American Journal of Intelligent Systems 2014, 4(1): 9-14
DOI: 10.5923/j.ajis.20140401.02

A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis

Esraa M. Hashem, Mai S. Mabrouk*

Biomedical Engineering, Misr University for Science and Technology (MUST University), 6th of October, Egypt

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. The liver has many essential functions, and liver disease presents a number of concerns for the delivery of medical care. Chronic liver disease (CLD) is common long-term conditions in the developed and developing world. Classification techniques are very popular in various automatic medical diagnosis tools. Early identification of the cancer has been often vital for the survival of the patients. Support vector machine (SVM) is supervised learning model with associated learning algorithms that analyze data and recognize patterns. In this work, Support vector machine is used for classifying liver disease using two liver patients datasetswith different features combinations such as SGOT, SGPT and Alkaline Phosphates, evaluating a support vector machine classifier by measuring its performance based on: accuracy, error rate, sensitivity, prevalence and specificity. Results show that the accuracy, error rate, sensitivity and prevalence at first 6ordered features are the best for ILPD dataset compared to BUPA dataset. This can be attributed to a number of useful attributes like Total bilirubin, direct bilirubin, Albumin, Gender, Age and Total proteins are available in the ILPD liver dataset compared to the BUPA dataset which can help in diagnosis of liver disease.

Keywords Chronic Liver Disease (CLD) and Support Vector Machine (SVM)

1. Introduction

The liver is an essential body organ that forms an important barrier between the gastrointestinal blood, which contains large amounts of toxins and antigens in the body. The liver produces a large amount of hormones, enzymes, and performs several functions essential to life. It is also the organ responsible for cleansing of toxins from the bloodstream, by turning them into removable substances. Liver diseaserefers to many diseases and disorders that may cause impaired liver function that can make liver decrease of its functions. The dysfunction may be primary, but the liver is often secondarily affected by disorders of other organ systems, since it is involved in many metabolic and detoxifying processes.

Hepatic fibrosis and its end stage cirrhosis are an increasing worldwide concern. Cirrhosis is the irreversible end result of fibrosis scarring and normal hepatic architecture is replaced by interconnecting bands of fibrosis tissue. The most common etiological factors resulting in cirrhosis are hepatitis B, hepatitis C, and excessive alcohol consumption [1].

Chronic HCV infection is normally a slow, progressive disease that may produce few or no symptoms for many years after infection. Some patients develop chronic infection and suffer no significant liver damage, while others progress quickly to liver cirrhosis and may develop hepatocellular carcinoma[2]. Patients with chronic liver diseases belong to a high-risk group for hepatocellular carcinoma and should be followed up regularly for early diagnosis.

Chronic HCV infection is the major cause of cirrhosis and hepatocellular carcinoma (HCC). In this condition, alpha fetoprotein levels may be elevated. The incidence of hepatocellular carcinoma is rising, and this trend is expected to continue for years[3]. Figure 1, show that liver cancer is the most cause of death in Egypt among other types of cancer.

According to the current studies, the majority of HCC patients contracted the disease from the accumulation of genetic abnormalities, probably induced by exterior etiological factors especially HBV and HCV infections[4]. These risk factors can induce mutations and damage in DNA sequences, such as p53 mutation induced by aflatoxin and DNA damage induced by the intrusion of the HBV genome[5]. The important thing in preventing liver cancer is to prevent hepatitis virus infection and eliminate hepatitis virus in chronic hepatitis patients.

Automatic classification tools may reduce the burden on doctors. Data classification is a two phase process in which first step is the training phase where the classifier algorithm

* Corresponding author:
msm_eng@k-space.org (Mai S. Mabrouk)

Published online at http://journal.sapub.org/ajis

Copyright © 2014 Scientific & Academic Publishing. All Rights Reserved

builds a classifier with the training set of dataset the second phase is classification phase where the model is used for classification and its performance is analyzed with the testing set of datasets[6].

Existing feature selection methods broadly fall into two categories, filter methods and wrapper methods. Filter methods select features based on some discriminate criteria that rely on the characteristics of data and are independent of any classification algorithms[7]. Wrapper methods use the predictive accuracy of predetermined classification algorithms as the criteria to determine the goodness of a subset of features[8, 9].

Most wrapper methods adopt sophisticated multivariate machine learning tools such as SVMs that take the combinatorial effects of features into account. These have been shown in many experiments to be more powerful in terms of classification accuracy than the filter methods[10].

