ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse272952020.pdf
https://doi.org/10.30534/ijatcse/2020/272952020

Heart Disease Prediction Using Extended KNN(E-KNN)

R.Sateesh Kumar[1], Dr.S.Sameen Fatima[2]
[1]Assistant Professor, Department of Computer Science and Engineering, Vasavi College of Engineering, Hyderabad(Telangana), INDIA , sateeshramatenki@gmail.com
[2]Registar,Anurag University, Hyderabad (Telangana), INDIA, sameenf@gmail.com

ABSTRACT

The WHO estimates that deaths due to heart disease are the number one cause worldwide, accounting for around 30% annually taking an estimated 1.5 crores died due to this disease . In this study we are using an extension of KNN algorithm known as E-KNN and compare the results with K-Nearest Neighbor(KNN) , Support Vector Machine(SVM) and Classification and Regression Trees (CART)[14] in the prediction of heart disease. To improve the efficiency of the proposed system the most important features are selected using chisquare test. The performance and efficiency of the algorithms are evaluated and compared on the basis of accuracy, recall, precision and F1 score. The results of the proposed algorithm was more accurate with lesser attributes than all 13. The performance of E-KNN by using 11 attributes. It has the accuracy value of 90.10 %. It is followed by SVM with 89 % accuracy

Key words : Heart disease, EKNN,KNN, SVM, Decision Tree, Classification

1. INTRODUCTION

WHO estimated that there are 17 million people die of this heart diseases (CVD) every year. These death rate can be reduced with early diagnosis and inform to the patients. Common cardiovascular diseases include coronary heart disease, cardiomyopathy, hypertensive heart disease, heart failure, etc. Smoking, diabetes, lack of physical activity, hypertension, high cholesterol diet are some of the common causes of heart diseases.

Research leading to the sciences of cardiovascular diseases using machine learning and data mining has been an endeavor encompassing optimum treatment plans, expedited and early disease prognosis, identification of risk factors. Numerous CVD surveys have been conducted with the primary dataset as Cleveland dataset.. Recommending the parameters from this dataset proposes a predictive system to apply logistic regression model, random forests, SVM, and KNN to obtain predictions from several learners which are in turn used in meta models . The results of each ensemble modeling is then compared on the basis of several evaluation parameters.

In this paper we are proposing extension of KNN algorithm namely Extended K-Nearest Neighbor algorithm(E-KNN). The results of this algorithm compared with K-Nearest Neighbor algorithm, Decision Tree algorithm and Support Vector Machine algorithm. The specified data set is used to train these algorithms . With enough data and time, it is good to all the attributes or features including the irrelevant features to train the classification algorithm. The problems of using the irrelevant features are that it leads to lesser accurate results and it may also lead to overfitting. Feature subset selection is a dimensionality reduction method which is used to select the most relevant attributes from the dataset [14]. In this paper the Chi-square test is used as the feature subset selection method. The selected features are fed to the algorithms as input.

2. LITERATURE REVIEW

Revathi et al [7], provides an analysis of various data mining methods for prediction of heart disease. Decision tree, Back-propagation algorithm and Naïve Bayes are used. The system used 14 parameters including blood pressure, chest pain, cholesterol and heart rate to enhance the system accuracy. A comparison of the performance of the three algorithms is done. Study shows that the neural network has the accuracy of 100%. It has outperformed the other two algorithms.

Shinde et al [10], using the data mining methods introduced a system for predicting heart disease. K-means clustering algorithm and Naïve Bayes were used in this system. For prediction, a combination of both techniques has been used. The K-means clustering algorithm is used to enhance overall system efficiency. It is used to group the different attributes present in the dataset and the prediction is done by the Naïve Bayes algorithm. The system used 10 attributes. Compared with the other algorithms, the system produces better results. Sateesh et al[14], using ensemble learning method to predict the heart disease. They used KNN,SVM,CART algorithms and design an ensembler.

