A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis

Bendi Venkata Ramana[1], Prof. M.Surendra Prasad Babu[2], Prof. N. B. Venkateswarlu[3]

[1]Associate Professor, Dept.of IT, AITAM, Tekkali, A.P. India.,
ramana.bendi@gmail.com
[2]Dept. of CS&SE, Andhra University, Visakhapatnam-530 003, A.P, India.,
drmsprasadbabu@yahoo.co.in
[3]Professor, Dept. of CSE, AITAM, Tekkali, A.P., India.
Venkat_ritch@yahoo.com

Abstract:
Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. Automatic classification tools may reduce burden on doctors. This paper evaluates the selected classification algorithms for the classification of some liver patient datasets. 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.

Keywords: *Classification Algorithms, Data Mining, Liver diagnosis*

1. INTRODUCTION

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 [1]. An early diagnosis of liver problems will increase patients survival rate. Liver disease can be diagnosed by analyzing the levels of enzymes in the blood [5]. Moreover, now a day's mobile devices are extensively used for monitoring humans' body conditions. Here also, automatic classification algorithms are needed. With the help of Automatic classification tools for liver diseases (probably mobile enabled or web enabled), one can reduce the patient queue at the liver experts such as endocrinologists.

Michael J Sorich [11] reported that SVM classifier produces best predictive performance for the chemical datasets. Lung-Cheng Huang reported that Naïve Bayesian classifier produces high performance than SVM and C 4.5 for the CDC Chronic fatigue syndrome dataset [14]. Paul R Harper [12] reported that there is not necessary a single best classification tool but instead the best performing algorithm will depend on the features of the dataset to be analyzed.

DOI: 10.5121/ijdms.2011.3207

In this paper, five Classification algorithms Naive Bayes classification (NBC), C 4.5 Decision Tree, Back Propagation, K-Nearest Neighbour (KNN) and Support Vector Machines (SVM) have been considered for comparing their performance based on the liver patient data [8].

Two Liver patient datasets were used in this study, one is from Andhra Pradesh state of India and the second one is BUPA Liver Disorders datasets taken from University of California at Irvine (UCI) Machine Learning Repository [2]. In this experimentation, 10-fold cross-validation methods for comparison of selected classification algorithms have been used. This paper concentrates on performance of classification algorithms with different features combinations such as SGOT, SGPT and ALP with the selected datasets.

2. CLASSIFICATION ALGORITHMS

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 [13].

2.1 Naive Bayes Algorithm:

Bayesian Classifiers are statistical classifiers based on bayes theorem. Bayesian classification is very simple and it shows high accuracy and speed when applied to large data bases. It works on one assumption that is the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence [3].

Bayesian classification can predict class membership probabilities, such as probability that a given tuple belongs to a particular class [6]. The Naïve Bayesian classification predicts that the tuple *X* belongs to the class $C_{i.}$ Using the formula

$$P\left(C_i/X\right) = \frac{P\left(X/C_i\right)P(C_i)}{P(X)}$$

Where P (C_i /*X*) is *maximum posteriori hypothesis* for the class C_i.

As *P*(*X*) is constant for all classes, only $P(X/\ C_i)\ P\ (C_i)$ needed to be maximized.

If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is,

$$P(C_1) = P(C_2) = \ldots.. = P(Cm).$$
$$\mathrm{P}(C_i\ /X) = P(X_j/\ C_i).$$

Otherwise

$$P(C_i\ /X) = \mathrm{P}(X/C_i)P(C_i).$$

Note that the class prior probabilities may be estimated by $P\ (C_i) = |C_i,\ D|/|\mathrm{D}|$, where $|C_i,\ D|$ is the number of training tuples of class C_i in *D*.

Given datasets with many attributes, it would be extremely computationally expensive to compute $P(X/C_i)$. In order to reduce computation in evaluating $P(X/C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple i.e., that there are no dependence relationships among the attributes.

