BAUSTJ
BAUST Journal

A Comparison Of Accuracy On Clinical Decision Support System Using Different Classifiers

Fatema Tuz Zohra[1*], Md. Mamunur Rashid[2] and Ashrafun Zannat[3]

[*1,2,3] Department of CSE, Bangladesh Army University of Science and Technology, Saidpur, Bangladesh

emails: *[*1]fatema@baust.edu.bd; [2]dmmr@baust.edu.bd; spzannat@baust.edu.bd ('*' - corresponding author)*

ARTICLE INFO

Article History:

Received: 10th September 2019
Revised: 25th July 2020
Accepted:27th July 2020
Published online: 25th August 2020

Keywords:

Clinical Decision Support System (CDSS)Classifier Instance Attribute

ABSTRACT

This paper presents a survey on accuracy of prediction based on different data mining classifications for Clinical Decision Support System (CDSS) for different diseases. Classifiers are very useful in different CDSS for analyzing diseases. There are several classification techniques that can be applied to build a CDSS. CDSS can help patients specially, rural patients know about diseases at an early stage detection. CDSS can analyze data and predict almost accurate class of predicted outcome. CDSS is being widely used in digital healthcare sector for initial disease detection reducing the hassel of a long waiting queue to get specified physician.

1. INTRODUCTION

Data mining is used to extract useful information from large datasets. Data mining classifiers can easily analyze large datasets and predict outcome which is based on a set of criteria through creating a pattern. Peoples are affected by diseases (heart, breast cancer, diabetes, liver disorder, lung cancer etc) at early ages now for their food habits, lifestyle, obesity, lack of exercise, stress and many more. Heart disease is the number one cause of death globally. 17.6 million deaths were caused due to heart disease in 2016 but within 2030 it can be more than about 23.6 million [American Herat Association, 2019]. Of the total death in Bangladesh diseases are: cancer-25%, heart-6%, diabetes-3% and liver-3% [wordatlas]. First leading cause of death in our country is cancer among which breast cancer is leading for women death. In every year 1.08 lakh die and 1.5 lakh develop cancer [newagebd]. Diabetes is the most growing disease now-a-days. About 7.1 million people have diabetes and almost equal number with undetected diabetes. This will be double in 2025 [A.K Mohiuddin et. al, 2019]. More importantly it leads to other diseases like heart attack, stroke, kidney disease etc. About 1.3 million people dies from liver diseases per year in our country. According to National Liver Foundation about 60-70% of the infected individuals are not aware about the existence of liver disease. Infants get easily affected by liver disease [dhakatribune]. If people get to know about the probable disease and their basic solution at early stage then the impact of disease can be reduced greatly. It can be done by CDSS. The main idea to raise awareness among patients. Besides it helps physicians to take decision.

2. DATA MINING

Data mining is the application of a combination of Artificial Intelligence, pattern recognition, statistics and database systems. In traditional data mining frequent patterns are discovered to analyze from large collection of data to get useful result. But in medical data mining each of minor features should be counted with common pattern for accurate prediction.

3. DATA MINING CLASSIFICATION

It is a data analysis task. It is a process of finding a model which describes and distinguishes data classes and concepts. Classifiers are of two types 1) descriptive, and 2) predictive. Here we will use predictive classifier. There are two steps of classification. Firstly, the datasets are trained and then testing is done on those trained data. There are many classification techniques available. But the same technique doesn't give the same result for different datasets. We will compare different techniques on different models of datasets below.

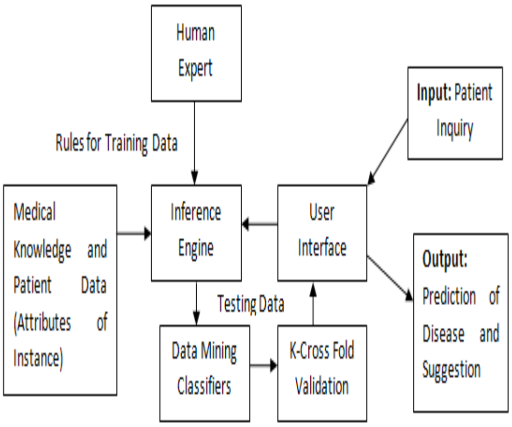

Fig. 1: Basic architecture of CDSS.