Support Vector Machines proved to be effective for a lot of classifications problems. For binary-class classification, SVM constructs an optimal separating hyper plane between the positive and negative classes with the maximal margin. It can be formulated as a quadratic programming problem involving inequality constraints[11, 12].

SVMs are one of the most promising machine learning algorithms and there are many examples, where SVMs are used successfully, e.g. text classification, face recognition, and Bioinformatics. On these data sets SVMs perform very well and often outperform other traditional techniques[13]. SVMs have gained an enormous popularity in statistics, learning theory, and engineering[14, 15], and the many references therein. With a few exceptions most support vector learning algorithms have been designed for binary problems. A few attempts have been made to generalize SVM to multiclass problems[16].

In this work, a support vector machine is used as a classification algorithm in order to compare its performance with different features combinations such as aspartate aminotransferase (SGOT), glutamic pyruvic transaminase (SGPT), andalkaline phosphatase (Alkphos) using two datasets. The first one is BUPA Liver Disorders datasets taken from the University of California at Irvine (UCI) Machine Learning Repository, and the second one is from ILPD (Indian Liver Patient Dataset), it was collected from north east of Andhra Pradesh, India.

2. Materials and Methods

The liver is one of the major targets for insulin and its count regulatory hormones, such as glucagon. HCC patients who abuse alcohol are more likely to develop cirrhosis than those who do not. The most common cause of liver disease is non-alcoholic fatty liver disease. Cirrhosis is the end-result of many liver conditions and involves severe scarring of the liver. It is associated with a progressive decline in liver function resulting in liver failure. Hepatocellular carcinoma is the most common primary cancer of the liver. There are factors that may impact progression include age, gender, chronic alcohol abuse, and quantity of virus of exposure. The disease appears to be more aggressive in patients that acquire HCV after age 40 and may be more progressive in men than women[2]. In this paper SVM classification algorithm has been applied to: BUPA liver disorders dataset and Indian Liver Patient Dataset for evaluating SVM performance with different features.

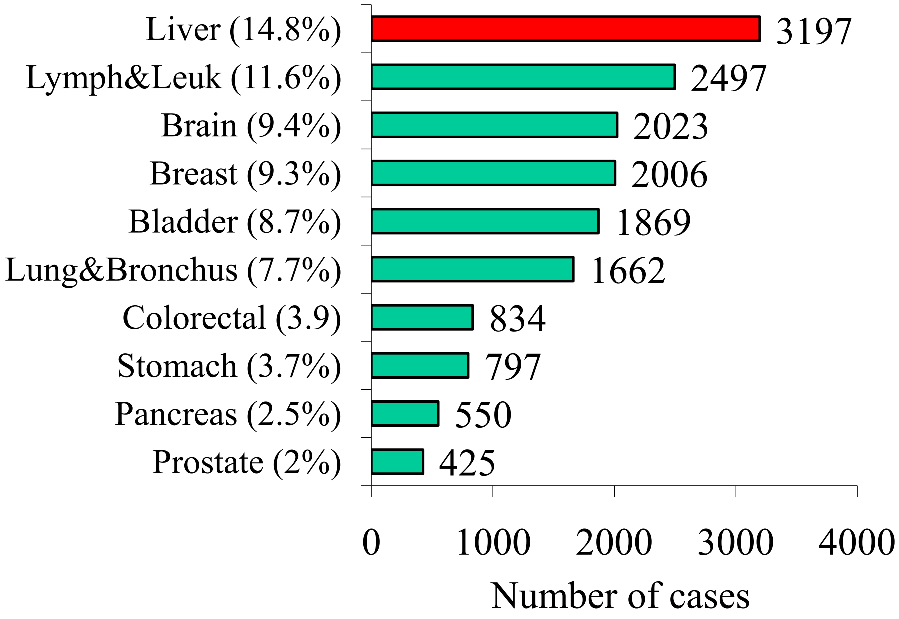

Figure 1. Egypt Mortality Statistics, Most common sites (The Cancer Database, 2001)

3. Dataset

3.1. BUPA Liver Disorders

BUPA liver disordershas 6 numeric Attributes, 345 Instances. Relevant information: The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise fromexcessive alcohol consumption, each line in the BUPA. Data file constitutes the record of a single male individual. It appears that drinks>5 is some sort of a selector on this dataset. University of California at Irvine (UCI) machine learning repository (WWW.UCI.Com).

