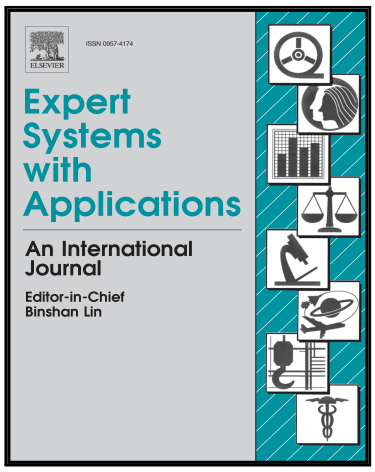

Performance analysis of classification algorithms on early detection of Liver disease

Moloud Abdar , Mariam Zomorodi-Moghadam , Resul Das , I-Hsien Ting

PII: S0957-4174(16)30464-X
DOI: 10.1016/j.eswa.2016.08.065
Reference: ESWA 10858

To appear in: *Expert Systems With Applications*

Received date: 31 March 2016
Revised date: 26 August 2016
Accepted date: 26 August 2016

Please cite this article as: Moloud Abdar , Mariam Zomorodi-Moghadam , Resul Das , I-Hsien Ting , Performance analysis of classification algorithms on early detection of Liver disease, *Expert Systems With Applications* (2016), doi: 10.1016/j.eswa.2016.08.065

Highlights

- In this research UCI Indian Liver Patient Dataset (ILPD) used.
- Boosted C5.0 and CHAID algorithms are used to identify liver disease risk factors.
- This research shows females have more chance of liver disease than males.
- Common risk factors of liver disease were extracted by data mining.
- This research were produced quite simple rules.

Performance analysis of classification algorithms on early detection of Liver disease

Moloud Abdar [a], Mariam Zomorodi-Moghadam [b], Resul Das [c,*], I-HsienTing [d]

[a] Department of Computer Engineering, Damghan University, Damghan, Iran

[b] Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

[c] Department of Software Engineering, Technology Faculty, Firat University, Elazig, Turkey

[d] Department of Information Management, National University of Kaohsiung, Kaohsiung, Taiwan

*Corresponding author: Tel: +90-424-2370000 Ext: 4305

e-mail addresses: m.abdar1987@gmail.com, m_zomorodi@um.ac.ir, rdas@firat.edu.tr, iting@nuk.edu.tw

Abstract

The human liver is one of the major organs in the body and liver disease can cause many problems in human life. Fast and accurate prediction of liver disease allows early and effective treatments. In this regard, various data mining techniques help in better prediction of this disease. Because of the importance of liver disease and increase the number of people who suffer from this disease, we studied on liver disease through using two well-known methods in data mining area.

In this paper, novel decision tree based algorithms is used which leads to considering more factors in general and predictions with high accuracy compared to other studies in liver disease. In this application, 583 UCI instances of liver disease dataset from the UCI repository are considered. This dataset consists of 416 records of liver disease and 167 records of healthy liver. This dataset is analyzed by two algorithms named Boosted C5.0 and CHAID algorithms. Until now there is no work in the literature that uses boosted C5.0 and CHAID for creating the rules in liver disease. Our results show that in both algorithms, the DB, ALB, SGPT, TB and A/G factors have a significant impact on predicting liver disease which according to the rules generated by both algorithms important ranges are DB = [10.900 – 1.200], ALB [4.00 – 4.300], SGPT = [34 – 37], TB = [0.600 – 1.200] (by boosted C5.0), A/G = [1.180 – 1.390], as well as in the Boosted C5.0 algorithm, Alkphos, SGOT and Age have significant impact in prediction of liver disease. By comparing the performance of these algorithms, it becomes clear that C5.0 algorithm via Boosting technique has an accuracy of 93.75 percent and this result reveals that it has a better performance than the CHAID algorithm which is 65.00 percent. Another important achievement of this paper is about the ability of both algorithms to produce rules in one class for liver disease. The results of our assessment show that Boosted C5.0 and CHAID algorithms are capable to produce rules for liver disease. Our results also show that boosted C5.0 considers the gender in liver disease, a factor which is missing in many other studies. Meanwhile, using the rules generated in boosted C5.0 algorithm, we obtained the important result about low susceptibility of female to liver disease than male. This factor is missing in other studies of liver disease. Therefore, our proposed computer-aided diagnostic methods as an expert and intelligent system have impressive impact on liver disease detection. Based on obtained results, we observed that our model had better performance compared to existing methods in the literature.

Keywords: Liver disease, C5.0 algorithm, Boosting technique, CHAID algorithm, Data mining, Classification

1. Introduction

In recent years, we have faced with an increasing number of data stored in various organizations such as banks, hospitals, universities and etc. that encourages us to find a way to extract knowledge from this large amount of data and to efficiently use them. Data mining is defined as a method to discover and extract knowledge from large volumes of data that is useful, practical and understandable (Han, Kamber, & Pei, 2011). It is also defined as a semi-automated way to find hidden patterns among data (Han & Kamber, 2001). One of the most important uses of data mining is the extraction of knowledge from data more accurately in a less time, less cost and possibly to have comprehensive and more complete results. This knowledge is used in various fields such as medical application, web mining, security, prevention of crime and many other fields (Witten & Frank, 2005). Medical science is one of the important areas where data mining is used. Since this branch of science deals with human life, it is highly sensitivities. In recent years, a lot of researches have been done on a variety of diseases using data mining. Looking more closely at the research done in recent years in this field, specifically, in the medical field, we can see many works that use data mining for forecasting, prevention and treatment of patients (Riganello , Candelieri, Quintieri , Conforti, & Dolce, 2010; Das, 2010; Gauthier, Nahar, Imam, Tickle, & Chen, 2013; Marateb, Mansourian, Faghihimani, Amini, & Farina, 2014; Kasabov & Capecci, 2015; Patidar, Pachori, & Acharya, 2015; Souillard-Mandar et al., 2015; Tomczak, & Zięba, 2015; Tanha, van Someren, & Afsarmanesh, 2015 Rodríguez-Jiménez et al. 2016; Alemayehu & Berger, 2016). In medical science, accuracy and speed are two important factors that should be considered chiefly in dealing with any disease. In this regard, data mining techniques can be of great help to physicians.

The organization of this paper is as follows. In section 2, some background on data mining, liver disease, classification algorithms, and related works are provided. Section 3 describes our method in the implementation of boosted C5.0 and CHAID classification algorithms for the early detection of liver disease. Finally, we conclude our paper in section 4 with some discussion and suggestion for future works.

2. Background

2.1. Data Mining

With advances in science, several machines have entered in our lives. One of the most famous areas where computers as the mostly used machines can be helpful is knowledge extraction with the help of a machine (machine learning). This approach that can be of great help to all scientific fields is called data mining or Knowledge Discovery of the Databases (KDD). Supervised and unsupervised learning are two main methods for machine learning (Han, Kamber, & Pei, 2011). The purpose of these methods is to learn by use of data mining approaches and to use part of data for training and the other part for test.

In fact, in supervised learning, the purpose of data mining is clear and researcher knows the desired result. While in the learning without supervision, the goal of data mining in general is not clear (Marinakis, Marinaki, & Matsatsinis, 2008). Data mining can be considered as a good way to reduce the costs in different areas of science

(Witten & Frank, 2005; Alizadehsani et al., 2013), such as healthcare, engineering, crime, security, etc.

There are three major categories of data mining including: *Clustering, Classification and Association Rule* (Deekshatulu & Chandra, 2013). There are many algorithms in each category, including (Gorunescu, 2011): Artificial Neural Network (Daş, Turkoglu, Sengur, 2009; Daş, Turkoglu, Sengur, 2009), C5.0 (Meng et al., 2013), SVM (Pazhanirajan, & Dhanalakshmi, 2016), Apriori (Khade, Lashier, & Yoon, 2016), CART and CHAID (Chen, 2016), Support Vector Regression (Xu, Liu, & Zhang, 2015), Linear Regression (Matusevich, Cabrera, & Ordonez, 2015), K-Means (Yadav, Bansal, & KumarSunkaria, 2015) and etc. By more investigation around researches conducted in recent years, we can find large number of studies performed in medical data.

The algorithms used in data mining have different powers in predicting, classification, and clustering which also depends on the type of data and how it is implemented. Research on the medical data is very important and small mistakes in the analysis and application of results causes risk at human life. To extracting knowledge from the data, data mining consists of 5 steps as shown in **Fig. 1** (Gullo, 2015).

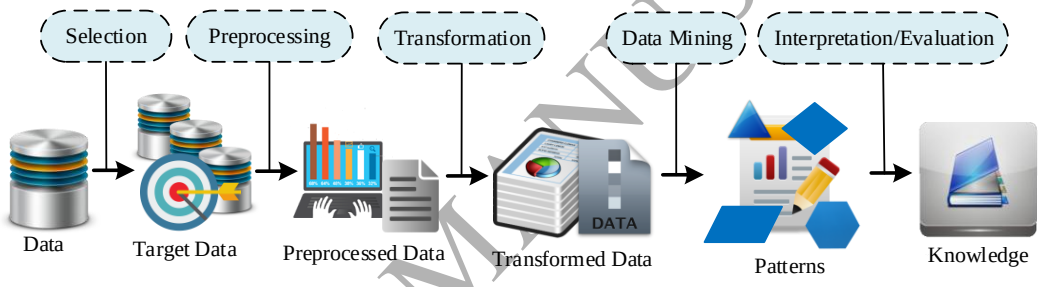

Fig.1. The Data mining or Knowledge Discovery in Databases (KDD) process (Gullo, 2015).

2.2. Liver disease

The liver is one of the largest organs of the human body (LDCR, 2015). Liver is in the right upper abdomen and below the diaphragm (LDCR, 2015) as it can be seen in Fig. 2.

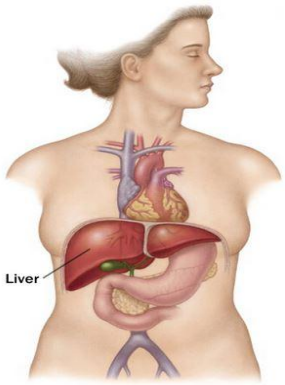

Fig.2. The location of the liver in the human body (LDCR, 2015).

The liver makes blood purification and identifies toxic substances like alcohol and excretes them. In other

words, the liver converts toxic substances into nutrients and then the body uses them to control the level of hormones in the body. The other important tasks of the liver are the production of hormones and proteins, controlling the blood sugar and to help in controlling blood clotting. There are many types of liver problems that causes more than 100 diverse diseases (LDCR, 2015; LDI, 2015) such as Neonatal Hepatitis, Primary Biliary Cirrhosis, Fatty Liver, Cirrhosis, Liver Cancer, Primary Sclerosing Cholangitis, Hepatitis, Porphyria, Reye's Syndrome, liver disease in pregnancy, Sarcoidosis, Toxic Hepatitis, Type 1 Glycogen Storage Disease, Tyrosinemia, and Wilson Disease, and among them the most common types of liver problems are (Rajeswari & Reena, 2010): fatty liver, hepatitis, cirrhosis, liver cancer.

The Canadian Liver Foundation (CLF) published a report in 2013 that shows during the last 8 years, about 30 percent of deaths in this country are due to liver disease. According to the estimates by this foundation, one person out of 10 suffers one of the various types of liver disease, that means about 3 million people in Canada have some kind of liver disease (LDCR, 2015). In (Lin, 2009) this issue has been emphasized and liver disease has been introduced among the 10 dangerous and deadly diseases in Taiwan. There are specific factors for the emergence of liver disease that most important ones are family history of liver disease, smoking, alcohol consumption, consumption of contaminated food, obesity, and diabetes (Vijayarani & Dhayanand, 2015**).** Each of these factors has a great effect in diagnosis of liver disease.

2.3 Classification algorithms

In this section, we present a brief introduction to the algorithms used in this paper. One of the best categories of data mining algorithms for creating rules is the decision trees algorithms. *Boosted C5.0* and *CHAID* algorithms are two important and useful algorithms which are based on the decision trees. These two algorithms are used in this paper to create rules from liver disease dataset. To the best of our knowledge, until now there is no work in the literature that uses Boosted C5.0 and CHAID for creating the rules in liver disease.

2.3.1 C5.0 algorithm

In the late 1970s, Ross Quinlan introduced and provided ID3 algorithm that operates based on decision trees (Quinlan, 1979; Quinlan, 1986). In 1993, in order to overcome to some drawbacks related to ID3 method, he expanded this algorithm and developed C4.5 algorithm based on principles of ID3 and as result, he showed that the C4.5 algorithm has better performance than ID3 algorithm (Quinlan, 1996; Quinlan, 2014). This algorithm was improved more, again by Ross Quinlan and finally he introduced the C5.0 algorithm as the last implementation of this family (Pang & Gong, 2009; Kuhn & Johnson, 2013; Chambers, & Dinsmore, 2014). This algorithm is the result of upgrading ID3 and C4.5 algorithms. It can be used to create a decision tree or a set of rules which are usually simpler and more accurate than trees. The basis of C5.0 algorithm is the sequential analysis. Each sub-sector is re-analyzed to small pieces until it cannot be parsed more. One of the important points about this algorithm is the ability of the algorithm to predict a target (Quinlan, 1986; Quinlan, 2014; Quinlan, 1996; Wu et al., 2008). One of the criteria of the C5.0 algorithm to test the features is *Gain Ratio* which is a modification of the

information gain that reduces its bias. If we consider the *S* as a set of training samples and also *X* includes *n* attributes and divides *S* into *n* subsets *S1, S2… Sn*, and, and n_i = The instance number of decision attribute (Hou et al., 2014),

H(X) = the information entropy of subset, and

For *S= {S1, S2, S3, …, Sn}*,

$|S|$ = Count of samples in *S*, and

If *i= {1, 2 , 3, …, N}* The number of sample which belongs to C_i is: Ci=$freq(Ci,S)$,

And the probability of a sample to belongs to Ci is C_i =$\log_2(freq(ci,s))/|S|$,

$$info(S) = -\sum_{n=1}^{N}\left(\frac{freq(Ci,S)}{|S|}\log_2\frac{freq(Ci,S)}{|S|}\right) \quad (1)$$

$$info_x(S) = -\sum_{j=1}^{n}\frac{|Si|}{S}\times info(Si) \quad (2)$$

Gain (X) = info (S) – info $_x$ (Si) (3)

$$H(X) = -\sum_{i=1}^{N} \mathrm{Pi}\log_2 \mathrm{Pi}\ ,\ \mathrm{Pi} = \frac{ni}{|X|} \quad (4)$$

$$Split\ Info(X) = -\sum_{i=1}^{N}\left(\frac{|Si|}{|S|}\times\log_2\frac{|Si|}{|S|}\right) \quad (5)$$

$$Gain\ ratio = \frac{\Delta info}{Split\ Info} \quad (6)$$

Thus to create a tree or a set of rules, this algorithm works by subsequent division of the sample so it can produce the maximum information about gain.

Each sample is divided into sub-sections and this process continues until samples cannot be divided into sub-sections. Eventually, lowest splits of the tree, which are the most worthless of them are removed or pruned. This algorithm is able to produce two types of models (Figueiredo, Rodrigues, Vale, & Gouveia, 2005):

1. A decision tree: This model is a straightforward and simple description of the splits which have been found by the algorithm.
2. A rule set: In fact, this model indicates a simplified version of available information in the decision tree.

As decision trees algorithm, C5.0 has several the strengths which some of these strengths are as follows (Lantz, 2013):

1. C5.0 works on various problems as an all-purpose classifier;
2. More productivity rather than existing methods;
3. Utilizes just the most important and significant features;
4. Works with both relatively few or a very large samples for training.

As a worth, we mention to the some drawbacks of the proposed method which are as follows (Lantz, 2013):

1. C5.0 and other decision tree methods are frequently biased toward splits that features have a large number of levels;
2. Some problems for model may occur such as over-fit or under-fit challenges;
3. Major changes to decision logic can outcome of small changes in the training data;
4. Because of relying on axis-parallel splits, C5.0 can have inconvenience modeling the number of relations.

Boosting technique: One of the important features of C5.0 algorithm is the capability to utilize boosting technique. Although it is possible that by using this technique the runtime of the algorithm is increased, at the same time in most cases, the accuracy of prediction will also increase. The class imbalance can be considered a challenge in data mining. Unbalanced classes of data means that the number of data in one class are higher than the other class or classes and C5.0 algorithm is an effective solution to deal with this challenge. This technique was implemented in C5.0 as an improvement of this version over C4.5. The first works for applying boosting technique to C4.5 was carried out in (Quinlan, 1996; Freund & Schapire, 1997). For more clarity, here is a brief description of boosting technique in C5.0 and more details can be found in the original papers. We rewrite the same steps mentioned in (Pang & Gong, 2009) for description of boosting technique.

It is based on following assumptions:
There is a samples set *S* of *N* samples and for sample number i, *i = 1, 2, ..., N,*
T = the number of decision trees, t = 1, 2, ..., T,
C^t = the decision tree by learning system generates in trail *t*,
C^* = the final decision tree,
P_i^t = the normalized factor of w_i^t,
β^t = the factor for weight adjustment,
$\theta^t(i)$ = 1if sample *i* is misclassified and otherwise is 0,

In general, the Boosting technique contains the following main steps as described in (Pang & Gong, 2009):
Step 1: Initialize the variables: set T=10 (value of T usually is 10) and also t=1 then w_t^1=1/n.
Step 2: When $\sum_{i=0}^{n}(p_i^t) = 1$ calculate$P_i^t = w_i^t / \sum_{i=0}^{n}(w_i^t)$.
Step 3: For each sample use of P_i^t as weight then create C^t under this distribution.
Step 4: Using the $\varepsilon^t = \sum_{i=0}^{n}(p_i^t\, \theta_i^t)$ for calculation the error rate of C^t
Step 5: If $\varepsilon^t < 0.5$, the trails are closed, put T = t + 1;
Else if $\varepsilon^t = 0$, the trails are closed, put T = t;
Else if $0 < \varepsilon^t < 0.5$, continue from step 6.
Step 6: β^t is calculated as $\beta^t = \varepsilon^t / (1 - \varepsilon^t)$
Step 7: Considering the error rate, Weight adjustment is calculated as follows:
$w_i^{t+1} = w_i^t \beta^t$ sample is misclassified and w_i^t sample is classified rightly
Step 8: If t = T, the trails are closed.
Else, put t = T + 1 and go to step 2 to start the next trail.

After doing all these steps, we generate Boosted C^* from the total votes in the decision trees (C^1, C^2,…., C^T). It should be noted that the vote for C^T is worth log ($1/\beta^t$) units. Also, C^* is calculated as follows:

$C^* = \sum_{t=1}^{T}(1/\beta^t)C^t$ (7)

In general, to carry out classification on a sample, the sample is initially, classified by C^T ($1 \leq t \leq T$) and accordingly results can be received. Finally, on the basis of the weight of C^T ($1 \leq t \leq T$) the final vote of each class can be calculated. The final result is the class with the highest vote.

2.3.2 CHAID algorithm

CHAID is a type of Decision Tree (DT) that was developed by Kass in 1980s (Kass, 1980). This algorithm stands for CHi-squared Automatic Interaction Detection that can be used for prediction and classification. Due to the use of categorical inputs (as predictors) and as well as targets by this algorithm, a chi-square test will be calculated between the target feature and each existing predictor and as result the best predictor will be utilized in order to partition the sample. The process will be continue with each part until do not remain any significant splits. (Chambers, & Dinsmore, 2014; Althuwaynee, Pradhan, & Lee, 2016). This algorithm is a statistical method which is appropriate for classification. Using statistical tests, CHAID evaluates all attribute values of predictor. In this algorithm the values that are considered statistically homogeneous are integrated due to the target variable and to retain all the values which are heterogeneous (dissimilar). Then this algorithm selects the best predictor of the branches of the tree, so that all child nodes will be selected with homogeneous values of attribute. The indicators used in CHAID are P-values that used to find the best attributes among data. The lowest value of P-values will be selected to divide each node. As mentioned, P-value is used as a criterion for investigating quantitative dependent variables. The great feature of this algorithm is that it can produce non-binary trees which means that the number of splits have more than two branches. This algorithm works by the following steps:

Step 1: In the first step, categorical predictors are generated of each continuous predictor which are obtained by dividing the respective continuous distributions into a number of categories nearly equal number of to observations.

Step 2: With more than two columns, by combining the column categories, it will find out the best appropriate formed.

Step 3: In this step, categories combined at step 2, can be divided into smaller parts.

Step 4: The optimal combination will be completed. After finding the best of them optimally, they are combined with explanatory variables. Then "Bonferroni" which its important objective is to account for multiple testing; is calculated to adjust chi-squared test for reducing table in each explanatory variable.

Step 5: The most significant values available in step 4 will be used to split the node with regards to the merged categories for that variable and this is done for all offspring.

There are several significant characteristics for the popularity of CHAID algorithm which are as follows (Ritschard, 2010; Althuwaynee, Pradhan, & Lee, 2016):

1. For each node, this algorithm specifies for each potential predictor the optimal n-ary split it would generate. Then chooses the predictor according to these optimal splits.

2. The characteristics for CHAID algorithm is the use of P-value by a *Bonferroni* correction as splitting criteria.
3. The CHAID algorithm is able to expand each node to two or more branches as well as to produce the final tree; there is no need for pruning.
4. Actually this algorithm makes use of categorical or ordinal data and continuous data automatically will be converted to ordinal.
5. In the case of distributed data, CHAID has better performance than regression to perform the clustering.
6. CHAID is not needed to Pre-processing of the relationship between dependent variable and conditioning factors.

The CHIAD algorithm uses two kinds of statistical tests for management of the data.

1. If the dependent variable is categorical, Pearson's chi-squared test is applied as follows:

Assume:
n_{ij} = the observed cell frequency,
m_{ij} = the estimated expected cell frequency for ($x_n = i$, $y_n = j$),
f_n = the frequency weight associated with case,
x_n and y_n = whole learning sample,
I = degrees of freedom,
D = represents the relevant data,
Then:

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}, \quad (8)$$

$$n_{ij} = \sum_{n\in D} f_n\, I(x_n = i \cap y_n = j), \quad (9)$$

The corresponding P-value is presented in Formula 10 (Baker & Cousins, 1984):

$$P = \Pr(x_d^e > x^2) \quad (10)$$

2. For scale-dependent variables, the F-test will be used as below:

$$F = \frac{\sum_{i=1}^{I}\sum_{n\in D} w_n f_n I(X_n=i)(\bar{y}_i-y)^2 / (I-1)}{\sum_{i=1}^{I}\sum_{n\in D} w_n f_n I(X_n=i)(y_n-\bar{y}_i)^2 / (N_f-I)} \quad (11)$$

Which for y_n, y and N_f are defined by the following formulas (12-14):

$$y_n = \frac{\sum_{n\in D} w_n f_n y_n I(x_n=i)}{\sum_{n\in D} w_n f_n I(x_n=i)} \quad (12)$$

$$y = \frac{\sum_{n\in D} w_n f_n y_n I}{\sum_{n\in D} w_n f_n I} \quad (13)$$

$$N_f = \sum_{n\in D} f_n \quad (14)$$

Finally, the corresponding P-value is calculated as follows:

$$P = P(F(I-1, N_f - 1) > F). \quad (15)$$

2.4 Related works

In recent years, numerous studies have been conducted on a variety of diseases around the world by using data mining techniques (Daş, Turkoglu, Sengur, 2009), each of which has helpful hints and are invaluable to the scientific community. In this section we review most important works in data mining related to liver disease.

Ramana, Babu and Venkateswarlu (2012) have compared and evaluated the data on liver disease in America and India. Their results show that there is a significant difference by combining all possible attributes in groups.

They also suggest that, there are no significant differences in SGPT features in the absence of liver disease in America and India. In another study conducted by Bahramirad, Mustapha and Eshraghi (2013) two types of data for liver patients from the reservoir of University of California Irvine (UCI) have been by using eleven classification algorithms. In their study, two datasets for liver disease named BUPA and AP were compared with various data mining techniques. Their results have shown that accuracy of AP dataset is slightly better than BUPA. The reason for this issue is the type and the number of data. But also in some comparisons, BUPA data have shown better performance than AP. Their investigations also showed that the accuracy of AP data is fewer than BUPA data. Hunt, Yuen, Stirnadel-Farrant and Suzuki (2014) also used data mining to investigate liver disease. They studied the relationship between age and liver disease and reactions to drugs for liver diseases. The percentage of liver patients in three age groups including 0-17, 18-64 and 65 years was 6%, 62% and 32% respectively. In fact according to the statistics, children are less prone to liver disease than adults. Anisha, Reddy and Prasad (2015) have examined the liver cancer by using image processing and data mining. Their method for diagnosis of liver cancer is MRI, CT and USG scan imagery. In their paper, K-means algorithm for clustering has been used. In the next stage by using the Haar Wavelet the threshold values have been calculated and the accuracy of this calculation is 82%.

Alfisahrin and Mantoro (2013) implemented and compared the performance of Naive Bayes, and NB Tree algorithms on the liver disease data to help physicians in accurate and timely diagnosis. Their results show that Naive Bayes algorithm is more accurate in prediction of liver disease; however, the NB Tree algorithm has less run time. Tiwari, Sharma and Krishna (2013) have also analyzed the performance of these algorithms in liver patients by using different data mining techniques such as BP, RBF, SOM, SVM that operate based on the Artificial Neural Network (ANN). The result of their study showed that SVM algorithm has high correctly classified success rate 99.76 percent and 97.70 percent related to men and women data. It also showed that the classification based on ANN can be used as an important method to predict liver disease.

In (Abdar, 2015) different algorithms in the IBM SPSS Modeler and Rapid Miner software has been compared. Eight algorithms were implemented on liver disease data and the results showed that the C5.0, had the best performance with 87.91 percent accuracy. In (Nagaraj & Sridhar, 2015) authors have introduced a hybrid algorithm called NeuroSVM model in order to increase the accuracy and it is developed using SVM and feed-forward artificial neural network (ANN). They reported 98.83 percent accuracy for NeuroSVM algorithm when 70% of data set used for training and 30% of data set used for testing.

In (Ramana, Babu, & Venkateswarlu, 2011; Jin, Kim, & Kim, 2014), authors examined the data related to liver disease by using WEKA in six algorithms. They obtained the best precision in Naïve Bayes with 95.10 percent but Random forest and Logistics algorithms had best recall and sensitivity while they had the worst specificity. The average in reports of the first paper was 96.552 and for the second paper that used the decision tree the best accuracy was 69.40 percent.

Montazeri, Montazeri, Beygzadeh and javad Zahedi (2014) studied the performance of QUEST, C5.0, CRT and CHAID on the data related to Indian liver disease data set. They used two techniques including INN

classifier and Naïve Bayes classifier. The results of their research shows that the best accuracy by using INN classifier technique in C5.0 is 71 percent and the best accuracy by using Naïve Bayes classifier in QUEST and CHAID is 55 percent. Weng, Huang & Han, (2016) have investigated four datasets, including the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, the Indian Liver Patient Dataset (ILPD) dataset, the Vertebral Column Data Set (VCDS) and Heart Disease Data Set (HDDS), using artificial neural network algorithm. Their study showed that the best accuracy in the Individual Classifier (IC) and Ensemble Classifier (EC) for the Indian Liver Patient Dataset (ILPD) has been 79.38 % when K=5 and classifier were EC–IC-3 and EC–IC-Best.

3. The implementation of classification algorithms

In this paper, we have used Boosted C5.0 and CHAID algorithms that are relevant to the Decision Trees in order to discover hidden knowledge in the liver disease dataset in UCI repository. In this regard, we benefited from IBM SPSS Modeler 14.2 software *(Firat University license)* and evaluated the algorithms. For our purpose, the data are divided into two groups: training and testing. For more clarity, all stages of this research is presented in **Fig. 3**:

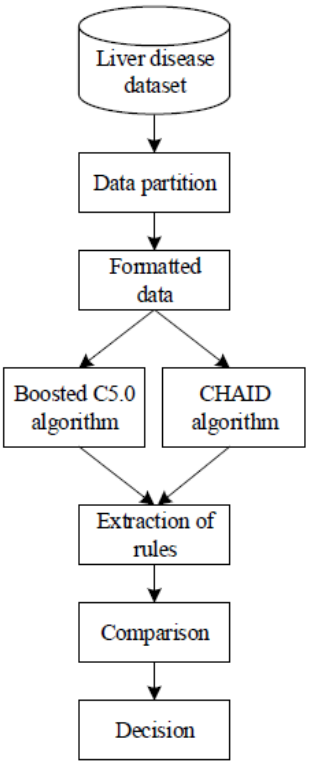

Fig.3. proposed system for performance analysis of classification algorithms.

In this regard, the implementation steps used in this paper are as follows:

1) In the first stage of implementation, we used liver disease data in IBM SPSS Modeler.

2) In the second stage, in order to predict, it is necessary for the algorithms to be trained using data and then training phase is run on the other part of data. The data were divided into three groups: training, testing, validation which are as follows: 60% for training, 30% for testing and 10% for validation. IBM SPSS Modeler has specific node which divides the data into these groups.

3) In this stage, data is cleared because some of them do not have value. Therefore, using the Most Common Attribute Value approach (Grzymala-Busse and Hu, 2000), for missing values in A/G ration which were include 4 numbers, 1.0 has been utilized as number that occurs more than other numbers for all the unknown values of the attributes.

4) In the final stage of the algorithm, we determined our data as the target data and inputs data. Therefore, at this stage, we consider the selector (class) factor to target and other factors as inputs.

3.1 Dataset

The dataset used in this paper is based on the liver disease patients from India in the data repository of the University of California, Irvine (UCI) in 2012 namely the Indian Liver Patient Dataset (ILPD) (UCI, 2012). These data include 416 records of patients with liver disease and 167 records of healthy persons without liver disease. In this dataset 441 records belong to male and 142 records are female patients all of them from different ages. One of the points to be mentioned in this dataset is that for all people who are older than 89 years, their age factor is equally considered 90. It can be seen from Table 1 that the available data has 10 features and feature number 11 indicates the state of liver disease:

Table 1. Attributes of liver disease as indicated in (UCI, 2012)

No.	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 *	[1-2]

* used to split the data into two sets (class 1: 416 liver patient records and class 2: 167 non-liver patient records) (UCI, 2012).

3.2 Rule extraction for Liver disease

To test the functionality of boosted C5.0 and CHAID for detecting the liver disease, we used these algorithms for extraction of rules among liver disease dataset and to test if these rules can predict appropriately. Below the rule extraction using these algorithm is described.

In this paper, we have used the classification methods for production of rules among liver disease dataset.

In the first step, both algorithms were implemented which then all of the rules and most important factors were obtained. The second step was dedicated to evaluating the performance of algorithms and also identifying and recommending sensitive areas in each of the important factors. Although there are many algorithms for classification, in this paper we have used two important algorithms in decision trees including boosted C5.0 and CHAID algorithms. Using the rules generated by these algorithms are fully understandable and accurate which results in timely and correct diagnosis of liver disease.

In some studies like: (Alfisahrin & Mantoro, 2013; Montazeri, Montazeri, Beygzadeh, & javad Zahedi, 2014; Jin, Kim, & Kim, 2014; abdar, 2015; Tanha, van Someren, & Afsarmanesh, 2015) the decision trees with different approaches have been used on liver disease dataset. But in our study, we use the C5.0 with boosting technique and CHAID algorithms to achieve the three objectives as follows:

1. Acceptable and good accuracy;
2. The understandable and simple rules;
3. Finally, we use mentioned algorithms in order to sort the factors based on the role of each factor in the emergence of liver disease from the most significant factor to the least significant one.

One of the key points in the rules generated by boosted C5.0 and CHAID algorithms is that the rules are very simple. Thus, these rules can be used very easily by physicians and also other individuals in hospital. In the following, more details about the implementation steps of algorithms are presented. The proposed methodology, which is illustrated in **Fig. 4**, boosted C5.0 and CHAID are implemented with the IBM SPSS Modeler 14.2 software enterprise miner streamlines the entire data mining process from data access to model evaluation.

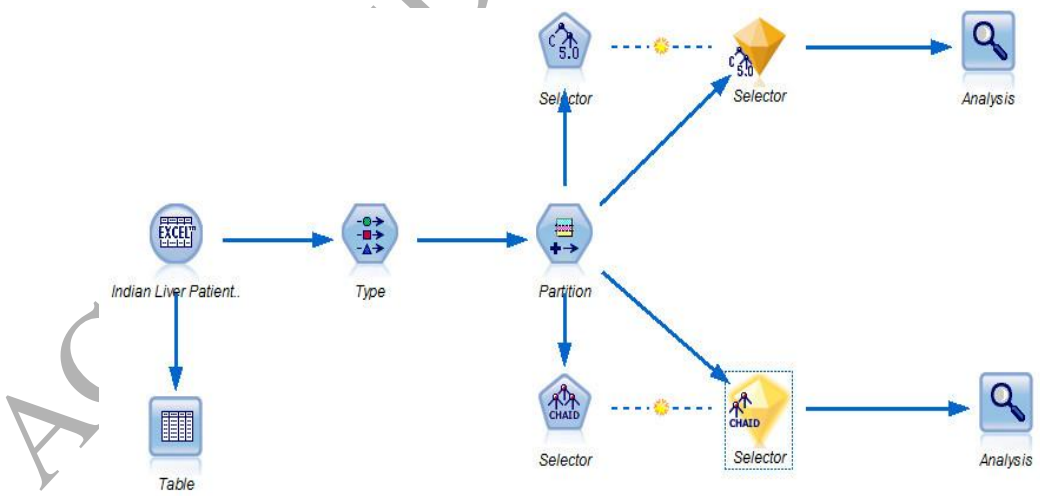

Fig. 4. Proposed system for liver disease recognition.

In the first part of implementation, CHAID algorithm was implemented. First of all, four records were filled due to the lack of values as earlier explained. Through this action, we achieved more realistic prediction and

the accuracy of algorithms on liver disease was increased. In this implementation, the depth of the tree generated by CHAID algorithm is 4. **Fig. 5** shows five relevant factors in the prediction of liver disease according to their importance.

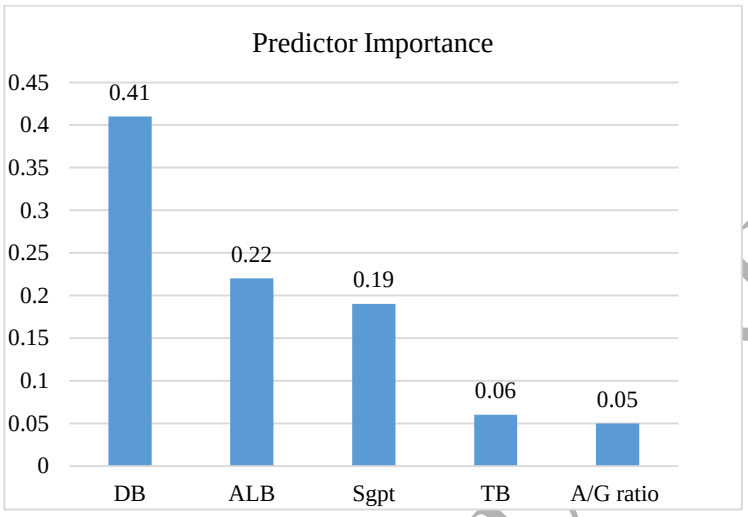

Fig. 5. The importance of the factors in the prediction of liver disease by using CHAID algorithm.

According to **Fig. 5.** predictor importance for DB, ALB, Sgpt, TB and A/G ratio are 0.41, 0.22, 0.19, 0.06 and 0.05 respectively. The rules generated by this algorithm is shown in Table 2. According to the Table 2, it can be seen that 9 rules have been produced by CHAID algorithm. It is also clear that CHAID 6 rules have generated for class 1 (class 1 is devoted to liver diseases) and 3 rules for class 2 (class 2 is devoted to non-liver diseases).

Table 2. Rules generated by the CHAID algorithm.

Num	Rules
1.	IF DB <= 0.900 and A/G <= 1.180 and Sgpt <= 35 and ALB <= 4.100 THEN class 1.0
2.	IF DB <= 0.900 and A/G <= 1.180 and Sgpt <= 35 and ALB > 4.100 THEN class 2.0
3.	IF DB <= 0.900 and A/G <= 1.180 and Sgpt > 35 THEN class 1.0
4.	IF DB <= 0.900 and A/G > 1.180 and ALB <= 3.100 THEN class 1.0
5.	IF DB <= 0.900 and A/G > 1.180 and ALB > 3.100 THEN class 2.0
6.	IF DB > 0.900 and ALB <= 2.900 THEN class 1.0
7.	IF DB > 0.900 and ALB > 2.900 and ALB <= 3.100 and TB <= 3.600 THEN class 2.0
8.	IF DB > 0.900 and ALB > 2.900 and ALB <= 3.100 and TB > 3.600 THEN class 1.0
9.	IF DB > 0.900 and ALB > 3.100 THEN class 1.0

The rules generated by CHAID algorithm are shows that this algorithm is suitable for the diagnosis of liver disease (class 1 related to liver disease).

In the second experiment boosted C5.0 algorithm was implemented where number of trials for boosting was 7 and number of folds for cross-validate was 30. The depth of trees produced by this algorithm was 14, which is very different compared to the CHAID algorithm. The importance of various factors in the prediction of liver patients using boosting in C5.0 algorithm has been shown in Fig. 6, and the rules generated by this algorithm can

be seen in Table 3:

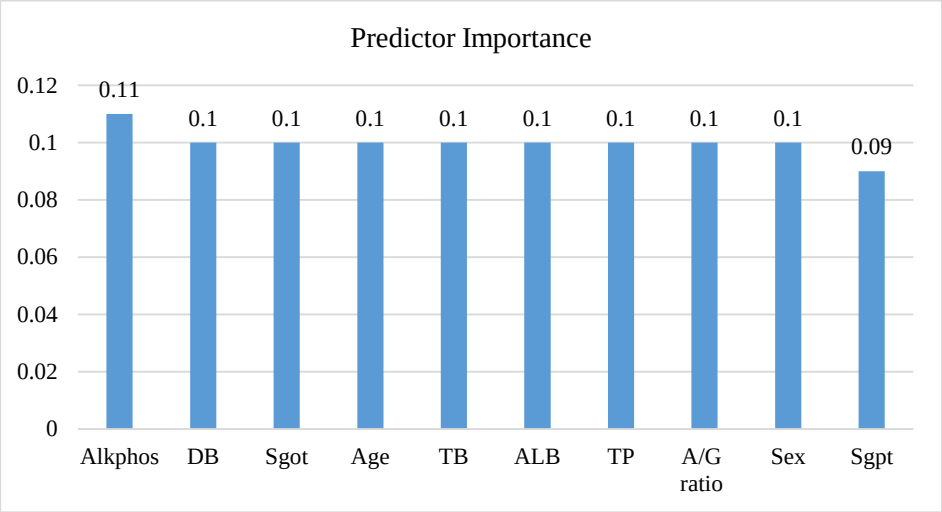

Fig. 6. The importance of the factors in the prediction of Liver disease by using boosted C5.0 algorithm.

According to **Fig. 6**, predictor importance for these factors, and Sgpt are in three groups which are as follows:

1. Predictor importance for Alkphos is 0.11;
2. Predictor importance for DB, Sgot, Age, TB, ALB, TP, A/G ratio and Sex is 0.10;
3. Predictor importance for Sgpt is 0.09

Table 3. Some rules generated by the boosted C5.0 algorithm.

Num	Rules
1.	IF DB <= 1.200 and Sgpt <= 65 and TB <= 1.600 and Alkphos <= 211 and DB <= 0.100 and SEX = Male THEN class 1.0
2.	IF DB <= 1.200 and Sgpt <= 65 and TB <= 1.600 and Alkphos <= 211 and DB <= 0.100 and SEX = Female and Alkphos <= 153 THEN class 2.0
3.	IF DB <= 1.200 and Sgpt <= 65 and TB <= 1.600 and Alkphos <= 211 and DB <= 0.100 and SEX = Female Alkphos > 153 and TB <= 0.600 and A/G <= 0.950 THEN class 2.0
4.	IF DB <= 1.200 and Sgpt <= 65 an d TB <= 1.600 and Alkphos <= 211 and DB <= 0.100 and SEX = Female and Alkphos > 153 and TB <= 0.600 and A/G > 0.950 THEN class 1.0
5.	IF DB <= 1.200 and Sgpt <= 65 and TB <= 1.600 and Alkphos <= 211 and DB <= 0.100 and SEX = Female and Alkphos > 153 and TB > 0.600 THEN class 1.0
6.	IF DB <= 1.200 and Sgpt <= 65 and TB <= 1.600 and Alkphos <= 211 and DB > 0.100 and AGE <= 21 THEN class 2.0
7.	IF DB <= 1 and Age > 17 and Sgot <= 1.30 and Age <= 65 and Age > 58 and A/G > 1.390 THEN class 1.0
8.	IF DB <= 3.600 and Sgot <= 64 and DB <= 0.300 and Alkphos > 130 and DB <= 0.100 and TB <= 0.700 and Age <= 61 and Alkphos > 156 and TB > 0.600 THEN class 1.0
9.	IF DB <= 3.600 and Sgot <= 64 and DB <= 0.300 and Alkphos > 130 and DB <= 0.100 and TB <= 0.700 and Age > 61 THEN class 1.0
10.	IF DB <= 3.600 and Sgot <= 64 and DB <= 0.300 and Alkphos > 130 and DB <= 0.100 and TB > 0.700 THEN class 2.0
11.	IF DB <= 3.600 and Sgot <= 64 and DB <= 0.300 and Alkphos > 130 and DB > 0.100 and TB <= 0.700 and ALB <= 2.300 THEN class 2.0
12.	IF DB <= 3.600 and Sgot <= 64 and DB <= 0.300 and Alkphos > 130 and DB > 0.100 and TB <= 0.700 and ALB > 2.300 and TB <= 0.600 THEN class 1.0
13.	IF DB <= 3.600 and Sgot <= 64 and DB <= 0.300 and Alkphos > 130 and DB > 0.100 and TB <= 0.700 and ALB > 2.300 and TB > 0.600 and ALB <= 4.200 and Sgpt <= 36 and Sgot <= 14 THEN class 1.0
14.	IF DB <= 3.600 and Sgot <= 64 and DB > 0.300 and TB <= 5.900 and Age <= 38 and SEX = Male and ALB > 4

	THEN class 2.0
15.	IF DB <= 3.600 and Sgot <= 64 and DB > 0.300 and TB <= 5.900 and Age <= 38 and SEX = Female THEN class 2.0
16.	IF DB <= 3.600 and Sgot <= 64 and DB > 0.300 and TB <= 5.900 and Age > 38 and DB <= 1 and TP <= 6.200 and Alkphos <= 314 THEN class 2.0
17.	IF DB <= 3.600 and Sgot <= 64 and DB > 0.300 and TB <= 5.900 and Age <= 38 and SEX = Male and ALB <= 4 and Sgpt <= 22 THEN class 2.0
18.	IF DB <= 3.600 and Sgot <= 64 and DB > 0.300 and TB <= 5.900 and Age > 38 and DB <= 1 and TP <= 6.200 and Alkphos > 314 THEN class 1.0
19.	IF DB <= 3.600 and Sgot > 64 and ALB > 2.200 and Sgot <= 298 and TP > 5.200 and Age <= 39 and TP <= 7.900 THEN class 1.0
20.	IF DB <= 3.600 and Sgot > 64 and ALB > 2.200 and Sgot <= 298 and TP > 5.200 and Age <= 39 and TP > 7.900 THEN class 2.0

Note that in overall 92 rules have been generated through using the boosted C5.0 but we have only presented 20 rules of them randomly. The rules in Tables 2 and 3 are simple and understandable for physicians and their colleagues in hospital. Thus, these findings can be useful as an appropriate solution to identify individuals with liver disease and without liver disease in the real environment. In other words, by using the results of this study, more effective rules for the diagnosis of liver disease can be achieved. Based our results and according to Table 3, we observed that boosted C5.0 creates more rules for liver disease than CHAID which are 92 rules. One of the important advantages of the rules generated by boosted C5.0 algorithm is the ability to represent more details. According to obtained outcomes and as well as Table indicates some generated rules by boosted C5.0, we observed that C5.0 has produced 50 rules for liver disease (class 1) while 42 rules have been produced for non-liver disease (class 2). With regard to the results and some rules in Table 3, it can be deduced that boosted C5.0 is more suitable for people in class 1. Generally, by comparing our results and as well as Tables 2 and 3, it can be said that CHAID and boosted C5.0 are more successful in identifying liver disease (class 1). But it should be noted that the rules produced by boosted C5.0 have shown more details about both classes.

Another result of this paper is related to the importance of gender (sex) in diagnosis of liver disease. As can be seen in Tables 2 and 3, gender of patients' factor has been used only in boosted C5.0 algorithm to produce the rules. In general, 92 rules have been produced by boosted C5.0 that there is gender (Sex) factor in 32 rules. By evaluating all generated rules as some of them presented in Table 3, it is revealed that there are 15 rules for male while 17 rules for female. According to Table 4, it can be said that female are more susceptible to liver disease than male about 1.13 times more likely. Another important point is that boosted C5.0 for non-liver patient has created more rules for female than male (9 rules for female and 8 rules for male). Furthermore, these results show that boosted C5.0 had a better performance to create rules for female.

Table 4. The number of each Sex in two classes by boosted C5.0 algorithm

	Sex	
Class	Male	Female
Class 1 (liver patient)	7	8
Class 2 (non-liver patient)	8	9

Generally, the two algorithms produced different rules, but according to the above results, we would argue that females are more at risk of developing liver disease (in our study 53.125 percent of all generated rules by boosted

C5.0 for female and 46.875 percent of all rules for male) but we do not have enough experimental results and for this reason in order to study on effect of Sex in liver disease through more studies and various data are required. This finding is more analyzed in the following sub-section.

3.2 The Experimental classification results

By comparing Tables 2 and 3 it can be seen that the boosted C5.0 algorithm has better performance than CHAID algorithm and examines the factors in the smaller intervals and this increases the accuracy of the algorithm. One of the important achievements of this paper is that there are only five effective factors in CHAID algorithm for prediction of liver diseases, while this number for boosted C5.0 algorithm, is 10 factors which indicates that this algorithm use the factors with greater caution. It should be noted that in medical science consider all aspects of diseases to keep of patients' lives. Since our proposed approach has used all factors, we believe that this is an achievement in the prediction of liver disease. Indicating the importance of DB factor in the prediction of the liver disease is one of the important aspects of creation of rules by boosted C5.0 and CHAID.

According to Table 2 it turns out that in the CHAID algorithm, five factors include DB, ALB, Sgpt, TB and A/G are important in the diagnosis of liver disease, while according to Fig. 6 and Table 3 of boosted C5.0 algorithm, all factors which are involved in the dataset have role in predicting of liver disease but the importance of each factor in both algorithms will vary and can be seen in Fig. 5 and Fig. 6, that DB, ALB, Sgpt, TB and A/G factors as joint (common) factors between boosted C5.0 and CHAID algorithms are important for diagnosing of liver disease in both algorithms. Our investigation of obtained rules about these factors shows that in the CHAID algorithm the number of cut-off points at each factor are 1, 3, 1, 1 and 1 respectively while cut-off point in the boosted C5.0 algorithm are 7, 5, 8, 8 and 5 respectively. In Table 5, more details are involved about cut-off points. Further investigation also shows that TB and SGPT factors are two other important factors for the prediction of liver disease.

Table 5. Cut-off points in the various factors

Algorithm	Factors				
	DB	**ALB**	**Sgpt**	**TB**	**A/G**
Boosted C5.0	0.100, 0.200, 0.300, 0.500, 1.00, 1.200, 3.600	2.200, 2.300, 4.00, 4.200, 4.300	16, 22, 30, 32, 34, 36, 37, 65	0.600, 0.700, 0.800, 0.900, 1.00, 1.200, 1.600, 5.900	0.750, 0.950, 1.250, 1.390, 1700
CHAID	0.900	2.900, 3.100, 4.100	35	3.600	1.180

It can be seen from Table 5 that boosted C5.0 and CHAID have different cut-off points. Each of these numbers are very effective in the prediction of liver disease and with the slightest change, there is the possibility of changing class of diseases to healthy and vice versa. For instance, consider the rule number 19 and 20 in Table 3, it can be seen that all parts of the rules are similar with the exception of TP, when TP <= 7.900 class is 1, while if TP > 7.900 class is 2. As it is observed from Table 3, when the range of factors in the rules are changed, the type of classes will also be changed. For this purpose and according to the important points in these two algorithms, we propose an applicable range for each factor to be used in identification of liver disease as follows:

1. DB = [0.900 – 1.200]
2. ALB = [4.00 – 4.300]
3. Sgpt = [34 – 37]
4. TB = [0.600 – 1.200] (just by boosted C5.0)
5. A/G = [1.180 – 1.390]

In the CHAID algorithm, where DB <= 0.900 the probability of liver disease is 58.130 %, while when the DB > 0.900, the probability of liver disease is 95.283% (P-value = 0.00 and Chi-square = 48.079). One of the important point about tree generated by CHAID is that when DB > 0.900 and ALB > 3.100 the probability of liver disease is 100 % while when ALB <= 2.900 and ALB = (2.900 – 3.100] the probability of liver disease are 96.721% and 70% respectively (P-value = 0.011 and Chi-square = 16.236). However, when DB > 0.900, ALB = (2.900 – 3.100] and TB > 3.600 the probability of liver disease is 100% (P-value = 0.038 and Chi-square = 4.286). Another important result is that when DB <= 0.900, A/G <= 1.180, Sgpt <= 35.00 and ALB > 4.100 the probability of non-liver disease is 100 % (P-value = 0.036 and Chi-square = 8.305). Also, according to the boosted C5.0 algorithm, if DB <= 1.200, the probability of the liver disease is 63.218 % while when DB > 1.200 the probability of the liver disease is 95.270 %. It should be noted that when DB > 1.200 and Sgpt > 32.000 the probability of the liver disease is 98.333 %, whereas when DB > 1.200, Sgpt <= 32.000 and Age > 42 the probability of the liver disease is 100 %. In another case, when DB <= 1.200, Sgpt <= 65.00, TB > 1.600, Sex = female and Age <= 40 the probability of the non-liver disease is 100 %. All of the mentioned examples are only part of the whole of the established rules that the evidence indicates that these algorithms have generated these rules with a lot of details in order to prediction of liver disease. Another significant result of this experiment is the low impact of gender (sex) in the detection and diagnosis of liver disease in boosted C5.0 algorithm. Even this factor has lower role compared to other factors for the diagnosis of liver disease but should not be overlooked.

To investigate the performance of two algorithms more accurately, we used the Confusion Matrix (Das, Sengur, 2010) which can be seen in Table 6. With this matrix the amount of each indicator is calculated and then results are compared (Fawcett, 2006). This matrix is a useful tool to analyze function clustering method and to identify different categories of data. Ideal situation is when the most relevant data are on the main diameter matrix and the rest matrix values are zero or near zero (Alizadeh, Ghazanfari, & Teimorpour, 2011; Jiawei Han & Pei, 2011). Various indicators such as specificity, sensitivity, precision and accuracy of assessment categories are calculated according to the formulas 16 to 22 (Choudhury & Bhowal, 2015; Bittel, Kaiser, Teichmann & Thoma, 2015; Boubekeur, Luo & Labidi, 2015; Weng, Huang & Han, 2016).

Table 6. Confusion matrix in our study

Actual	Predicted	
	Disease (positive)	No-disease (negative)
Positive	TP	FP
Negative	FN	TN

Specificity = TNR = TN / TN+FP (16)
Sensitivity = TPR = TP / TP+FN (17)

Precision = TP / TP+FP (18)
FPR = FP / FP + TN = 1 - TNR (19)
FNR = FN / FN + TP = 1 – TPR (20)
F_1 = 2TP / (2TP + FP + FN) (21)
Accuracy = TP+TN / TP+TN+FP +FN (22)

Where:
TP = the number of positive examples correctly classified.
FP = the number of positive examples misclassified as negative
FN = the number of negative examples misclassified as positive
TN = the number of negative examples correctly classified.

By using the confusion matrix and the above formulas, the values of the mentioned indicators for two algorithms boosted C5.0 and CHAID can be specified. The results are shown in Table 7. Fig. 7, shows a comparative representation of these indicators for two algorithms.

Table 7. The measured metrics for algorithms (%)

Algorithm	Specificity	Sensitivity	Precision	FPR	FNR	F_1	Validation	Accuracy (training)	Accuracy (testing)
Boosted C5.0	91.42	94.40	97.52	8.58	5.60	95.93	91.55	92.33	93.75
CHAID	24.24	75.59	79.33	75.76	24.41	77.41	69.01	76.14	65.00

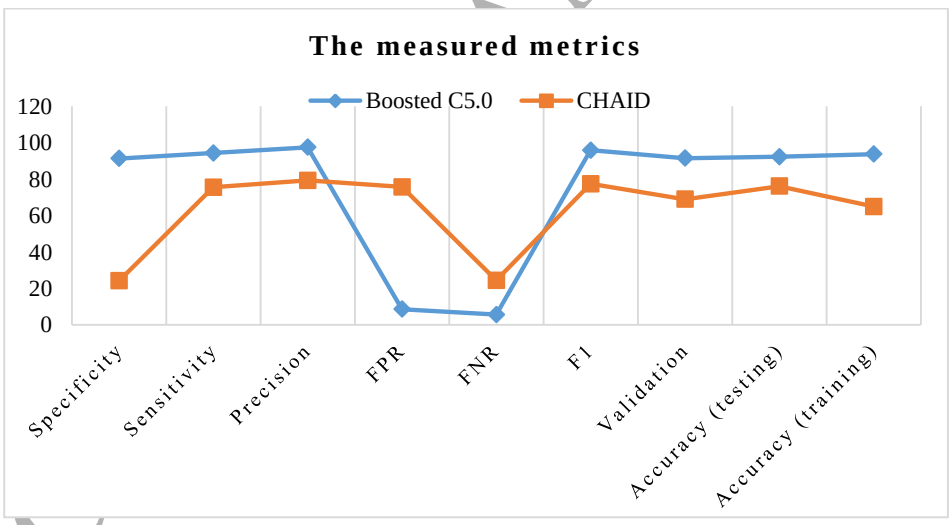

Fig. 7. A comparison between the Performance of C5.0 and CHAID algorithms using different indicators.

By more analysis of Table 7 and Fig. 7, it can be seen that in all cases, the boosted C5.0 algorithm has shown better performance on liver disease data set. The false positive rate (FPR) and the false negative rate (FNR), are referred to the types of errors. According to (Weng, Huang & Han, 2016), FPR is more important for medical centers to determine higher risks than FNR. Therefore, according to Table 7 and Fig. 7, boosted C5.0 has lower error rate compared to CHAID algorithm and for this reason, it can be concluded that it is more appropriate for the diagnosis of liver disease.

The unbalanced classes (Class Imbalance) have become one of the major challenges in data mining for classification. To address this challenge, there are different methods that one of the most important ones is boosting. In data mining, high accuracy is very important particularly when it is used in sensitive areas such as medicine science. Boosting technique helps to improve the accuracy of prediction by creating and combining multiple classifications. Boosting works with a structure of several continuing models. The first model is usually made from this method, and then the second model focuses on the records that were made in the first model. The third model is created based on the errors of the second model and the process continues. In last stage the sample is classified by counting the final vote of each class. The class with highest vote is selected for that sample. Boosting dramatically improves the accuracy of C5.0 model. But it takes more time for training addition sieve auto classifieds attributes in C5.0 and this ultimately leads to smaller and more accurate predictions (Fürnkranz, 2002; Fürnkranz, 2002; Pang & Gong, 2009; Pashaei, Ozen, & