A Study of support vector machine algorithm for liver disease diagnosis

Esraa M. Hashem¹, Mai S. Mabrouk²
1, 2 Biomedical Engineering, Misr University for Science and Technology
(MUST University), 6th of October, Egypt
1 bioeng_esra@hotmail.com
2 msm_eng@yahoo.com

Abstract:
Diabetes and chronic liver disease (CLD) are common long-term conditions in the developed and developing world. 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. 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) are supervised learning models with associated learning algorithms that analyze data and recognize patterns. In this work, Support vector machine is used for classifying liver disease using diabetes disease dataset and two liver patents datasets , 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 and specificity at first 8 ordered features are the best for Diabetes diagnosis dataset compared to other two datasets. The sensitivity and prevalence at first 8 ordered features are the best for AP Liver dataset compared to other two datasets.

Keywords: Diabetes, chronic liver disease (CLD), and Support vector machine.

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 disease refers 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. The presence of chronic liver disease (CLD) is associated with significant impairment in glucose homeostasis. Glucose intolerance is seen in up to 80% of patients with CLD, and frank diabetes is present in 30–60% [1]. Depending on its etiology, CLD has a significant impact on hepatic glucose metabolism.

Liver disease is an important cause of death in some types of diabetic, the management of diabetes in patients with liver disease is theoretically complicated by liver-related alterations in drug metabolism, potential interactions between the drugs, and the incidence of hepatotoxicity.

Diabetes is a complex group of diseases with a variety of causes. People with diabetes have high blood glucose, also called high blood sugar or hyperglycemia. Diabetes develops when the body doesn’t make enough insulin or is not able to use insulin effectively, or both. As a result, glucose builds up in the blood instead of being absorbed by cells in the body. The body’s cells are then starved of energy despite high blood glucose levels. Over time, high blood glucose damages nerves and blood vessels, leading to complications such as heart disease, stroke, kidney disease, blindness, dental disease, and amputations. Other complications of diabetes may include increased susceptibility to other diseases, loss of mobility with aging, depression, and pregnancy problems.

Some types of diabetes are associated with a large number of liver disorders including elevated liver enzymes, fatty liver disease, cirrhosis, hepatocellular carcinoma, and acute liver failure.

Hepatic fibrosis and its end stage cirrhosis are an increasing worldwide concern. Cirrhosis is the irreversible end result of fibrous scarring and normal hepatic architecture is replaced by interconnecting bands of fibrous tissue. The most common etiological factors resulting in cirrhosis are hepatitis B, hepatitis C, and excessive alcohol consumption [2].

Hepatitis C virus (HCV) is a blood-borne, single stranded RNA flavivirus encoding for a capsid protein, two envelope proteins, and some nonstructural proteins [3]. Patients who are transplanted for HCV are more likely to develop diabetes than those who are transplanted for other liver diseases [4]. Taken together, these observations suggest that HCV may play a pathogenic role in some types of diabetes.

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 [5].

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 [6]. Figure 1, show that liver cancer is the most cause of death in Egypt among other types of cancer.

![Figure 1: Egypt Mortality Statistics, Most common sites (The Cancer Database, 2001)](image)

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Number of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver (14.8%)</td>
<td>3197</td>
</tr>
<tr>
<td>Lymph&amp;Leuk (11.6%)</td>
<td>2497</td>
</tr>
<tr>
<td>Brain (9.4%)</td>
<td>2023</td>
</tr>
<tr>
<td>Breast (9.3%)</td>
<td>2006</td>
</tr>
<tr>
<td>Bladder (8.7%)</td>
<td>1869</td>
</tr>
<tr>
<td>Lung&amp;Bronchus (7.7%)</td>
<td>1662</td>
</tr>
<tr>
<td>Colorectal (3.9%)</td>
<td>834</td>
</tr>
<tr>
<td>Stomach (3.7%)</td>
<td>797</td>
</tr>
<tr>
<td>Pancreas (2.5%)</td>
<td>550</td>
</tr>
<tr>
<td>Prostate (2%)</td>
<td>425</td>
</tr>
</tbody>
</table>

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 [7]. 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 [8].