3.2. Indian Liver Patient Dataset (ILPD)

Indian Liver Patient Dataset (ILPD) has 9 attribute, 483 Instances. 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). This data set contains 441 male patient records and 142 female patient records .this data downloaded from (WWW.UCI.Com).

4. Support Vector Machine (SVM)

Classification algorithms are widely used in various medical applications. Classification aims to build an effective model for predicting class labels of unknown data. The model is built on the training data, which consists of data points chosen from input data space and their class labels. 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 classical multilayer perceptron neural networks.

A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

There 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, the unconstrained minimization problem as in standard neural network training[17], There are many possible kernel functions and the most common kernel are: Linear, polynomial, sigmoid and radial basis function (RBF). In this paper we use linear kernel function shows in equation .1:

$$K\,(xi, xj) \;=\; xiTxj \qquad \text{eq.(1)}$$

Depending on the kernel type we choose the kernel parameters have to be set. Which kernel type performs best, depends on the application and can be determined by using cross-validation.

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[18]. Here, choosing the most suitable representation can be taken as feature selection. A set of features that describes one case is called a vector. The goal of this modeling is to find the optimal hyperplane which separates clusters of vector in such a way those cases with one category of the targetVariable are on one side of the plane and cases with the other category are on the other side of the plane. The vectors near to the hyper plane are the support vectors[19] as in figure 2.

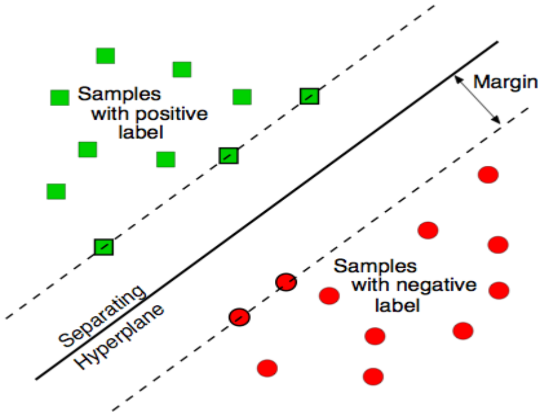

Figure 2. Maximum margin, the vectors on the dashed line are the support vectors[20]

5. Evaluation

To evaluate performance of SVM, accuracy, error rate, sensitivity, prevalence and specificity are calculated for each dataset. They are defined as follows:

– Error rate: The error rate of a classifier is the percentage of the test set that are incorrectly classified by the classifier.

$$\text{Error rate} = \frac{\text{Incorrectly Classified Samples}}{\text{Classified Samples}} \times 100$$

– Sensitivity: Sensitivity is referred as True positive rate.

$$\text{Sensitivity} = \frac{\text{Correctly Classified Positive Samples}}{\text{True Positive Samples}} \times 100$$

– Prevalence: Prevalence is defined as the proportion of the true positives against the entire samples results.

$$\text{Prevalence} = \frac{\text{True Positive Samples}}{\text{Total Number of Samples}} \times 100$$

– Specificity: Specificity is the correctly negative rate that is the proportion of true negative samples[20].

$$\text{Specificity} = \frac{\text{Correctly Classified Negative Samples}}{\text{True Negative Samples}} \times 100$$

– Accuracy: Accuracy is the percent of correct classifications.

$$\text{Accuracy} = 1 \text{ - Error rate}$$

Feature ranking is useful to gain knowledge of data and identify relevant features, also it helps reduce the number of features a learner has to examine and reduce errors from irrelevant features. Figure 3 summarize overall the process of this work.

