LIVER DISEASE PREDICTION BY USING DIFFERENT DECISION TREE TECHNIQUES

Nazmun Nahar[1] and Ferdous Ara[2]

[1]Department of Computer Science & Engineering, University of Chittagong
[2] Lecturer, Department of Computer Science & Engineering
BGC Trust University Bangladesh

ABSTRACT

Early prediction of liver disease is very important to save human life and take proper steps to control the disease. Decision Tree algorithms have been successfully applied in various fields especially in medical science. This research work explores the early prediction of liver disease using various decision tree techniques. The liver disease dataset which is select for this study is consisting of attributes like total bilirubin, direct bilirubin, age, gender, total proteins, albumin and globulin ratio. The main purpose of this work is to calculate the performance of various decision tree techniques and compare their performance. The decision tree techniques used in this study are J48, LMT, Random Forest, Random tree, REPTree, Decision Stump, and Hoeffding Tree. The analysis proves that Decision Stump provides the highest accuracy than other techniques.

KEYWORDS

Data Mining, Decision Tree, Liver Disease

1. INTRODUCTION

The liver plays an important role in many bodily functions from protein production and blood clotting to cholesterol, glucose (sugar), and iron metabolism. It has a range of functions, including removing toxins from the body, and is crucial to survival. The loss of those functions can cause significant damage to the body. When liver is infected with a virus, injured by chemicals, or under attack from own immune system, the basic danger is the same – that liver will become so damaged that it can no longer work to keep a person alive. Liver disease caused by hepatotrophic viruses imposes a substantial burden on health care resources. Persistent infections from hepatitis B virus (HBV), hepatitis C virus, and hepatitis delta virus result in chronic liver disease. The most basic classification of liver disease is as acute and chronic. The definition of acute liver disease is based on duration, with the history of the disease does not exceed six months. Acute viral hepatitis and drug reactions account for the majority of cases of acute liver disease.

Liver disease is also referred to as hepatic disease. Usually nausea, vomiting, right upper quadrant abdominal pain, fatigue and weakness are classic symptoms of liver disease. Symptoms of liver patient include jaundice, abdominal pain, fatigue, nausea, vomiting, back pain, abdominal swelling, weight loss, fluid in abnormal cavity, general itching, pale stool, enlarged spleen and gallbladder [1]. Symptoms of liver disease can vary, but they often include swelling of the abdomen and legs, bruising easily, changes in the colour of your stool and urine, and jaundice, or yellowing of the skin and eyes. Sometimes there are no symptoms. Tests such as imaging tests and liver function tests can check for liver damage and help to diagnose liver diseases.

DOI: 10.5121/ijdkp.2018.8201

The purpose of this study is to compare the decision tree algorithms such as J48, LMT, Random Tree, Random Forest, REPTree, Decision Stump and Hoeffding Tree in diagnosis liver disease. The liver dataset are analyzed using above decision tree algorithms and compare their performance with respect to seven performance metrics (ACC%, MAE, PRE, REC, FME, Kappa Statistics and runtime). The paper is organized as follows: In Section 2, the Related Works is presented. In Section 3, Methodology used in this paper is given. Experiments and result is presented in Section 4. The paper is concluded in Section 5.

2 RELATED WORKS

Machine learning has attracted a huge amount of researches and has been applied in various fields in the world. In medicine, machine learning has proved its power in which it has been employed to solve many emergency problems such as cancer treatment, heart disease, dengue fever diagnosis and so on. Among several outstanding methods, Decision Tree algorithms have been employed for many researches.

