Cumhuriyet Üniversitesi Fen Fakültesi
Fen Bilimleri Dergisi (CFD), Cilt:36, No: 3 Ozel Sayı (2015)
ISSN: 1300-1949

Cumhuriyet University Faculty of Science
Science Journal (CSJ), Vol. 36, No: 3 Special Issue (2015)
ISSN: 1300-1949

# A Survey and Compare the Performance of IBM SPSS Modeler and Rapid Miner Software for Predicting Liver Disease by Using Various Data Mining Algorithms

Moloud ABDAR[1,*]

[1]*Department of Computer Engineering, Damghan University, Iran*

Received: 01.02.2015; Accepted: 05.05.2015

**Abstract.** Today, with the development of industry and mechanized life style, prevalence of diseases is rising steadily, as well. In the meantime, the number of patients with liver diseases (such as fatty liver, cirrhosis and liver cancer, etc.) is rising. Since prevention is better than treatment, early diagnosis can be helpful for the treatment process so it is essential to develop some methods for detecting high-risk individuals who have the chance of getting liver diseases and also to adopt appropriate solutions for early diagnosis and initiation of treatment in early stages of the disease. In this study, we tried to use common data mining techniques that are used nowadays for diagnosis and treatment of different diseases, for the diagnosis and treatment of liver disease. For this purpose, we used Rapid Miner and IBM SPSS Modeler data mining tools together. Accuracy of different data mining algorithms such as C5.0 and C4.5, Decision tree and Neural Network were examined by the two above tools for predicting the prevalence of these diseases or early diagnosis of them using these algorithms. According to the results, the C4.5 and C5.0 algorithms by using IBM SPSS Modeler and Rapid Miner tools had 72.37% and 87.91% of accuracy respectively. Further, Neural Network algorithm by using Rapid Miner had the ability of showing more details.

**Keywords:** Data mining techniques, Liver diseases, Rapid Miner, IBM SPSS Modeler.

## 1. INTRODUCTION

Liver is one of the largest internal organs in the human body that pumps averagely 1.4 liters of blood per minute [1]. The liver purifies the blood and identifies and disposes toxic substances such as alcohol or some way, converts them to beneficial nutrients in body and is also used to control the body hormone levels. Some other functions of the liver are producing hormones and proteins, controlling the FBS and also helping blood coagulation. Liver diseases are more than 100 different types. According to published statistics by Canadian Liver Foundation in 2013, the morality rate due to liver diseases has been increased by 30% during 8 years. According to a Rong-Ho Lin study which was carried out on liver diseases using intelligent techniques, the liver disease is introduced as one of the top 10 dangerous diseases [2]. Currently, early diagnosis of liver diseases in the early stages of prevalence and performing essential and appropriate cares for treatment of patients has become as an important and a also a challenging issue. This article consists of 5 chapters. Chapter II discusses the key principles related to this article. Chapter III discusses the related studies in this field. Chapter IV focuses on our own proposed method, and compares the accuracy and performance of several data mining techniques for diagnosis of liver disease. Finally, chapter V pluralizes and concludes the considered discussions.

* Corresponding author. *m.abdar1987@gmail.com*

*Special Issue:* The Second National Conference on Applied Research in Science and Technology

http://dergi.cumhuriyet.edu.tr/ojs/index.php/fenbilimleri ©2015 Faculty of Science, Cumhuriyet University

## 2. BASIC CONCEPTS

### 2.1. Data mining

Data mining has become as one of the most important scientific topics all over the world which is useful in most of scientific fields and is a useful technique for extracting knowledge from a mass of stored raw data [3]. By using various models in data mining, it was attempted to greatly reduce the human errors.

Data mining is finding the hidden patterns in different data in semi-automatic form [4] that discovers the information using methods and models such as analytical models and classification and finally presents the results using different data mining tools. For performing data mining on records, a pre-processing is performed in two forms: data reduction and data generalization. Data reduction is performed in order to produce a smaller set among basic data that the obtained results are almost unchanged and identical with the results of data mining which is performed for basic data. By data reduction, it was attempted to remove the attributes which are not associated or have less association with the basic data [5].

