Heart Disease Prediction using Ensemble Learning Method

Ramatenki Sateesh Kumar, S.Sameen Fatima, Anna Thomas

***ABSTRACT:** The human heart is the very important organ in our body. The World Health Organization estimates 31% of deaths are due to heart disease taking an estimated 1.79 crore lives. Unhealthy lifestyle, family history of heart problems, stress, etc. are few risk factors for heart disease. In this paper we are proposing an ensembling classifier using K-NN[17] , SVM[18], MK-NN and CART[19] (Decision Tree algorithm) for the efficient prediction of heart disease. The performance and efficiency of the algorithms and ensembling classifier are evaluated. The results indicate that the proposed system was more accurate to determine the existence or non-existence of heart disease. Out of these algorithms, ensemble classifier predicts heart disease more accurately. The accuracy is above 93%.*

***Keywords:** Heart disease, K Nearest Neighbor, Suport Vector Machine, Decision Tree, Classification*

I. INTRODUCTION

Across all the domains there is an enormous amount of data being generated every single minute. Especially in the healthcare industry after the shift from the written records to the electronic health records. There is a large amount of patient data, diagnosed test results and electronic health records available. With data at hand, datamining methods can now be used to extract unknown knowledge and patterns [12]. This extracted knowledge can be used to determine the existence of a disease, recognize various factors on which it is dependent etc. One such disease is heart diseases. Since heart or cardiac disease is a significant cause of death globally, early detection and proper medication are very essential.

In this paper an ensembling classifier is proposed using KNN,SVM,CART and MKNN algorithms. These algorithms are trained and then used to predict or determine the existence or non-existence of heart disease. After the algorithms have been trained their performance is evaluated and compared by using accuracy, recall, precision, and F1 score measures[1].

Manuscript received on May 18, 2020.
Revised Manuscript received on May 27, 2020.
Manuscript published on May 30, 2020.

* Correspondence Author

R.Sateesh Kumar*, M.Tech from JNTU, Hyderabad

Dr.S.Sameen Fatima, Department of Computer Science and Engineering of Osmania

Anna Thomas, M.Tech in Computer Science & Engineering, Department of Computer Science & Engineering, Vasavi College of Engineering, Hyderabad.

II. LITERATURE REVIEW

Krishnaiah et al [1], reviewed the various data mining algorithms used for predicting heart disease. They studied the various approaches used in different papers. Their findings showed different accuracy by taking different number of features and methods which are used for implementation. It was observed from their study that the Fuzzy Intelligent Techniques increased the accuracy of the prediction system.

Chitra et al [2], for predicting heart disease used Supervised Learning Algorithm. The results were compared with SVM. A Cascaded Neural Network (CNN) classifier is used to classify the record of the patient. To calculate the risk of the disease 13 attributes are given to the CNN classifier as input. The dataset contained records collected from 270 patients. The results showed a better accuracy for the CNN classifier than the known SVM classifier.

Medhekar et al [3], used the classifier method for the identification of heart disease. For the classification they used Naive Bayes classifier. The system divides the data into five groups. For the new sample that comes, the system predicts the risk level of the patient. They have experimented with different ratio of train and test set.

Purushottam et al [4], using data mining proposed Heart disease prediction system. The research was to support the non-specialized doctors in making the right decision regarding the risk level of heart disease. The rules produced by the proposed system are ranked. The study was made on Cleveland dataset. 13 input attributes were used in the study. The system was trained and tested using the 10-fold method. The accuracy found in training phase was 87.3% and in the testing phase, it was found to be 86.3%.

S.Dangare et al [5], studied three algorithms for predicting heart disease. Naive Bayes algorithm, Decision Tree algorithm and Neural Networks were used. They also considered two more attributes, namely smoking and obesity along with the other 13 known attributes from the dataset. These attributes are added to get more accurate results on disease prediction. The system has a 100% accuracy for Neural Networks, 90.74% accuracy for Naive Bayes algorithm and 99.62% accuracy for Decision Trees algorithm. It has showed that out of these three algorithms used for prediction Neural Networks has the highest accuracy. AL-Milli [6], developed a system for predicting heart disease using Neural Networks. The dataset used in this research was obtained from the Cleveland database. It had a total of 166 records. The dataset was subdivided into two sets, namely the training set which consisted of 116 records and the testing set which consisted of 50 records. It used the back-propagation algorithm which is a technique for creating a neural multilayer network.