Liver disease of the patients has been continuously increasing because of inhale of harmful gases, intake of contaminated food, different kinds of drugs and excessive consumption of alcohol. Automatic classification tools may reduce burden on doctors [2]. The classification algorithms based on classification of some liver patient datasets. For the algorithm he considered Naïve Bayes classifier, C4.5, Back propagation Neural Network algorithm, and Support Vector Machines which evaluated based on four criteria: Accuracy, Precision, Sensitivity and Specificity. On the other hand, Aneeshkumar [3] used a methodology to effective classification of liver and non-liver disease dataset. Pre-processing method is used to cleansing the data for effective classification, after cleansing the data. 15 attributes of real medical data are collected from dataset. C4.5 and Naive Bayes are the two algorithms used in his study. He divided datasets into three different types of ratio based on average and standard deviation of each factor of both class and evaluated the accuracy. The result in his study after evaluate the accuracy, he said C4.5 is gives better accuracy than Naive Bayes, because it gives more accuracy with the minimum time taken. Naïve Bayes is sometimes better than FT growth algorithm with the use of machine learning for detection of liver disease [4]. He compared among 29 datasets with 12 different attributes. By comparing two decision tree algorithms which are FT growth and Naïve Bayes and found that Naïve Bayes is better than FT growth algorithm with the use of machine learning because, Naïve Bayes (75.54%) gives more accuracy than FT growth algorithm (72.66) using WEKA Tool. Whereas, in comparison of Ft tree Naïve Baiyes and Kstar to predict the liver disease disorder with evaluate using 10-fold cross validation, Rajeswari [5], identified FT tree gives better role for increasing the accuracy of the dataset in classification technique algorithm. The computer aided diagnosis (CAD) system consists of the segmentation of liver and lesion, extraction of features from alesion and characterization of liver diseases by means of a classifier.

In an experiment Gunasundari [6] found conversional image processing operations, neural networks and Genetic algorithm gives successful result for liver disease disorder diagnosis. In future liver disease disorder diagnosis extended in many directions. Such as using effective algorithms and more texture feature technique algorithms. CART uses a purity-based measure, and the algorithm splits the training data set based on how probably the subsets become purer for a class, and it spends more time to generate smaller trees. CART and C4.5 both algorithms are gives good result with oversampling for liver disease disorder dataset. These algorithms could reduce the minor class increment to smaller percentage [7]. Decision tree algorithm does not give high priority for minor classes for that reason using duplication in BUPA liver disease disorder

dataset, increase the number of instances of minor class and proceed with two decision tree algorithms and both algorithms gives good result in insufficiency of liver disease disorder data. Case Based Reasoning (CBR) and Classification and Regression Tree (CART) techniques could be useful to detect the liver disease [8]. Feature selection plays a vital role in text categorization. A range of different methods have been developed, each having unique properties and selecting different features. We show some results of an extensive study of feature selection approaches using a wide range of combination methods. Bendi [9], proposed a Modified Rotation Forest algorithm to calculate the accuracy of the liver classification techniques in UCI liver dataset using the combo of feature selection technique and selected classification technique algorithm. Over the past few years, the increasing attention on severe challenges in medical diagnosis process such as sharply increased elderly patients, limited medical personnel, has led to a number of contributions in the areas of the intelligent medical diagnosis methods. The early contributions can be found on the neural networks, it provides a new significant way for intelligent medical diagnosis. A model proposed by Kiruba [10] on intelligent agent based system to hike a precise and accurate of diagnosis system. C4.5 decision tree algorithm and Random tree algorithm are used to predict. Two different types of liver disease disorder dataset are combined and predict the accuracy of the disease. And then conclude these both algorithms gives very good accuracy for diagnosing liver disease disorder. Liver abscess is the commonest cause of hepatomegaly and it is due to amoebiasis, followed by fatty liver, congestive cardic failure, hepatocellular carcinoma, and viral hepatitis seen only in few patients [11].

3. METHODOLOGY

Main objective of this study is to identify that whether the patient has liver disease or not. Some of the parameter are used for predicting the liver disease and compare the performance of the various decision tree techniques. Weka is a data mining tool which is written in java and developed at Waikato. WEKA is a very efficient data mining tool to classify the accuracy by applying different algorithmic approaches and compare on the basis of datasets [12].It is also a good tool for build new machine learning schemes. The result found from the liver disease dataset by using Weka tool are in section 4.10-fold cross validation performed on the dataset.

The objective of this study is liver disease prediction using data mining tool. The main task in this study is:

- Various decision tree techniques are used for the Prediction of the liver disease.
- Comparing different decision tree techniques.
- Finding best decision tree for the liver disease prediction.