Thus,

$$P(X / C_i) = \prod_{k=1}^{n} p(x_k / c_i) \qquad (5)$$

$$= P\ (X_1|C_i) \times P\ (X_2|C_i) \times \ldots\ P\ (X_n|C_i)$$

Probabilities $P(X_1/C_i)$, $P(X_2/C_i)$,…. are easily estimated from the training tuples. Recall that that here X_k refers to the value of attribute A_k for tuple X which may be categorical or continuous-valued.

2.2 C4.5 Algorithm:

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. 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 statistical classifier. 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,... of already classified samples. Each sample S_i = X_1, X_2,... is a vector where X_1,X_2,... represent attributes or features of the sample. The training data is augmented with a vector C = C_1, C_2,.. Where C_1, C2, represent the class to which each sample belongs. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

2.3 Back propagation Algorithm:

The Back Propagation Algorithm is a multi-layered Neural Networks for learning rules [4], credited to Rumelhart and McClelland. It produces a prescription for adjusting the initially randomized set of synaptic weights such that to maximize the difference between the neural network's output of each input fact and the output with which the given input is known (or desired) to be associated. Back propagation is a supervised learning algorithm and is mainly used by Multi-Layer- perceptron to change the weights connected to the net's hidden neuron layer(s).

The back propagation algorithm uses a computed output error to change the weight values in backward direction[12]. To get this net error, a forward propagation phase must have been done before. The neurons are being activated using the sigmoid activation function while propagating in forward direction.

2.4 K-Nearest Neighbor Algorithm:

K-Nearest neighbor algorithm (KNN) is one of the supervised learning algorithms that have been used in many applications in the field of data mining, statistical pattern recognition and many others. It follows a method for classifying objects based on closest training examples in the feature space.

An object is classified by a majority of its neighbors. K is always a positive integer. The neighbors are selected from a set of objects for which the correct classification is known. The K-nearest neighbors' algorithm is as follows:

1. Determine the parameter K i.e., number of nearest neighbors beforehand.
2. Distance between the query-instance and all the training samples is calculated using any distance measure algorithm.
3. Distances for all the training samples are sorted and nearest neighbor based on the K-th minimum distance is determined.
4. Since the K-NN is supervised learning, get all the Categories of your training data for the sorted value which fall under K.
5. The prediction value is measured by using the majority of nearest neighbors.

2.5 Support Vector Machines (SVM) Algorithm:

A Support Vector Machine (SVM) separates the data into two categories of performing classification and constructing an N-dimensional hyper plane. These models are closely related to neural networks. In fact, this model uses a sigmoid kernel function which is equivalent to a two-layer, perceptron neural network.

These models are closely related to classical multilayer perceptron neural networks. By using a kernel function, these are an alternative training method for polynomial, radial basis function and multi-layer perceptron classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training.

In the SVM literature, a predictor variable which is called an attribute and a transformed attribute that is used to define the hyper plane is called a feature[11]. Here, choosing the most suitable representation can be taken as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. The goal of this modeling is to find the optimal hyper plane which separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyper plane are the support vectors.

3. RESULTS AND DISCUSSION:

Performance of Selected classification algorithms were evaluated with two datasets. First dataset contains 751 liver patient records from Andhra Pradesh, India with 12 attributes as shown in Table 1. Second dataset is taken from University of California at Irvine (UCI) Machine Learning Repository [2], which contains 345 records with 5 attributes as shown in Table 2. For the purpose of experimentation, Weka© Data Mining open source machine learning software [7].
It is used on i7 processor with 4 GB RAM. With Each algorithm, we have observed Accuracy, Precision, Sensitivity and Specificity which can be defined as follows:

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

$$\text{Accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

Sensitivity: Sensitivity is also referred as True positive rate i.e the proportion of positive tuples that are correctly identified.

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

Precision: precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

$$\text{Precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}}$$

Specificity: Specificity is the True negative rate that is the proportion of negative tuples that are correctly identified [9].

