CLASSIFICATION OF LIVER BASED DISEASES USING RANDOM TREE

Parminder Kaur[1]and Aditya Khamparia[2]
[1,2]Department of Computer Science and Engineering,
Lovely Professional University, Phagwara, India

ABSTRACT
The liver is the largest organ of our body .It is a cancer that begins in the cells of your liver. In this paper, we discuss the Classification of Liver Based Diseases which contains the general information on liver. With neurological, Psychiatic, pathological, physical, cognitive disordering there occur some inflammation, fibrosis, cirrhosis, *and disruptions in the liver that lead some problems like failing voice, memory loss, enlarged liver, weight loss etc. In this work, random tree algorithm is deployed for the classification of Liver based diseases. On the basis of analysis derived from weka tool we conclude that the disease is fatty liver, Wilson, Inherited, Autoimmune, and Cholestatic with the help of weka tool. In order to provide easy access to user, decision tree is used for generating the tree. The results are shown in terms of random decision tree, true positive rate, and false positive rate and ROC curve.*

KEYWORDS: *liver based diseases, confusion matrix, data mining, random tree, ROC curve.*

I. INTRODUCTION

The liver is the largest abdominal organ and is the sixth most common cancer in the world. It is triangular in shape. Liver can be divided into two parts right and left hemi liver. Liver is a single organ. It is essential to the proper functioning of our complete body. It is the Primary organ which maintains balance of many nutrients and chemicals like glucose, fat, cholesterol, vitamins, and hormones Primary liver cancer is a cancer which starts in the liver. Secondary liver cancer is a cancer which spreads from another part of the body. If the liver fails it causes many problems like jaundice, malnutrition. Liver based diseases such as fatty liver, inherited, Wilson, autoimmune, Cholestatic [www.cancer.org].

1.1. Data Mining

Data Mining is a term used to turn raw data into useful information. Data Mining has two basic operations predictive and descriptive. Predictive such as regression, classification, collaborative filtering and descriptive such as clustering, deviation detection.

Random tree is used for both classification and regression problems. It is a collection of tree predictors that is called forest. There is no need for any accuracy estimation procedures in random trees, such as cross-validation or bootstrap. Random tree is a class used for constructing a tree that considers k randomly chosen attributes at each node.

Random tree provides a powerful technique for classification .Various random tree are available to classify the data, including ID3, C4.5, C5, J48, CART. In this paper we have chosen random tree to establish the model. Use training set is used to prepare training and test data. After data pre-processing (CSV format), the random tree is employed on the dataset using WEKA. Fig1. Shows the flow of the research conducted to construct the model [weka.sourceforge.net].