A. DECISION TREE

Data mining is a process where intelligent methods are used to find out data patterns. It is an important process of discovering pattern and knowledge from large volume of data. Now a day, the usefulness of the methods has been proven in medical field by trying different algorithms. One of the algorithm is data classification is the process of finding a model that can explain different data classes. Some classification algorithms are Decision tree, Support vector Machine, K-NN, Neural networks , Association rule, Bayesian networks, etc. However, for this work decision tree is used because it provides the more accurate results than other algorithms. The large datasets easily classified by the decision tree which is easy to understand by the human. The structure of the Decision tree is looks like a tree structure. Decision tree is made of a root, leaf

nodes and internal nodes. Seven decision tree techniques have been used in this study. They are J48, LMT, Random Forest, Random tree, REPTree, Decision Stump, and Hoeffding. Their performance was analyzed using Accuracy(ACC),Precision(PRE),Recall(REC),Mean Absolute Error(MAE),F-Measure(FME),Kappa Statistic and Run time. The descriptions of decision tree technique that are used in this study are given below:

J48:

J48 is advance version of C4.5. The technique of this algorithm is to use divide-and-conquer method. It uses pruning method to construct tree. It is a common method which is used in information gain or entropy measure. Thus it is like tree structure with root node, intermediate and leaf nodes. Node holds the decision and helps to acquire the result.

REP Tree:

REP Tree is a fast decision tree learner. Builds a decision/regression tree using entropy as impurity measure and prunes it using reduced-error pruning. It only sorts values for numeric attributes once.

Random Tree:

Random Tree is a group learning algorithm that creates many individual learners. It is an algorithm for build a tree that treats K random features at each node. It involves a bagging idea to create a random set of data for building a decision tree. For building a standard tree each node is split using the best split among all variables.

Decision Stump:

Decision stumps are basically decision trees with a single label. A stump is opposed to a tree which has multiple layers. It basically stops after the first split. Decision stumps are usually used in large data. Hardly, they also help to make simple yes/no decision model for smaller dataset.

LMT:

LMT means logistic model tree. LMT is a classification model with an associated supervised training algorithm. It combines decision tree learning and logistic prediction. Logistic model trees use a decision tree that has linear regression models at its leaves to provide a section wise linear regression model.

Random Forest:

Random forests are a group learning method for regression, classification and other works that operate by building a multitude of decision trees at training period and outputting the class that is the mode of the classification or mean prediction of the individual trees. Random forests average multiple deep decision trees, which are trained on various parts of the same training set, with the aim of minimizing the variance.

Hoeffding Tree

Hoeffding Tree is known as the streaming decision tree induction. The name is derived from the Hoeffding bound that is used in the tree induction. The basic idea is, Hoeffding bound provides

particular level of confidence on the best attribute to split the tree, hence we can construct the model based on particular number of instances that has used.

B. Dataset

The data are collected from UCI Machine Learning Repository [13] and it predicts liver disease based on the given attributes. The data set has eleven attributes which predict the liver disease. The attributes description is given below. Based on data types the attributes are given. The data set is built on both numerical and nominal data types. In our study the dataset contains the attribute such as total bilirubin, direct bilirubin, age, gender, total proteins, albumin, albumin and globulin ratio which is the symptoms of liver disease. The data sets consist of 538 liver and non-liver instances. The Dataset used in the study consist of 167 negative tested for liver disease and 416 are positively tested. Class value "Yes" means having liver disease and "No" means negative liver disease. The entire attribute and their types are given in Table 1.A portion of liver dataset is shown in Fig1 which is loaded in the Weka tool.

Table 1: Attribute Description

Attribute Name	**Possible value**
Age of the patient	Numeric
Gender of the Patient	Nominal
Total Bilirubin	Numeric
Direct Bilirubin	Numeric
Alkphos Alkaline Phospotase	Numeric
Sgpt Alamine Aminotransferase	Numeric
Total proteins	Numeric
Albumin	Numeric
Albumin and Globulin Ratio	Numeric
Class	Nominal

Relation: Indian Liver Patient Dataset (ILPD)