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

Table 1: AP Liver dataset and attributes

Attribute	Type
Gender	Categorical
Age	Real number
Total_bilirubin	Real number
Direct_ bilirubin	Real number
Indirect_ bilirubin	Real number
Total_protiens	Real number
Albumin	Real number
Globulin	Real number
A/G ratio	Real number
SGPT	Integer
SGOT	Integer
ALP	Integer

Table 2: UCLA Liver dataset and attributes available

Attribute	Type
Mcv	Integer
Alkphos	Integer
SGPT	Integer
SGOT	Integer
Gammagt	Real number

Our previous Experimentations motivates us to use 10 fold cross validation in our analysis with each of the selected algorithms. That is, each dataset was divided into ten parts out of which nine parts were used as training set and the remaining part is used as testing set. Repeating these ten folds ensures that each part is used for training and testing thus minimizing the sample bias.

The performance of Naive Bayes, C 4.5, Back Propagation, K-NN and SVM Classification Algorithms are analyzed with AP dataset. The features are ranked based on priority using the ranking algorithm available in Weka tool. Ordering of the attributes are Total Bilirubin, Direct Bilirubin, Indirect Bilirubin, Albumin, SGOT, SGPT, Gender, Age, Total Proteins, Globulin, A/G Ratio and ALP given in table 3. Performances of Classification Algorithms for the different feature set combination are presented in the following Tables 3-11. The Sensitivity is high for Naïve Bayes for the different feature set combination that indicates Naïve Bayes identifies high true positive rate. Specificity is higher than Accuracy, Precision and Sensitivity that indicates true negative rate is high. Sensitivity increases upto first 8 ordered features and from 9th feature it decreases and constant for the remaining features that indicates true positive rate is high upto first 8 ordered features and from 9 th feature onwards they are unimportant. SVM classifier shows more precision for almost all set of features that indicates percentage of tuples labeled as “Liver damage” that actually are “Liver damage tuples. We can find the parameters accuracy etc all improving with the addition of new attribuites. This is observed with all the selected algorithms. However K-NN algorithm is identified to be giving better accuracies with all feature set combination.

Table 3: Ordering of attributes using ranking algorithm

Attribute	Rank
Total_bilirubin	1
Direct_ bilirubin	2
Indirect_ bilirubin	3
Albumin	4
SGOT	5
SGPT	6
Gender	7
Age	8
Total_protiens	9
Globulin	10
A/G ratio	11
ALP	12

Table 4: Performance of Classification Algorithms for first 4 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	95.33	87.17	94.44	95.62
C 4.5	96.93	97.57	89.44	99.29
Back Propagation	98.002	99.4	92.22	99.82
K-NN	97.203	95.42	92.77	98.59
SVM	97.73	98.22	92.22	99.47

Table5: Performance of Classification Algorithms for first 5 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	95.2	86.36	95	98.27
C 4.5	96.13	94.67	88.88	98.42
Back Propagation	97.73	97.66	92.77	99.29
K-NN	97.33	94.44	94.44	98.24
SVM	97.86	98.8	92.22	99.64

Table 6: Performance of Classification Algorithms for first 6 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	94.94	85.14	95.55	94.74
C 4.5	96.27	95.78	88.33	98.77
Back Propagation	97.73	97.66	92.77	99.29
K-NN	97.2	94.41	93.88	98.24
SVM	97.07	97.02	90.55	99.12

Table 7: Performance of Classification Algorithms for first 7 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	95.07	85.22	96.11	97.74
C 4.5	96.67	95.85	90	98.77
Back Propagation	97.73	98.22	92.22	99.47
K-NN	97.2	94.41	93.88	98.24
SVM	97.73	98.22	92.22	99.47

Table 8: Performance of Classification Algorithms for first 8 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	95.07	85.22	96.11	94.74
C 4.5	96.67	95.85	90	98.77
Back Propagation	97.73	97.66	92.77	99.29
K-NN	97.86	96.06	95	98.77
SVM	96.93	98.15	88.88	99.47

Table 9: Performance of Classification Algorithms for first 9 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	95.2	86.36	95	95.27
C 4.5	96.67	95.85	90	98.77
Back Propagation	97.73	97.66	92.77	99.29
K-NN	97.73	95.53	95	98.59
SVM	96.93	98.15	88.88	99.47

