Minor Project Synopsis

" Effective analysis and diagnosis of Liver disorder using clustering and association rules in data mining "

Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Bawana Road, Delhi-110042

Submitted to: -

Ms. Indu Singh
Assistant Professor
Deptt. Of Computer Sc. & Engg.
Delhi Technological University

Submitted by:-

Ajit Singh Kushwaha (2K14/CO/006)
Bablu Singh (2K14/CO/020)
Tabrez Alam (2K14/CO/131)

Approval Certificate

THIS IS CERTIFYING THAT I APPROVED THE PROPOSAL OF MINOR PROJECT ON TOPIC " **Effective analysis and diagnosis of Liver disorder using clustering and association rules in data mining** ".

Project Guide

Ms. Indu Singh

Assistant Professor

Department of CSE DTU

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Bawana Road,
Delhi-110042

ABSTRACT

There are many disorders of the liver that require clinical care by a physician or other healthcare professional. In this project we will present classification mode of liver patient dataset in order to predict liver diagnosis. Medical data mining is one of the critical aspect of automated disease diagnosis and disease prediction. Medical Data Diagnosis involves developing data mining algorithms and techniques to analyze medical data. In recent years, liver disorders have excessively increased and liver disease are becoming one of the most fatal diseases in several countries.

In this study, one real patient dataset was investigated for building classification models in order to predict liver diagnosis. Five data mining classification algorithms were applied to the dataset and the performance of all classifiers are compared against each other in terms of accuracy, precision, and recall. Several investigations have also been carried out to improve performance of the classification models. Finally, the results shown promising methodology in diagnosing liver disease during the earlier stages.

This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector.field 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. Any patient whose age exceeded 89 is listed as being of age "90". It contains 11 attributes which are :

- Age — Age of the patient
- Gender — Gender of the patient
- TB — Total Bilirubin
- DB — Direct Bilirubin
- Alkphos — Alkaline Phosphotase
- Sgpt Alamine — Aminotransferase
- Sgot — Aspartate Aminotransferase
- TP — Total Protiens
- ALB — Albumin
- A/G Ratio — Albumin and Globulin Ratio
- Selector field — used to split the data into two sets (labeled by the experts)

In this Dataset Selector field is made as a factor field on which all the algorithms make their decisions. Every dataset contains atleast one factor field which have some levels in this dataset levels are two in numbers. 1stands for the liver patient and 2 for the non-liver patient.

INTRODUCTION

The liver is the largest glandular organ of the body. It weighs about 3 lb (1.36 kg). It is reddish brown in colour and is divided into four lobes of unequal size and shape. The liver lies on the right side of the abdominal cavity beneath the diaphragm. Blood is carried to the liver via two large vessels called the hepatic artery and the portal vein. The heptic artery carries oxygen-rich blood from the aorta (a major vessel in the heart). The portal vein carries blood containing digested food from the small intestine. These blood vessels subdivide in the liver repeatedly, terminating in very small capillaries. Each capillary leads to a lobule. Liver tissue is composed of thousands of lobules, and each lobule is made up of hepatic cells, the basic metabolic cells of the liver. This paper describes disorder of liver can cause an acute or chronic inflammation and weakness of the liver and may even harm other organs in the body, alcohol induced liver disease remains a major problem. This paper also describes the blood test taken when a person is affected to liver disorder such as alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, gamma-glutamyl transpeptidase.

II. CAUSES OF LIVER DISEASE

When the liver becomes diseased, it may have many serious consequences. Liver disease (also called hepatic disease) is a broad term describing any single number of diseases affecting the liver. Many are accompanied by jaundice caused by increased levels of bilirubin in the system. The bilirubin results from the breakup of the hemoglobin of dead red blood cells; normally, the liver removes bilirubin from the blood and excretes it through bile.

1) Disease of Liver
   Several diseases states can affect the liver. Some of the diseases are Wilson's disease, hepatitis (an inflammation of the liver), liver cancer, and cirrhosis (a chronic inflammation that progresses ultimately to organ failure). Alcohol alters the metabolism of the liver, which can have overall detrimental effects if alcohol is taken over long periods of time. Hemochromatosis can cause liver problems.