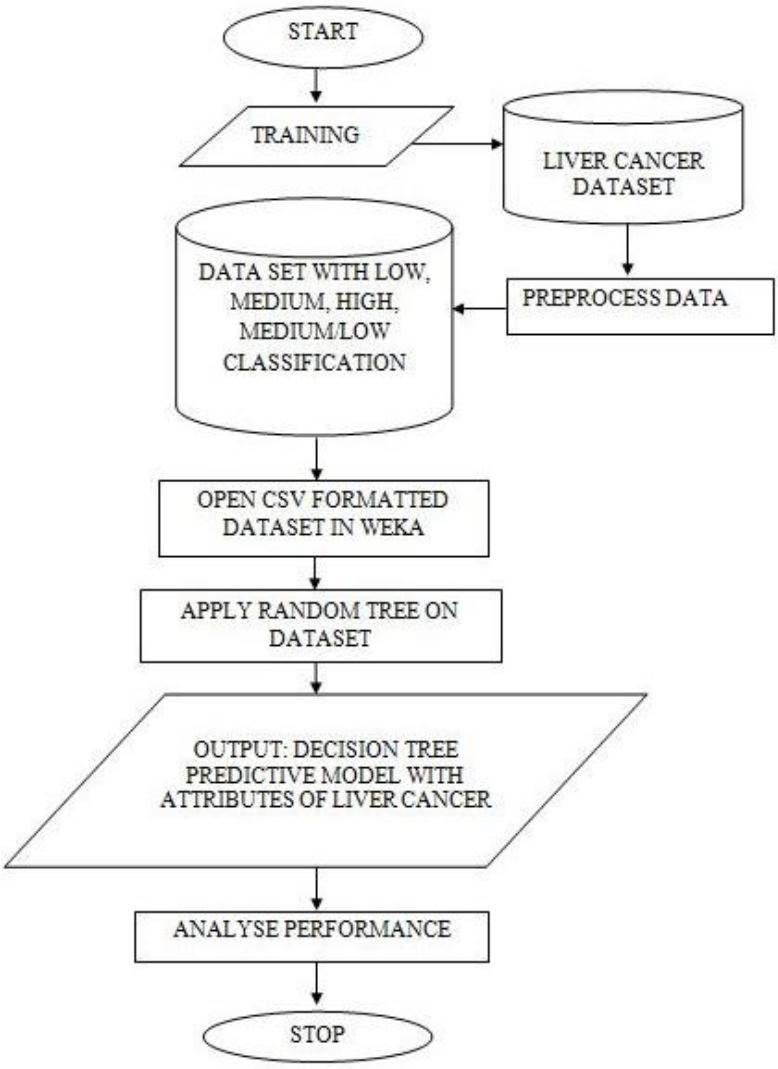


Fig 1: Flow diagram for Liver based diseases

The different types of liver based diseases are:-

Fatty Liver Diseases (FLD) Fatty liver builds up of fat in the liver. Fatty liver is caused by alcoholism, drugs, viruses, excess body weight. The sign and symptoms of FLD are FG, AM, WTL, WK, PRA, LOC.

Wilson Diseases (WD) Wilson is a inherited disorder; it strikes in between the ages of 12 and 23.The sign and symptoms of WD are SS, ML, FV, BC, DP, FG.

Inherited Liver Diseases (IHD) Inherited is one of the most common systematic disorders. It is a pathologic impact on the liver. The sign and symptoms of IHD are JP, FG, WTL, AP, DI, CC, JA, SAF.

Autoimmune Liver Diseases (AID) Autoimmune is a disease in which your immune system attacks your liver cells .The sign and symptoms of AID are EL, DU, ABVS, AD, FG, VM, JP.

Cholestatic Liver Diseases (CLD) Basic kinds of Cholestatic are primary biliary cirrhosis (PBC) and primary sclerosing cholangitis (PSC). The sign and symptoms of CLD are FG, IS, JA, SC, RAP, FV.

Fig2. Shows the hierarchical correlation of sign and symptoms of liver based diseases