Table 10: Performance of Classification Algorithms for first 10 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	95.07	85.92	95	95.09
C 4.5	94.4	94.73	90	98.42
Back Propagation	97.6	97.09	92.77	99.12
K-NN	97.73	95.53	95	98.59
SVM	96.93	98.15	88.88	99.47

Table11: Performance of Classification Algorithms for first 11 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	95.6	87.69	95	95.79
C 4.5	96.27	94.7	89.44	98.42
Back Propagation	96.8	94.82	91.66	98.42
K-NN	97.47	95.48	93.88	95.59
SVM	97.47	97.07	92.22	99.12

Table12: Performance of Classification Algorithms for first 12 ordered features of AP dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	95.07	85.92	95	95.09
C 4.5	96.27	94.7	89.44	98.42
Back Propagation	96.93	95.37	91.66	98.59
K-NN	97.47	95.48	93.88	98.59
SVM	97.07	97.02	90.55	99.12

Figure 1-4 shows accuracy, precision, sensitivity and specificity of the selected algorithms with AP dataset.We have carried out similar experiments with UCLA dataset. Table 15 shows observed parameters with all the available attributes. Always accuracy is high for both AP Liver dataset compared to UCLA Liver dataset [10]. Kenal Polat [10] also reported their observations of UCLA dataset which all matching our observations. We decided to explore why we are getting high accuracy rate around 98 with attribute set. Similar high accuracy levels are also observed with Taiwan dataset[1]. Table 13 shows the attributes available in Taiwan dataset.

Table 13: TAIWAN Liver dataset and attributes

Attribute	Type
Gender (Sex)	Categorical
Age	Integer
Total_bilirubin	Real number
Direct_ bilirubin	Real number
Total_protiens	Real number
Albumin	Real number
Globulin	Real number
A/G ratio	Real number
SGPT	Integer
SGOT	Integer
ALP	Integer
Gamma-glutamyl transpeptidase)	Integer
Alpha- fetoprotein	Integer
Blood type	Categorical
HBsAg	Categorical
HBeAg	Categorical
Anti-HBs	Categorical
Anti-HBe	Categorical
Anti-HBc	Categorical
Anti-HCV	Categorical

The common attributes for AP and Taiwan Liver datasets are Age, Sex, SGOT, SGPT, ALP, Total Bilirubin, Direct Bilirubin, Total Proteins and Albumin and where as the common features for AP and UCLA dataset are SGOT, SGPT and ALP.