The results that this method obtained better results that the other known algorithms. Soni et al [7], used the Weighted Association rule-based classifier and developed a GUI interface for the predicting heart disease. This classifier assigns different weights to the attributes based on their predicting capability. After the consultation with an expert Doctor the weights were assigned to the different attributes. The WAC stores significant patterns from the dataset.
The significant patterns are stored in the form of Prediction rules in the rule base. The patient is predicted the existence or non-existence of heart disease based on these rules which are stored in the rule base. Instead of using 5 class labels for the target attribute only two of them were used in this study. The system achieved a maximum accuracy of 81.51 % by using the support value as 25 % and the confidence value as 80%.
Shouman et al. [9], applied voting with K-Nearest Neighbor for the disease prediction. The results show that KNN achieved the highest accuracy than any other algorithms applied on the dataset. It is also found that voting could not improve KNN's accuracy.
David et al. [11], used the following three classification algorithms for heart disease prediction they are Random Forest, Decision Tree and Naïve Bayes. It was conducted on the benchmark database found at UCI repository. 14 attributes were used. The performance of algorithms is compared using precision, recall, F1 score, ROC area, and PRC area.

III. METHODOLOGY

The approach used in this study is indicated in Fig. 1.

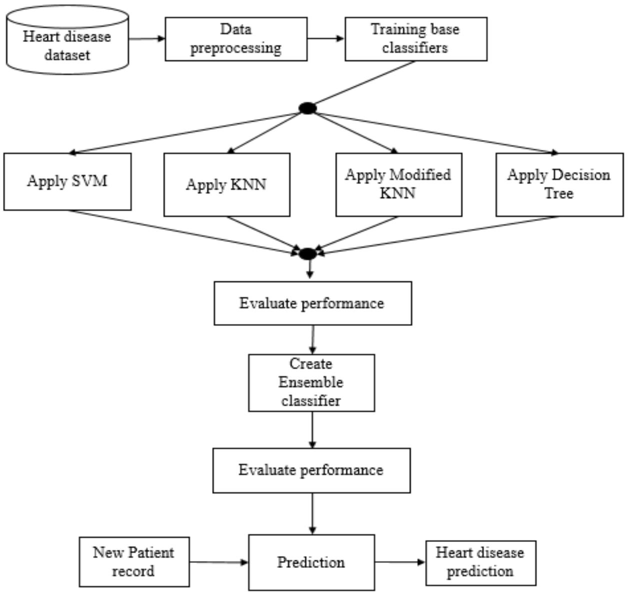

Fig.1 Methodology

The dataset contains 303 records of patients based on 13 attributes. Now the records are divided into two parts. One part is used to train the algorithms, and the other to evaluate the performance. The above-mentioned data mining algorithms are implemented.

Dataset description
This research makes use of the Cleveland Clinic Foundation dataset [8. There are 76 attributes present in the dataset. In this study 13 attributes from them are used and the description is given in the table below. The suffering from heart disease or suffering from heart disease is shown by the target attribute called disease.

Table 1. Dataset description

Attribute No.	Attributes used	Attribute description
A1	age_years	This value in years indicate the age of the patient
A2	gender	It indicates the gender 1 indicates Male 0 indicates Female
A3	cpain_type	It shows the type of chest pain 0 indicates typical angina 1 indicates atypical angina 2 indicates non-anginal pain 3 indicates asymptomatic
A4	rest_bp	It gives the value of resting blood pressure represented in mm Hg
A5	serum_chol	It gives the serum cholesterol value represented in mg/dl
A6	fb_sugar	It gives the value of fasting blood sugar 1 indicates > 120 mg/dl 0 indicates < 120 mg/dl
A7	rest_ecg	It indicates the resting ECG value 0 indicates normal 1 indicates having a ST-T wave abnormality 2 indicates showing possible or definite left ventricular hypertrophy
A8	max_heartrate	It shows the maximum heart rate achieved
A9	ex_angina	It gives the value of angina induced by exercise 1 indicates yes 0 indicates no
A10	ex_ST_depression	It shows the value of the ST depression caused by exercise
A11	peak_slope	It shows the value of peak ST segment during exercise 0 indicates upsloping 1 indicates flat 2 indicates down sloping
A12	num_vessels	It gives the count of the major vessels which are colored by fluoroscopy. It has a value from 0 – 3
A13	defect_type	It gives the type of defect of heart 1 indicates normal 2 indicates fixed defect 3 indicates reversable defect