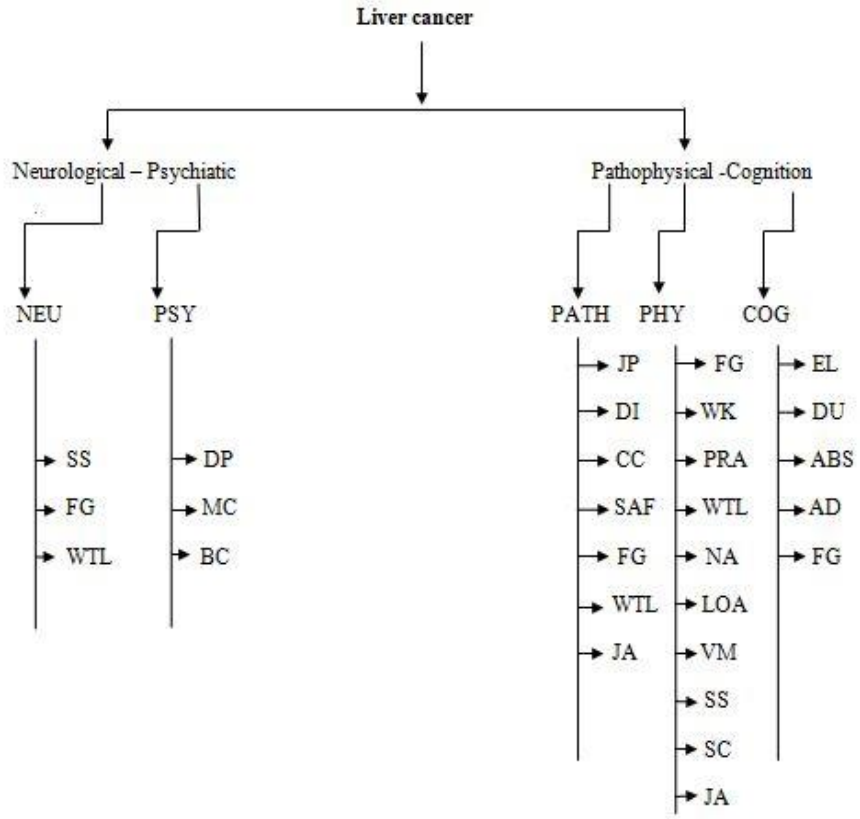


Fig 2: Hierarchical correlation of sign and symptoms of the liver cancer disease

The paper has been structures as, apart from introduction in Section 1, Section 2 presents the related works and section 3 is our problem description of the various liver based diseases with their neurological, psychiatric, pathological, physical, cognitive parameters. Section 4 is our observation and discussion on each of these liver cancer diseases.
Section 5 shows the computation and results for diagnosis of liver based disease using Weka tool (random tree). Section 5 deals with conclusion and section 6 are our future scope.

II. RELATED WORKS

Many research groups have developed different approaches for liver cancer. Sammouda.et.al uses histogram, physically, neighborhood based segmentation. It uses HNN artificial neural networks for classification [6]. Upadhyay & Wasson uses genetic algorithm, region growing, threshold based, level set method, statistical model, and histogram based approach for liver cancer detection [7]. Selvaraj & S.based on particle swarm optimization on liver cancer diagnosis [8]. Massieh.et.al uses 3-D consistency check based on knowledge based constraints [9]. Kundra & Pandey uses j48 algorithm for the classification of diseases. The paper facilities with neurological and physiological disordering that occurs inflammation and disruptions in the brain that creates some problems which is based on two parameters: Psycho- physical and EEG characteristics. J48 can also handle the data with missing attribute value [11].Sumblay .et.al discussed the diagnosis of breast cancer and classification of breast cancer whether the patient had benign or malignant tumor. It also generates a decision tree and confusion matrix. The paper works with different approaches that include neural networks, digital mammography, Naïve Bayes [12]. Alie.et.al uses supervised and un-supervised machine learning algorithms to find the accurate results in classification. It also generate true positive rate and false positive rate [13].Qiu.et.al uses three methodologies to find the accuracy of three classifiers:-K-means, back propagation neural network and SVM. SVM find the accurate result as compared to other methods [14].

III. PROBLEM DESCRIPTION

As shown in table1, it describes diseases with their two important parameters.Pathological-physical cognition and Neurological Psychiatic. Pathological-physical cognition parameter are further divided into three parts 1) pathological (PATH) parameter are joint pain (JP), diabetes (DI), chronic cough (CC), swelling of the ankles and feet (SAF). 2) The physical (PHY) parameter are fatigue(FG), weakness(WK), pain in the right abdomen(PRA), weight loss(WTL), nausea(NA), loss of appetite(LOA), vomiting(VM), jaundice(JD).3) The cognition parameters are enlarged liver(EL), dark urine(DU),abnormal blood vessels on skin(ABVS), abdominal distention(AD). Neurological-Psychiatic characteristics is divided into two parts: 1) Neurological is observed in different ways such as slurred speech (SS), failing voice (FV).2) Psychiatic is observed in Behavioral changes (BC), memory loss(MC), depression(DP). In Table1 row represents the disease and column represents their respective parameters. The subcolumns of neurology, psychiatric, pathology, physical, cognitive parameter contains "Y" if the symptom is present in the disease shown in respective row. For example, liver cancer has neurology symptoms such as: failing voice. Therefore these columns contain "Y" as in Table 1. Similarly, the pain in the right abdomen and loss of appetite sub-column of physical contains "Y". Finally the table is formed with five rows corresponding to five diseases and 24 columns corresponding to the parameters of the diseases.