2) Common Liver Disorder
   Fatty liver (also known as steatorrhoeic hepatosis or steatosis hepatitis) is a reversible condition where large vacuoles of triglyceride fat accumulate in liver cells via the process of steatosis. It can occur in people with a high level of alcohol consumption as well as in people who never had alcohol. Hepatitis (usually caused by a virus spread by sewage contamination or direct contact with infected body fluids). Cirrhosis of the liver is one of the most serious liver diseases. It is a condition used to denote all forms of diseases of the liver characterized by the significant loss of cells. The liver gradually contracts in size and becomes leathery and hard. The regenerative activity continues under liver cirrhosis but the progressive loss of liver cells exceeds cell replacement. Liver cancer. The risk of liver cancer is higher in those who have cirrhosis or who have had certain types of viral hepatitis; but more often, the liver is the site of secondary (metastatic) cancers spread from other organs.

3) Symptoms of Liver Disorder

The external signs include a coated tongue, itchy skin, excessive sweating, offensive body odor, dark circles under the eyes, red swollen and itchy eyes, acne rosacea, brownish spots and blemishes on the skin, flushed facial appearance or excessive facial blood vessels. Other symptoms include jaundice, dark urine, pale stool, bone loss, easy bleeding, itching, small, spider-like blood vessels visible in the skin, enlarged spleen, and fluid in the abdominal cavity, chills, pain from the biliary tract or pancreas, and an enlarged gallbladder.

The symptoms related to liver dysfunction include both physical signs and a variety of symptoms related to digestive problems, blood sugar problems, immune disorders, abnormal absorption of fats, and metabolism problems. Nervous system disorders include depression, mood changes, especially anger and irritability, poor concentration and "foggy brain", overheating of the body, especially the face and torso, and recurrent headaches (including migraine) associated with nausea. The blood sugar problems include a craving for sugar, hypoglycaemia and unstable blood sugar levels, and the onset of type 2 diabetes.Abnormalities in the level of fats in the blood stream, whether too high or too low levels of lipids in the organism. Hypercholesterolemia: elevated LDL cholesterol, reduced HDL cholesterol, elevated triglycerides, clogged arteries leading to high blood pressure heart attacks and strokes, build up of fat in other body organs (fatty degeneration of organs), lumps of fat in the skin (lipomas and other fatty tumors), excessive weight gain (which may lead to obesity), inability to lose weight even while dieting, sluggish metabolism, protuberant abdomen (pot belly), cellulite, fatty liver, and a roll of fat around the upper abdomen (liver roll) etc. Or too low levels of lipids: hypocholesterolemia: low total cholesterol, low LDL, VLDL cholesterol and low triglyderides.

Symptoms may include:

- Jaundice
- Tendency to bruise or bleed easily
- Ascites
- Impaired brain function
- General failing health

Analysis

Hepatitis is an inflammation of the liver that can be caused by a virus, inherited disorders, and sometimes by certain medications or toxins such as alcohol and drugs. Scientists have identified four main types of viral hepatitis: hepatitis A, hepatitis B, and hepatitis C, and hepatitis D. A fifth type, hepatitis E, is generally not found in North America.

Hepatitis A is waterborne and spread mainly via sewage and contaminated food and water. Hepatitis B is transmitted by contact with infected semen, blood, vaginal secretions and from mother to newborn. Hepatitis B is most commonly spread by unprotected sex and by sharing of infected needles (including those used for tattooing, acupuncture, and ear piercing). Hepatitis C spreads via direct blood-to-blood contact. Hepatitis D is spread by infected needles and blood transfusions.

Improved screening of donated blood has greatly reduced the risk of catching hepatitis B or C from blood transfusions. Both hepatitis B and C can be spread through sharing of razors, toothbrushes, and nail clippers.