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 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 [9].

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 [10]. Wrapper methods use the predictive accuracy of predetermined classification algorithms as the criteria to determine the goodness of a subset of features [11, 12]

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 [13].
Support Vector Machines proved to be effective for a lot of classifications problems. For binary-class classification, SVM constructs an optimal separating hyperplane between the positive and negative classes with the maximal margin. It can be formulated as a quadratic programming problem involving inequality constraints [14, 15].

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 [16]. SVMs have gained an enormous popularity in statistics, learning theory, and engineering [17, 18], 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 [19].

In this work, a support vector machine is used as a classification algorithm in order to compare its performance using three datasets. The first one is from the Diabetes diagnosis datasets, the binary-valued diagnostic variable presented in this dataset relates whether a patient showed signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The second one is ILPD (Indian Liver Patient Dataset), it was collected from north east of Andhra Pradesh, India, and the third one is BUPA Liver Disorders datasets taken from the University of California at Irvine (UCI) Machine Learning Repository.

Materials and methods

The liver is one of the major targets for insulin and its count erregulatory 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 [5]. Diabetes can increase the risk of liver cancer, especially in those who drink heavily or have viral hepatitis. In this paper SVM classification algorithm has been applied to: Diabetes diagnosis dataset, BUPA liver disorders dataset and Indian Liver Patient Dataset for evaluating SVM performance.

1. dataset:

1.1 Diabetes diagnosis dataset
Diabetes diagnosis dataset, it has 8 numeric Attributes, Nominal (binary) Class, and 768 Instances. The binary-valued diagnostic variable presented in this dataset relates whether a patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.
University of California at Irvine (UCI) Machine Learning Repository (WWW.UCI.Com).

1.2 BUPA liver disorders
BUPA liver disorders has 7 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 from excessive 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).

1.3 Indian Liver Patient Dataset (ILPD)
Indian Liver Patient Dataset (ILPD) has 10 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).

2. 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 multilayer 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 [20]. 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:

\[ K(x_i, x_j) = x_i^T x_j \]  

Comment [M1]: Eq. editor

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 [21]. 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 target Variable 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 [22] as in figure 2.

![Figure 2. Maximum margin, the vectors on the dashed line are the support vectors[23].](image)

3. 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 [23].

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

- **Accuracy**: Accuracy is the percent of correct classifications.

\[\text{Accuracy} = 1 - \text{Error rate}\]

### 4. 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 diabetes diagnosis dataset, BUPA liver patient’s dataset and ILPD Liver dataset using support vector machine classification algorithm.

### Result and discussion:

Chronic liver disease (CLD) is often associated with glucose intolerance and diabetes. 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 three datasets (Diabetes diagnosis, BUPA liver disorders, ILPD Indian Liver Dataset) as shown in table 1, table 2 and table 3. Table 1 shows the attributes of Diabetes dataset, table 2 show the attributes of BUPA liver disorders and table 3 show the attributes of Indian Liver Patient datasets.
Table 1. Diabetes diagnosis dataset and available attributes.

<table>
<thead>
<tr>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pregnancies</td>
</tr>
<tr>
<td>PG Concentration:</td>
</tr>
<tr>
<td>Diastolic BP</td>
</tr>
<tr>
<td>Tri Fold Thick</td>
</tr>
<tr>
<td>Serum Ins</td>
</tr>
<tr>
<td>BMI</td>
</tr>
<tr>
<td>DP Function</td>
</tr>
<tr>
<td>Age</td>
</tr>
</tbody>
</table>

Table 2. BUPA liver disorders dataset and available attributes.