Table1. Problem Description Table (PDT)

DISEASES	NEUROLOGICAL PSYCHIATIC		PATHOPHYSICAL-COGNITIVE																					
	NEUROLOGY (NEU)		PSYCH IATIC (PSY)			PATHOLOGICAL (PATH)				PHYSICAL (PHY)										COGNITIVE (COG)			
	SS	FV	DP	BC	MC	J P	DI	CC	SAF	FG	WK	PRA	WL	NA	LOA	VM	JA	SS	SC	EL	DU	AVS	AD
FLD	N	N	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N
WILSON	Y	Y	Y	Y	Y	N	N	N	N	Y	Y	N	N	N	N	N	Y	N	N	N	N	N	N
CHOLESTATIC	N	N	N	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	N	N	N	N	N	N
INHERITED	N	N	N	N	N	Y	Y	Y	Y	Y	Y	N	Y	N	N	N	N	N	N	N	N	N	N
ALD	N	N	N	N	N	N	N	N	N	Y	N	N	N	N	N	Y	Y	N	N	Y	Y	Y	Y

IV. OBSERVATION AND DISCUSSION

From detailed study of liver cancer it has been observed that according to doctors of Fortis hospital every year due to liver disease Over 2 lakh people lose their lives in India and close to 20,000 people need liver transplantation to survive [www.gastroliverspecialist.com].

Fig 3. Shows percentage usage of different liver diseases are autoimmune liver disease is of 25%, Cholestatic liver disease is of 13%, fatty liver disease is of 22%, Wilson disease is of 33%, inherited liver disease is of 7%. Fatty liver include symptoms such as pain and swelling in abdomen, decreased appetite, nausea, fatigue, vomiting, dry mouth & increased thirst, yellow color in the skin or mucus membrane or eyes, small, red spider-like veins on skin, itching and redness on feet or hands. Autoimmune liver disease is four times more common in women than in men. Cholestatic liver disease is of two types: primary sclerosing cholangitis (PSC) and primary Biliary Cirrhosis (PBC). Females mostly suffer from PBC and males more effects from male population. Wilson's disease is a long term liver disorder resulting in excess copper deposition in liver, brain, kidney and cornea of eye. It strikes in between the ages of 5-40 years [www.kucancercenter.org].

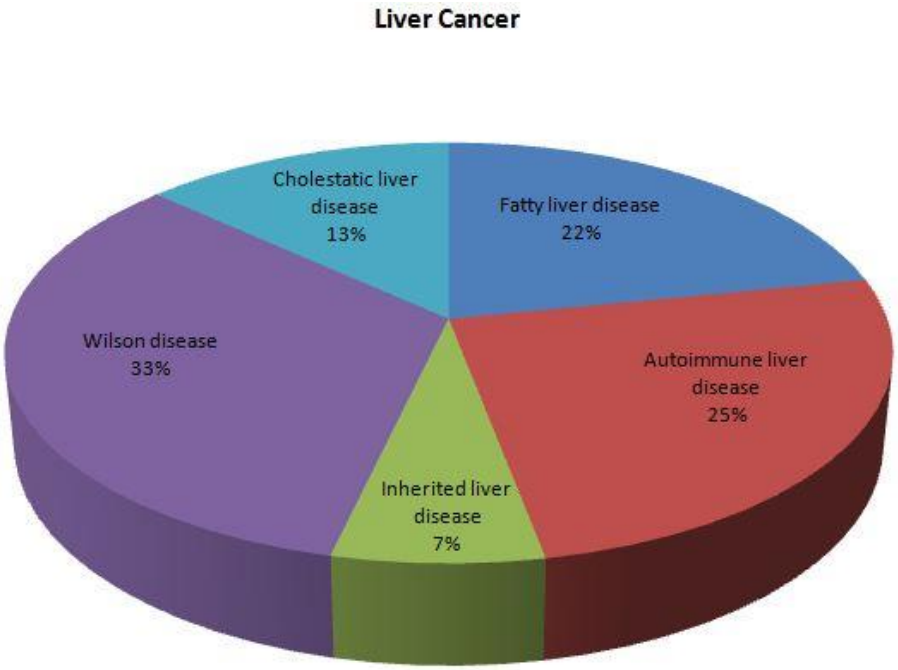


Fig 3: Liver Cancer Diseases

V. RESULT

The liver cancer diseases are classified using random tree algorithm of Weka tool. Data which is used for the experiment is Fig 4: random tree generated by Weka collected from hospital. The dataset consist of many patients and 5 attributes. Detailed description of data set is given in table 2.