No.	1: Age	2: Gender	3: TB	4: DB	5: Alkphos	6: Sgpt	7: Sgot	8: TP	9: ALB	10: A/G Ratio	11: Class
	Numeric	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	65.0	Female	0.7	0.1	187.0	16.0	18.0	6.8	3.3	0.9	yes
2	62.0	Male	10.9	5.5	699.0	64.0	100.0	7.5	3.2	0.74	yes
3	62.0	Male	7.3	4.1	490.0	60.0	68.0	7.0	3.3	0.89	yes
4	58.0	Male	1.0	0.4	182.0	14.0	20.0	6.8	3.4	1.0	yes
5	72.0	Male	3.9	2.0	195.0	27.0	59.0	7.3	2.4	0.4	yes
6	46.0	Male	1.8	0.7	208.0	19.0	14.0	7.6	4.4	1.3	yes
7	26.0	Female	0.9	0.2	154.0	16.0	12.0	7.0	3.5	1.0	yes
8	29.0	Female	0.9	0.3	202.0	14.0	11.0	6.7	3.6	1.1	yes
9	17.0	Male	0.9	0.3	202.0	22.0	19.0	7.4	4.1	1.2	no
10	55.0	Male	0.7	0.2	290.0	53.0	58.0	6.8	3.4	1.0	yes
11	57.0	Male	0.6	0.1	210.0	51.0	59.0	5.9	2.7	0.8	yes
12	72.0	Male	2.7	1.3	260.0	31.0	56.0	7.4	3.0	0.6	yes
13	64.0	Male	0.9	0.3	310.0	61.0	58.0	7.0	3.4	0.9	no
14	74.0	Female	1.1	0.4	214.0	22.0	30.0	8.1	4.1	1.0	yes
15	61.0	Male	0.7	0.2	145.0	53.0	41.0	5.8	2.7	0.87	yes
16	25.0	Male	0.6	0.1	183.0	91.0	53.0	5.5	2.3	0.7	no
17	38.0	Male	1.8	0.8	342.0	168.0	441.0	7.6	4.4	1.3	yes
18	33.0	Male	1.6	0.5	165.0	15.0	23.0	7.3	3.5	0.92	no
19	40.0	Female	0.9	0.3	293.0	232.0	245.0	6.8	3.1	0.8	yes
20	40.0	Female	0.9	0.3	293.0	232.0	245.0	6.8	3.1	0.8	yes
21	51.0	Male	2.2	1.0	610.0	17.0	28.0	7.3	2.6	0.55	yes
22	51.0	Male	2.9	1.3	482.0	22.0	34.0	7.0	2.4	0.5	yes
23	62.0	Male	6.8	3.0	542.0	116.0	66.0	6.4	3.1	0.9	yes
24	40.0	Male	1.9	1.0	231.0	16.0	55.0	4.3	1.6	0.6	yes
25	63.0	Male	0.9	0.2	194.0	52.0	45.0	6.0	3.9	1.85	no
26	34.0	Male	4.1	2.0	289.0	875.0	731.0	5.0	2.7	1.1	yes
27	34.0	Male	4.1	2.0	289.0	875.0	731.0	5.0	2.7	1.1	no
28	34.0	Male	6.2	3.0	240.0	168...	850.0	7.2	4.0	1.2	yes
29	20.0	Male	1.1	0.5	128.0	20.0	30.0	3.9	1.9	0.95	no
30	84.0	Female	0.7	0.2	188.0	13.0	21.0	6.0	3.2	1.1	no
31	57.0	Male	4.0	1.9	190.0	45.0	111.0	5.2	1.5	0.4	yes
32	52.0	Male	0.9	0.2	156.0	35.0	44.0	4.9	2.9	1.4	yes
33	57.0	Male	1.0	0.3	187.0	19.0	23.0	5.2	2.9	1.2	no
34	38.0	Female	2.6	1.2	410.0	59.0	57.0	5.6	3.0	0.8	no
35	38.0	Female	2.6	1.2	410.0	59.0	57.0	5.6	3.0	0.8	no
36	30.0	Male	1.3	0.4	482.0	102.0	80.0	6.9	3.3	0.9	yes

