Comparative Analysis Of Early Detection Techniques For Liver Diseases Through Python Algorithms

Hemlata Sinha[1]*, Sumit Kumar Roy[2]

[1*]Associate Professor, Department of Electronics and Telecommunication Engineering Shri Shankaracharya Institute of Professional Management and Technology, Raipur, Chhattisgarh, 492001, India. Email: Sinha.hemlata552@gmail.com
[2]PhD Scholar, National Institute of Technology, Raipur, Chhattisgarh, 492001, India. Email: sumit.roy97@gmail.com

***Corresponding Author:** Hemlata Sinha
*Associate Professor, Department of Electronics and Telecommunication Engineering Shri Shankaracharya Institute of Professional Management and Technology, Raipur, Chhattisgarh, 492001, India. 9827986985
Email: Sinha.hemlata552@gmail.com

Abstract—
Early diagnosis of liver disease is very important in order to save human lives and take appropriate measure to control the disease. In several fields, especially in the field of medical science, the ensemble method was successfully applied. This research work uses different ensemble methods to investigate the early detection of liver disease. The selected data set for this analysis is made up of attributes such as total bilirubin, direct bilirubin, age, sex, total protein, albumin, and globul in ratio. This research mainly aims at measuring and comparing the efficiency of different ensemble methods. Ada Boost, Logit Boost, BeggJ48 and Random Fore stare the ensemble method used in this research. The study shows that Logit Boost is the most accurate model than other ensemble approaches.

Keywords— Data Mining, Ensemble Method, Bagging, Boosting, Stacking, Liver Disease

I. INTRODUCTION

With an average weight of 1.6 kg (3.5 pounds), Liver is the largest internal organ within the human body. The principal function is to transport the blood from the digestive tract before it is transferred to the rest of the body. In addition, the liver produces proteins essential for blood clotting and other functions. Carbohydrates are processed in the liver where they are divided on to glucoses i phoned into the blood stream to maintain normal levels of glucose. The liver is the only regenerating visceral organ.

Truth be told, this vitalorgan carries out more than 500 roles in human body. As it does so much task, it's not surprising that it has a list of countless diseases. Though a healthy liver functions very efficiently, the consequences of a diseased liver can be dangerous and even deadly. Liver disease may be inherited(genetically)or caused by a number of factors including viruses and the consumption of alcohol that damage the liver. Gradually, liver damage causes scarring, resulting in liver failure. Diseases that are caused by viruses are hepatitis a, hepatitis b, hepatitis c. In addition to anunusual gene inherited from one or both parents, variety of substances may cause liver damage in the liver. Our own immune system can also cause disease if it target so there part so four body(autoimmune)that can damage the liver. Various kinds of cancer are a reason toget concerned as well.

Liver disorders are noticed only when it is often too late because the liver continues to work, even if it is weakened partly [1]. Early diagnosis of liver disease could save lives. Although not detectable even by experienced medical practitioners, early symptoms of this disease can be observed. Thus it would be of great benefit in the medical field to build a device that would improve the diagnosis ofthe disease. Such tools can allow physician to make specific decisions about patients and also with the aid of automatic liver disease classification tools.

Ensemble learning methods in the area of predictive modeling have become more interesting today. The use of multiple learning algorithms is successful in increasing overall prediction accuracy[2]. Ensemble methodologies have been successfully used in a wide range of fields, including healthcare, economics, manufacturing, bioinformatics etc. In this research work, ensemble method is used for predicting the liver disease. The paper makes two contributions. Firstly, the output of the ensemble classifier with accuracy rate and RMSE in relation to different sets is evaluated. The second one is to compare and analyzed the performance in terms of Receiver Operating Characteristics(ROC) curve, True Positive Rate (TPR) and False Positive Rate(FPR) of the ensemble classifier.

In the following section, the other parts of the paper are organized. Related work is shown in section 2. Section 3describe the ensemble method. Methodology will be defined in the section 4. The experimental results are discussed in section 5. In the last section, conclusion and future work is shown.