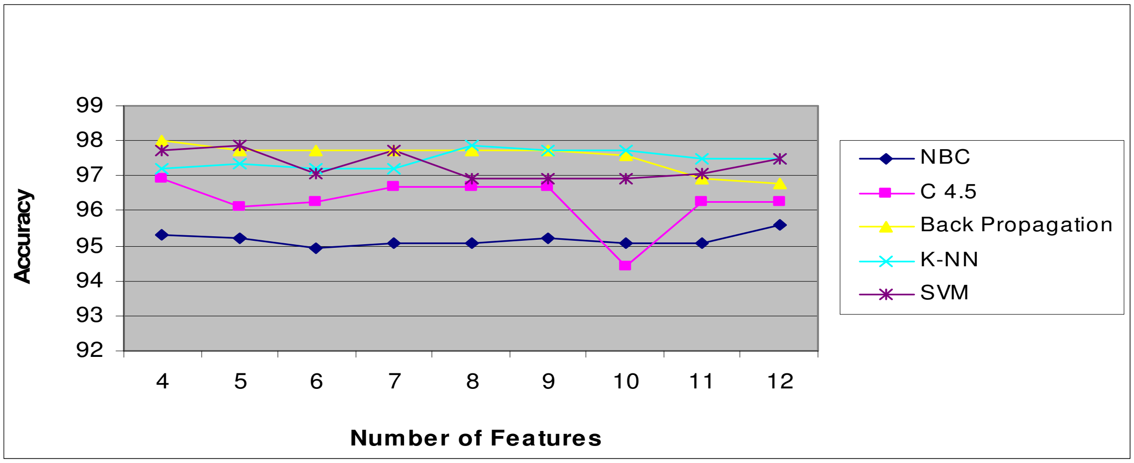

Fig 1: Accuracy for selected Classification Algorithms

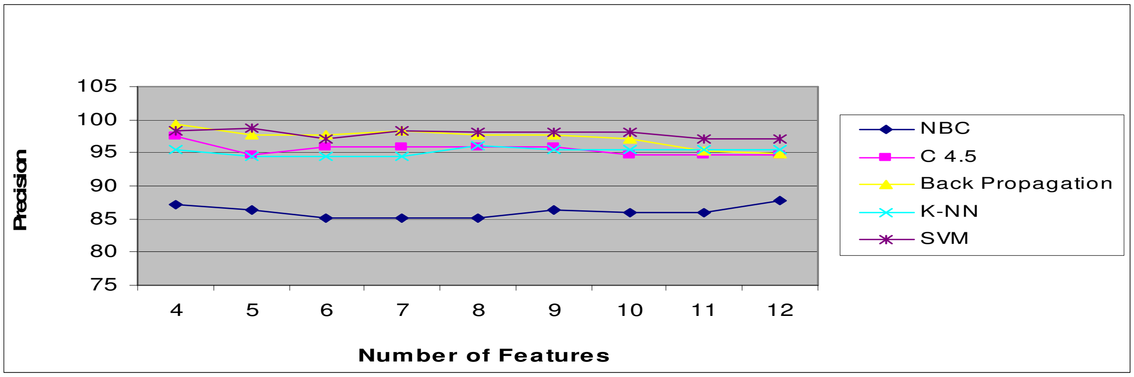

Fig 2: Precision for selected Classification Algorithms

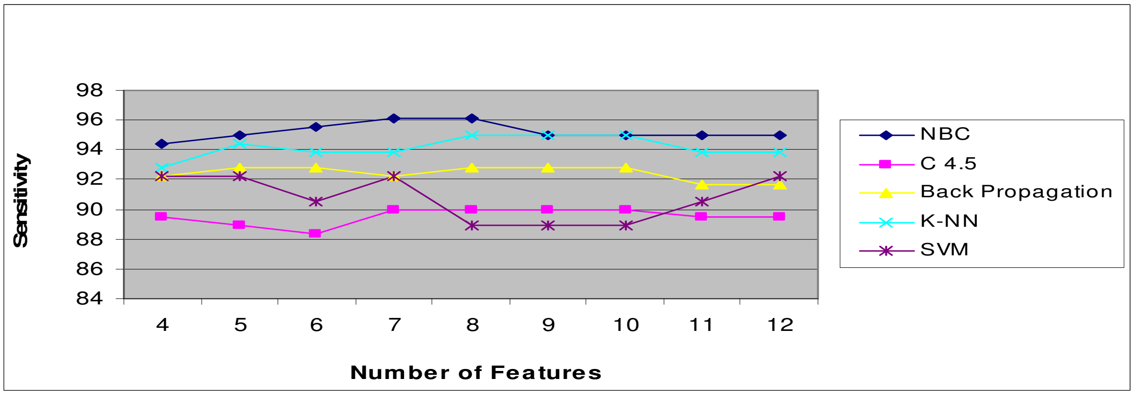

Fig 3: Sensitivity for selected Classification Algorithms

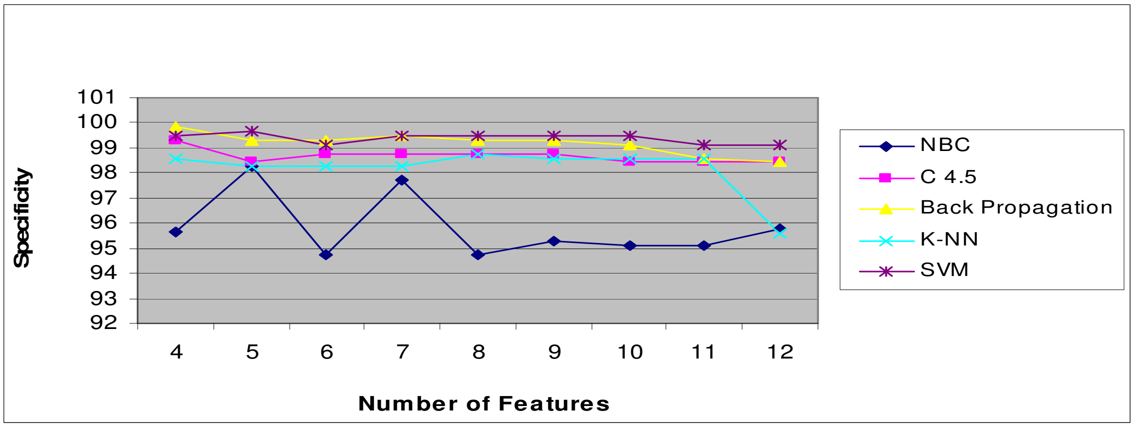

Fig 4: Specificity for selected Classification Algorithms

We have taken common attributes (SGOT, SGPT, ALP) of both AP and UCLA datasets and carried out our experimentation. Table 14 shows the observations with AP dataset, observed parameters are very good. Which indicates the three common attributes are vital in diagnosis of liver diseases. However, poor results with UCLA dataset can be attributed to the limited number of samples compared to AP dataset.

Table 14: Performance of Classification Algorithms

Classification Algorithms	**Accuracy**		**Precision**		**Sensitivity**		**Specificity**	
	AP Liver Dataset	**UCLA Liver Dataset**	**AP Liver Dataset**	**UCLA Liver Dataset**	**AP Liver Dataset**	**UCLA Liver Dataset**	**AP Liver Dataset**	**UCLA Liver Dataset**
NBC	**95.6**	**51.59**	**86.56**	**45.17**	**96.66**	**71.03**	**95.27**	**37.5**
C 4.5	**96.4**	**55.94**	**95.8**	**47.71**	**88.88**	**50.34**	**98.77**	**60**
Back Propagation	**95.73**	**66.66**	**92.52**	**62.71**	**89.44**	**51.03**	**97.72**	**78**
K-NN	**97.47**	**57.97**	**95.48**	**0**	**93.88**	**0**	**98.59**	**1**
SVM	**96.27**	**62.6**	**92.22**	**55.47**	**92.22**	**55.86**	**97.54**	**67.5**

Figure 5 to 8 shows performance of selected algorithms with both datasets with the common attributes.