A. NaiveBayes(NB):

Naïve Bayes classification is the simplest classification technique based on Bayes theorem and it is fast. But in complex system with large dependency sometimes it does not give accurate result [S. Joshi et. al, 2018]. It can be done through calculating the probability for each class, assuming conditional independence of the attributes of class. The NB technique has performed remarkably in medical diagnosis and system performance measurement [Y. Kumar et. al, 2013]. Laryngeal cancer based CDSS is developed on the basis of Bayesian Network having 1000 variables with about 1300 dependencies. A subsystem of 303 variables reached 100% correct predictions [M.A Crypko et. al, 2019]. NaïveBayesUpdateable is the update version of Naïve Bayes. It is used to classify here. According to Bayes theorem of probability:

$$P\ (C_k|\{A_1,A_2,\ldots A_n\}) = [\ \prod^{n}_{i=1} P(\{A_i\}|C_k)*P(C_k)]/P\ (\{A_1,A_2,\ldots A_n\ \})$$

All possible attributes $A_1,A_2,\ldots A_n$ are class conditionally independent and this refers to as Naïve Bayes.

B. Artificial Neural Network (ANN):

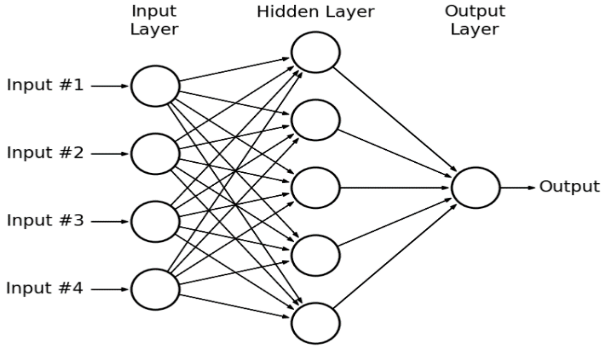

Fig. 2: Architecture of MLP

It is used in adaptive learning and fast prediction. It can learn in the presence of noise. But the processing time is comparatively high of ANN. MultilayerPerceptron (MLP), is an example of ANN. Each ANN has input, hidden and output layer. Fig. 2 is the architecture of MLP. It allows the system to learn from existing knowledge and experiences. ANN is used to predict fetal delivery to either has to be done in normal or by surgery [R.R. Janghel et al, 2009]. ANN is used to develop the prediction and diagnosis of Celiac disease with 84.2% accuracy [J.M Tenorio et. al, 2011].

C. K-nearest neighbor (K-NN):

It is very simple and easy to implement. But it requires large storage space and gets easily affected by irrelevant attributes. In the case of prediction, it finds the closest training point to unknown point measured in distance vector. A diagnostic software tool is developed to obtain correct diagnosis of Skin disease using K-NN [S. Joshi et. al, 2018].

D. Support Vector Machine (SVM):

SVM classifies by finding a separating boundary called hyper-plane. Sequential minimal optimization (SMO) is an optimization technique of SVM used to train a dataset. SVM model has highest accuracy of 77.63% in prediction of heart disease [Y.J Son et. al, 2010]. Two models were analyzed with different number of attributes. SVM got higher accuracy than Radial basis function network and Naïve Bayes in heart, breast cancer and diabetes on the dataset of UCI [P. Jannardhanan et. al, 2015]. Fig. 3 is the architecture of SMO.

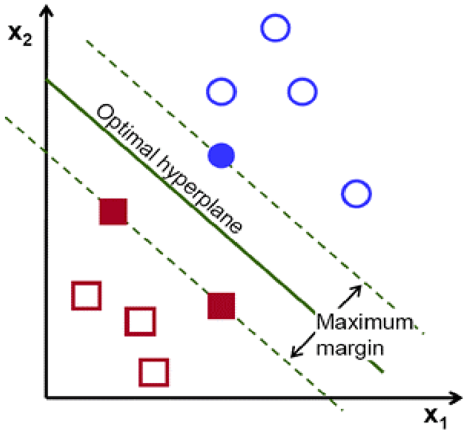

Fig. 3: Architecture of SMO

E. Decision Tree (DT):

The output of J48 is decision trees. These tree also have root nodes, intermediate nodes and leaf nodes. Each node leads to a decision and that helps CDSS in the case of clinical DSS. It can easily predict in Boolean class format. Decision tree can be visualized. Decision tree is most accurate for Chronic Renal Failure [Dr. N. Bhargava et. al, 2013]. Fig. 4 is the sample architecture of a decision tree whether a 1 is red or blue and then the red one's are underlined or not [towardsdatascience].