II. LITERATUREREVIEW

A recent study used a vector machine algorithm to diagnosis liver disease [3]. In their research, support vector machine(SVM) was used to identify liver disease using diabetes data set and two liver patients' datasets. In another

paper the author used supervised machine learning for automated identification of white regions for liver biopsies [4]. In the paper they have said that the liver images were analyzed using supervised machine learning classifier based on an notations given by two expert pathologists.
In a study the author uses Bayesian Classification for liver disease prediction [5]. In that article, they used Bayesian classification methodology, which is one of the main models of classification.
A study of 2012 used the data mining method for the diagnosis of fatty liver disease in ultrasound [6]. The aim of this work was to establish one of the computer-aided diagnostic (CAD) techniques to accurately identify healthy livers and unhealthy livers affected by fatty liver disease(FLD).
In 2016, a survey was conducted regarding data mining classification techniques to analyze liver disease disorder[7].The process of data mining is the extraction of meaningful information from the large databases. This paper describes application of data mining classification technique for the study of liver disease disorder.
Another paper "Pruned" Machine Learning Models and Public Data have been used for Predicting Mouse Liver Microsomal Stability[8].There, they discussed the design of machine learning models that can increase the likelihood of MLM stability recognition compounds.
Hepatitis is the diseases of liver caused by virus. A study was performed with Artificial Neural Networks to diagnose liver disease caused by hepatitis virus. [9]. Here the paper also used the results of the previous studies on the diagnosis of hepatitis which use the same database.
In another paper the author uses case-based reasoning(CBR) uses a variety of common data mining classification methods to build efficient model for diagnosing the liver disease earlier[10].
The amount of liver patients is increasing day by day. That's what in another study; published in 2014 Liver Patient Classification using Intelligence Techniques has been done[11].
In a study of 2012, the authors howed Automatic Identification of Ultrasound Liver Cancer Tumor Using Support Vector Machine [12]. There they approached totally automatic machine learning system for identifying the liver cancer tumor from ultrasonic images.

III. ENSEMBLEMETHOD

Ensemble techniques combine several classifiers of to generate better predictive efficiency than a single classifier[2]. The primary concept behind the ensemble model is to bring together a group of weak learners to create a power full earner, thereby improving the model's precision. When we use any machine learning method to estimate the target variable, the primary causes of the difference in real and expected values are noise, variance and bias. Ensemble helps to decrease these variables (with the exception of noise, an irreducible mistake).

The introduced first word, bagging [13], is shorthand for boot strapping and aggregating mixture. Bootstrapping [14]is a technique to assist lower the classifier's variance and reduce over fitting by resampling the training set information with the same cardinality as the initial set. The generated model should be less over-fitted than a single model. An elevated variance is not useful for a model, implying that its output is vulnerable to the information supplied for practice. So, the model may still perform badly even if more training data is supplied. And, maybe not even decrease our model's variance. Boot strap relates to replacement by random sampling. Bootstrap enables us to better comprehend the dataset's bias and variance. Bootstrap includes random sampling of the dataset's tiny subset of information. It is possible to substitute this subset. Choosing the entire instance in the dataset is equally likely. This technique can assist make the mean and standard deviation from the dataset better comprehend. Boosting[15][16] refers to a group of algorithms that use weighted average to turn weak learners into strong learners. Unlike bagging, each model runs independently and then aggregates the inputs at the end without any choice of model. Boosting is a
technique that involves fitting sequentially multiple weak learners in a very adaptive manner. Intuitively, the new model focuses its attention on the most challenging findings to match upto now, so we get a good learner with lower bias at the end of the process. Boosting can be used for regression as well as identification issues, such as bagging.