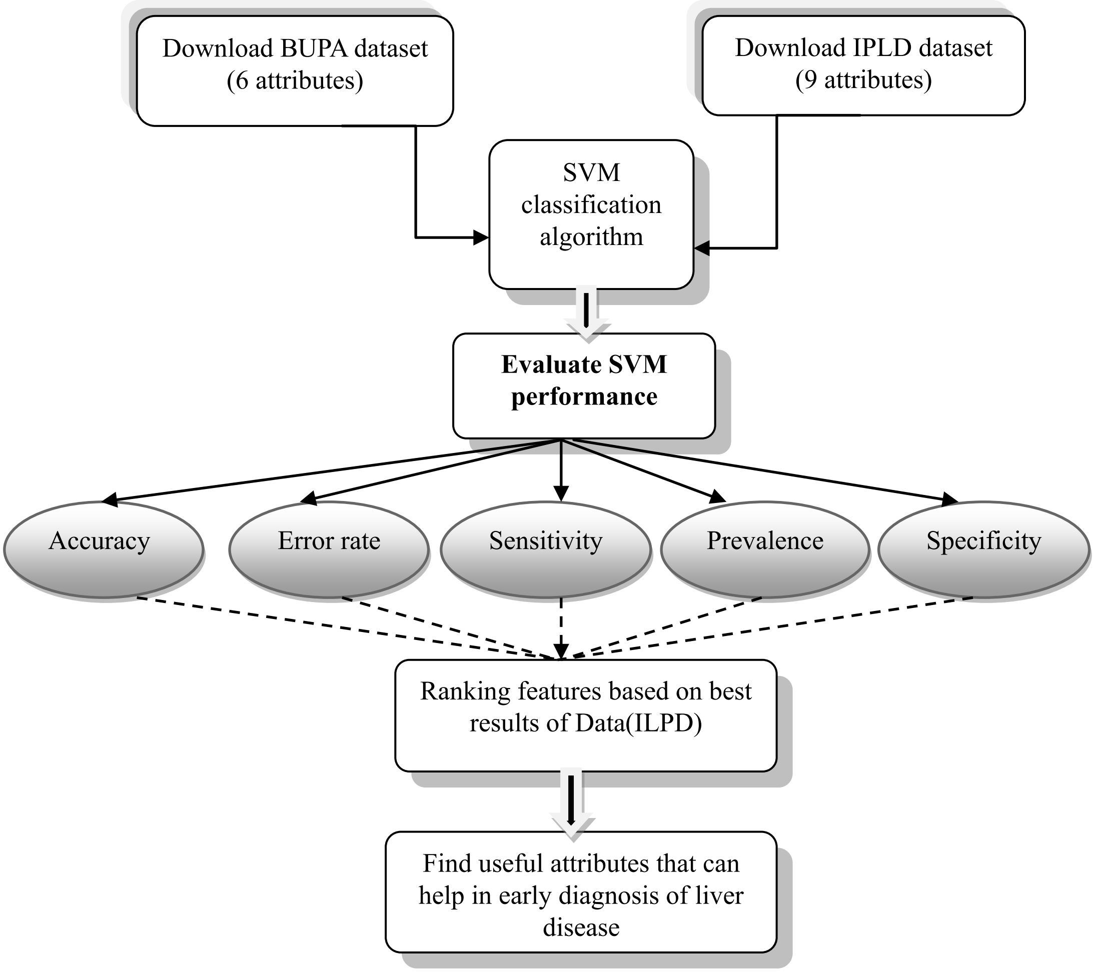

Figure 3. The overall process

6. Implementation

SVM is a new promising non-linear, non-parametric classification technique, which already showed good results in the medical diagnostics, optical character recognition, electric load forecasting and other fields. The SVM algorithm is written and implemented using MATLAB and it is also tested on the WINDOWS platform using MATLAB version 7.0 with its powerful Bioinformatics and statistics for machine learning Toolbox. The Accuracy, Prevalence, Sensitivity, Error rate and Specificity are calculated for classifying BUPA liver patient's dataset and ILPD Liver dataset using support vector machine classification algorithm.The features are ranked based on priority using the ranking algorithm available in MATLAB for each dataset.

7. Result and Discussion

Liver tumors are divided into two categories: benign and malignant. HCC is a malignant tumor derived from hepatocytes that belong to primary malignant epithelial tumors of the liver. An early diagnosis of liver problems will increase the patient's survival rate. Performance of Support vector machine classification algorithm is tested and evaluated using two datasets (BUPA liver disorders, ILPD Indian Liver Dataset) as shown in table 1, and table 2. Table 1show the attributes of BUPA liver disordersand table 2 show the attributes of ILPD IndianLiver Patient datasets.

Table 1. BUPA liver disorders dataset and available attributes

Attribute
Mcv
Alkphos
Sgpt
sgot
gammagt
Drinks of alcoholic

Table 2. ILPD Liver dataset and available attributes

Attribute
Age
TB
DB
Alkphos
Sgpt
Sgot
TP
Albumin
A/G Ratio

Table 3. Performance of SVM for number of features of BUPA Liver dataset

Num of features	**Error Rate**	**Sensitivity**	**Prevalence**	**Accuracy**	**Specificity**
first 4 ordered features	37.7%	80%	58%	63%	37.5%
first 6 ordered features	30%	75%	58%	70%	61%

Table 4. Performance of SVM for number of features of ILPD dataset

Num of features	**Error Rate**	**Sensitivity**	**Prevalence**	**Accuracy**	**Specificity**
first 4 ordered features	29%	95%	71%	71%	10.8%
first 6 ordered features	27%	96.6%	71%	73%	12%
first 8ordered features	26.8%	90%	71%	73.2%	30%

Previous experimentations motivates us to use cross validation in this analysis with SVM by randomly divide data to 50/50 training set equal to testing set. The error rate, specificity, accuracy, prevalence and sensitivity are calculated for BUPA Liver dataset as shown in table3 and the ILPD dataset as shown in table 4.