Fig1: Dataset for Liver Disease prediction

4. EXPERIMENTS AND RESULT

The comparison of various decision tree algorithms performed on liver disease data is shown in Table 2.Decision Stump hast the highest accuracy rate. The accuracy rate of this algorithm is 70.67%.It is seen that Decision Stump is the most powerful classifier for this example. This result shows that the liver disease of a new patient is predicted successfully with an acceptable ratio 70.67%.Random Tree, LMT and Hoeffding Tree has the accuracy rate as 69.47%, 69.30% and 69.75%.It is also seen that J48 has the worst accuracy rate with 65.69%.The comparison of decision tree algorithm with respect to accuracy shown in Fig2. The comparison of decision tree algorithm with respect to kappa statistics and runtime is shown in Fig3 and Fig4 respectively. REPTree, Random Tree and Decision Tree are faster than other algorithms.LMT algorithm takes a long time even though a small dataset is used. A decision tree is shown in Fig5 which is generated by J48 algorithm.

Table 2: Comparing the various decision tree algorithms carried out liver dataset

Techniques	Tree Size	ACC (%)	MAE	PRE	REC	FME	Kappa Statistics	Time
J48	65	65.69	0.3678	0.651	0.657	0.654	0.158	0.11
LMT	1	69.47	0.4116	0.632	0.695	0.628	0.065	0.88
Random Forest		69.30	0.3464	0.667	0.693	0.674	0.186	0.5
Random Tree	267	66.55	0.3382	0.662	0.666	0.663	0.183	0.01
REPTree	27	66.13	0.3800	0.630	0.691	0.629	0.067	0.03
Decision Stump	Single Level	70.67	0.4392	0.499	0.707	0.585	0.379	0.01
HoeffdingTree	1	69.75	0.4091	0.634	0.700	0.619	0.0501	0.12

.

