ISSN:0975-9646

Liver Patient Classification Using Intelligent Techniques

Anju Gulia[#] , Dr. Rajan Vohra[*] , Praveen Rani[#]

[#] Students, Computer Science & Engineering Deptt.,
P.D.M College Of Engineering, Bhadurgarh,Haryana(India)
[*] Head of Deptt., Computer Science & Engineering Deptt.,
P.D.M College Of Engineering, Bhadurgarh , Haryana(India)

Abstract— **Classification techniques have been widely used in the medical field for accurate classification than an individual classifier. This paper presents computational intelligence techniques for Liver Patient Classification. This paper evaluates the selected classification algorithms (J-48, Multi Layer Perceptron, Support Vector Machine, Random Forest and Bayesian Network) for the classification of liver patient datasets.**

This paper implements hybrid model construction and comparative analysis for improving prediction accuracy of liver patients in three phases. In first phase, classification algorithms are applied on the original liver patient datasets collected from UCI repository. In second phase, by the use of feature selection, a subset (data) of liver patient from whole liver patient datasets is obtained which comprises only significant attributes and then applying selected classification algorithms on obtained, significant subset of attributes. SVM algorithm is considered as the better performance algorithm, because it gives higher accuracy in respective to other classification algorithms before applying feature selection. But, Random Forest algorithm is considered as the better performance algorithm after applying feature selection. In third phase, the results of classification algorithms with and without feature selection are compared with each other. The results obtained from our experiments indicate that Random Forest algorithm outperformed all other techniques with the help of feature selection with an accuracy of 71.8696%.

KEY WORDS: Classification, J-48, Multi Layer Perceptron, Support Vector Machine, Random Forest, Bayesian Network, Feature Selection, Weka tool.

I. INTRODUCTION

Liver is the largest internal organ in the human body, playing a major role in metabolism and serving several vital functions. The liver is the largest glandular organ of the body. It weighs about 3 lb (1.36 kg) .The liver supports almost every organ in the body and is vital for our survival. Liver disease may not cause any symptoms at earlier stage or the symptoms may be vague, like weakness and loss of energy. Symptoms partly depend on the type and the extent of liver disease. Liver diseases are diagnosed based on the liver functional test [1].

Classification techniques are very popular in various automatic medical diagnoses tools. Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged [2]. An early diagnosis of liver problems will increase patient's survival rate. Liver disease can be diagnosed by analyzing the levels of enzymes in the blood [3].

Moreover, now a day's mobile devices are extensively used for monitoring human's body conditions. Here also, automatic classification algorithms are needed [4].

In this paper, five Classification algorithms J-48, Multi Layer Perceptron, Support Vector Machine, Random Forest and Bayesian Network algorithms have been considered for comparing their performance based on the ILPD (Indian Liver Patient Dataset).

A. Significance of the problem

The questions this research work can provide the solutions to, can be given as follows:

1) How hybrid model construction is performed?
2) How feature selection applied on liver datasets?
3) How Comparative analysis of classification algorithms is performed for improving prediction accuracy of liver patients with or without Feature Selection?

This paper finds answers to these questions which can help to know the various aspects about classification of liver patients. By performing this work, it is shown that feature selection has a great significance as the process of selecting a subset of relevant features for use in model construction. By using feature selection on ILPD before a classification algorithm can be applied, performance of classification algorithm increases.

B. Data Description

Databases of 583 records/entries are taken from the ILPD(Indian Liver Patient Dataset)Data setfor the purpose of solving problem of this paper. This dataset is downloaded from UCI machine Learning Repository (http://archive.ics.uci.edu/ml/). Entire ILPD dataset contains information about 583 Indian liver patients. In which 416 are liver patient records and 167 non liver patient records .The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not).

II. RESEARCH BACKGROUND

P.Rajeswari and G.Sophia Reena[5][2010]. In this paper, Authors perform data classification which is based on liver disorders This paper deals with the results in the field of data classification obtained with Naive Bayes algorithm, FT Tree algorithm and KStar algorithm.

Bendi Venkata Ramana ,Prof. M.Surendra Prasad Babu and Prof. N. B. Venkateswarlu, [4][2011]. The classification algorithms considered here are Naïve Bayes classifier, C4.5, Back propagation Neural Network algorithm, and Support Vector Machines. These algorithms are evaluated based on four criteria: Accuracy, Precision, Sensitivity and Specificity.