To analyze the factors and early predictions of diseases based on the analysed factors in initial stages of liver. Various data mining algorithms are implemented on the records which were recorded from previous patients and some non-patients, as the result from the implemented algorithms new patients were discovered at a fast rate. By finding the disease early there are more chances to cure them immediately, this will cause a grate loss in the diseased patients.

The chosen categories of selected classification algorithms are tree-based, statistical-based, rule-based. Various machine learning algorithms have been explored for liver disease diagnostic classification, such as Naive-Bayes, Random forest, K-means, C5.0 and K-Nearest Neighbors(KNN) .

REFERENCES

1) **Duggan JM, Duggan AE**. Systematic review: The liver in celiac disease. *Aliment Pharmacol Ther* 2005; 21: 515-518.

2) **Cassagnou M, Boruchowicz A, Guillemot F et al**. Hepatic steatosis revealing celiac disease: a case complicated by transitory liver failure. *Am J Gastroenterol* 1996; 91 : 1291-1292.

Related Literature Survey

There are various authors have worked in the field of classification of liver patient data. **H. Jin et.al**, (2014) proposed "Decision Factors on Effective Liver Patient Data Prediction" and evaluated on Indian Liver Patient data (ILPD) dataset . The WEKA data mining tool which is most widely known as a useful tool for data mining is used for validating the model. This research work analyses the classification algorithms such as Naïve Bayes, Decision Tree, Multilayer Perceptron, k-NN, Random Forest and Logistic. It compares the performance of different classification algorithms. These algorithms were compared in several kinds of evaluation criteria like precision, recall, sensitivity, specificity. For comparison of selected classification algorithms, cross validation method was used. Experimental results show that Logistic and Random Forest gives highest and second highest precision and recall value respectively as compared to other previous four algorithms which are Naïve Bayes, Decision Tree, Multilayer Perceptron and KNearest Neighbor. In case of sensitivity and specificity Naïve Bayes shows the highest value and Random Forest and Logistic show relatively higher than others.

P. Saxena et.al, (2013) proposed "Analysis of Various Clustering Algorithms of Data Mining on Health Informatics" . They worked on clustering technique of data mining. The evaluation took place on ILPD dataset. They used WEKA tools to predict the result. In this research work they used different clustering algorithms such as COBWEB clustering algorithm, DBSCAN clustering algorithm, Hierarchical clustering algorithm and K-means clustering algorithm on ILPD dataset. Experimental results show that k-means clustering algorithm is simplest and fastest algorithm as compared to other clustering algorithms which are COBWEB, DBSCAN and Hierarchical clustering algorithm.

A. Guliaet.al, (2014) proposed "Liver Patient Classification using Intelligent Techniques" . They used J-48 classifier, Multilayer Perceptron classifier, Random Forest classifier, Support Vector Machine classifier and Bayesian Network classifier for classification of liver patient data. ILPD dataset available on UCI Repository is used for evaluation. Performance analysis is done using WEKA tool. Feature selection strategy also known as Variable selection or attribute selection is used to reduce the number of features with maintaining good result. This research work compares the result of classification algorithms with and without the application of feature selection. The results obtained show that Support Vector Machine algorithm gives better performance with an accuracy of 71.3551% as compared to other algorithms when evaluated without feature selection and Random Forest algorithm gives better performance with an accuracy of 71.8696% as compared to other algorithms when evaluated after feature selection. They stated that the future work will include the construction of hybrid model for classification of health care data. They also proposed the use of new algorithms to gain improved performance than the techniques used.

J. Pahareeya et.al, (2014) proposed "Liver Patient Classification using Intelligence Techniques". They used J-48, Multilayer Perceptron, Random Forest, Multilayer Regression, Support Vector Machine and Genetic Programming (GP) for classification of liver patient data. The ILPD dataset available on UCI Repository is used for evaluation. They employed under sampling and over sampling approach. They also used 10-fold cross validation in this research work. Experimental results show that Random Forest over sampling model proved to be better technique than all the other techniques. It also shows that Genetic Programming for the original data and Genetic Programming for the under sampling stood second with an accuracy of 84.75%.