The performance of SVM Classification Algorithm is analyzed with BUPA and ILPD datasets, the Specificity at first 6 ordered features are best for BUPA dataset compared to other dataset. The Sensitivity, Error rate, Accuracy and Prevalence at first 6 ordered features are best for ILPD Liver dataset compared to BUPA dataset.

The features are ranked based on priority for each dataset. Table 5 shows ordering of the attributes of BUPA liver disorders which are: aspartate aminotransferase (SGOT), gamma-glutamyl transpeptidase (gammagt), mean corpuscular volume (mcv), alkaline phosphotase (alkphos), alamine aminotransferase (SGPT) and drinks. Ordering of the attributes of ILPD datasets as: A/G Ratio Albumin and Globulin Ratio, Direct Bilirubin (DB), Total Bilirubin (TB), Alkphos, SGPT, SGOT, Albumin, Age and Total Proteins (TP) given in table.6

Poor results with BUPA dataset can be attributed to the limited number of samples compared to ILPD dataset, However, We have taken common attributes (SGOT, SGPT, Alkphos) of both BUPA and ILPD datasets and implemented the experimentation.

SGOT aspartate aminotransferase or (AST) test is part of an initial screening for liver disease. AST is normally found in red blood cells, liver, heart, muscle tissue, pancreas, and kidneys. AST formerly was called serum glutamic oxaloacetictransaminase (SGOT). this test is done to check for liver damage and Help identify liver disease, especially hepatitis and cirrhosis.[21]

SGPT isan alanine aminotransferase (ALT) test measures the amount of this enzyme in theblood[21]. ALT is found mainly in the liver, ALT was formerly called serum glutamic pyruvic transaminase (SGPT). ALT is measured to see if the liver is damaged or diseased.

Alkphos is An alkaline phosphatase (ALP) test measures the amount of the enzyme ALP in the blood. ALP is made mostly in the live and used to help detect liver disease or bone disorders.[22]

Table 4 shows the observations with ILPD dataset, observed parameters were very good which indicates that these three common features are important for (1) detect the presence of liver disease, (2) distinguish among different types of liver disorders, (3) gauge the extent of known liver damage, and (4) follow the response to treatment.

Table 5. Ordering of attributes using ranking algorithm of BUPA dataset

Attribute	Rank
SGOT	1
gammagt	2
mcv	3
alkphos	4
SGPT	5
drinks	6

Table 6. Ordering of attributes using ranking algorithm of ILPD dataset

Attribute	Rank
A/G Ratio	1
DB	2
TB	3
Alkphos	4
SGPT	5
SGOT	6
Albumin	7
Age	8
TP	9

8. Conclusions

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An important task in cancer research is to separate healthy patients from cancer patients and to distinguish patients of different cancer subtypes, based on their cytogenetic profiles. This is known as the classification problem. These tasks help successful cancer diagnosis and treatment. Machine learning is about designing algorithms that allow a computer to learn. Support vector machine has proved to be a powerful approach for classifier design. It has become an increasingly popular tool for machine learning tasks involving classification, regression or novelty detection. The key idea of SVM is to find a hyper plane that maximizes the margin separating the two classes of instances. In this study, the SVM classification algorithm were considered in evaluating its classification performance in terms of Accuracy, Prevalence, Sensitivity, Error rate and Specificity in classifying BUPA liver patient dataset and ILPDIndian Liver dataset.

The Specificity at first 6 ordered features are best for BUPA dataset compared to ILPD dataset. The Sensitivity, Error rate, Accuracy and Prevalence at first 6 ordered features are best for ILPD Liver dataset compared to BUPA dataset, so the three common attributes (SGOT, SGPT, Alkphos) are important in diagnosis of liver diseases. This conclusion can be attributed to more number of useful attributes like Total bilirubin, direct bilirubin, Albumin, Gender, Age and Total proteins are available in the ILPD liver dataset compared to the BUPA dataset. So as to increase the number of features it improves the performance in classification algorithm that can help in early diagnosis and treatment of liver cancer.

