ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 1, January 2013

Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease

Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg

***Abstract* – Medical errors are both costly and harmful. Medical errors cause thousands of deaths worldwide each year. A clinical decision support system (CDSS) offers opportunities to reduce medical errors as well as to improve patient safety. One of the most important applications of such systems is in diagnosis and treatment of heart diseases (HD) because statistics have shown that heart disease is one of the leading causes of deaths all over the world. Data mining techniques have been very effective in designing clinical support systems because of its ability discover hidden patterns and relationships in medical data. This paper compares the performance and working of six CDSS systems which use different data mining techniques for heart disease prediction and diagnosis. This paper also finds out that there is no system to identify treatment options for HD patients.**

Keywords- data mining; heart disease prediction and diagnosis systems.

I. INTRODUCTION

Medical errors are both costly and harmful [1]. Medical errors cause tens of thousands of deaths in U.S. hospitals each year, more than from highway accidents, breast cancer, and AIDS combined [2]. Based on a study of 37 million patient records, an average of 195,000 people in the U.S. died due to potentially preventable, in-hospital medical errors [3]. Statistics also show that cardiovascular disease is one of the leading causes of death all over the world [4]. Hence reliable and powerful clinical decision support systems (CDSSs) are required to reduce the time of diagnosis and increase diagnosis accuracy especially for heart disease diagnosis [5]. Clinical decision support systems have evolved from statistical algorithms to complex artificial neural networks. The early decision support systems, also, were based on Bayesian statistical theory [6], probability diagnoses based on essential variables [7].

***Syed Umar Amin**, Department of Computer Science & Engg, Integral University, Lucknow, India.*
***Kavita Agarwal**, Departmentt of Computer Science & Engg, Integral University, Lucknow, India.*
***Dr. Rizwan Beg**, Departmentt of Computer Science & Engg, Integral University, Lucknow, India.*

The use of data mining tools has become widely used in clinical applications for disease diagnosis more effectively. Various data mining techniques such as decision trees, artificial neural networks, Bayesian networks, support vector machines kernel density, bagging algorithm have been actively used in clinical support systems for diagnosis of heart disease [8-10]. Although there have been promising results in applying data mining techniques in heart disease diagnosis and treatment, the study done in finding out treatment options for patients and particularly heart patients is comparatively elemental. It has been suggested by researchers that application of data mining techniques for proposing suitable treatments options for patients would not only improve patient care but would also reduce investigation time, errors and would also improve the performance of medical practitioners [11]. There has been a lot of investigation for applying different data mining techniques in the diagnosis of heart disease to find out the most accurate technique but there is no study to find out the data mining technique which can increase reliability and accuracy in finding out effective treatment for heart disease patients.

The remainder of this paper is divided as follows: Various clinical decision systems for heart disease are presented in section 2 followed by comparison and analysis of the presented systems in section 3 and section 4 has conclusion.

II. CLINICAL DECISION SUPPORT SYSTEMS FOR HEART DISEASE USING DATA MINING

Heart disease refers to various ailments that affect the heart and the blood vessels in the heart. Heart attack Coronary artery disease, heart failure, Angina are some examples, which have different symptoms and causes [12]. The detection of heart disease is a complex procedure because of availability of incomplete data and its dependence on several diverse factors. Therefore, intelligent systems using data mining techniques are required for increasing the accuracy of diagnosis. A large number of clinical

decision support systems have been built specially for the diagnosis of various kinds of heart diseases. Six of these systems are discussed here to analyze their performance based on types of heart diseases diagnosed, their strengths and shortcomings.

A. A multilayer perceptron-based medical decision support system for heart disease diagnosis (2006)

This [13] is a multiplayer perceptron (MLP) based decision support system for the diagnosis of heart diseases in which the input layer comprises of 40 input variables, categorized into four groups and then encoded. The number of nodes in the hidden layer is calculated based on learning. The output layer has 5 nodes each corresponds to a particular heart disease.