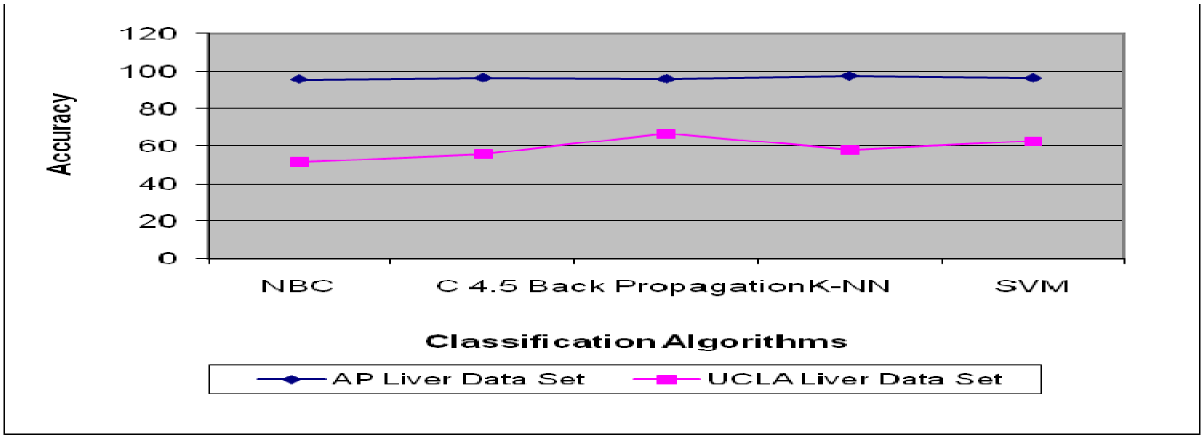

Fig 5: Accuracy for selected classification algorithms

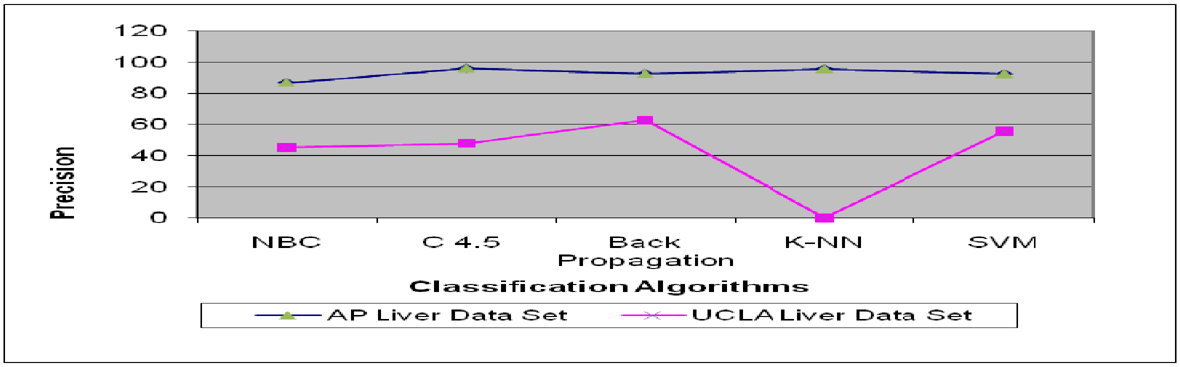

Fig 6: Precision for selected Classification Algorithms

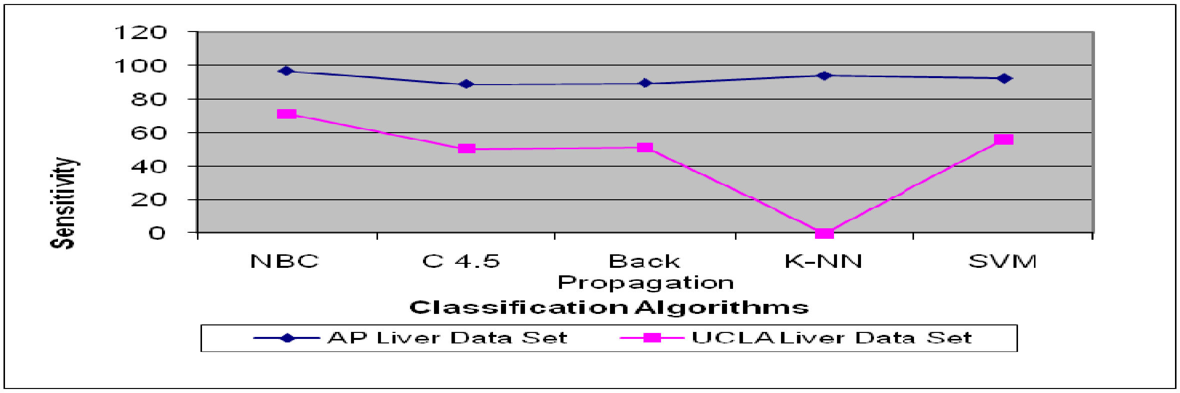

Fig 7: Sensitivity for selected Classification Algorithms

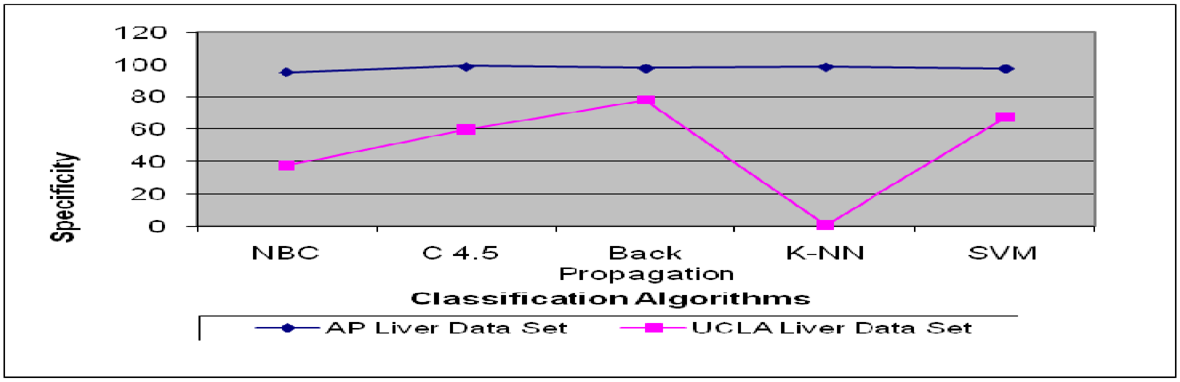

Fig 8: Specificity for selected Classification Algorithms

Table 15: Performance of Classification Algorithms with all features of UCLA Liver Dataset

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
NBC	56.52	48.91	77.93	41
C 4.5	68.69	65.81	53.1	80
Back Propagation	71.59	69.74	57.24	82
K-NN	62.89	55.78	56.51	67.5
SVM	58.26	1	68.9	1

4. CONCLUSIONS:

In this study, popular Classification Algorithms were considered for evaluating their classification performance in terms of Accuracy, Precision, Sensitivity and Specificity in classifying liver patients dataset. Accuracy, Precision, Sensitivity and Specificity are better for the AP Liver Dataset compared to UCLA liver datasets with all the selected algorithms. This can be attributed to more number of useful attributes like Total bilirubin, Direct bilirubin, Indirect bilirubin, Albumin, Gender, Age and Total proteins are available in the AP liver dataset compared to the UCLA dataset. The common attributes for AP liver data and Taiwan data are Age, Sex, SGOT, SGPT, ALP, Total Bilirubin, Direct Bilirubin, Total Proteins and Albumin are crucial in deciding liver status. . With the selected dataset, KNN, Back propagation and SVM are giving better results with all the feature set combinations.

5. ACKNOWLEDGEMENTS

We take this opportunity with much pleasure to thank Dr. Bevera. Lakshmana Rao for his help during the collection of data and in labeling samples.