Stacking [17] is another ensemble model in which a fresh model is educated from two (or more) prior models mixed results. For each sequential layer, the model predictions are used as inputs and combined to create a fresh set of predictions. Ensembles tacking can be called mixing because all figures are mixed to create a forecast or classification. This model improves performance over learning algorithms individually. The best-of–the art methods are the stacking of probability distributions (PDs)and linear multi-response (MLR)[18]. It is common knowledge that sets of various base-level classifiers (with predictions that are weakly correlated) yield good performance. Merz [19] suggests a stacking method called SCANN that uses correspondence analysis to identify associations between base-level classifier predictions

IV. METHODOLOGY

In this research work, we divided the methodology into four steps. The proposed methodology is shown in Fig 1 with block diagram. These steps are described in the following.
The first step is data collection. After that, models are created using Bagging, Boosting and Stacking using training data. Then model are evaluated using testing data. In the last step, models are compared to find the best model among bagging, boosting and stacking.

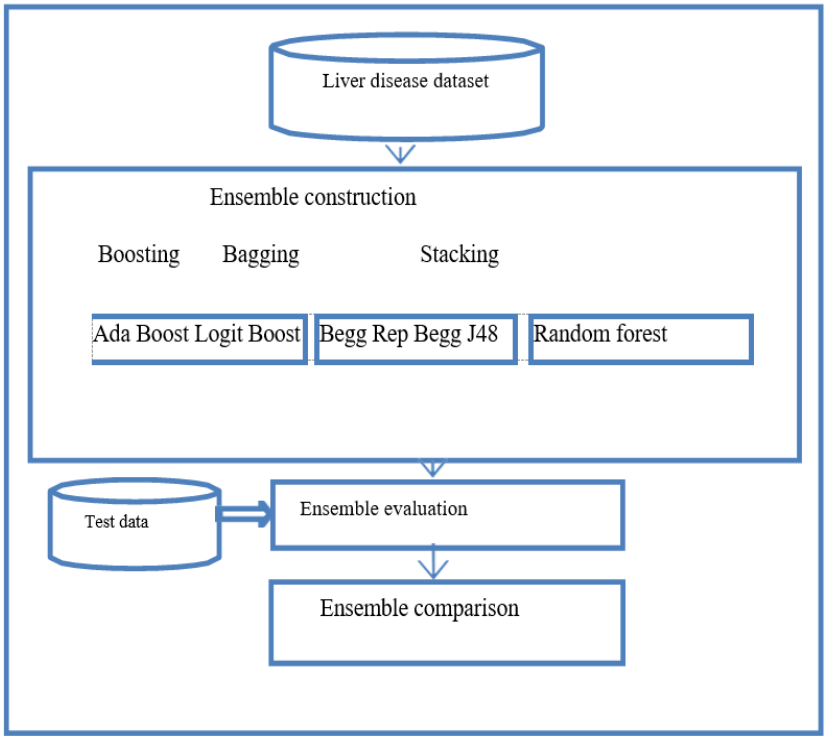

Fig1: Proposed Methodology

A) Data Collection

For the analysis, the ILPD (Indian Liver Patient Dataset) are used which is publicly vailable on the UCI Machine Learning Repository [20]. The dataset has ten attributes and one class attribute. The information on the attribute is shown in Table 1. This set of data contains 416 records of patients with liverand167records of patients without liver. The total number of patients is 583. Selector is category label used for splitting the patient into classes (patient or non-patient). This set of data comprises 441 records of men and142 records of women. The class value "No" means that the patient has no liver disease and "Yes" means that the patient has liver disease.

Table1: Description of the Attributes

S.N	Attribute Name	Description	Possible values
1	Age	The Patient Age	Numeric
2	Gender	The patients Sex	Nominal
3	Total Bilirubin	Bilirubin is a yellow pigment present in blood and urine.	Numeric
4	Direct Bilirubin	Bilirubin is of two forms. The direct bilirubin is flowing into the blood directly.	Numeric
5	Alkaline Phospotase	Alkaline phosphatase is a blood-related enzyme which helps break down proteins.	Numeric
6	Alamine Aminotransferase	The enzyme is present in the blood and is good indicator of testing if a liver is weakened by cirrhosis or he patitis in general.	Numeric
7	Aspartate Aminotransferase	This enzyme is present inside blood with low level. Higher level impliesd amage to the organ such as heart Disease and liver.	Numeric
8	Albumin	Albumin is protein that stops the blood fluid from flowing into the tissue.	Numeric
9	Total Proteins	The body's main proteins are globulin and albumin. These are symptoms of liver disease.	Numeric
10	Albumin and Globul in Ratio	It's a healthy liver status indicator. The standard A/ Gratioisaround0.8to2.0	Numeric
11	Class	Select or which classify the liver disease.	Nominal