A14	disease	Identification of heart disease 1 indicates suffering from of disease 0 indicates not suffering from disease

A.KNN

KNN[16] is one of the most useful classification techniques proposed by Hodges and Fix [16]. It is a supervised learning algorithm that can be used both for classification and regression problems in data mining. KNN classifies a new input on the basis of the similarity measures e.g. Euclidean distance [13].

This algorithm consists of no explicit training phase. The KNN classifies a new instance by finding its nearest k neighbors that are obtained by the similarity measures. It assigns the class label which has the maximum count in the nearest Neighbors.

KNN algorithm steps
1 Load the training set and the testing set
2 Choose a k value
3 For every data point present in the test set perform the following
3.1 Evaluate the Euclidean distance between the training data points in each row and the test data point
3.2 Calculated Euclidean distance taken in the non-decreasing order
3.3 Retrieve the top k rows from the above sorted distance array
3.4 The most suitable class label is obtained from these sorted rows will be assigned to the test data
3.5 Now return the class value obtained from the above step for the test data point
4 Stop

B.Modified K-Nearest Neighbor(MKNN)

A possible scenario where the KNN algorithm might misclassify a data point is identified. In order to overcome this problem, the modified KNN is implemented. The possible scenario is when a data point belonging to class X is surrounded by class Y data points. Now when the top k Neighbors are considered, the new point is misclassified as class Y because of the majority of class Y points. To overcome this problem, the selection of top k neighbors is varied. To reduce the class Y neighours, it skips 3 Neighbors and considers the next 3 Neighbors (k+3 Neighbors). Now class X has the possibility of becoming the majority class among both the classes. With this approach, the new point can be correctly as class X, thereby reducing the misclassification.

Modified KNN algorithm steps
1 Load the training set and the testing set
2 Choose a k value
3 For every data point present in the test set perform the following
3.1 Evaluate the Euclidean distance between the training data points in each row and the test data point
3.2 Calculated Euclidean distance taken in the non-decreasing order
3.3 To obtain the first k rows from the above distance array, the following two value are used
start = k / 2, end = k + s
The top k rows are obtained by adding
k = sorted_distance[0: start] + sorted_distance[k-1: end]
3.4 The most frequent class from these sorted rows will be assigned to the test data
3.5 Now return the class value obtained from the above step for the test data point
4 Stop

C.SVM

SVM[17] is a supervised learning algorithm. In SVM, each data point from the dataset is plotted in a n-dimensional region, here n indicates the number of features. A hyperplane is a decision plane that divides the set of data points into distinct classes. Data points which lie nearest to the decision plane are called as support vectors [15]. These data points help in defining the decision boundary. The distance between the two decision boundaries form the margin. The width of the margin denotes the error in the classifier. Therefore, a classifier with a larger margin will have lower classification error. The main objective of SVM is to identify a maximum margin hyperplane. There can many hyperplanes that can be found. Thus, during the training phase, it iteratively constructs hyperplanes and then search for the hyperplane with maximum margin.

D.Decision tree(CART)

A Decision tree algorithm is an example of supervised learning algorithm. Both classification and regression problems can be solved by it. Decision tree node indicates a test performed on that attribute, whereas the branch indicates the test result and the leaf nodes contains the class label. ID3, CART, C4.5, CHAID, and J48 are the different decision tree algorithms [7]. In this system CART algorithm also known as Classification & Regression Trees is used. Gini index is used as the metric [13] to identify the best splitting criteria. In CART, a binary tree is constructed where every internal node will have only two output for that attribute. The Gini Indices are calculated for all the attributes. The attribute with the minimum Gini Index will be considered as the splitting attribute [14]. By recursively selecting the minimum Gini index attribute, a tree is constructed. Gini Index is calculated as shown below

$$Gini\,(D) = 1 - \sum_{i=1}^{n} p_i^2$$

here p_i is the probability of the samples that belongs to class C_i for that particular node and n represents number of classes.

