighboring distance based outlier detection approach. It can give better results in terms of accuracy. This can help patients in getting a quick diagnosis with a lot less cost.

REFERENCES

[1] C. Ordonez, “Improving Heart Disease Prediction using Constrained Association Rules,” Tech. Semin. Present. Univ. Toyo, 2004.

[2] M. C. and P. M. Franc Le Duff, Cristian Munteanb, “Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method,” Stud. Health Technol. Inform., vol. Vol. 107, no. 2, p. No. 2, pp. 1256–1259, 2004.

[3] W.J. F. and G. Piatets y-Shapiro, “Knowledge Discovery in Databases: An Overview,” AI Mag., vol. Vol. 13, N, no. 3, pp. 57–70, 1996.

[4] Y. N. and R. HeonGyu Lee, “Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV,” Proc. Int. Conf. Emerg. Technol. Knowl. Discov. Data Min., p. pp. 56–66, 2007.

[5] B. J. L. and H. R. iyong Noh, HeonGyu Lee, Ho-Sun Shon, “Associative Classification Approach for Diagnosing Cardiovascular Disease,” Intell. Comput. Signal Process. Pattern Recognit., vol. 345, pp. 721–727, 2006.

[6] L. P.and R. Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm,” Int. J. Biol. Biomed. Med. Sci., vol. Vol. 3, no. No. 3, pp. 1-8, 2008.

[7] A. D. and N. R. Niti Guru, "Decision Support System for Heart Disease Diagnosis using Neural Networ ," Delhi Bus. Rev., vol. Vol. 8, no. 1, pp. 1–6.

[8] S. P. and R. Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques," Int. J. Comput. Sci. Netw. Secur., vol. Vol. 8, no. No. 8, p. pp. 1–6, 2008.

[9] Shanta umar B. Patil and Y.S. umaraswamy, "'Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Networ'," Eur. J. Sci. Res., vol. Vol. 31, no. No. 4, p. pp. 642-656, 2009.

[10] X. Y. et Al., "Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease," Proc. Int. Conf. Converg. Inf. Technol., pp. 868–872, 2007.

[11] E. Y. and C. ilicier, "Determination of Patient State from Cardiotocogram using LSSVM with Particle Swarm Optimization and Binary Decision Tree," Master Thesis, Dep. Electr. Electron. Eng., no. Uludag, 2013.

[12] N. S. and D. Singh, "Performance Evaluation of k-means and Hierarchal Clustering in Terms of Accuracy and Running Time," Dep. Comput. Sci. Eng. Bar atullah Univ. Inst. Technol., no. Ph.D Dissertation, 2012.

[13] S. G. Manoj B, umar G, Ramesh G, "Emerging risk factors for cardiovascular diseases: Indian context.," Indian J Endocrinol Metab, vol. 17, pp. 806–814, 2013.

[14] ElamaZannatul F., "Combination of Naive Bayes classifier and K-NN in the classificationbased classification models.," Comput. Inf. Sci, no. 6, pp. 48–56,2013.

[15] Halaudi Daniel M., "Prediction of heart disease using classification algorithms.," WCSECS, pp. 22–24, 2014.

[16] A. Uma ND, "Extraction of action rules for chronic idney disease using Naive Bayes classifier.," IEEE Int Conf. ComputIntell. Comput Res, 2016.