B) Ensemble Construction

These steps involves the ensemble learning which is consists of Bagging, Boosting and stacking. Bagging method is performed using2 base learners, Reptree (BeggRep) and J48(BeggJ48). Boosting method is performed using Ada Boost and Logit Boost classifier along with Dicision Stumpas base classifier. Stackingisd one using random forest. The Bagging, Boosting and Stacking methodare run for 1, 2, 5, 10, 15, 50 100 iteration and they also run with their default parameter.

C) Ensemble Evaluation

After constructing the ensemble model, evaluation is performed on the test dataset. Accuracy rate and Root MeanSquareError(RMSE)areusedtocheckthecorrectpercentageofthemodel.

D) Ensemble Comparison

The last step of the methodology is the comparative analysis of the entire ensemble which is applied in this research.

V. EXPERIMENTAL RESULT

In this research, WEKA toolkit is used to build and evaluate the model. WEKA is an open source platform that has different algorithm for machine learning and data mining. We concentrate on the ensemble approach in our analysis. Ada Boost, Logit Boost, and Bagging ensemble techniques are available in the meta algorithms whereas Random forest is in the tree algorithm on WEKA tool. In the WEKA tool, the Explorer interface is used for various ensemble model constructions and the Knowledge Flow interface is used for building multiple ROC curve for model comparisons.

$$FPR = \frac{FP}{FP + TN}$$

Accuracy of the all ensemble method is shown in Table 2. The result for one single model construction is indicated by iteration number1.Iteration number 2,5,10,15,50,100 represents result of respective number model construction. From the table, it can be noticed that Logit Boost algorithm has the highest accuracy which is achieved in 10th iteration. The accuracy of the Random Forest algorithm varies from 65% to 69%. Bagging algorithm accuracy are 66.03% to 69.93% for J48 algorithm and 67.07% to 70.15%. The accuracy of the Ada Boost algorithm is range from 65.95%to70.25%.It is clear from the above accuracy that the accuracy is improved when the iteration number is increased.

Fig2 depicts the accuracy with regard to the ite ration number of all described ensemble methods. The X axis displays the iteration number and the Y axis indicates the accuracy of the ensemble method. From the Fig 1, it can be observed that Logit Boost, Ada Boost and bagging Rep tree algorithm have good accurate rate but random forest and Bagging J48 accuracy rate are relatively lower than the other in this work.

Table2: Accuracy of Ensembles

S.N	No of Iteration	Ada Boost	Logit Boost	BeggJ48	Begg Rep	RF
1	1	67.92	68.09	66.03	67.07	65.52
2	2	65.95	69.98	68.78	66.72	69.46
3	5	67.92	68.98	68.87	68.75	68.43
4	10	69.98	71.53	69.93	70.15	69.12
5	15	67.75	70.15	68.17	69.46	68.45
6	50	69.25	69.39	69.46	68.27	68.45
7	100	70.25	70.53	68.72	69.98	69.26

The error which is found in the prediction of the ensemble method is shown in the Table 3. This error is known as Root Mean Squared Error (RMSE). Table3 shows that Log it Boost algorithm has the lowest error rate and the highest error rate is found in Random Forest algorithm. Fig3 displays the RMSE graphical representation of the above ensemble method. The X axis shows the serial number of the iteration and the Y axis indicates the errors of the RMSE.