In fact, data mining finds the hidden information among the existing data and then predicts the relationships that are hidden and unknown [6]. For performing data mining, different algorithms are proposed and applied such as Kohonen, K means, COX, SMO, SVM, CHAID, C5.0, Neural Network, and KNN.

#### 2.1.1. Classification

Classification can be considered as one of the most common operations in data mining. Classification is a process that divides the dataset into specified sections and then classifies the data which is a two-phase process. In the first phase, it develops a model based on educational datasets of databases and then creates ab educational dataset including records, samples, examples and things with a collection of attributes and aspects. Each sample has a specific class label. In the second phase, the developed model in the previous phase is used to classify new samples. In a general point of view, regression and classification are two types of predictive factors that regression is used for prediction of continuous data and classification is used for prediction of discrete and nominal data.

#### 2.1.2. Clustering

In fact, clustering is considered as unsupervised operations. We use clustering when we are looking for groups with similar type while there are no available predictions about the similarities already.

#### 2.1.3. Correlation rules

Correlation rules are one of the main techniques of data mining and it can be considered as the most important form of discovery and extraction of patterns. This method retrieves all possible patterns of databases. In recent years, each of these algorithms has been implemented on different records, and each of them has indicated different functions according to implementation conditions and also data types. Each method has strengths and weaknesses which are detected while running on different records.

### 2.2. Liver and liver diseases

Liver is one of the largest internal organs in the human body, which can analyze about 500 types of physical activities. Averagely, the liver pumps about 1.4 liters of blood per minute [1]. Location of the liver is shown in the following figure.

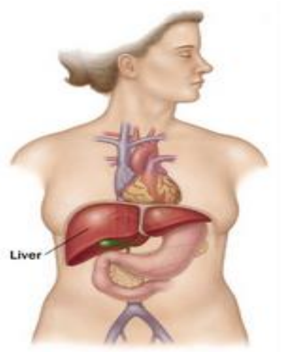


**Figure 1.** The location of the liver in the human body [1].

There are more than 100 different types of liver diseases. Some of the most important types are as the following [1]:

- Alagille Syndrome
- Alpha 1 Anti-Trypsin Deficiency
- Autoimmune Hepatitis
- Biliary Atresia
- Cirrhosis
- Cystic Disease of the Liver
- Fatty Liver Disease
- Galactosemia
- Gallstones
- Gilbert’s Syndrome
- Hemochromatosis
- Liver Cancer
- Liver disease in pregnancy
- Neonatal Hepatitis
- Primary Biliary Cirrhosis
- Primary Sclerosing Cholangitis
- Porphyria
- Reye’s Syndrome
- Sarcoidosis
- Toxic Hepatitis
- Type 1 Glycogen Storage Disease
- Tyrosinemia
- Viral Hepatitis A, B, C
- Wilson Disease

The importance of the liver as an important part of the body is obvious, but it is noteworthy that in recent years, the morality rate due to liver disease had a rising trend. Therefore, in this article, the early diagnosis methods of liver diseases using different data mining models are studied and examined.

### 2.3. Dataset

For performing data mining on liver disease data, different types of data are used that are recorded in different parts of the world. For this purpose, in this article the data is for Indian patients in 2012and is available in archive of University of California Irvine (UCI). The data included 416 records of patients with liver disease and 167 records of normal people.

441 records were for males and 142 of them were for females. These people were in different age range. It is noteworthy that for all the people who were over 89 years old, equally the age field has written 90.

According to the table below, the available records have 10 attributes [7] :

**Table 1.** Attribute Information

| number | Attribute name | Range |
|---|---|---|
| 1. | Age : Age of the patient | [4-90] |
| 2. | Gender :Gender of the patient | [male-female] |
| 3. | TB :Total Bilirubin | [0.4-75] |
| 4. | DB:Direct Bilirubin | [0.1-19.7] |
| 5. | Alkphos: Alkaline Phosphotase | [63-2110] |
| 6. | Sgpt Alamine :Aminotransferase | [10-2000] |
| 7. | Sgot Aspartate: Aminotransferase | [10-4929] |
| 8. | TP: Total Protiens | [2.7-9.6] |
| 9. | ALB: Albumin | [0.9-5.5] |
| 10. | A/G Ratio: Albumin and Globulin Ratio | [0.3-2.8] |
| 11. | Selector field :used to split the data into two sets (class 1: 416 liver patient records and class 2: 167 non liver patient records.) | [1-2] |