Table 2: Detail of Dataset

Attributes	Data Type
Neurological	Categorical(high, medium, low, medium/low)
Psychiatic	Categorical(high, medium ,low, medium/low)
Pathological	Categorical(high ,medium, low, medium/low)
Physical	Categorical(high, medium, low, medium/low)
Cognitive	Categorical(high, medium, low, medium/low)

Random tree generate rules as shown in Figure 4 and 5

Random Tree
==========
NEU = HIGH
| PSY = HIGH: WD (1/0)
| PSY = MEDIUM/LOW: NOD (1/0)
NEU = LOW
| PHY = HIGH: FLD (1/0)
| PHY = MEDIUM/LOW: CLD (1/0)
NEU = MEDIUM
| COG = HIGH: AID (1/0)
| COG = MEDIUM/LOW: NOD (1/0)
NEU = MEDIUM/LOW
| PATH = HIGH: IHD (1/0)
| PATH = MEDIUM/LOW: NOD (1/0)
Size of the tree: 13

Fig 4: Random tree generated by Weka

Rules generated by random tree

Rule1: If neurological (NEU) symptoms and Psychiatic (PSY) symptoms are high, then disease is "Wilson".
Rule 2: If neurological (NEU) symptoms are high and psychiatric (PSY) symptoms are medium/low, then disease is "No disease".
Rule 3: If neurological (NEU) symptoms are low and physical (PHY) symptoms are high, then disease is "Fatty Liver".
Rule 4: If neurological (NEU) symptoms are low and physical (PHY) symptoms are medium/low, then disease is "Cholestatic".
Rule 5: If neurological (NEU) symptoms are medium and cognitive (COG) symptoms are high, then disease is "Autoimmune".
Rule 6: If neurological (NEU) symptoms are medium and cognitive (COG) symptoms are medium/low, then disease is "No disease".
Rule 5: If neurological (NEU) symptoms are medium/low and pathological (PATH) symptoms are high, then disease is "Inherited".
Rule 6: If neurological (NEU) symptoms are medium/low and cognitive (COG) symptoms are medium/low, then disease is "No disease".