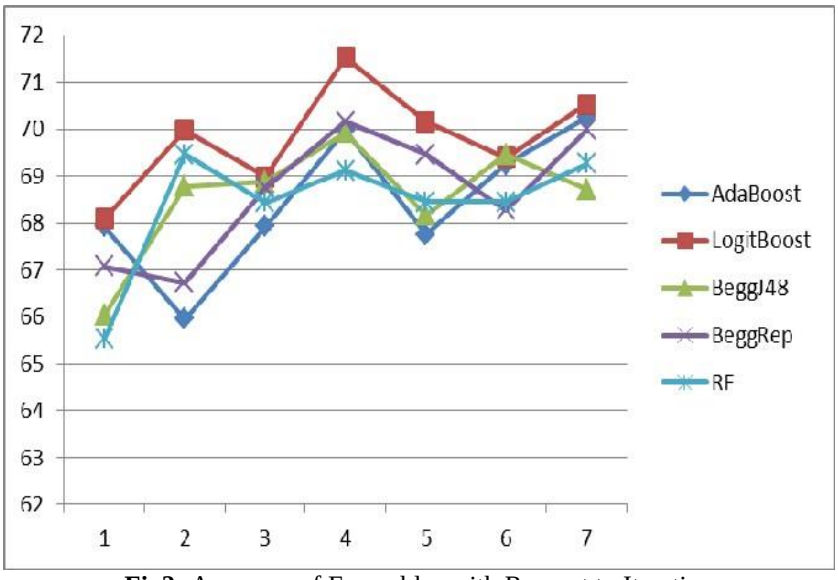

Fig2: Accuracy of Ensembles with Respect to Iteration

Table3: RMSE Error of Ensemble Method

S.N	No of Iteration	Ada Boost	Logit Boost	BeggJ48	Begg Rep	RF
1	1	0.4447	0.4259	0.5557	0.5121	0.5872
2	2	0.4428	0.4271	0.4842	0.4771	0.5036
3	5	0.4421	0.4259	0.4512	0.4431	0.4602
4	10	0.4378	0.4271	0.4401	0.4367	0.4477
5	15	0.4374	0.4255	0.4374	0.4342	0.4381
6	50	0.4369	0.4205	0.4256	0.4321	0.4260
7	100	0.4245	0.4180	0.4283	0.4251	0.4271

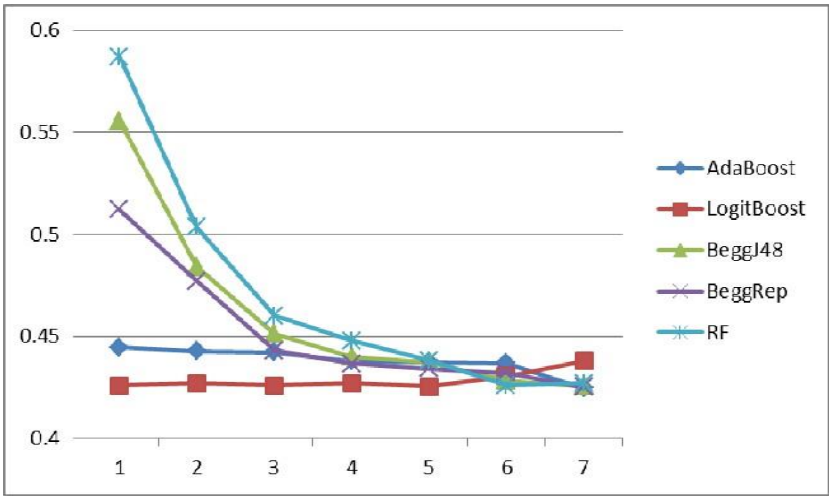

Fig3: RMSE of Ensembles with Respect to Iteration

The last phase of the methodology is the comparative analysis of all ensemble classifier which is used in this research work. These ensemble models are compared for the10th iteration. 10 fold cross validation method is used for the model evaluation. Generally, accuracy rate is the most usable measurement scale for the model comparison. But itis not always enough when there is imbalance data in the dataset. Because it is does not separate the numbers from the respective classes of correctly and incorrectly categorized examples and may result in erroneous conclusion [22]. For this reason, True Positive Rate (TPR), False Positive Rate (FPR and ROC curve are also used for model evaluations.