### 2.4. Methods

#### 2.4.1. C4.5 algorithm

This algorithm is one of the types of decision tree that was introduced after upgrading the ID3 algorithm. This algorithm can classify the records with noisy and continuous amplitude. When the records are with discrete amplitude, this algorithm operates like ID3 algorithm but when the data amplitude is continuous, it will consider a threshold for all selectable modes and an effective standard is assessed for the threshold and then, the threshold with the highest rate is chosen as the decision index of that node. The most important attribute which distinct this method from previous method is the possibility of pruning the tree after it is being fully formed and this is usually implemented in this way: a threshold is considered and if the occurrence probability rate of a leaf of the tree was less than this threshold in comparison to the adjacent leaves, then this leaf will be eliminated from the tree or it will be combined with other leaves if necessary. The purpose of this process is to reduce the height of the tree in order to prevent over fitting and to remove the noise data [8, 9, and 10].

### 2.4.2. C5.0 algorithm

Decision tree is one of the most important and widely used data mining algorithms. One of the algorithms which belongs to decision tree and has great importance in data mining is C5.0 algorithm which is the developed form of C4.5 algorithm and this algorithm itself is the developed form of ID3 algorithm. This algorithm has the ability to be used for classifying as a decision tree or a set of rules. In many applications, because set of rules can be understood easily, so they are preferred in comparison to other rules. Some of the strengths of this algorithm are managing the missing values, controlling high input number and less learning time [8, 9, 10, 11].

### 2.4.3. SVM algorithm

Support Vector Machine (SVM) algorithm was introduced and presented by Vapnik in 1995 as a supervisory algorithm. The instruction of this algorithm is using accuracy for generalization of errors. The algorithm operates by forming a hyper plane and divides the data into classes. The method of this division is in this way: all the samples that are belonged to a class are placed in one side and other classes are placed in the other side. For performing the SVM classifier operation, a linear classification of the data is defined and in division process, it tries to choose a line that has the greatest margin of safety [12, 13].

### 2.4.4. KNN algorithm

K-Nearest Neighbor is an algorithm that the basis of classification in it is based on similarity with other items. The items which are similar to each other are called neighbors. Once a new item is found, its distance from other items in the model is calculated. This classification partitions the item to the nearest neighbor which is also the most similar one; so places the item in a group that includes the nearest neighbors. This algorithm is also able to obtain the values for continuous targets. In this case, the average or median target value of the nearest neighbor is used to obtain the prediction value of new items [14].

### 2.4.5. Neural Network algorithm

Artificial Neural Network which is inspired from the brain is considered as a data processing system. In this algorithm, many microprocessors are responsible for data processing and they are acting as an interconnected and parallel network with each other to solve a problem. In these networks with the help of programming science, a data structure is designed that can act as a neuron and this data structure is called *neuron*. By setting a network between the neurons and applying a learning algorithm, the network is trained. In this Neural Network, neurons are divided into two enable (ON or 1) or disable (OFF or 0) modes and each edge (synapses or connections between nodes) has a weight. Edges with positive weight, stimulate or enable the next disable nodes and edges with negative weights, disable or inhibit the next connected nodes (if they are enabled) [15, 16, 17, 18].

### 2.4.6. CHAID algorithm

This algorithm is one of the types of decision tree that was introduced by Kass in 1980. CHAID is abbreviation of CHi-squared Automatic Interaction Detection that can be used for prediction, classification and also connection between different factors. Decision trees usually present simple and apprehensible results. One of the advantages of this algorithm is also simplicity of the results for understanding and interpreting. CHAID algorithm can also be utilized for grouped qualitative and quantitative variables. This algorithm by using three steps, integration, division and stopping, which is performed repeatedly, moves downward from the root node of tree, using these three steps in each node [19]. CHAID algorithm selects the best choice for prediction in each step and the best choice is continued until reaching to the end of the tree. In this algorithm, p-values are used for finding the best attribute on each node and variables which have lower p-value amounts, will be considered for divisions on the nodes in the first step.