REFERENCES

[1] K. Golla, J B. Epstein,and J. Robert. "Liver disease: Current perspectives on medical and dental management". *Medical management update,* vol. 98 , No. 5 , November 2004.

[2] T.J. Liang, B .Rehermann, L.B. Seeff, J.H. Hoofnagle. "Pathogenesis, natural history, treatment, and prevention of hepatitis C.".*Ann Intern Med*, pp132:296,vol.305 ,2000

[3] P.J. Johnson. "Hepatocellular carcinomaa: is current therapy really altering outcome". *Gut*, pp51:459, vol.62, 2002.

[4] M. S. Mabrouk, E. M. Hashem, A. Sharawy, ''Statistical Approaches for Hepatocellular Carcinoma (HCC) Biomarker Discovery", *American Journal of Bioinformatics Research*, Vol. 2 No. 6, pp. 102-109, 2012.

[5] M. S. Mabrouk, E. M. Hashem , A. Sharawy. "Discrete Stationary Wavelet Transform of Array CGH Data on Hepatocellular Carcinoma'', *Journal of Bioinformatics and Intelligent Control,*vol.1,No 2,2013.

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

[7] C.Ding and H.Peng. ''Minimum redundancy feature selection from microarray gene expression data''. In CSB '03: *Proceedings of the IEEE Computer Society Conference on Bioinformatics,* pp 523, 2003.

[8] I. Guyon, J. Weston, S. Barnhill, and V.Vapnik. "Gene selection for cancer classification using support vector machines." *Machine Learning*,vol.46(1-3),pp 389–422, 2002.

[9] K. B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje. "Multiple svmfor gene selection in cancer classification with expression data." *IEEE Trans Nanobioscience,* vol.4(3), pp228–234, September 2005.

[10] H. Chai and C.Domeniconi."An evaluation of gene selection methods for multi-class microarray data classification."*In Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics,* pp3:10, 2004.

[11] C. J. Burges. "A tutorial on support vector machines for pattern recognition." *Data Mining and Knowledge Discovery*, vol 2(2), pp121:167, 1998.

[12] N. Cristianini and J.S. Taylor. "Support Vector Machines and other Kernel-based Learning Methods". *Cambridge University Press*, 2000.

[13] K. P. Bennett and C. Campbell. "Support vector machines: hype or hallelujah" *SIGKDD Explor. Newsl*, vol.2, pp 1:13, 2000.

[14] V. N. Vapnik. ,Statistical Learning Theory. Wiley, 1998 .

[15] B. Sch¨olkopf, C. Burges, and A. Smola.", Advances in Kernel Methods,Support Vector Learning," 1998.

[16] J.Weston and C. Watkins. "Support vector machines for multi-class pattern recognition." *In Proceedings of the Seventh European Symposium on Artificial Neural Networks*, pp219:224, April 1999.

[17] B. V. Ramana, M.S. Prasad , N. B. Venkateswarlu," A Critical Study of Selected Classification algorithms for Liver Disease Diagnosis'', *International Journal of Database Management Systems (IJDMS),* Vol.3, pp 111:114, May 2011.

[18] M.J. Sorich, J. O. Miners, R.A. McKinnon, D. A. Winkler, F. R. Burden, P.l A. Smith. " Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucurono syltransferase Isoforms." *Journal of Chemical Information and Computer Sciences* ,vol.43(6).pp2019:2024,2003

[19] F. Markowetz. "Klassifikationmit support vector Machines". http://lectures.molgen.mpg.de/statistik03/docs/Kapitel 16.pdf, 2003.

[20] W. W. Chapman, M.Fizman, B. E. Chapman, and P. J. Haug, "A Comparison of Classification Algorithms to Automatically Identify Chest X-Ray Reports That Support Pneumonia *"journal of biomedical informatics*,vol.34,pp 4: 14,2001.

[21] J. David "Special Considerations in Interpreting Liver Function Tests". *American Family Physician* .vol.59 (8),pp. 2223:2230,1999.

[22] T. Mazda & W.L. Gyure "Assay of alkaline phosphatase isoenzymes by a convenient precipitation and inhibition methodology." *Chem Pharm Bull (Tokyo);* vol.36 (5), pp.1814:1818, 1988.