RandomForest consists of a large number of individual decision trees which operate as an ensemble. Each tree results in a class prediction and the class with most occurring becomes final model's prediction. It can also thought as a nearest neighbor predictor. It's prediction is accurate more than any individual tree. In Fig.5 sample architecture of RandomForest is shown [towardsdatascience].

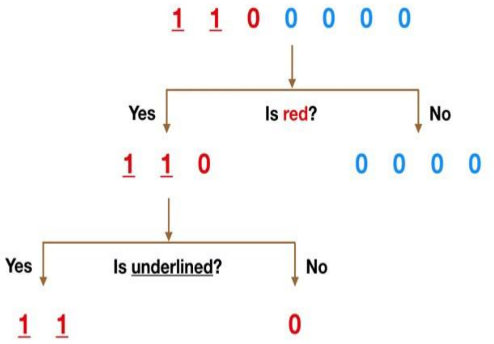

Fig. 4: Architecture of J48

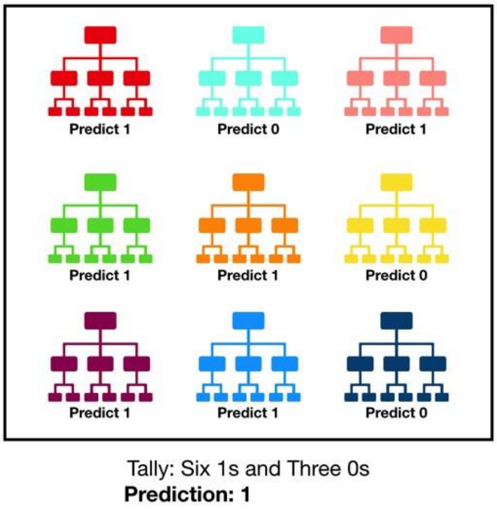

Fig. 5: Architecture of RandomForest (RF)

F. Logistic Regression (LR):

It is a powerful statistical method to analyze containing one or more independent variables determining an outcome. It uses a logistic function to estimate probabilities. It is used to identify risk factors associated with Type 2 Diabetes Risk facors [S. Joshi et. al, 2018]. It is a fast and simple technique. Simple logistic regression (SL) is used here. It can be used for prediction of occurrence or non-occurrence of an event.

Classifiers	Algorithm
Bayesian Classifier	NaiveBayesUpdateable (NB)
Decision Tree (DT)	J48, RandomForest (RF)
Support Vector Machine(SVM)	Sequential Minimal Optimization (SMO)
Artificial Neural Network (ANN)	MultilayerPerceptron (MLP)
K-nearest neighbors (K-NN)	K-Star
Logistic Regression	SimpleLogistic(SL)

Table 1: The applied algorithms of different classifiers in this paper.

4. TOOLS

There are many types of data mining tools can be used such as Weka, R, RapidMiner, MATLAB. Waikato Environment for Knowledge Analysis (WEKA) 3.8.3 is used for experiment of different classification techniques on different datasets. It is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-supported for developing a new machine learning algorithm.

5. EXPERIMENTAL ANALYSIS

We have used the datasets of machine learning laboratory at University of California, Irvine. The presence of the disease is identified through this classification.

Cross Validation is used to test the effectiveness of a system. It evaluates the accuracy of predictive models. Here we used 10 fold and 5 fold cross validation model to analysis for each classifier. K fold cross validation indicates a method which divide a dataset into K subsamples and first K-1 subsamples is used for training and the final rest subsample is for testing.

Attributes can be selected with AttributeEvaluator in WEKA. At the same time Ranker can be chosen as search method to rank the attributes for selection of high ranked attributes.

Accuracy of a model is measured by the percentage of correctly classified instances.