TPR can be described as a percentage of positively classified tuples in a given class [23]. It is known as sensitivity and hit ratio.

$$TPR = \frac{TP}{TP + FN}$$

FRP can be described as a percentage of negatively classified tuples in a given class [19]. It is also known as false alarm.

The ROC curve known as Receiver Operating Characteristics curve, is visualizing tool which is used to compare several classifier. The ROC curve represents the difference of one particular model between true positive and false positive rate [19]. The X-axis shows the true positive rate and Y axis shows false positive rate. Area under curve(AUC) is used for measuring the accuracy of the model for ROC curve. If the AUC value approaches 1.0implies that the model is more reliable while the AUS valueapproaches0.5, then the model is considered to be lea accurate and the model is considered to be good if the AUC value is1.0.

Table 4: Comparison of Ensemble Result

Classifier	Class	TPR	FPR	ROCcurve
Ada Boost	yes	0.854	0.702	0.702
	no	0.298	0.146	0.701
Logit Boost	yes	0.930	0.801	0.722
	no	0.199	0.070	0.722
BeggJ 48	Yes	0.830	0.643	0.720
	No	0.357	0.170	0.720
Begg Rep	yes	0.876	0.719	0.712
	No	0.281	0.124	0.712
Random Forest	Yes	0.886	0.749	0.706
	no	0.251	0.114	0.706

Table4 shows the comparison o fall ensemble method result. Table4 represents the comparison results of ensemble method in terms of TP rate, FP rate and ROC curve for both class value of 'yes' and 'no'. From the Table 4 it can be noticed that the AUC of the Logit Boost algorithm is 0.722 which is highest among all the ensemble method. The TP rate of the Log it Boost algorithm is also high than all other method. Fig4 shows the graphical representation of the TPR and FPR for all five ensemble technique. For class value 'yes' the range of TP rate is greater than 80% and FP rate is greater than 60%. similarly, for the class value of 'no' the range of TP rate is less than 40% and FP rate is less than 20%.

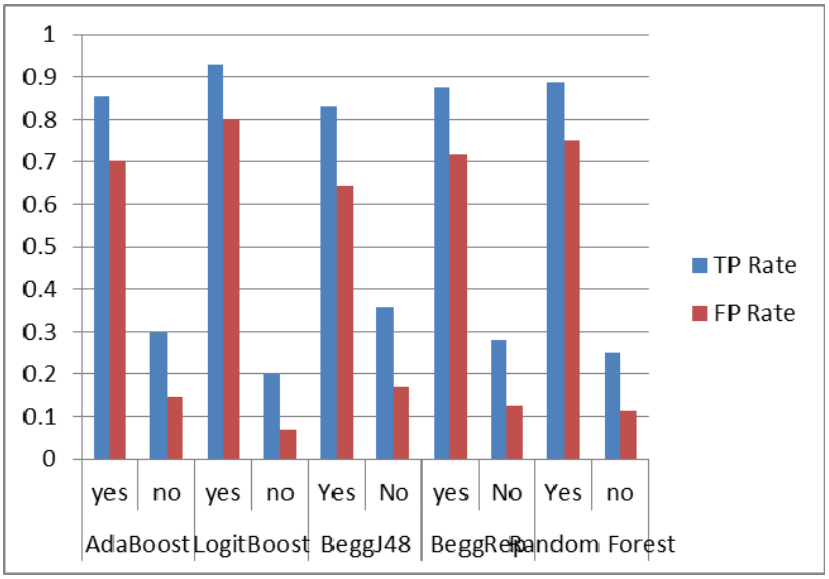

Fig4: TPR and FPR for all Ensembles Method

depicts the comparisons of ROC curve for all the ensemble method which are used in this work for the class value of 'no'. An optimal ROC curve value should be approaching to the 1.0.