The MLP system uses back propagation as a learning algorithm. The number of hidden nodes is determined as 15 using cascade learning process. The heart disease database used for testing and tuning the system in this study consists of 352 cases gathered from the Southwest Hospital and the Dajiang Hospital, both located in Chongqing, P. R. China.

Three different evaluation methods have been adopted to assess the performance of the proposed MLP-based decision support system. The cross-validation classification accuracy determined was 91.5% which is high and process accuracy for each individual process was also found out to be >90%. From the results reported that the system has a high capability to accurately recognize all the five heart diseases (>90%) with comparable small intervals (5%).

B. Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm (2007)

This system [14] is based on coactive neuro-fuzzy inference system (CANFIS) which combines the neural network adaptive capabilities and the fuzzy logic qualitative approach that is then integrated with genetic algorithm to diagnose the presence of the heart disease.

This system works like a three-layer feed forward neural network. The first layer represents input variables, the hidden layer represents fuzzy rules and the third layer represents output variables. The system comprises of fuzzy axon which applies MFs to the inputs. Then the second major element is the modular network which applies functional rules to the inputs. Two fuzzy structures are mainly used: the Tsukamoto model and the Sugeno (TSK) model. Then a combiner is used to apply the MF outputs to the modular network outputs. The outputs are then calculated and, the error is back propagated to both the MF and the modular network. A genetic algorithm is used for optimization. The GA combines selection, crossover, and mutation operators with the goal of finding the best solution to a problem by searching until the specified criterion is met.

The simulation of the model was implemented using NeuroSolution software and the publicly available Cleveland heart-disease database consists of 303 cases was used where the disorder is one of four types of heart-disease or its absence. The results shown by CANFIS showed very high performance and accuracy.

C. AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction (2007)

This clinical decision support system (CDSS) [15] for diagnosis of cardiovascular heart disease (CVD) is based on classifier ensemble method, hence it does overcome the performance limitations of conventional single classifier based methods. It combines a set of four different classifiers with ensembles. Support vector machines (SVM), neural networks (ANN), Decision trees (DT) and Bayesian networks (BN) are used as classifiers. The different classifiers in the ensemble are used to analyze patient's serum microarray chip data. This system, Apta-CDSS-E (Aptamer biochip-based CDSS – ensemble), provides diagnosis information to the physicians and it also helps clinicians by providing a set of biomarker candidates which can be used for CVD diagnosis. In AptaCDSS-E, the patient's blood sample is collected and an Aptamer biochip is created with the serum separated from the patient blood and protein expression levels are scanned. Then, a new work list is created by the scanner interface and analyzed by the decision engine of AptaCDSS-E trained with prior sample sets. The CVD levels of patient are classified into four classes: normal (NM), stable angina (SA), unstable angina (UA), and 14 myocardial infarction (MI). At last the system provides integrated analysis results and clinical analysis facts. The physician's, decision results can then be saved into the system database for future model updates.

AptaCDSS-E adopted the ensemble approach to generate enhanced results by grouping a set of classifiers of each SVM, ANN, DT, and BN. Each ensemble is constituted with several classifier models of each classification method. The training data are augmented by bagging and fed to each classifier

member of ensembles. Final decision is decided by weighted majority vote of each ensemble t s decision with respect to their training accuracies. The system was trained with four different disease data sets consisting of 242 cases including cardiovascular disease and the data sets were augmented by bagging for classifier ensemble training. The system predicts the level of CVD with an accuracy (>94%).

D. Intelligent Heart Disease Prediction System Using Data Mining Techniques (2008)

This Intelligent Heart Disease Prediction System (IHDPS) [16] is developed using Decision Trees, Naïve Bayes and Neural Network techniques. There are total of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Data Mining Extension (DMX), query language, is used for building and accessing the contents of models. The models are trained and validated against a test dataset using Lift Chart and Classification Matrix methods to find out the model providing the highest correct prediction. Tabular and graphical visualization methods are used in IHDPS for better interpretation of results.