3.APPROACH

The methodology used showed in the fig1

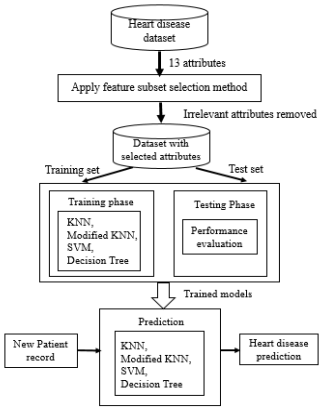

Fig.1 Approach

The Fig1 shows the approach used in the work.The study is using the dataset has 300 records of patients. The original dataset contains thirteen attributes. From the dataset identify the required attributes using the statistical measure (Chi-square test).

3.1 Dataset

It is an openly available dataset and can be found at the UCI Machine Learning Repository. Table 1 describes the data set. The original data set consists of 76 features . We used 13 features from them. The features are described in the following table The disease attribute is the class label which specifies the patient is suffering from the heart disease or not .

Table 1. Dataset description[14]

Feature No.	Name of Feature	Feature description
1.	age_years	Age of the patient
2.	Gender	1 Male 0 Female
3.	cpain_type	Chest pain {0,1,2,3}
4.	rest_bp	This feature represents the value of the blood sugar in rest
5.	serum_chol	Measure the Serum Cholesterol
6.		
7.	fb_sugar	Measure fasting blood sugar { 0 if fb_sugar<120, 1 if fb_sugar>120}
8.	rest_ecg	ECG in resting {0,1,2}
9.	max_heartrate	maximum heart rate
10.	ex_angina	It takes {0,1}
11.	ex_ST_depression	ST depression during exercise
12.	peak_slope	Takes the value s{0,1,2}
13.	Number of vessels	Indicates number of vessels blocked {01,2,3}
14.	defect_type	Heart defect type {1,2,3}
15.	disease	Identification of heart disease 1 Suffering from disease 0 No disease

3.2 Chi-squared value

Sometimes our dataset may contain both relevant and irrelevant features. It is very important to find and discard the unimportant features that does not contribute to the outcome to be predicted by the system. There are various methods to identify the relevant features from our input attributes. The methods to select the relevant features are called as Feature selection methods. Chi-squared test applies the statistical method to identify the relation between the input variable and the target variable.

Steps to determine the Chi-squared value

1. Create a contingency table for two attributes. The values present in the table are called as observed values (O)
2. Calculate the expected frequency (E) value for each of the cell present in the table.
   $E = (row\ total * column\ total)/overall\ sum$
3. Calculate the chi-squared value
   $$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$
4. Arrange the chi-squared value in the descending order.
5. Select the top-ranked attributes from the above list.

3.3 Extended K-Nearest Neighbor

The KNN classifier uses the polling to assign the class label the new data item . This leads to the miss classification. To avoid this problem, the E-KNN is implemented. In KNN a new data item belongs to certain class C1 is surrounded by classes C2 data points. The new data point is misclassified as class C2 because of the majority of class. To avoid this problem in E-KNN , the selection of top k neighbors is varied. To reduce the class C2 neighbors, it skips 3 Neighbors and select the next 3 Neighbors i.e 3+k Neighbors and also assign the weight to the each neighbor based on the distance from the new data item . Class label to the new data item is selected based on the k+3 neighbors and their weights

E-KNN Algorithm steps

1.Divide the data set in training and testing
2. Choose the K values based on the requirements
3. For every new data item compute the following
   3.1 Find the Distance from the new data item to all the data items in the training set using distance measure
   3.2 Sort the distances in descending order and identify the K nearest data points
   3.3 Assign the weights to all the K nearest neighbors (Weight w=1/d where d is the distance from the new data point to its neighbor)
   3.4 To obtain the top k rows from the above by , calculating the two values start = k / 2, end = k + start
   3.5 The top k rows are obtained by adding k = sorted_distance(0: start) + sort ed_distance (k-1: end) and based on the weights
   3.6 The most frequent class from the sorted rows and their weights
4.Stop

3.4 .K-Nearest Neighbor

1.Divide the data set in training and testing
2. Choose the K values based on the requirements
3. For every new data item compute the following
   3.1 Find the Distance from the new data item to all the data items in the training set using distance measure
   3.2 Sort the distances in descending order and identify the K nearest data points
   3.4 The most frequent class from the sorted rows and their weights
4.Stop

4. RESULTS AND DISCUSSION

This section provides the results of implementing the various above discussed algorithms on the heart disease dataset. Table 2 indicates the top 11 attributes ranked from the Chi-square test.