<table>
<thead>
<tr>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mcv mean corpuscular volume</td>
</tr>
<tr>
<td>Alkphos alkaline phosphotase</td>
</tr>
<tr>
<td>Sgpt alamine aminotransferase</td>
</tr>
<tr>
<td>sgot aspartate aminotransferase</td>
</tr>
<tr>
<td>gammagt gamma-glutamyl transpeptidase</td>
</tr>
<tr>
<td>Drinks number of half-pint equivalents of alcoholic beverages drunk per day</td>
</tr>
<tr>
<td>selector field used to split data into two sets</td>
</tr>
</tbody>
</table>
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 Diabetes diagnosis dataset as shown in table 4, BUPA Liver dataset as shown in table 5 and the ILPD dataset as shown in table 6. The Error rate, Accuracy, and Specificity in Diabetes diagnosis dataset give best percentage compared to other two datasets as shown in table 4. Sensitivity increases up to first 8 ordered features in the diabetes diagnosis dataset in table 4. Error rate decrees and specificity increase up to first 8 ordered features in a BUPA Liver dataset as shown in table 5. The Sensitivity and Prevalence in ILPD dataset give best percentage compared to other two datasets as shown in table 6.

<table>
<thead>
<tr>
<th>Num of features</th>
<th>Error Rate</th>
<th>Sensitivity</th>
<th>Prevalence</th>
<th>Accuracy</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>first 4 ordered features</td>
<td>27%</td>
<td>45%</td>
<td>35%</td>
<td>73%</td>
<td>88%</td>
</tr>
<tr>
<td>first 6 ordered features</td>
<td>21%</td>
<td>59%</td>
<td>35%</td>
<td>79%</td>
<td>89%</td>
</tr>
<tr>
<td>first 8 ordered features</td>
<td>21.5%</td>
<td>62%</td>
<td>35%</td>
<td>78.5%</td>
<td>87.6%</td>
</tr>
</tbody>
</table>
Table 5: Performance of SVM for number of features of BUPA Liver dataset

<table>
<thead>
<tr>
<th>Num of features</th>
<th>Error Rate</th>
<th>Sensitivity</th>
<th>Prevalence</th>
<th>Accuracy</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>first 4 ordered features</td>
<td>37.7%</td>
<td>80%</td>
<td>58%</td>
<td>63%</td>
<td>37.5%</td>
</tr>
<tr>
<td>first 6 ordered features</td>
<td>30%</td>
<td>75%</td>
<td>58%</td>
<td>70%</td>
<td>61%</td>
</tr>
<tr>
<td>first 8 ordered features</td>
<td>30%</td>
<td>75%</td>
<td>58%</td>
<td>70%</td>
<td>61%</td>
</tr>
</tbody>
</table>

Table 6: Performance of SVM for number of features of ILPD dataset

<table>
<thead>
<tr>
<th>Num of features</th>
<th>Error Rate</th>
<th>Sensitivity</th>
<th>Prevalence</th>
<th>Accuracy</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>first 4 ordered features</td>
<td>29%</td>
<td>95%</td>
<td>71%</td>
<td>71%</td>
<td>10.8%</td>
</tr>
<tr>
<td>first 6 ordered features</td>
<td>27%</td>
<td>96.6%</td>
<td>71%</td>
<td>73%</td>
<td>12%</td>
</tr>
<tr>
<td>first 8 ordered features</td>
<td>26.8%</td>
<td>90%</td>
<td>71%</td>
<td>73.2%</td>
<td>30%</td>
</tr>
</tbody>
</table>

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 diabetes diagnosis, BUPA liver patient dataset and ILPD Indian Liver dataset. The
Accuracy, Error rate and specificity at first 8 ordered features are best for Diabetes diagnosis dataset compared to other two datasets. The Sensitivity and Prevalence at first 8 ordered features are best for ILPD Liver dataset compared to other two datasets. 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
[18] B. Schölkopf, C. Burges, and A. Smola ,“, Advances in Kernel Methods,Support
Vector Learning,” 1998
recognition.” In Proceedings of the Seventh European Symposium on Artificial Neural
Classification algorithms for Liver Disease Diagnosis”, International Journal of
Smith. “Comparison of Linear and Nonlinear Classification Algorithms for the
Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase
[22] F. Markowetz. “Klassifikation mit support vector
Classification Algoritms to Automatically Identify Chest X-Ray Reports That Support