## 3. RELATED STUDIES

Chang No Yoon et al. [20] studied the liver diseases using logistic regression algorithms, decision trees and Neural Networks for data mining. The results of their study showed that the Neural Network having 72.55% of Accuracy and 78.62% of Sensitivity using the estimation of growth curve is able to detect the liver disease. Rong-Ho Lin [2] by using intelligent models predicted the chance of getting liver diseases. In this research, the CART model was utilized based on Case - Based Reasoning (CBR) and the results of their study showed that the Accuracy of the CART algorithm was 92.94% while the Accuracy of the model was 90% by considering the CBR. A.S.Aneeshkumar and C.Jothi Venkateswaran [21] examined the chances of liver disease in ectopic pregnancy using data mining. Their study showed that finding a relationship between the two mentioned topics is difficult and requires a lot of study and analysis by specialists on liver diseases. However, this study showed liver diseases occur in some ectopic pregnancies. Aida Mustapha et al. [22] studied on two types of data about Liver patients that were available in University of California Irvine (UCI) archive using 11 different classification algorithms. In their research, BUPA and AP datasets, which are related to liver diseases, were compared with each other using data mining techniques. The results of this comparison showed that the Accuracy of the BUPA dataset was slightly better than AP dataset and the reasons of this difference were the number of datasets and the type of used datasets. In some of comparisons, BUPA dataset had better performance. Accuracy and Recall rate in AP dataset were less than BUPA dataset, as reported. Sa'diyah Noor Novita Alfisahrin and Teddy Mantoro [23] identified people with liver diseases using data mining. They chose decision tree, Naive Bayes and NBTree algorithms for their research. The results of their study showed that the NBTree algorithm had the most Accuracy in identifying patients. The Naive Bayes algorithm was faster than other algorithms. Finally, it was pointed out that the NBTree algorithm had developed the simplest tree.

Anil Kumar Tiwari et al. [24] utilized several data mining models such as SVM, SOM, RBF and BP which they are operating based on Artificial Neural Network (ANN), studied the performance of these algorithms on liver patients. Based on the results of this research, SVM algorithm had the best performance with an Accuracy of 99.7%. Their research showed that ANN-based classification can be used as an important method for prediction of patients.

In Jankisharan Pahareeya et al. study [25], they examined classification of liver patients using intelligent techniques. They utilized different method such as Multilayer, Support Vector Machine, Linear Regression, viz.Multiple, Feed Forward, Neural Network, J-48, Random Forest and Genetic Programming. Random Forest algorithm had 84.75% Accuracy which has shown better results in comparison to other algorithms. Jinhong Kim et al. [26] studied the effective factors of in identifying liver patients using data mining techniques. They compared Multi-Layer Perceptron, Decision Tree techniques Naive Bayes and KNN techniques. The results of their study showed that the Naive Bayes algorithm had better Precision than other algorithms. Manuel Cruz-Ramirez et al. [27], focused on survival of patients with liver transplantation and they used multi-objective artificial Neural Network technique in order to examine their longevity and survival. They considered all the relevant factors to performance; factors such as correct classification rate (C), minimum sensitivity (MS), area under the receiver operating characteristic curve (AUC), root mean squared error (RMSE) and Cohen's kappa (Kappa). According to the results of their study, it was characterized that multi-objective evolutionary had better performance in comparison to single-objective evolutionary algorithms. Michele Berlingerio et al. [28] carried out a case study on liver transplanted patients using clinical data mining with a temporal dimension. The purpose of their study was to investigate the effectiveness of extracorporeal photopheresis which is a method for treating and preventing liver transplant rejection in the body. The results of their study suggested that the use of a method for creating frequent associative patterns in the direction of qualifying various biochemical variables in a temporal dimension is an important step. Christine M. Hunt et al. [29] studied the relationship between the age of liver patients and their body reaction to liver drugs intake using data mining. They reported the percentage of liver patients based on their age in three groups: 6% for 0-17 years, 62% for 18-64 years and 32% for 65 years and over and according to statistics, the risk of liver disease is low in children. The results of their study require further studies to confirm the obtained information so if these results are confirmed, they could be available for practical use of physicians.