Table 2. Top 11 attributes selected by chi-squared test

Attribute	Rank
A1	max_heartrate
A2	ex_ST_depression
A3	num_vessels
A4	cpain_type
A5	ex_angina
A6	serum_chol
A7	Age
A8	rest_bp
A9	peak_slope
A10	Gender
A11	defect_type

Comparison of algorithms according to the number of attributes used show in Table 3 and Table 4.

Table 3. Comparison of algorithm result with 13 attributes

	SVM	KNN	Modified KNN	Decision Tree
Accuracy	89.10 %	86 %	87 %	79 %
Precision	89 %	86 %	86.5 %	79 %
Recall	89 %	86 %	86.5 %	79 %
F1 score	89 %	86 %	86.5 %	79 %

Table 4. Comparison of algorithm result with selected attributes

	EKNN	KNN	SVM	Decision Tree
Accuracy	90.10 %	88 %	89 %	81 %
Precision	90 %	87.5 %	89 %	81 %
Recall	90 %	88 %	88.5 %	81 %
F1 score	90 %	87.5 %	89 %	81 %

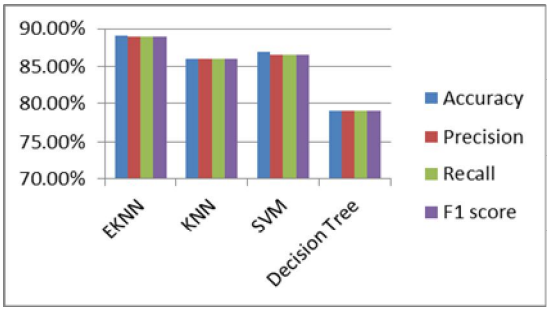

Fig.2 Shows the results with 13 attributes

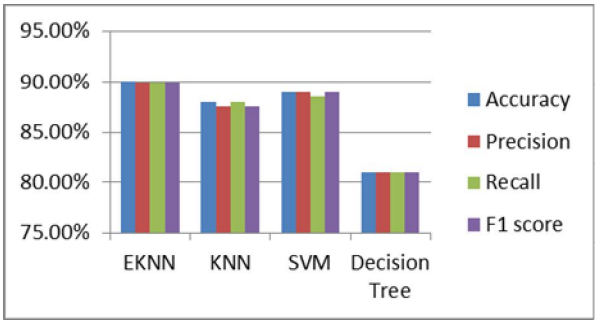

Fig.3 Shows the results with selected attributes

Table 8. Comparison of accuracy according to the number of attributes used

Algorithm	Accuracy value with	
	13 attributes	Selected attributes
EKNN	89.10 %	90.10 %
KNN	86 %	88 %
SVM	87 %	89 %
Decision Tree	79 %	81 %