IV. RESULTS

This section shows the results ofe discussed algorithms on the heart disease dataset.
Accuracy, recall, precision and F1 score are used to assess the algorithm efficiency. They are determined from the confusion matrix. Table 2 shows the Confusion matrix obtained for each algorithm.

Table 2. Confusion matrix[17] of the algorithms
Confusion matrix of SVM

	x	Y
x	38	5

y	2	46

Confusion matrix of KNN

	X	Y
x	33	7
y	1	50

Confusion matrix of the Modified KNN

	x	y
x	38	4
y	3	46

Confusion matrix of Decision Tree

	x	y
x	32	9
y	2	48

Confusion matrix of Ensemble

	x	y
x	35	5
y	1	50

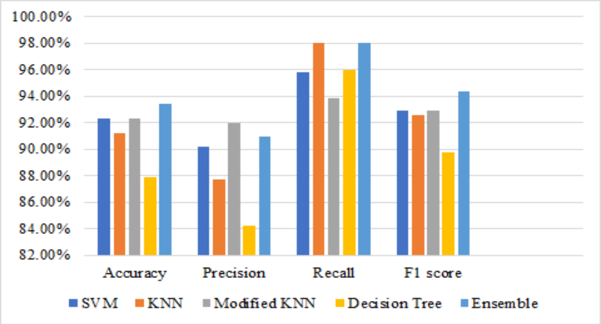

Fig.2 Comparison of algorithm results with 13 attributes

Comparison of algorithms and ensembling classifier is show in Table 3.

Table 3. Results of algorithms and ensembling classifier result

	SVM	KNN	Modified KNN	Decision Tree	Ensemble
Accuracy	92.31 %	91.21 %	92.31 %	87.91 %	93.41 %
Precision	90.20 %	87.72 %	92 %	84.21 %	90.91 %
Recall	95.83 %	98.04 %	93.88 %	96 %	98.04 %
F1 score	92.93 %	92.59 %	92.93 %	89.72 %	94.34 %

Table 7. Misclassification rate of KNN and Modified KNN

KNN	Modified KNN
8.79 %	7.69 %

The accuracy of the algorithms is plotted on a graph as shown below

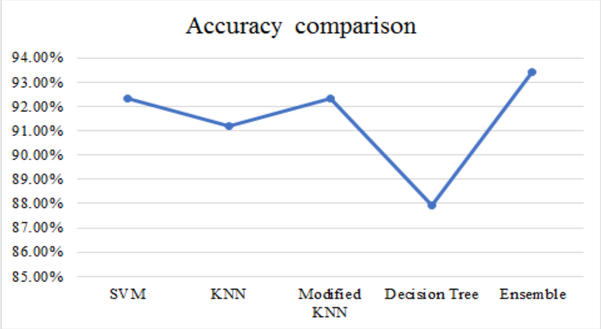

Fig.3 Accuracy comparison plot of each algorithm

V. CONCLUSION

The main purpose of this study was to more accurately predict the existence or non-existence of heart disease. Four algorithms were used they are KNN , SVM, Modified KNN and Decision Tree algorithm. Four performance evaluation measures were considered for the classifiers. It is observed from the result that the ensemble achieved greater accuracy than the base classifiers. From the results of comparison of all the individual classifiers, it is seen that Support Vector Machine (SVM) and modified KNN achieved the highest value in accuracy. There is also a decrease in the misclassification rate in Modified KNN when compared to KNN. Modified KNN performed better than the KNN. The overall study showed that heart disease can be predicted more accurately using an ensemble than the individual base classifiers.

FUTURE SCOPE

The system can use more input attributes present in the dataset and use AI methods. Any other data mining techniques can also be used for prediction. Text mining can be used to extract large quantities of unstructured data available in the database of the healthcare sector. Combining different algorithms to create ensemble for better prediction. Advanced ensemble classifiers can also be used.