A total of 909 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records). The most effective model to predict patients with heart disease appears to be Naïve Bayes followed by Neural Network and Decision Trees.

All three models could answer complex queries, each having its own strength based on accuracy, ease of interpretation.

E. Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network (2009)

IEHPS [17] presents methodology for the extraction of significant patterns from the heart disease warehouses for heart attack prediction. Initially, the data warehouse is preprocessed to make the mining process more efficient. The preprocessed data warehouse is then clustered using the K-means clustering algorithm. Frequent Itemset Mining (FIM) is performed using MAFIA (MAximal Frequent Itemset Algorithm) for the extraction of association rules from the clustered dataset. Weightage is then calculated and the patterns vital to heart attack prediction are selected according to the weightage. The neural network is trained with the selected patterns. Multi-layer Perceptron model is used with Back-propagation as the training algorithm. The significant patterns are extracted with the aid of the significance weightage greater than the pre-defined threshold.

IEHAPS uses heart attack dataset obtained from online Machine Learning Repository (UCI). The values corresponding to each attribute in the significant patterns are given as : blood pressure range is greater than 140/90 mm Hg, cholesterol range is greater than 240 mg/dl, maximum heart rate is greater than 100 beats/ minute, abnormal ECG and unstable angina. Some more parameters significant to heart attack with their weightage and the priority levels are advised by the medical experts. The MLPNN is trained according to these attributes and patterns and provides the different risk levels. The experimental results show high efficiency and the overall effectiveness of the system.

F. Clinical Decision Support System: Risk Level Prediction of Heart Disease Using Decision Tree Fuzzy Rules (2012)

It [18] is a weighted fuzzy rule-based clinical decision support system (CDSS) for the diagnosis of heart disease. Fuzzy rules are automatically generated to and are weighted in accordance with their importance using the attribute weightage. These weighted fuzzy rules are applied on the rule base of the fuzzy system before carrying out prediction on the designed fuzzy-based CDSS.

After data pre-processing, the input training dataset used for prediction is classified into two subsets of data based on two class labels, in which ‘0’ indicates that the disease status is less than 50% and ‘1’ represent that the risk level for heart disease is more than 50%. To choose the most relevant and important attributes, the frequent attribute category is mined from the input datasets. Then, based on the frequency of attribute category and the weightage of attributes, the fuzzy rules are generated automatically. Mining is done for one length attribute category by finding the frequency in the database and then, the attribute category of the attributes within that class are arranged in accordance with their frequency. Then, for every attributes, a set of attribute category are selected from the sorted list based minimum support. The selected attribute category is then stored in a two vector, having maximum and minimum for each class, in which one vector contains the minimum value corresponds to the attribute category of every

attribute and second vector contains the maximum value corresponds to the attribute category of every attribute. The deviation range is then identified for the two maximum vectors of two classes. The deviation vector, and obtained from the previous step is employed here to generate the decision rules that specified the risk level of heart patients in terms of numerical variables. The rules are generated automatically from the two deviation vectors that contain the deviation of each attributes comparing two classes. From the equal size deviation vector, three decision rules are generated from every element by comparing the corresponding elements of both vectors. For each rule generated, the number of patients which satisfy these rules is found then the weightage of the rule is calculated. The decision rules obtained contain IF and THEN part, in which IF part specifies the numerical variable and THEN part specifies the class label. At first, the numerical variable specified in the IF part of the decision rules is converted into the linguistic variable according to the fuzzy membership function and THEN part of the fuzzy rules is similar to that of decision rules. A group of fuzzy IF-THEN rules is thus obtained. These fuzzy rules is used as the input for the defuzzification process and membership functions based mapping of fuzzy sets to a crisp output is used to obtain a single number as the output.