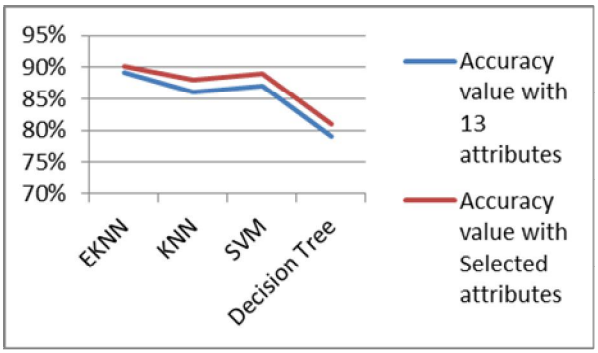

Fig.4 Shows the accuracy of all the algorithms

5.CONCLUSION

The Chi-squared test was used as a feature selection method this help in identifying the suitable attributes for this study. In this spaper we used E-KNN,KNN, SVM, and Decision Tree algorithms. For Support Vector Machine algorithm and Decision Tree algorithm, 11 attributes out of the 13 attributes were used and for K-Nearest Neighbor algorithm and E-KNN algorithm 10 attributes were used. Decision Tree algorithm, K-Nearest Neighbor algorithm and modified K-nearest Neighbor performed better with fewer number of attributes however Support Vector Machine algorithm performed the same even with fewer number of attributes. Support Vector machine achieved the highest value in accuracy, precision, recall and F1 score and outperformed all the other three algorithms. Finally we are concluding that the algorithms efficiently predict heart disease with less number of attributes. We want to extend this work of predicting heart disease by analyzing ECG[16] image.

ACKNOWLEDGEMENT

Authors would like to express gratitude towards Vasavi College of Engineering, Hyderabad for their constant supervision as well as for providing necessary information regarding this research and also their support in completing this endeavor.

REFERENCES

1. Abhishek, T. (2013). **Heart disease Prediction System Using Data Mining Techniques**. *Oriental Journal of Computer Science & Technology*, 6(4), 457-466.
2. Purushottam, Saxena, K., & Sharma, R. (2016). **Efficient Heart Disease Prediction System**. *Procedia Computer Science*, *85*, 962–969.
3. Revathi, T., & Jeevitha, S. (2015). Comparative Study on Heart Disease **Prediction System Using Data Mining Techniques**. *International Journal of Science and Research (IJSR)*, 4(7), 2120-2123
4. Shinde, R.M., Arjun, S., Patil, P., and Waghmare, P.J. (2015). **An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm**. *International Journal of Computer Science and Information Technologies*, 6(1), 637-639.
5. Soni, J., Ansari, U., and 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.
6. Ishtake, S. H., Sanap, S.A. (2013**). Intelligent Heart Disease Prediction System Using Data Mining Techniques**. *International J. of Healthcare & Biomedical Research*, 1(3), 94-101.
7. Jabbar, M. A., Deekshatulu, B. L., and Chandra, P. (2013). **Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. Procedia Technology**, 10, 85–94.
8. Ghumbre, S., Patil, C., and Ghatol, A. (2011). **Heart Disease Diagnosis using Support Vector Machine**. *International Conference on Computer Science and Information Technology*, 84-88.
9. Fida, B., Nazir, M., Naveed, N., and Akram, S. (2011). **Heart disease classification ensemble optimization**

using Genetic algorithm. *2011 IEEE 14th International Multitopic Conference*.

10. Kalaiselvi, C. (2016). **Diagnosing of heart diseases using average k-nearest neighbor algorithm of data mining**, *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 3099-3103.
11. David, H.B., and Belcy, S.A. (2018). **Heart Disease Prediction Using Data Mining Techniques.** *ICTACT Journal on Soft Computing*, 9(1), 1817- 1823.
12. Kavitha, R., and Kannan, E. (2016). **An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining**. *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS).*
13. Jiawei Han et al, **Data Mining Concepts and Techniques**, third edition, ISBN 978-0-12-381479-1, 2012.
14. R.Sateesh Kumar, Ms.Anna Thomas(2020) ,**Heart Disease Prediction using Ensemble Learning Method**,*International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878, Volume-9Issue-1, May2020
15. Siti Salasiah Mokri, M Iqbal Saripan, Abdul Jalil Nordin, Mohammad Hamiruce, Noraishikin, Zulkarnain(2019) **Level Set Based Whole Heart Segmentation in Non-Contrast Enhanced CT Images**, *International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.1.6, 2019
16. Edward B. Panganiban, Arnold C Paglinawan, Wen Yaw Chung, Gilbert Lance S. Paa(2019), **A Novel Techniques in classifying Heart Diseases based on Electrocardiogram (ECG) Signals using Deep Learning and Spectrogram Image Analysi**s, *International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.4, July – August 2019