S. Bahramiradet.al, (2013) proposed "Classification of Liver Disease Diagnosis: A Comparative Study". They used two liver patient dataset. Both datasets are taken from UCI Repository. The first dataset is AP dataset and the Second dataset is BUPA dataset. They worked with different classification algorithms such as Logistic, Bayesian Logistic Regression, Logistic Model Trees (LMT), Multilayer Perceptron, K-star, RIPPER, Neural Net, Rule Induction, Support Vector Machine (SVM) and CART. Accuracy, Precision and Recall parameters are used to evaluate the performance of the proposed method. They also used Brute Force Optimization and Bayesian Boosting for improving the result. The result obtained shows that AP proved to be slightly better than the BUPA dataset in terms of accuracy while BUPA dataset proved to be better than AP dataset in terms of precision and recall.

E. M. Hashemet.al, (2013) proposed "A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis". They used Support Vector Machine (SVM) classification technique for classification of liver patient data to achieve improved performance. Two dataset are used for performance evaluation. Both dataset are obtained from UCI Repository. Error Rate, Sensitivity, Prevalence, Specificity and Accuracy are used to evaluate the performance of Support Vector Machine (SVM). Feature Ranking is also used to reduce the number of features. MATLAB is used to write and implement the Support Vector Machine (SVM) algorithm. This research work also used cross validation. The result obtained shows that the Specificity at first 6 ordered features are best for BUPA dataset compared to ILPD dataset while the Sensitivity, Error Rate, Accuracy and Prevalence at first 6 ordered features are best for ILPD dataset as compared to BUPA dataset.

C. Liang et.al, (2013) proposed "An Automated Diagnosis System of Liver Disease using Artificial Immune and Genetic Algorithms" . They used a combination of two methods to diagnose the liver disease which are artificial immune and genetic algorithm. The experiments and assessments of the proposed method were performed with ILPD dataset and Liver Disorder dataset both taken from UCI Repository. F-measure, Accuracy, Sensitivity, Specificity and Precision are parameters that are used for performance evaluation. This research work also performed 20-fold cross validation on datasets. According to experimental results, the predicting accuracy obtained by this system is highest as compared to other classification methods such as Support Vector Machine (SVM), Naïve Bayes classifier, k-nearest neighbor, C4.5 and Backpropogation.Suryakantet.al, (2015) proposed "An Improved K-means Clustering with Atkinson Index to Classify Liver Patient Dataset" . They worked on clustering technique of data mining to classify liver patient dataset. The dataset being used is ILPD dataset taken from UCI Repository. Two evaluation matrices are used. The first matrix is Gold Standard and the second matrix is F-score. They used k-means clustering algorithm with Atkinson Index which is the technique for measuring the inequality. The result obtained shows that K-means clustering algorithm with Atkinson index gives better result as compared to K-means clustering algorithm for both evaluation matrices. They stated that the future work can be carried out by applying k-nearest neighbor (K-NN) to the data received by Atkinson Index which will improve the result further. Further work can be carried out by using Information Gain technique instead of Atkinson Index for selecting initial centroids.

Dr. S. Vijayaraniet. al, (2015) have proposed "Liver Disease Prediction Using SVM and Naïve Bayes Algorithm" . They have used Support Vector Machine (SVM) and Naïve Bayes algorithms for classification of liver patients. Indian Liver Patient Dataset(ILPD) available on UCI repository has been used for performance evaluation. This work is implemented using Matlab 2013 tool. These classification algorithms are evaluated on the basis of accuracy and execution time. Experimental

result shows that Support Vector Machine (SVM) is better than Naïve Bayes algorithm in classifying liver patient dataset.