Each input fuzzy set defined in the fuzzy system includes four membership functions (VL, L, M and H) and an output fuzzy set contains two membership functions (L and H).The experimentation is carried out on the proposed system using the datasets obtained from the UCI repository and the performance of this system has significantly improved the risk prediction as compared with the neural network-based clinical support system.

III. ANALYSIS AND COMPARISON

The various models discussed use different data mining techniques. Some use only single technique while others use multiple or hybrid data mining techniques. There are some other systems that are based on ensemble methods in which more than one classification techniques are used then results are combined. The models using multiple techniques or hybrid models have a better performance and are more stable as compared to the single technique models. The ensemble technique models show improved accuracy over other types of models as they have an ability to reduce errors [19]. Ensemble models combine multiple models to achieve better prediction accuracy than any of the individual models. The idea behind ensemble technique is that if one technique makes an error, others could correct that. Research also shows that complex and sophisticated data mining techniques have better heart disease diagnosis accuracy [20].

Almost all the systems discussed above face problems in learning and training due to poor quality and the inherent properties of the available data hence the effectiveness of the system is greatly reduced.

The MLP based system discussed is a single technique model which offer a number of advantages, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms but at the same time it suffers from its "black box" nature, greater computational burden, proneness to overfitting, and the empirical nature of model development. The AptaCDSS has poor generalization ability because it is highly dependent on the dataset which is quite small and limited in diversity because of which AptaCDSS had biased decision and classification result. But this system is cost effective though it has a risk of overfitting classifiers due to extra learning. Coactive neuro fuzzy is a suitable method to find the nonlinear relationships between attributes. ANN makes learning easy and has good generalization ability. Genetic algorithm makes the ANN less complicated. Hence the overall model has high accuracy and reliability than other models.

But the ANN and Fuzzy implementation again have low understandability. IHDPS use Decision Tree, ANN and NB techniques but the results are not combined to form a final decision from these three classifiers. The advantage is that it answers complex queries that other systems were not able to answer. This system not complex and can be extended to add new queries. But it is highly dependable on unassociated nature attributes because of use of NB classifier and thus could give a biased result if

TABLE I: SUMMARY OF COMPARISON OF CDSS FOR HEART DISEASE PREDECTION

CDSS System	Type of heart Disease	Data Mining Technique	Accuracy	Reliability	Generalization Ability	Ease of Interpretation	Overall Cost	Rule Extraction
MLP	All Heart Diseases	ANN-MLP,BP	High	Medium	Medium	Low	Low	Nil
CANFIS-GA	Four types of Heart Disease	Neuro-Fuzzy; ANN; Genetic algorithm	High	Medium	High	Low	Medium	Nil
AptaCDSS-E	CVD	SVM, ANN, DT, BN	High	Low	Low	High	Medium	Nil
IHDPS	Heart Disease in general	DT,ANN, NB	Medium	Medium	Poor	High	Low	Medium
IEHAPS	Risk of Heart Attack	K-means, ANN, MAFIA	High	Low	Low	High	High	High
DT-Fuzzy	Risk Level of Heart Disease	DT, Fuzzy	Medium	Low	Low	High	Medium	High

dependencies exist. Other models like AptaCDSS do consider this dependency requirement. IEHAPS has high prediction accuracy but unfortunately it has poor generalization ability and low reliability.

The Decision tree Fuzzy System the fuzzy system is constructed in accordance with the weighted fuzzy rules and chosen attributes. It has an automated system of generating fuzzy rules and is more accurate then ANN implementation. The Fuzzy approach reduces the uncertainty due to decision tree but the disadvantage is that the system has very limited classes and thus it incorporates vagueness.

The summary of the overall comparisons of the models discussed above is provided in Table 1 based on certain key parameters which are Accuracy, Reliability, Generalization Ability, and Ease of Interpretation, Overall Cost, Rule Extraction. These all models perform prediction and recognition of different heart diseases but all have certain strengths and weaknesses due to the different techniques used and the data set applied. But almost all the models lack in generalization capability which is due to unknown data, limited number of training sets and risk of over training. So we have to insure the quality as well as the completeness of medical data to be able to build efficient decision support systems.