A. Breast Cancer Data [M. Zwitter et. al, 1988]

Number of Instances: 286

Attributes: 10

5 fold CV gives better result except J48 in Fig.6. MLP takes time long and poor result. Simple Logistic classifier with 75.87% accuracy gives best result with 5 fold CV which is better than previous work with decision tree [C. Vaghela et. al, 2015].

B. Diabetes Data [V. Sigillito et. al, 1990]

Number of Instances: 768

Attributes: 9

Simple Logistic classifier with 77.86% accuracy gives best result with 5 fold CV which is better than previous work with Naïve Bayes [C. Vaghela et. al, 2015] in Fig. 7.

C. Liver Disorder [R.S Forsyth et. al, 1990]

Number of Instances: 345

Attributes: 7

RandomForest gives the best result than others with 74.20% accuracy than Neural Network with 73.91% [P. Rajeswari et. al, 2010] in Fig. 8

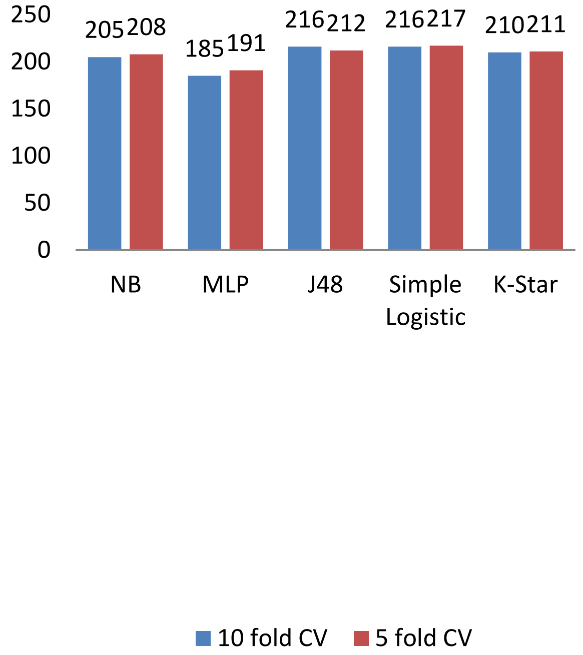

Fig. 6: Correctly classified number of instances of Breast Cancer dataset

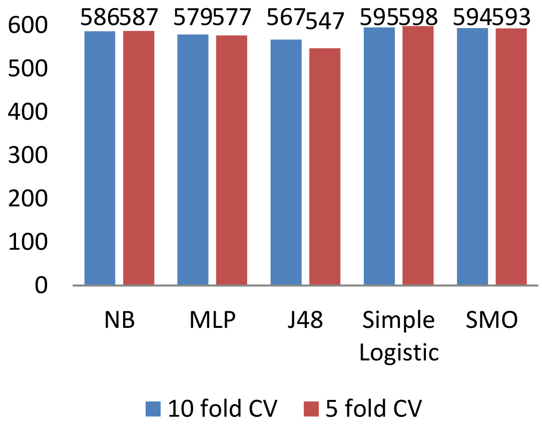

Fig. 7: Correctly classified number of instances of Diabetes dataset

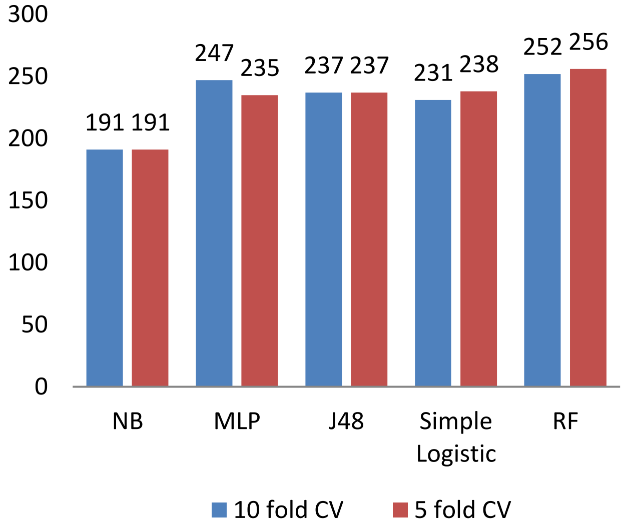

Fig. 8: Correctly classified number of instances of Liver Disorder dataset

6. CONCLUSIONS (COPYRIGHT AND LICENSE)

CDSS is a framework that uses patient information with medical knowledge to derive models so that it can predict the possible disease and give patients instant decision based on predictions. Besides it can assist health professionals while taking decision. It can be modified as e-health monitoring system. It also can be modified in any specific hospital usage. So, it has a great impact in healthcare system. The accuracy should be high as much as possible. Accuracy is measured using different classifiers here. From this survey Simple Logistic algorithm with 5 fold cross validation gives higher accuracy for Breast cancer and diabetes dataset while RandomForest algorithm with 5 fold cross validation for liver disorder dataset gives higher accuracy.

REFERENCES

Heart Disease and Stroke Statistics-2019 At a Glance, 2019 American Heart Association.

https://www.worldatlas.com/articles/the-10-leading-causes-of-death-in-bangladesh.html.

http://www.newagebd.net/article/50603/108-lakh-people-die-of-cancer-in-bangladesh-a-year.

A K Mohiuddin, “Internationale Journal of Diabetes Research,” Vol 2, No 1, 24 February 2019, pp 14-20.

https://www.dhakatribune.com/bangladesh/2018/09/21/unicef-under-5-mortality-rate-falls-sharply-in-bangladesh.