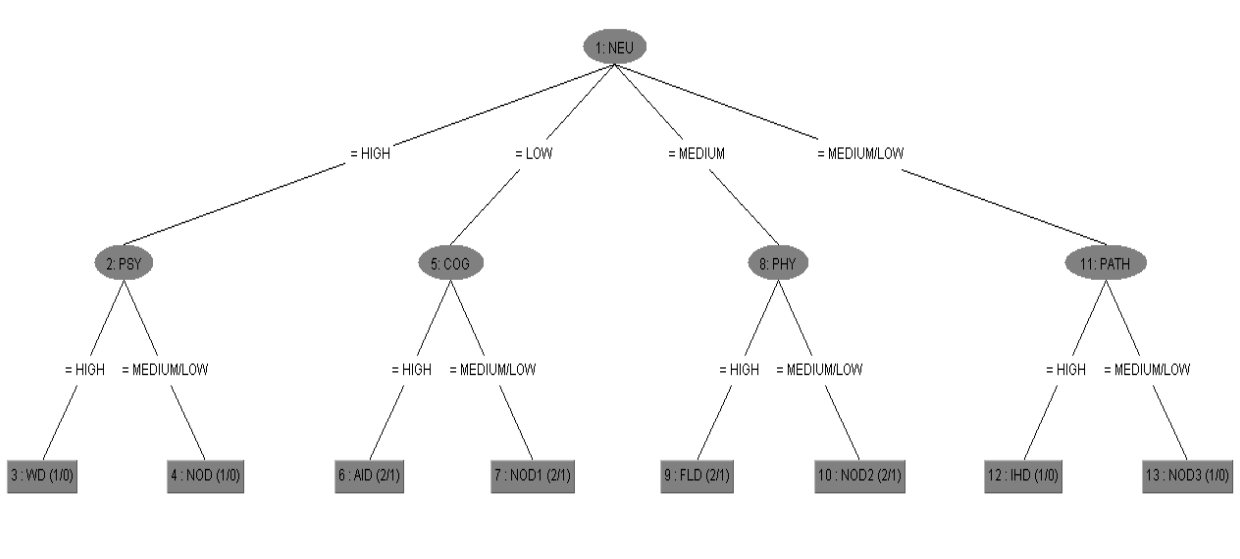


Fig 5: random tree generated by Weka

Different attribute have been chosen randomly from the dataset. Random tree is applied on the dataset and a confusion matrix is generated as shown in Fig.6

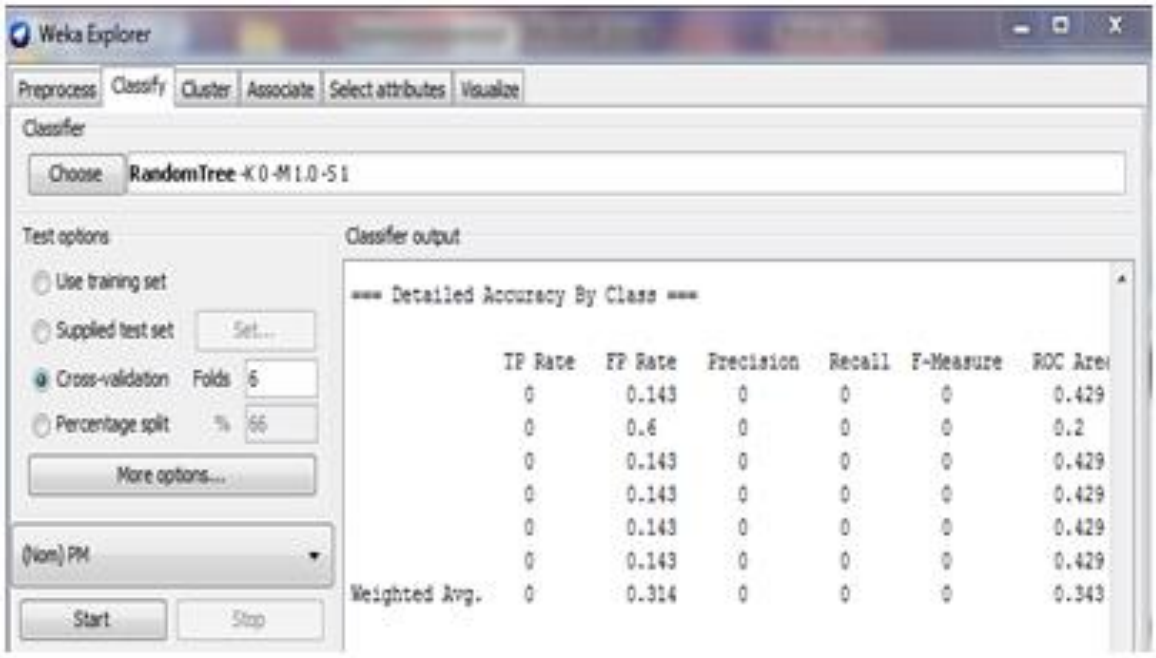
```
=== Confusion Matrix ===

 a  b  c  d  e  f   <-- classified as
 0  1  0  0  0  0 |  a = WD
 1  0  0  0  1  1 |  b = NOD
 0  0  0  1  0  0 |  c = FLD
 0  0  1  0  0  0 |  d = CLD
 0  1  0  0  0  0 |  e = AID
 0  1  0  0  0  0 |  f = IHD
```

Fig 6: Confusion Matrix

From above confusion matrix, true positive for class a ="IHD" is 1 while false positive values are 0, 0, 0, 0. Whereas for class b = "NOD", true positive is 1 while false positive values are 0,0,0,0. For class c= "FLD", true positive is 1 while false positive values are 0,0,0,0.For class d="CLD" true positive is 1 while false positive values are 0,0,0,0.For class e="IHD" true positive is 1 while false positive values are 0,0,0,0. In this diagonal elements of the confusion matrix represents the positive

(TP) values and rest of the values represents the false (FP) values. Different operative characteristics are defined as follows:
True positive (TP) = When test outcome is positive and condition is positive.
False positive (FP) = When test outcome is positive and condition is negative.
True negative (TN) = When test outcome is negative and condition is negative.
False negative (FN) = When test outcome is negative and condition is positive.