REFERENCES

[1] Rong-Ho Lin. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine 2009;47:53—62.

[2] BUPA Liver Disorders Dataset. UCI repository of machine learning databases. Available from ftp://ftp.ics.uci.edu/pub/machine-learningdatabases/ liverdisorders/bupa.data, last accessed: 07 October 2010.

[3] Prof.M.S.Prasad Babu, Bendi Venkata Ramana, Boddu Raja Sarath Kumar, New Automatic Diagnosis of Liver Status Using Bayesian Classification

[4] Paul R. Harper, A review and comparison of classification algorithms for medical decision making

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

[6] P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29 (2–3) (1997) 103–130.

[7] Weka-3-4-10jre : data mining with open source machine learning software © 2002-2005 David Scuse and University of Waikato

[8] 16th Edition HARRISON’S PRINCIPLES of Internal Medicine

[9] Wendy Webber Chapman,* Marcelo Fizman,† Brian E. Chapman,‡ and Peter J. Haug†, A Comparison of Classification Algorithms to Automatically Identify Chest X-Ray Reports That Support Pneumonia.

[10] Kemal Polat, Seral Sahan, Halife Kodaz and Salih Gunes, Breast Cancer and Liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism.

[11] Michael J. Sorich,† John O. Miners,*,‡ Ross A. McKinnon,† David A. Winkler,§ Frank R. Burden,| and Paul A. Smith‡ Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP- Glucuronosyltransferase Isoforms

[12] Paul R. Harper, A review and comparison of classification algorithms for decision making

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

[14] Lung-Cheng Huang, Sen- Yen Hsu and Eugene Lin, A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data (2009).