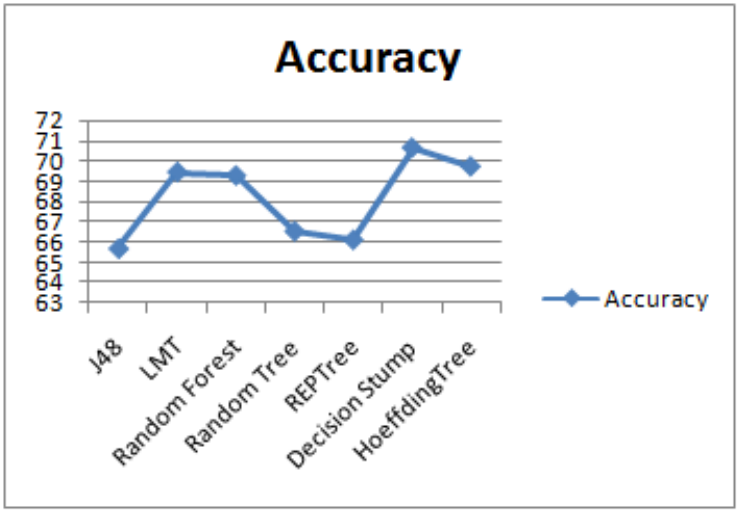

Fig2: Comparison of the decision tree algorithms according to Accuracy

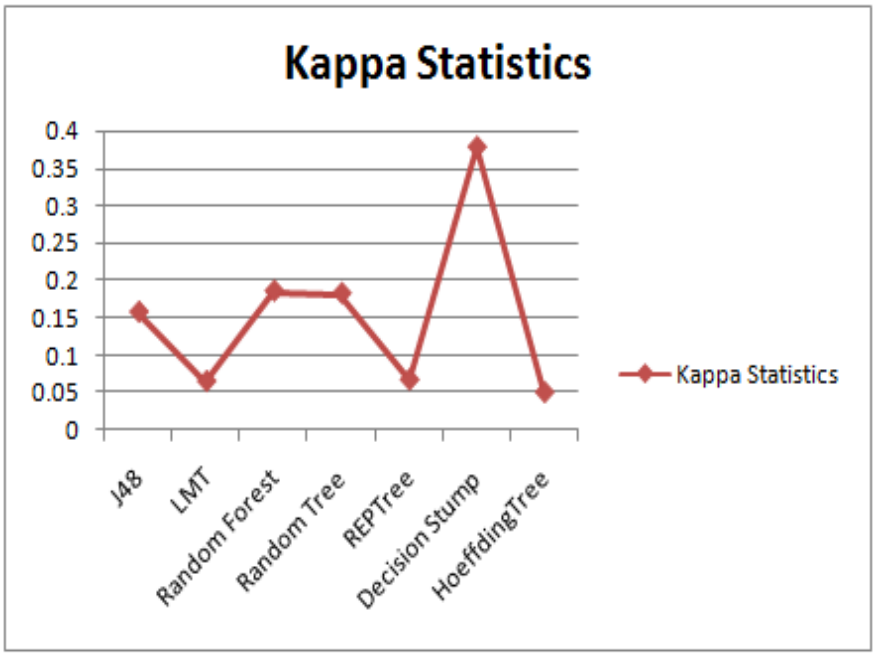

Fig3: Comparison of the decision tree algorithms according to Kappa statistics

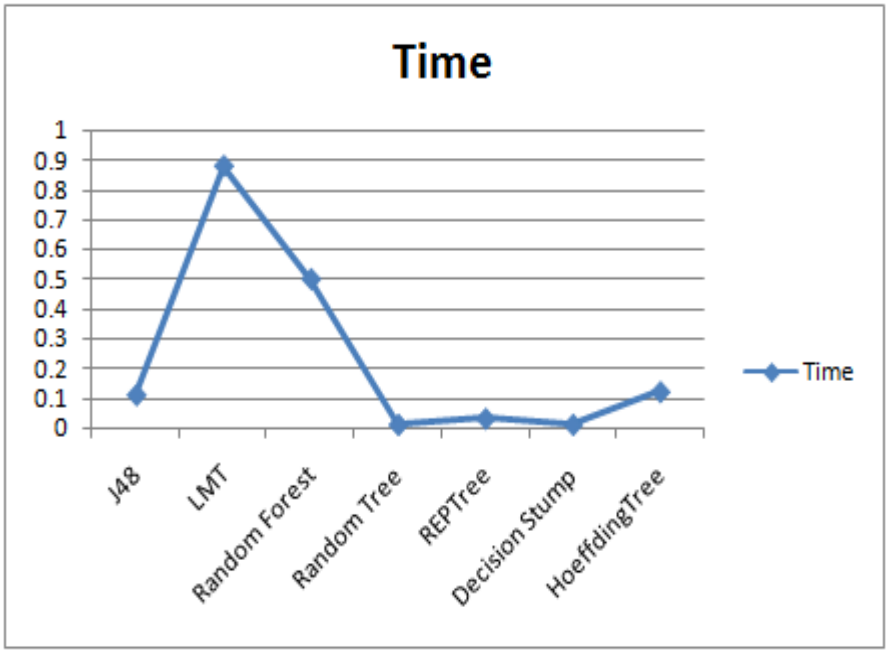

Fig4: Comparison of the decision tree algorithms according to runtime

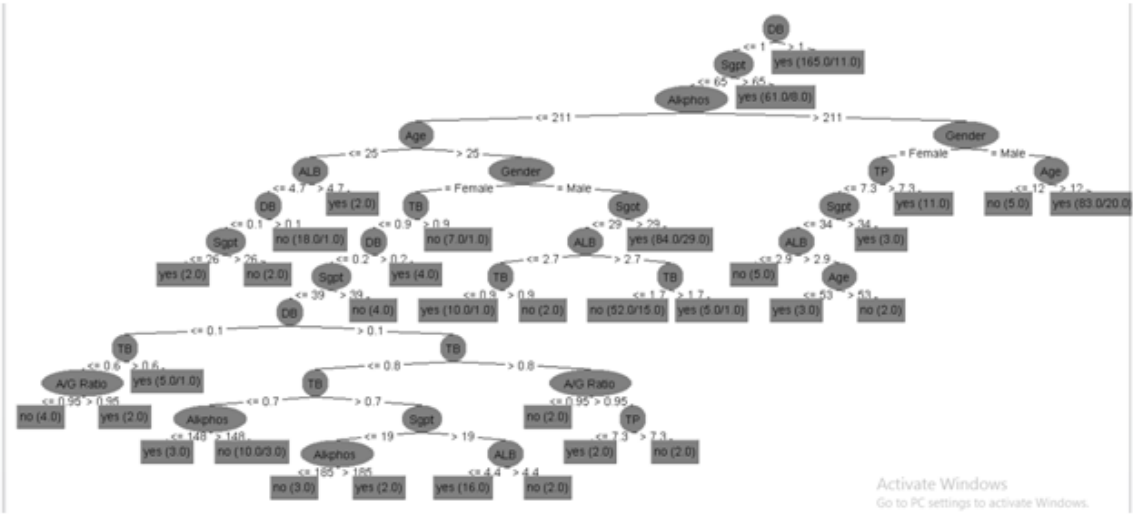

Fig5: Decision Tree Generated by J48

5. CONCLUSIONS

The study employed some decision tree algorithm such as J48, LMT, Random Forest, Random tree, REPTree, Decision Stump and Hoeffding Tree to predict the liver disease at an earlier stage. These algorithm gives various result based on Accuracy, Mean Absolute Error, Precision, Recall, Kappa statistics and Runtime. These techniques were evaluated and their performance was compared. From the analysis, Decision Stump outperforms well than other algorithms and its achieved accuracy is 70.67%. The performance measure used for comparison are listed in the table (Table 2) The application of Decision tree in predicting liver disease will benefit in managing the health of individuals. However, in future, we will collect the very recent data from various regions across the world for liver disease diagnosis. The results of this study will encourage us to continue developing other advanced decision trees such as CART.

REFERENCES

[1] D. Sindhuja and R. J. Priyadarsini, "A survey on classification techniques in data mining for analyzing liver disease disorder", International Journal of Computer Science and Mobile Computing, Vol.5, no.5 (2016), pp. 483-488.

[2] B. V. Ramana, M. R. P. Babu and N.B. Venkaeswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", International Journal of Database Management Systems (IJDMS), Vol.3, no.2, (2011) , pp. 101-114.

[3] A.S.Aneeshkumar and C.J. Venkateswaran, "Estimating the Surveillance of Liver Disorder using Classification Algorithms", International Journal of Computer Applications (0975 –8887) , Vol. 57, no. 6, (2012), pp. 39-42.