VI. CONCLUSION ANDFUTURE WORK

In this research, we have analyzed a novel and efficient approach of ensemble learning for the liver disease classification. Here, five ensembles algorithm known as Ada Boost, Logit Boost, Begg Rep, BeggJ48 and random Forest are applied and compared in terms of accuracy, RMSE TPR, FPR and ROC curve. Log it Boost algorithm outperforms well than the other ensemble method and the accuracy of the Log it Boost algorithm is 71.53%. The use of ensemble technique in liver disease prediction would improve people's health. In the future, we will however obtain the most current data from different regions worldwide for the treatment of liver disorders and the prevention of liver disease. We will also apply some methodology which can eliminate the uncertainty like Belief-Rule Based(BRB)[24-36].

ACKNOWLEGEMENT

First of all, we would like to give thanks to our department of Electronics and Telecommunication Engineering of Shri Shankaracharya Institute of Professional Management and Technology Raipur for giving me the opportunity to work in this esteemed organization, which not only has increased our awareness about latest fields but also taught us the importance of team building. with the deep sense of gratitude, we express our sincere thanks to Chairman sir Shri Nishant Tripathi sir, Principal sir Dr. Alok Jain sir for their active support and continuous guidance without them it would have been difficult for us to complete this research.

REFERENCE

1. Lin, R. Ho.2009, "An intelligent model for liver disease diagnosis." Artificial Intelligence in Medicine 47.1(2009):53-62.
2. Dietterich, G. Thomas."Ensemble methods in machine learning." International workshop on multiple classifier systems. Springer, Berlin, Heidelberg,2000.
3. E.M Hashem, and M.S. Mabrouk. "A study of support vector machine algorithm for liver disease diagnosis." American Journal of Intelligent Systems4, no.1(2014):9-14.
4. S. Vanderbeck, J. Bockhorst, R. Komorowski, D.E. Kleiner, and S. Gawrieh. "Automatic classification of white regions in liver biopsies by supervised machine learning." Human pathology 45, no. 4(2014): 785-792.
5. S. Dhamodharan "Liver disease prediction using bayesian classification." In 4th National Conference on Advanced computing, applications & Technologies, pp.1-3.2014
6. U.R. Acharya,, S.V. Sree, R Ribeiro, G Krishnamurthi, R.T. Marinho J. Sanches, and J. S. Suri. "Data mining framework for fatty liver disease classification in ultrasound: a hybrid feature extraction paradigm."Medical physics 39, no.7Part1(2012):4255-4264.

7. D. Sindhuja., and R.J Priyadarsini."A survey on classification techniques in data mining for analyzing liver disease is order." International Journal of Computer Science and Mobile Computing 5, no.5(2016):483-488.
8. A. Perryman, L. Stratton, T.P. Ekins, J.S Freundlich ("Predictingmouse liver micro somal stability with "pruned" machine learning models and public data". Pharmaceutical research, 33(2),433-449,2016).
9. S. Ansari, I. Shafi, A. Ansari,. J Ahmad,S. I. Shah. "Diagnosis of induced by hepatitis virus using artificial neural networks." In 2011 IEEE 14th International Multitopic Conference, pp.8-12. IEEE,2011.
10. Chuang, C. Ling."Case-base dreasoning support for liver disease diagnosis. "Artificial Intelligence in Medicine 53, no.1(2011):15-23.
11. Pahareeya, Jankisharan, R. Vohra, J. Makhijani, and S. Patsariya. "Liver patient classification using intelligence techniques." International Journal of Advanced Research in Computer Science and Software Enaineering 4, no.2 (2014).
12. V. Ulagamuthalvi., and D. Sridharan."Automatic identification of ultra sound liver cancer tumor using support vector machine." In International Conference on Emerging Trends in Computer and Electronics Engineering, pp.41-43.2012.
13. L.Breiman,"Baggingpredictors,"MachineLearning,vol.24,no.2,pp.123–140,1996.