Another major concern is that no system addresses the issue of prescription of a suitable treatment option. There is considerable success in the prediction and diagnosis of disease using data mining techniques but if could use these techniques to successfully prescribe a treatment plan then it would not only save cost but also but reduce unnecessary burden of testing and any rule out human errors.

IV. CONCLUSION

With the help of this study we can conclude that in spite of having a large amount of medical data, it lacks in the quality and the completeness of data because of which highly sophisticated data mining techniques are required to build up a efficient decision support system. Even then the overall reliability and generalization capability is still in question. We have to build systems which not only

are accurate and reliable but reduce cost of treatment and increase patient care. At the same time we have to build systems which are understandable and which could enhance human decisions.

Another major concern which was observed was there insignificant study for proposing treatment plans for patients. Though there is some research but we need a significant amount of study dedicated to this because data mining techniques have shown significant success in prediction and diagnosis of diseases and especially heart diseases, hence we could hopefully use these techniques

REFRENCES

1. Hall, J., First, make no mistakes. The New York Times. (2009).
2. SoRelle, R., Reducing the rate of medical errors in the United States. (2000).
3. Patient Safety in American Hospitals Study Survey by HealthGrades, 2004.
4. CDC's report, http://www.cdc.gov/nccdphp/overview.htm.
5. Yan, H.-M., Jiang, Y.-T., Zheng, J., Peng, C.-L., & Li, Q.-H. A multilayer perceptron-based medical decision support system for heart disease diagnosis. (2006)
6. http://astrosun.tn.cornell.edu/staff/loredo/bayes/
7. Miller, RA. Medical Diagnosis Decision Support Systems-Past, Present, and Future. JAMIA. 1994:1; 8-27.
8. Das, R., I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.
9. Srinivas, K., B.K. Rani, and A. Govrdhan, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (IJCSE), 2010. Vol. 02, No. 02: p. 250-255.
10. Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.
11. Garg, A.X., et al., Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: a Systematic Review. Journal of the American Medical Association, 2005. 293: p. 1223–1238.
12. "Heart diseases" from http://en.wikipedia.org/wiki/Heart_disease [13] H. Yan et al., "A multilayer perceptron-based medical decision support system for heart disease diagnosis", Expert Systems with
13. Applications 30 (2006) 272–281
14. Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", World Academy of Science, Engineering and Technology 5 2007
15. Hong, J. Kim, S. & Zhang, B., "AptaCDSS-E: A classifier ensemble based clinical decision support system for cardiovascular disease level prediction," Expert Syst. Appl. Vol 34, No. 1, 2008.
16. Awang, R. & Palaniappan, S., "Intelligent heart disease predication system using data mining technique". IJCSNS International Journal of Computer Science and Network Security. Vol. 8, No. 8, 2008.
17. Patil, S. & Kumaraswamy, Y., "Intelligent and effective heart attack prediction system using data mining and artificial neural network, "European Journal of Science Research. Vol 31, No. 4.2009.
18. P.K. Anooj, "Clinical Decision Support System : Risk Level Prediction of Heart Disease Using Decision Tree Fuzzy Rules" , IJRRCS, Vol. 3, No. 3, pp. 1659-1667, June 2012
19. L. Rokach, "Ensemble-based classifiers," Artif. Intell. Rev., vol.33, pp. 1– 39, 2010.
20. Shouman, M., T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients". 9th Australasian Data Mining Conference 2011. 121.
21. Razali, A.M. and S. Ali, Generating Treatment Plan in Medicine: A Data Mining Approach. American Journal of Applied Sciences, 2009. 6 (2): 345-351.
22. Saad Ali, S.N., et al., Developing treatment plan support in outpatient health care delivery with decision trees technique. Springer-Verlag Berlin Heidelberg, 2010. Part II, LNCS 6441, pp.475–482.