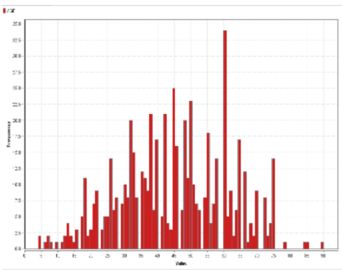

**Figure 1.** Histogram for Age.

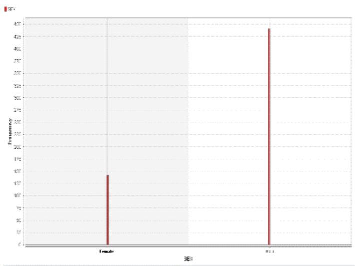

**Figure 2.** Histogram for SEX.

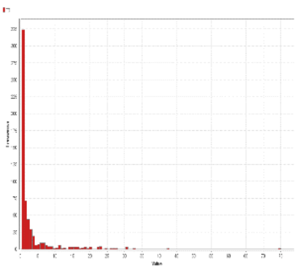

**Figure 3.** Histogram for TB.

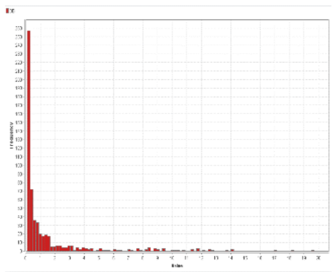

**Figure 4.** Histogram for DB.

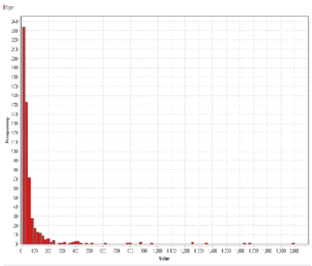

**Figure 5.** Histogram for Sgpt.

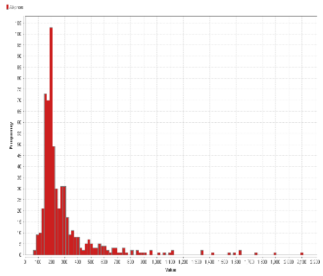

**Figure 6.** Histogram for Alkphos.

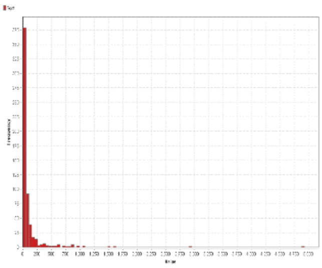

**Figure 7.** Histogram for Sgot.

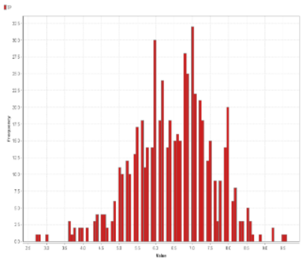

**Figure 8.** Histogram for TP.

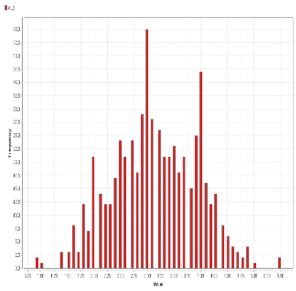

**Figure 9.** Histogram for ALB.

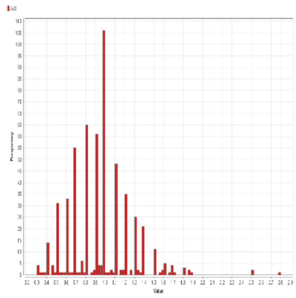

**Figure 10.** Histogram for A/G.

## 4. OUR PROPOSED APPROACH

In this stage, data analysis showed that 4 records were with a null value, so these records had been removed in order to increase the accuracy of the study. After removing the outliers, the data had entered into the RapidMiner software. Following the study, dispersion of each of these attributes was obtained as shown in Figures 1 to 10. To calculate the indicator values, perturbation matrix can be used. There are different indicators such as Accuracy and Precision for evaluation of classification methods that are calculated as equations 1 and 2. This matrix is a useful tool for analyzing the performance of classification method in identifying the data or observations of different classes. The ideal case is that most of relevant data to observations to be placed on the main diagonal of the matrix, and the remaining values of the matrix to be zero or near zero [30, 31, 32].