[4] S.Dhamodharan, "Liver Disease Prediction Using Bayesian Classification", 4th National Conference on Advanced Computing, Applications & Technologies, Special Issue, May 2014.

[5] P.Rajeswari and G.S. Reena, "Analysis of Liver Disorder Using data mining Algorithms", Global Journal of Computer Science and Technology, Vol.10, no. 14 (2010), PP. 48- 52.

[6] G. Selvara and S. Janakiraman, "A Study of Textural Analysis Methods for the Diagnosis of Liver Disease from Abdominal Computed Tomography", International Journal of Computer Applications (0975-8887), Vol. 74, no.11 (2013), PP.7-13.

[7] H. Sug, " Improving the Prediction Accuracy of Liver Disorder Disease with Oversampling", Applied Mathematics in Electrical and Computer Engineering, American-MATH 12/CEA12 proceedings of the 6th Applications and proceedings on the 2012 American Conference on Appied Mathematics (2012), PP. 331-335.

[8] R.H.Lin, "An Intelligent model for liver disease diagnosis", Artificial Intelligence in Medical, Vol. 47, no. 1 (2009), PP. 53-62.

[9] B. V. Ramanaland and M.S. P. Babu, "Liver Classification Using Modified Rotation Forest", International Journal of Engineering Research and Development ISSN: 2278-067X, Vol. 1, no. 6 (2012), PP.17-24.

[10] H.R. Kiruba and G. T. arasu, "An Intelligent Agent based Framework for Liver Disorder Diagnosis Using Artificial Intelligence Techniques", Journal of Theoretical and Applied Information Technology, Vol. 69 , no.1 (2014), pp. 91-100.

[11] C.K. Ghosk, F. Islam, E. Ahmed, D.K. Ghosh, A. Haque and Q.K. Islam, "Etiological and clinical patterns of Isolated Hepatomegaly" Journal of Hepato-Gastroenterology, vol.2, no. 1, PP. 1-4.

[12] S. S. Aksenova , 'Machine Learning with WEKA -WEKA Explorer Tutorial for WEKA Version 3.4', 2004.

[13] Machine Learning Repository, Center for Machine Learning and Intelligent Systems https://archive.ics.uci.edu/ml/index.php