Table 3: TPR and FPR values



The ROC curve obtained from Weka (random algorithm) tool is shown in Fig 7. It compares two operating characteristics that are true positive rate and false positive rate value. Receiver operating characteristics curve (ROC) and area under curve (AUC) are calculated by threshold curve class in Weka.

Fig7. ROC Curve

VI. CONCLUSION

The automatic classification of liver cancer is an important real world medical problem. In this work, random tree is deployed for the classification of liver based diseases such as: Wilson, fatty liver, Cholestatic, inherited, autoimmune .The data is collected from hospitals. This paper shows decision trees are used to model actual diagnosis of liver cancer for surgical and non-surgical treatment.

Random tree algorithm generates rules and decision trees for the classification of liver based diseases. Random algorithm in weka generate decision tree that help us to diagnose the disease on the basis of attributes and symptoms.

VII. FUTURE SCOPE

The future scope of classification of liver based diseases using random tree is that this work can be carried forward by using other parameters also apart from the two parameters and diseases and random tree that are already discussed. New tree can be developed for this purpose and input and output would also change accordingly.

REFERENCES

[1]. http://www.cancer.org/cancer/livercancer/overviewguide/liver-cancer-overview-what-is-liver-cancer
[2]. http://www.kucancercenter.org/cancer-information/specialties-and-treatment/liver-cancer
[3]. http://weka.sourceforge.net/doc.stable/weka/classifiers/trees/package-summary.html
[4]. Machine Learning, Wald I, Apr 21 2014July 2002. http://statwww.berkeley.edu/users/breiman /wald2002-1.pdf
[5]. http://www.gastroliverspecialist.com/liver-disease.html
[6]. Sammouda.et.al (1999)" Segmentation and Analysis of Liver Cancer Pathological Color Images based on Artificial Neural Networks", IEEE.
[7]. Upadhyay & Wasson (2014)" Application of Genetic Algorithm for Liver Cancer Detection", International Journal of Research in Engineering & Technology, volume 03, Issue 05.
[8]. Selvaraj & S., (2013)" Improved Feature Selection Based on Particle Swarm Optimization for Liver Disease Diagnosis", Springer International Publishing Switzerland, Part II, pp 214-225.
[9]. Massieh.et.al, (2010)" A Novel Fully Automatic Technique for Liver Tumor Segmentation from CT Scans with Knowledge- based constraints" ,IEEE.
[10]. Vivek vij, Liver transplants in India from: http://www.livertransplantsindia.com/liver-liver diseases.php#1.
[11]. Kundra & Pandey (2014) "Classification of EEG based Diseases using Data Mining", International Journal of Computer Applications, Volume 90.
[12]. Sumbaly.et.al (2014)"Diagnosis of Breast Cancer using Decision Tree Data Mining Technique", International Journal of Computer Applications, Volume 98- No.10.
[13]. Ali.et.al (2014)" Intelligent Image Processing Techniques for Cancer Progression Detection, Recognition and Prediction in the Human Liver", IEEE.
[14]. Qiu.et.al," Research on Fuzzy Enhancement in the Diagnosis of Liver Tumor from B-mode Ultrasound Images", IEEE.

BIBLIOGRAPHY

Parminder Kaur received his B.A. in Computer Science and Economics from Punjab University, Chandigarh. She received a post graduate degree MCA from Punjabi university, Patiala. She is pursuing M.Tech Computer Science & Engineering from Lovely Professional University, Phagwara. Her research interest includes Artificial Intelligence.

Aditya khamparia received his B.Tech in Computer Science & Engineering from RGTU, Bhopal. He received a post graduate degree M.Tech Computer Science & Engineering from VIT, Vellore. He is currently an Assistant professor in Lovely Professional University. His research interest includes semantic web, Artificial Intelligence, E- learning and Cognitive Science.