TP = number of positive labeled data that are correctly classified,

FP = number of negative labeled data that are incorrectly classified positive,

FN = number of positive labeled data that are incorrectly classified negative,

TN = number of negative labeled data that are correctly classified.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

After this calculation, according to equations 1 to 2, statistical table of indicators was obtained as shown in Table 2:

**Table 2.** Dispersion.

| Attribute | Min | Max | Average | Devation |
|---|---|---|---|---|
| Age | 4 | 90 | 44.746 | 16.190 |
| SEX | Female(142) | Male(441) | - | - |
| TB | 0.400 | 75 | 3.315 | 6.228 |
| DB | 0.100 | 19.700 | 1.494 | 2.816 |
| Alkphos | 63 | 2110 | 291.366 | 243.562 |
| Sgpt | 10 | 2000 | 81.126 | 183.183 |
| Sgot | 10 | 4929 | 110.415 | 289.850 |
| TP | 2.700 | 9.600 | 6.482 | 1.085 |
| ALB | 0.900 | 5.500 | 3.139 | 0.794 |
| A/G | 0.300 | 2.800 | 0.947 | 0.320 |

In this article, data were used in an equal condition by using Rapid Miner and IBM SPSS Modeler software. Linear regression, KNN, C4.5, C5.0, Naïve Bayes CHAID, SVM, Neural net and Random forest algorithms were implemented and run which all of them are available in Rapid Miner software. In this study the records were divided into two training and testing sets that 70% of them were allocated for training and 30% of them were allocated for testing. The results of utilized algorithms' performance are as shown in the Table 3:

**Table 3.** Comparison of different algorithms using Rapid Miner software.

| Algorithm | Accuracy | Precision |
|---|---|---|
| SVM | 72.54 % | 100 % |
| C4.5 | 72.37 % | 85.71 % |
| Random forest | 71.85 % | 100 % |
| CHAID | 71.51 % | unknown |
| Neural net | 70.81 ± 9.30 % | 48.36 % |
| Linear regression | 70.81 ± 4.79 % | 37.50 % |
| Naive Bayes | 66.92 % | 45.13 % |
| KNN | 64.94 % | 38.67 % |

By comparing the results in Table 3, it is obvious that the SVM algorithm with having 72.54% Accuracy and 100% Precision had better performance among all the utilized algorithms by Rapid Miner software.

Then, in a same conditions, CHAID, Logistic Regression, Bayesian net, SVM, Neural net, KNN, C5.0 and Decision list algorithms were implemented and run by IBM SPSS Modeler software. The records were also divided into two subsets: 70% for training and 30% for testing.

**Table 4.** Comparison of different algorithms using IBM SPSS Modeler Software.

| Algorithm | Accuracy | Precision |
|---|---|---|
| C5.0 | 87.91 % | 4.76 % |
| KNN | 78.411 % | 22.22 % |
| Neural net | 73.057 % | 45.16 % |
| SVM | 74.957 % | 38.63 % |
| Bayesian net | 74.266 % | 40.24 % |
| Logistic Regression | 73.921 % | 40.78 % |
| CHAID | 71.503 % | 29.29 % |
| QUEST | 72.021 % | - |

Comparing the results of the algorithms run by IBM SPSS Modeler software in equal conditions with Rapid Miner software showed that the Accuracy of C5.0 algorithm was obtained 87.91% which accordingly, had better performance than other algorithms.

## 5. DISCUSSION

After comparison of algorithms in this study, we found out that averagely, the Accuracy of three decision tree algorithms in Rapid Miner software was obtained 71.910% and also the average Accuracy of decision tree algorithms in IBM SPSS Modeler software was obtained 77.144%. The CHAID algorithm had almost a similar performance in both software. QUEST and Random forest algorithms with having 0.341% of difference, had also a close Accuracy percentage. However, C5.0 and C4.5 algorithms showed a significant difference of 15.37%. The reason of this difference is that the C5.0 algorithm was introduced after C4.5 algorithm and it was attempted to fix and improve the problems of C4.5 algorithm, so in this study, C5.0 algorithm had better Accuracy than C4.5 algorithm [8, 9, 10]. Another point about C4.5 algorithm is that the depth of tree in this algorithm was lower in comparison to C5.0 algorithm and the tree in C5.0 algorithm was considered more modes and thus the depth of the tree and the number of different modes was more. Then, by comparing the Bayesian net and naïve Bayes algorithms in two software we found out that the Bayesian net algorithm with having 7.346% of difference in Accuracy had a better performance in IBM SPSS Modeler software. Also KNN algorithm with having 13.471% of Accuracy in IBM SPSS Modeler software, performed better in comparison with its Accuracy in the Rapid Miner software. The rest of the algorithms had almost an equal performance in both of the software. But one of the significant points about the Neural net algorithm was that although it had a low difference in Accuracy percentage in Rapid Miner software but the developed Neural Network by this software had stratification with more details and explanations (for inner and outer layers) in comparison with the Neural Network developed by the IBM SPSS Modeler software.

## 6. CONCLUSION

In this study, the performance of two data mining software, IBM SPSS Modeler and Rapid Miner, on the records of liver patients was compared. For this purpose, the data for Indian patients which were 583 records and were available in University of California Irvine (UCI) archive were used. Among all the algorithms, the C5.0 algorithm had the best performance with having 87.91% of Accuracy. After implementation and running of the algorithms, the C5.0 algorithm showed a better performance in comparison to C4.5 algorithm in IBM SPSS Modeler software. CHAID, QUEST, Random Forest algorithms had almost a similar performance in both software. The results of this study showed that in IBM SPSS Modeler software, the Bayesian net algorithm with having 7.346% of difference in Accuracy had a better performance in comparison to naïve Bayes algorithm in Rapid Miner software. In the rest of the study, we found out that although the Neural Network algorithm in Rapid Miner software had 2.247% less Accuracy percentage, but the type of the developed Neural Network by this algorithm in Rapid Miner software was more accurate. Another significant point about the Neural Network algorithm in Rapid Miner software was that all the medial and output layers were developed in details but, these details were not available in the obtained diagram by IBM SPSS Modeler software.

## REFERENCES

[1] Ramona Tarba." Liver Disease in Canada Report. http://www.liver.ca/support-liver-foundation/advocate/Liver_Disease_in_Canada_Report.aspx. [accessed April 2015].

[2] Rong-Ho Lin, (2009). An intelligent model for liver disease diagnosis. Elsevier B.V, 0933-3657,pp. 47, 53-62.

[3] Silvia Sookoian, Carlos J. Pirola. (2012). The Genetic Epidemiology of Nonalcoholic Fatty Liver Disease. Elsevier Inc, 1089-3261/12, pp. 467–485.

[4] Han, J. and Kamber, M.( 2006). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, pp. 1-7.

[5] An Introduction to Data Mining: http://www.thearling.com/, [ accessed September 2014].

[6] Renza Campagni, Donatella Merlini, Renzo Sprugnoli, Maria Cecilia Verri. (2015). Data Mining models for student careers. Published by Elsevier ,S0957-4174(15)00159-1, pp. 1-21.

[7] Bendi Venkata Ramana, M. Surendra Prasad Babu, N. B. Venkateswarlu. ILPD (Indian Liver Patient Dataset) Data Set.

[8] Quinlan J R. (1986). Induction of decision trees. Machine Learning. pp. (4): 81–106.

[9] Quinlan J R. (1994). C4.5: Programs for machine learning. Machine,Learning. pp. (3): 235–240.

[10] Quinlan J R.(1996). Bagging, Boosting and C4.5. Proceedings of 14th National Conference on Artificial Intelligence, pp. 725–730.

[11] Xindong Wu , Vipin Kumar , J. Ross Quinlan , Joydeep Ghosh, Qiang Yang, Hiroshi Motoda , Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, ichael Steinbach, David J. Hand, Dan Steinberg. (2008). Top 10 algorithms in data mining. Springer, DOI 10.1007/s10115-007-0114-2.

[12] Sumit Bhatia, Praveen Prakash, and G.N. Pillai. (2008). SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features. WCECS. Proceedings of the World Congress on Engineering and Computer Science. ISBN: 978-988-98671-0-2.

[13] Vapnik, V. N. (1995). The nature of statistical learning theory. IEEE, VOL. 10, NO. 5, pp. 988- 999.