S. Karthik, A. Priyadarishini and J. Anuradha and B. K. Tripathy," [6][2011]. In first phase, ANN classification is applied for classifying the liver disease. In second phase rough set rule induction using LEM (Learn by Example) algorithm is applied to generate classification rules. In third phase fuzzy rules are applied to identify the types of the liver disease.

Bendi Venkata Ramana and Prof. M.Surendra Prasad Babu[7][2012]. Modified rotation forest algorithm was proposed with multi layer perception classification algorithm and random subset feature selection method for UCI liver data set.

A.S.Aneeshkumar and C.Jothi Venkateswaran[8][2012]. In this paper authors are using classification. The overall performance of C4.5 decision tree is better than Naive Bayesian.

Jankisharan Pahariya, Jagdeesh makhijani and sanjay patsariya [9][2014]. This paper presents computational intelligence techniques for Liver Patient Classification. The efficacy of the techniques viz. Multiple Linear Regression, Support Vector Machine, Multilayer Feed-Forward Neural Network, J-48, Random Forest and Genetic Programming has been tested on the ILPD Data Set. Authors employed under sampling and over sampling for balancing it. The results obtained from experiments indicate that Random Forest over sampling with 200% outperformed all the other techniques.

III. Research Methodology

For solving problems of this paper some research techniques and methodologies are used for obtaining the desired result. Some tools and algorithms are required for obtaining the result. Main steps under the research methodologies are:-

Review literature or research papers – first of all literatures and research papers were reviewed for getting more information about the problem and knowing which type of work was done by others on this topic and by which method.

Identify tools – then tools required for solving the problem were identified and the best tool – "WEKA" was selected from all

Study database attributes and data structure – attributes and structure of the database was thoroughly studied for finding out useful attributes from the liver patient database.

Determine nature and definition of research problem and work flow of the problem for getting accurate and desired result. A study of datasets taken from ILPD (Indian Liver Patient Dataset)-UCI Repository.

Organize the database with useful attributes and populate it, then perform data analysis using WEKA tool in order to generate the result.

IV. Conceptual Framework

Classification algorithms are widely used in various medical applications. Data classification is a two phase process in which first step is the training phase where the classifier algorithm builds classifier with the training set of tuples and the second phase is classification phase where the model is used for classification and its performance is analyzed with the testing set of tuples [10]. Classification is done to know the exactly how data is being classified. The Classify Tab is also supported which shows the list of machine learning algorithms. These algorithms in general operate on a classification algorithm and run it multiple times manipulating algorithm parameters or input data weight to increase the accuracy of the classifier. Two learning performance evaluators are included with WEKA. The first simply splits a dataset into training and test data, while the second performs cross-validation using folds. Evaluation is usually described by the accuracy [11].

The following techniques are applied to classify the Liver Patient:

1) *J-48 classifier:*

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [12]. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a classifier. It induces decision trees and rules from datasets, which could contain categorical and numerical attributes. The rules could be used to predict categorical values of attributes from new records.C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s_1, s_2, \dots$of already classified samples. Each sample $s_i$consists of a p-dimensional vector$(x_{1,i}, x_{2,i}, \dots, x_{p,i})$, where the$x_j$represent attributes or features of the sample, as well as the class in which $s_i$falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision [13, 14].

2) *MLP (Multilayer Perceptron) classifier:*

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network[15] .MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable[16].

The multilayer perceptron consists of three or more layers (an input and an output layer with one or more *hidden layers*) of nonlinearly-activating nodes. Each node in one layer connects with a certain weight w_{ij} to every node in the following layer. Some people do not include the input layer when counting the number of layers and there is disagreement about whether w_{ij} should be interpreted as the weight from i to j or the other way around.

3) *Random Forest (RF) classifier:*

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees[17]. It is unexcelled in accuracy among current algorithms. It runs efficiently on large data bases. It can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes) [18].

4) *Support Vector Machine (SVM) classifier:*

SVM or sequential minimal optimization (SMO) is a learning system that uses a hypothesis space of linear functions in a high dimensional space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory [19]. SVM uses a linear model to implement non-linear class boundaries by mapping input vectors non-linearly into a high dimensional feature space using kernels. The training examples that are closest to the maximum margin hyper plane are called support vectors. All other training examples are irrelevant for defining the binary class boundaries. The support vectors are then used to construct an optimal linear separating hyper plane (in case of pattern recognition) or a linear regression function (in case of regression) in this feature space. Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.

5) *Bayesian Network classifier:*

Bayesian networks are a powerful probabilistic representation, and their use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute Ai given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A1…..An and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes. In particular, the naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent.

6) *Feature Selection :*

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction [20, 21]. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction, and how these features are related. Subset selection evaluates a subset of features as a group for suitability [22, 23].

This paper gives solution of three problems which are faced in classification/prediction of liver disease patients. These three problems are:

A. *Applying Classification Algorithm without Feature Selection*

Applying selected classification algorithms on the original Indian Liver Patient Datasets (ILPD), this comprised of all relevant and irrelevant attributes without feature selection of liver patients. The result of all these techniques are obtained and analyzed in the form of accuracy of these classification algorithms.

B. *Applying Classification Algorithm after Feature Selection*

In this, attribute or feature selection is done with the help of greedy stepwise approach. The whole datasets of liver patients is comprised of all relevant or irrelevant attributes. By the use of feature selection, a subset (data) of liver patient from whole liver patient datasets will be obtained which comprises only significant attributes.

Applying selected classification algorithms on the obtained significant subset of attributes after feature selection of ILPD datasets. The result of all these techniques are obtained and analyzed in the form of accuracy of these classification algorithms.

C. *Comparative Analysis for Improving Prediction Accuracy*

In this, the results of classification algorithms with and without feature selection are compared with each other. A particular classification algorithm is identified by comparative analysis of all algorithm accuracies which improves prediction accuracy of liver patients. The figurative approach for performing these tasks is shown in figure 1.

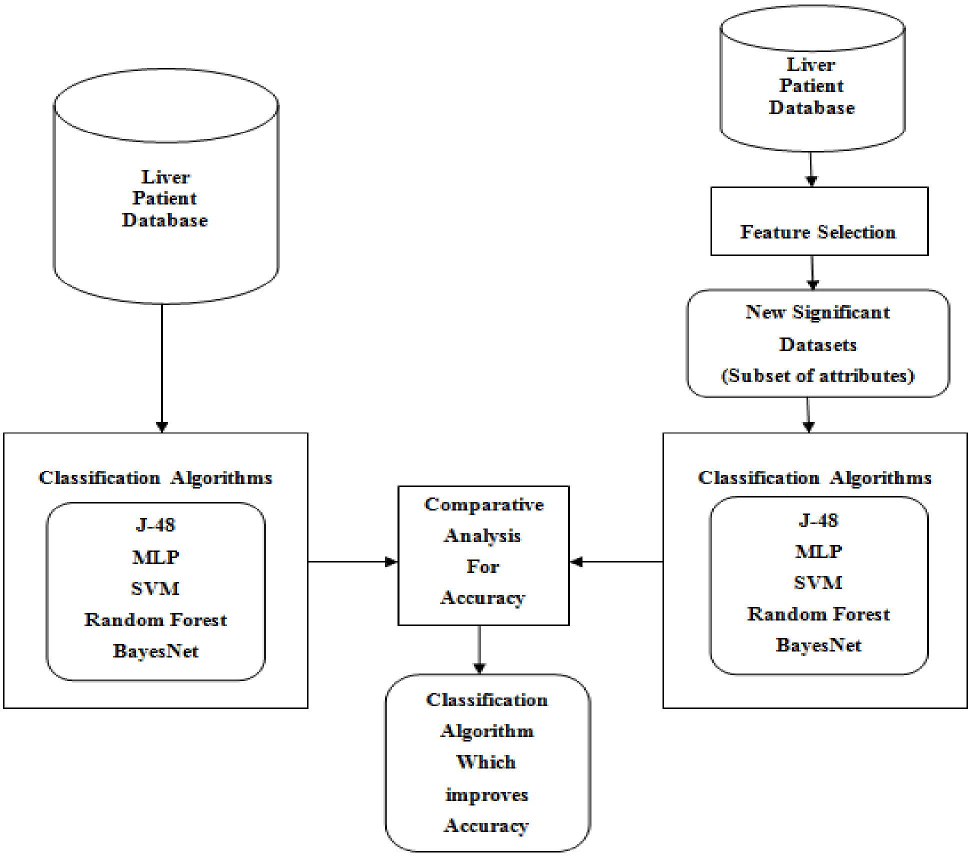


Figure 1 Hybrid Model Construction and Comparative Analysis for Improving Prediction Accuracy

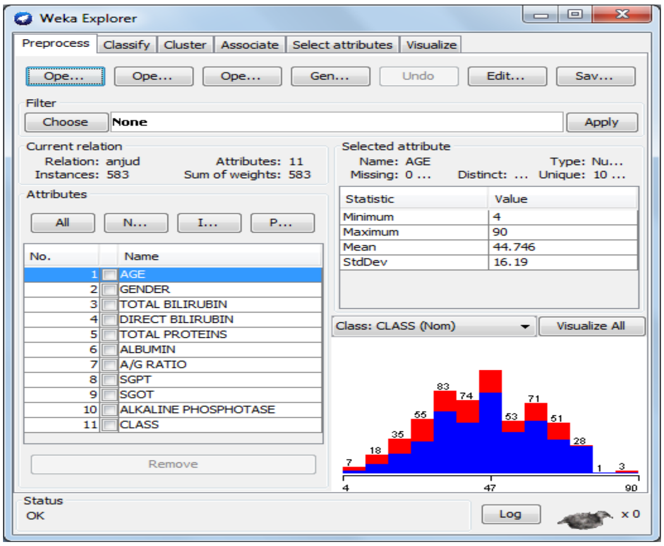
V. Results And Discussion

Preparing The Database [24] - For obtaining the result, this study used liver patient data sets from ILPD (Indian Liver Patient) Data Set. It has 583 samples with 10 independent variables and one dependent variable. Independent Variables are: Age, Gender, Total Bilirubin, Direct Bilirubin, Total Proteins, Albumin, SGPT (serum glutamic-pyruvic transaminase), SGOT (serum glutamic oxaloacetic transaminase), Alkaline Phosphatase and one dependent variable is Class.

A. Applying Classification Algorithm without Feature Selection

Applying various classification algorithms such as J-48, Multi Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest and Bayesian Network on the original Indian Liver Patient Datasets (ILPD), this comprised of all relevant and irrelevant attributes without feature selection of liver patients as shown in figure 2.

Table 1 consists of values of different Classification algorithms. According to these values the accuracy is calculated and analyzed. Performance can be determined based on the Correctly Classified Instances, Incorrectly Classified Instances, Mean absolute error and Accuracy. Comparison is made among these classification algorithms out of which SVM algorithm is considered as the better performance algorithm. Because it gives higher accuracy in respective to other classification algorithms without feature selection: with an accuracy of 71.3551%.



Filter: None

Current relation
Relation: anjud Attributes: 11
Instances: 583 Sum of weights: 583

Selected attribute
Name: AGE Type: Nu...
Missing: 0 ... Distinct: ... Unique: 10 ...

Statistic	Value
Minimum	4
Maximum	90
Mean	44.746
StdDev	16.19

Attributes

No.	Name
1	AGE
2	GENDER
3	TOTAL BILIRUBIN
4	DIRECT BILIRUBIN
5	TOTAL PROTEINS
6	ALBUMIN
7	A/G RATIO
8	SGPT
9	SGOT
10	ALKALINE PHOSPHOTASE
11	CLASS

Class: CLASS (Nom)

Status
OK

Figure 2 Hybrid model constructions before applying Feature Selection

TABLE 1
ACCURACY FOR CLASSIFICATION ALGORITHM BEFORE APPLYING FEATURE SELECTION

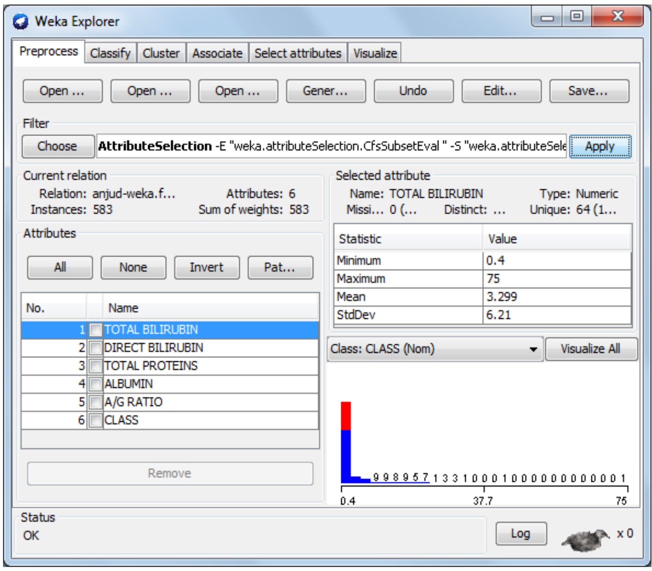
Classification Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Mean absolute error	Accuracy
J48	401	182	0.3292	68.7822
MLP	398	185	0.3458	68.2676
SVM	**416**	**167**	**0.2864**	**71.3551**
Random Forest	410	173	0.3341	70.3259
BayesNet	392	191	0.346	67.2384

B. Applying Classification Algorithm after Feature Selection

Attribute or feature selection is done with the help of greedy stepwise approach. The whole datasets of liver patients is comprised of all relevant or irrelevant attributes. By the use of feature selection, a subset (data) of liver patient from whole liver patient datasets will be obtained which comprises only significant attributes.

Applying feature selection or attribute selection using Greedy Stepwise Technique on 11 attributes. This results in the selection of 6 significant attributes as shown in figure 3.

Table 2 consists of values of different Classification algorithms. Comparison is made among these classification algorithms out of which Random Forest algorithm is considered as the better performance algorithm. Because it gives higher accuracy in respective to other classification algorithms after applying feature selection: with an accuracy of 71.8696%.



Filter

AttributeSelection -E "weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSele

Current relation

Relation: anjud-weka.f... Attributes: 6
Instances: 583 Sum of weights: 583

Selected attribute

Name: TOTAL BILIRUBIN Type: Numeric
Missi... 0 (... Distinct: ... Unique: 64 (1...

Attributes

No.		Name
1	□	TOTAL BILIRUBIN
2	□	DIRECT BILIRUBIN
3	□	TOTAL PROTEINS
4	□	ALBUMIN
5	□	A/G RATIO
6	□	CLASS

Statistic	Value
Minimum	0.4
Maximum	75
Mean	3.299
StdDev	6.21

Class: CLASS (Nom)

Status
OK

Figure 3 Hybrid model constructions after applying Feature Selection

TABLE 2
ACCURACY FOR CLASSIFICATION ALGORITHM AFTER APPLYING FEATURE SELECTION

Feature Selection		Correctly Classified Instances	Incorrectly Classified Instances	Mean absolute error	Accuracy
Classification Algorithm	J48	412	171	0.3885	70.669
	MLP	413	170	0.3455	70.8405
	SVM	416	167	0.2864	71.3551
	RandomForest	**419**	**164**	**0.3372**	**71.8696**
	Bayes Net	403	180	0.3443	69.1252

C. Comparative Analysis for Improving Prediction Accuracy

The results of classification algorithms before and after applying feature selection are compared with each other which are obtained from Table 1 and Table 2.Thus, a particular classification algorithm is identified by comparative analysis which improves prediction accuracy of liver patients.

Table 3 consists of values of different Classification algorithms. According to these values the accuracy is calculated and analyzed. Performance can be determined based on Accuracy. Comparison is made among these classification algorithms before and after applying feature selection, out of which Random Forest algorithm outperformed all other techniques with 71.8696% accuracy after applying Feature Selection.

TABLE 3
PREDICTION ACCURACY IMPROVES FOR CLASSIFICATION ALGORITHM AFTER APPLYING FEATURE SELECTION

Classification Algorithm	Accuracy	
	Before Feature Selection	After Feature Selection
J48	68.7822	70.669
MLP	68.2676	70.8405
SVM	71.3551	71.3551
RandomForest	70.3259	**71.8696**
BayesNet	67.2384	69.1252

VI. Future Work

This paper presents an approach that will be used for hybrid model construction of community health services. These classification algorithms can be implemented for other dominant diseases also like cardiac and diabetes prediction and classification. An another scope is to seeing weather by applying new algorithms will made any improvements over techniques which are used in this paper in future.

Acknowledgement

Authors would like to thanks to her head Dr. Rajan Vohra, HOD of CSE & I.T department, PDMCE, Bahadurgarh, for their valuable support and help. The authors would also like to thank the UCI Machine Learning Repository for providing the ILPD (Indian Liver Patient Dataset) Data Set.

References

[1] S. Karthik, A. Priyadarishini and J. Anuradha and B. K. Tripathy," Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types", Pelagia Research Library ,Advances in Applied Science Research, 2011.

[2] Kun-Hong Liu and De-Shuang Huang: "Cancer classification using Rotation forest". In Proceedings of the Computers in biology and medicine, 38, pages 601-610, 2008.

[3] Schiff's Diseases of the Liver, 10th Edition Copyright ©2007 Lippincott Williams & Wilkins by Schiff, Eugene R.; Sorrell, Michael F.; Maddrey, Willis C.

[4] BendiVenkataRamana, Prof. M.Surendra Prasad Babu and Prof. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", International Journal of Database Management Systems (IJDMS), Vol.3, No.2 (2011), PP.101-11.

[5] P.Rajeswari and G.SophiaReena," Analysis Of Liver Disorder Using Data Mining Algorithm", Global Journal Of Computer Science And Technology, Vol. 10 Issue 14 (Ver. 1.0) November 2010.

[6] S. Karthik, A. Priyadarishini and J. Anuradha and B. K. Tripathy," Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types", Pelagia Research Library ,Advances in Applied Science Research, 2011.

[7] BendiVenkataRamana and Prof. M.Surendra Prasad Babu," Liver Classification Using Modified Rotation Forest", International Journal of Engineering Research and Development, ISSN: 2278-067X, Volume 1, Issue 6 (June 2012), PP.17-24.

[8] A.S.Aneeshkumarand C.JothiVenkateswaran,"Estimating the Surveillance of Liver Disorder using Classification Algorithms", International Journal of Computer Applications (0975 – 8887), Volume 57– No.6, November 2012.

[9] Jankishran Pahariyavohra, Jagdeesh makhijani and sanjay patsariya, "Liver patient classification using intelligence techniques ", International journal of advanced research in computer science and software engineering, Volume 4,Issue 2,Pages 295-299.

[10] Mitchell TM. Machine learning. Boston, MA: McGraw-Hill, 1997.

[11] P.Rajeswari and G.SophiaReena," Analysis Of Liver Disorder Using Data Mining Algorithm", Global Journal Of Computer Science And Technology, Vol. 10 Issue 14 (Ver. 1.0) November 2010.

[12] Veronica S. Moertini," Towards the Use of C4.5 Algorithm For Classifying Banking Dataset",Integral, Vol.8 No.2,October2003.

[13] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

[14] Jung-Woo Ha,Classification using Weka (Brain, Computation, and Neural Learning)

[15] F.Rosenblatt, "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms", Spartan Books, Washington DC, 1961.

[16] G. Cybenko, "Approximation by superpositions of a sigmoidal function", Mathematics of Control, Signals, and Systems, Vol.2 (1989), PP. 303-314.

[17] Leo Breimen, "Random Forests", Machine Learning ,Vol. 45(2001), PP.5-32

[18] Akin Ozcift and Arif Gulten:" Classifier Ensemble Construction With Rotation Forest To Improve Medical Diagnosis Performance Of Machine Learning Algorithms". In Proceedings of the Computer Methods and Programs in Biomedicine, pages 443-451, 2011.

[19] John C. Platt," Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", Technical Report ,April 21, 1998

[20] Jihoon Yang and VasantHonavar, "Feature Subset Selection Using Genetic Algorithm", Artificial Intelligence Research Group.

[21] Isabelle Guyon and Andr´eElisseeff," An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1157-1182.

[22] Huan Liu,Hiroshi Motoda and Rudy Setiono,"Feature Selection: An Ever Evolving Frontier in Data Mining",JMLR: Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining.

[23] Andreas G. K. Janecek ,Wilfried N. Gansterer and Michael A. Demel,"On the Relationship Between Feature Selection and Classification Accuracy",JMLR: Workshop and Conference Proceedings 4: 90-105.

[24] Database collected from ILPD(Indian Liver Patient Dataset)Data set using UCI machine Learning Repository:https://archive.ics.uci.edu/ml/datasets/ILPD+ (Indian+Liver+Patient+Dataset)