REFERENCES

1. Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2015). Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach. *Advances in Intelligent Systems and Computing*, 371–384.
2. Chitra, R., & Seenivasagam, V. (2013). Heart Disease Prediction System Using Supervised Learning Classifier. Bonfring International Journal of Software Engineering and Soft Computing, 3(1), 01–07.
3. Medhekar, D.S., Bote, M.P., & Deshmukh, S. (2013). Heart Disease Prediction System Using Naive Bayes. *International Journal of Enhanced Research in Science Technology & Engineering*, 2(3) 1-5.
4. Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. *Procedia Computer Science*, *85*, 962–969.
5. S.Dangare, C., & S. Apte, S. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *International Journal of Computer Applications*, *47*(10), 44–48. https://doi.org/10.5120/7228-0076
6. AL-Milli, N.R. (2013). Backpropogation Neural Network for Prediction Of Heart Disease. *Journal of Theoretical and Applied Information Technology*, 56(1), 131-135.

7. Soni, J., Ansari, U., & Sharma, D.M. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. *International Journal on Computer Science and Engineering,* 3(6), 2385- 2392.
8. UCI Machine Learning Repository: Heart Disease Data Set. (n.d.). Retrieved September 2, 2019, from https://archive.ics.uci.edu/ml/datasets/Heart+Disease
9. Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-Nearest Neighbor in Diagnosing Heart Disease Patients. *International Journal of Information and Education Technology*, 220–223.
10. Bhargava, N., Dayma, S., Kumar, A., & Singh, P. (2017). An approach for classification using simple CART algorithm in WEKA. *2017 11th International Conference on Intelligent Systems and Control (ISCO)*.
11. David, H.B., & Belcy, S.A. (2018). Heart Disease Prediction Using Data Mining Techniques. *ICTACT Journal on Soft Computing*, 9(1), 1817- 1823.
12. https://doi.org/10.1109/icetets.2016.7603000
13. Kleinberg, J., Papadimitriou, C. & Raghavan, P. A Microeconomic View of Data Mining. *Data Mining and Knowledge Discovery 2*, 311–324 (1998). https://doi.org/10.1023/A:1009726428407
14. Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques.* Maarssen, Netherlands: Elsevier Gezondheidszorg. Breiman L., Friedman J.H., Olshen R.A. and Stone C. J. (1984). *Classification and Regression Trees (2nd Ed.).* Pacific Grove, CA; Wadsworth.
15. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory -COLT ’92,* 144 -152. https://doi.org/10.1145/130385.130401
16. Fix, E., and Hodges, J. (1951). Dicriminatory analysis ,non parametric discrimination:consistency properties. *Technical report 4,USA,School of aviation medicine Randolph field Texas.*
17. B. Sun, J. Du and T. Gao, "Study on the Improvement of K-Nearest-Neighbor Algorithm," *2009 International Conference on Artificial Intelligence and Computational Intelligence*, Shanghai, 2009, pp. 390-393, doi: 10.1109/AICI.2009.312.
18. Dengming Peng, F. C. Lee and D. Boroyevich, "A novel SVM algorithm for multilevel three-phase converters," *2002 IEEE 33rd Annual IEEE Power Electronics Specialists Conference. Proceedings (Cat. No.02CH37289)*, Cairns, Qld., Australia, 2002, pp. 509-513 vol.2, doi: 10.1109/PSEC.2002.1022504.
19. S. Gey and E. Nedelec, *"Model selection for CART regression trees*," in IEEE Transactions on Information Theory, vol. 51, no. 2, pp. 658-670, Feb. 2005, doi: 10.1109/TIT.2004.840903.

AUTHORS PROFILE

R.Sateesh Kumar He completed M.Tech from JNTU, Hyderabad in 2011. Currently pursuing Ph.D from OU, Hyderabad. he got more than 15 yrs experience and currently working with Vasavi college of Engineering, Hyderabad. He got more than 10 publications to his Credit.

Dr.S.Sameen Fatima She has over 33 years of experience in teaching, research and administration in India, USA and UAE. She joined as a faculty in the Department of Computer Science and Engineering of Osmania University during its formative years in 1984. She took over as Principal in July 2016, and holds the distinction of being the first lady Principal, in the history of the College of Engineering, Osmania University. Currently she is a Professor at the Department of Computer Science and also the Director, Centre for Women’s Studies at Osmania University. She is guiding more than 20 research scholars. She got more than 50 publications in her credit

Anna Thomas She is currently pursuing her final year M.Tech in Computer Science & Engineering, Department of Computer Science & Engineering, Vasavi College of Engineering, Hyderabad.