S. Joshi and M. K. Nair, “Survey of Classification Based Prediction Techniques in Healthcaere” in Indian Journal of Science and Technology, Vol 11(15), April 2018, DOI: 10.17485/ijst/2018/v11i15/121111.

Y. Kumar and G. Sahoo, “Prediction of different types of liver diseases using rule based classification model, ” Technology and Health Care 21(2013), pp417-432, 2013, DOI: 10.3233/THC-10742.

M. A. Crypko and M. Stoehr, “Digital patient models based on Bayesian networks for clinical treatment decision support,” Minimally Invasive Therapy and Allied Technologies, 27 February,2019. DOI: 10.1080/13645706.2019.1584572.

R. R. Janghel, Anupam shukla and Ritu Tiwari, IEEE 2009, “Clinical Decision Support System for Fetal Delivery using artificial neural network”.

J. M. Tenorio, .D. Hummel, F. M. Cohrs, V. L. Sdepanian, I. T. Pisa and Heimar, “Artifical intelligence technique applied to the development of a decision-support system for diagnosing celiac disease,” Int J Med Inform, 2011 november, 80(11):793-802, DOI: 10.1016/j.ijmedinf.2011.08.001.

Y. J Son, H.G. Kim, E.H. Kim, S.Choi, S.K.Lee,“Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients,” Healthcare Informatics Research, 16(4):253-259, December 2010.

P. Jannardhanan, Heena L. and F. Sabika, “ Effectiveness of Support Vector Machines in Medical Data mining,” JOURNAL OF COMMUNICATIONS SOFTWARE AND SYSTEMS, VOL11,NO 1, MARCH 2015.

Dr. N. Bhargava, G. Sharma, Dr. R. Bhargava, M. Mathuria, “ International Journal of Advanced Researchin Computer Science and Software Engineering,” Volume 3, Issue 6, June 2013.

https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

M. Zwitter and M. Soklic, UCI Machine Learning Repository [Internet]; 1988. Avaiable from: https://UCI_breast-cancer.arff%20_%20TunedIT.html.

C. Vaghela, N. Bhatt, D. Mistry, “A Survey on Various Classification Techniques for Clinical Decision Support System,” International Journal of Computer Applications(0975-8887), Volume 116-No 23, April 2015.

V. Sigillito, Pima Indians Diabetes Dataset from National Institue of Diabetes and Digestive and Kindney Diseases, 1990. Available from https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff.

R. S. Forsyth, UCI Machine Learning Repository [Internet]; 1990. Avaiable from: https://UCI_liver-disorders.arff%20_%20TunedIT.html.

P. Rajeswari, G. S. Reena, “Analysis of Liver Disorder Using Data Mining Algorithm,” Global Journal of Computer Science and Technology, Vol 10, Issue 14 (Version 1.0), November 2010.