[14] Yazdani A, Ebrahimi T, Hoffmann U. (2009). Classification of EEG signals using Dempster Shafer theory and a K-nearest neighbor classifier. IEEE. In: Proc of the 4th int EMBS conf on neural engineering, pp. 327–30.
[15] Daubechies I.(1990). The wavelet transform, time-frequency localization and signal analysis. IEEE. Trans Inform Theor pp. 36:961–1005.
[16] Demuth H, Beale M, Hagan M. (2009). Neural network Toolbox™ user's guide. The MathWorks, Inc, pp. 1-901.
[17] Leng, G., McGinnity, T.M., Prasad, G. (2006). Design for self-organizing fuzzy neural networks based on genetic algorithms. IEEE. Trans. Fuzzy Syst, Vol 14, No. 6, pp. 755–766.
[18] Leung, F.H.F., Lam, H.K., Ling, S.H., et al.(2003). Tuning of the structure and parameters of a neural network using an improved genetic algorithm. IEEE. Trans. Neural Networks , Vol 14, No. 1, pp. 79–88.
[19] Kass GV. (1980). An Exploratory Technique for Investigating Large Quantities of Data. Appl Stat, Vol 29, No. 2, pp.119.
[20] Young Sun Kim, SoYoung Sohn, Chang No Yoon. (2003). Screening test data analysis for liver disease prediction model using growth curve. Éditions scientifiques et médicales Elsevier SAS, doi:10.1016, pp.482–488.
[21] A.S.Aneeshkumar, C.Jothi Venkateswaran, (2012)," An Approach of Data Mining for Predicting the Chances of Liver Disease in Ectopic Pregnant Groups", Special Issue of International Journal of Computer Applications, (0975 – 8887), pp. 19-22.
[22] Sina Bahramirad, Aida Mustapha, Maryam Eshraghi, (2013). Classification of Liver Disease Diagnosis: A Comparative Study. IEEE ,ISBN: 978-1-4673-5256-7/13, pp. 42-46.
[23] Teddy Mantoro, Sa'diyah Noor Novita Alfisahrin, (2013). Data Mining Techniques For Optimatization of Liver Disease Clasification. IEEE, 978-1-4799-2758-6/13, DOI 10.1109/ACSAT.2013.81, pp.379-384.
[24] Anil Kumar Tiwari, Lokesh Kumar Sharma, G. Rama Krishna, (2013). Comparative Study of Artificial Neural Network based Classification for Liver Patient. Journal of Information Engineering and Applications, Vol.3, No.4, ISSN 2224-5782 (print) ISSN 2225-0506 (online), pp. 1-5.
[25] Jankisharan Pahareeya, Rajan Vohra, Jagdish Makhijani, Sanjay Patsariya, (2014)," Liver Patient Classification using Intelligence Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, pp. 295-299.
[26] Hoon Jin, Seoungcheon Kim, Jinhong Kim, (2014). Decision Factors on Effective Liver Patient Data Prediction, International Journal of Bio-Science and Bio-Technology, Vol.6, No.4, pp.167-178.
[27] Manuel Cruz-Ramirez, César Hervas-Martinez, Juan Carlos Fernandez,Javier Brice˜no, Manuel de la Mata, (2013). Predicting patient survival after liver transplantation using evolutionary multi-objective artificial neural networks. Elsevier B.V, 0933-3657, pp. 37-49.
[28] Michele Berlingerio, Francesco Bonchi, Fosca Giannotti, Franco Turini, (2007). Mining Clinical Data with a Temporal Dimension: a Case Study. IEEE International Conference on Bioinformatics and Biomedicine, 0-7695-3031-1/07, DOI 10.1109/BIBM.2007.42, pp. 429-436.
[29] Christine M. Hunt, Nancy A. Yuen, Heide A. Stirnadel-Farrant, Ayako Suzuki. (2014). Age-related differences in reporting of drug-associated liver injury: Data-mining of WHO Safety Report Database. Elsevier Inc, 0273-2300, pp. 519-526.
[30] Alizadeh S, Ghazanfari M.(2011). Teimorpour B .DataMining and Knowledge Discovery. Publication of IranUniversity of Science and Technology .2nd ed.[Persian].
[31] Han J. Kamber M.(2006) . chapter 1: introdution :DataMining: Concepts and Techniques. Morgan Kaufman Publisher. 2nd ed.
[32] David M W Powers. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. School of Informatics and Engineering, Flinders University • Adelaide • Australia, Technical Report SIE-07-001, pp. 1-24.