B. V. Ramanaet. al, (2012) have proposed "Liver Classification Using Modified Rotation Forest" . They have used 10 different classification algorithms from 5 different categories. They are J48 and simple cart classification algorithm from tree based algorithm, Naïve Bayes and Bayes Net classification algorithms from statistical classification algorithm, MLP and SMO classification algorithms from multilayer perceptron based algorithm, IBK and K Star classification algorithm from lazy learner and PART and zero classification algorithm from rule based algorithm. Two different datasets are used for evaluation. One is Indian Liver Dataset (ILPD) and second one is BUPA liver disorder dataset. They have used feature selection strategy on dataset to reduce the number of features. The four feature selection techniques used are PCA, CFS, Random Projections and Random subset. Results obtained show that multilayer perceptron classification algorithm with Random subset gives highest accuracy of 74.7826% for BUPA liver disorder dataset and nearest neighbor with CFS gives highest accuracy of 73.0703% for ILPD dataset.

After studying and analysing the above mentioned work we have selected **S. Bahramiradet.al**, (2013) proposed "Classification of Liver Disease Diagnosis: A Comparative Study" research paper to work on it for further studies. In this paper two type of liver patients datasets are used and compared in form of there accuracy, recall, precision by using different Data mining algorithms. We implemented algorithms on only ILPD dataset and compared the results of applied algorithms and calculated which algorithms gives high accuracy.

References

1. **H. Jin, S. Kim and J. Kim** , Decision Factors on Effective Liver Patient Data Prediction. International Journal of Bio Science and Bio-Technology. VOL.6, No.4, pp. 167-178, 2014.
2. **P. Saxena and S. Lahre**, Analysis of Various Clustering Algorithms of Data Mining on Health Informatics. International Journal of Computer & Communication Technology, Vol.4, Issue-2, pp.108- 112, 2013.
3. **A.Gulia, Dr. R. Vohra, P. Rani**, Liver Patient Classification using Intelligent Techniques, International Journal of Computer Science and Information Technologies, Vol.5 (4), pp.5110- 5115,2014.
4. **J. Pahareeya, R. Vohra, J. Makhijani and S. Patsariya**, Liver Patient Classification using Intelligence Techniques, International Journal of Advanced Research in Computer science and Software Engineering, Vol. 4, Issue 2, pp.295-299,2014.

PROPOSED WORK

Medical data mining is one of the critical aspect of automated disease diagnosis and disease prediction. Medical Data Diagnosis involves developing data mining algorithms and techniques to analyze medical data. In recent years, liver disorders have excessively increased and liver disease are becoming one of the most fatal diseases in several countries.

In this study, one real patient dataset was taken from UCI repository is investigated for building classification models in order to predict liver diagnosis. Five data mining classification algorithms were applied to the dataset and the performance of all classifiers are compared against each other in terms of accuracy, precision, and recall. Several investigations have also been carried out to improve performance of the classification models. Finally, the results shown promising methodology in diagnosing liver disease during the earlier stages.

In the Dataset taken from the UCI repository there are some missing values which have to operated before implementing rules on that data. These missing values are removed by the finding the mean in that column in which they are present , the mean would be founded after removing the missing values field and then this mean is inserted to the missing value field. After removing all of the missing values from the dataset, data should be Normalized or scalled. It is done to make the data in a particular range(i.e. between 0 and 1).

All the above preparation done on the dataset is known as data preprocessing. Data Preprocessing is done in R language using R-Studio software. Coding portion of the project is done in the R language. After preprocessing of the data implementations of the algorithms is done one by one and there results are recorded. Data mining algorithms used in our work are different from the algorithms used by the research paper on which we are working. This is done to check the effect of other algorithms present in data mining on the used dataset and study the new outcoming result.

Here Algorithms used are :

- **K-Nearest Neighbor Algorithm :** The KNN or k-nearest neighbors algorithm is one of the simplest machine learning algorithms and is an example of instance-based learning, where new data are classified based on stored, labeled instances. More specifically, the distance between the stored data and the new instance is calculated by means of some kind of a similarity measure. This similarity measure is typically expressed by a distance measure such as the Euclidean distance, cosine similarity or the Manhattan distance. In other words, the similarity to the data that was already in the system is calculated for any new data point that you input into the system.

- **K-means :** K-Means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly *k* different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

- **Naive-bayes :** It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the

other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

- **C5.0 :** C5.0 algorithm is used to build either a **decision tree** or a **rule set**. A C5.0 model works by splitting the sample based on the field that provides the maximum **information gain**. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or **pruned**.
- **Random forest :** Random Forest consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the depemdent variables. Alternatively, for regression problems, the tree response is an estimate of the dependent variable given the predictors.The Random Forest algorithm was developed by Breiman. A Random Forest consists of an arbitrary number of simple trees, which are used to determine the final outcome. For classification problems, the ensemble of simple trees vote for the most popular class.In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy (i.e., better ability to predict new data cases).

Assessment Criteria

For each algorithm, we measure the accuracy, precision and recall in order to compare the performance of different classification algorithms against each other. The measurement metrics are defined as follows:

- **Accuracy** – Accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

  Accuracy = TP +TN / TP+FP+TN+PN

- **Precision** – Precision is defined as the proportion of the true positives (TP) against all the positive results, which are both true positives (TP) and false positives(FP).

  Precision = TP / TP+FP

where TP is true positive rate, TN is true negative rate, FP is false positive rate and FN is false negative rate.

- **Recall** – Recall or Sensitivity is also referred as true positive (TP) rate i.e. the proportion of positive tuples that are correctly identified.

recall = TP TP + FN

where TP is true positive rate, TN is true negative rate, FP is false positive rate and FN is false negative rate.

RESULTS

On implementing the mentioned algorithms we got the results :

Classification Algorithm	Accuracy	Precision	Recall
KNN	70.68%	73.38%	93.81%
C5.0	71.43%	93.81%	73.98%
K-MEANS	69.47%	97.36%	70.80%
NAÏVE-BAYES	66.17%	1.00%	53.61%
RANDOM FOREST	72.18%	75.00%	92.78%
C5.0 with adaptive Boosting	75.19%	90.72%	78.57%

Conclusion

Finally, as the results in table shows that highest accuracy in given by the **Random forest** algorithm when applied all algorithms without any changes. But after adaptive boosting the **C5.0** algorithm it gives a more accurate result. The accuracy gained by our work is greater than the research paper which we have taken in study.

REFERENCES

1. **B. V. Ramana, M. S. P. Babu and N. B. Venkateswarlu**, Liver Classification Using Modified Rotation Forest. International Journal of Engineering Research and Development, 6(1):17-24, 2012.

2. **R. H. Lin**, An Intelligent Model for Liver Disease Diagnosis. Artificial Intelligence in Medicine, 47(1):53-62, 2009.

3. **P. Rajeswari and G. Reena**, Analysis of Liver Disorder using Data mining Algorithm. Global Journal of Computer Science and Technology, 10(14):48-52, 2010.

4. **B. Chen and H. X. Zhang**, An Approach of Multiple Classifiers Ensemble based on Feature Selection. In Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 2, pp. 390-394, 2008.

5. **S. Karthik, A. Priyadarishini, J. Anuradha and B. K. Tripathy**, Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and Its Types. Advanced Applied Scientific Research, 2(3):334345, 2011.

6. **S. Ansari, I. Shafi, A. Ansari, J. Ahmad and S. I. Shah**, Diagnosis of Liver Disease Induced by Hepatitis Virus using Artificial Neural Networks. In 2011 IEEE International Multitopic Conference (INMIC), pp. 8-12, 2011.

7. **A.Frank and A. Asuncion**, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

8. **ILPD (Indian Liver Patient Dataset) Dataset**. UCI Repository of Machine Learning Databases. Available in http://archive.ics.uci.edu/ml/datasets/ILPD, last accessed: 15 December 2012.

9. **BUPA Liver Disorders Dataset**. UCI Repository of Machine Learning Databases. Available in http://archive.ics.uci.edu/ml/datasets/Liver+Disorders, last accessed: 15 December 2012.

10. **I Mierswa, M. Wurst, R. Klinkenberg, M. Scholz and T. Euler, YALE**: Rapid Prototyping for Complex Data Mining Tasks. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2006), pp. 935-940, 2006.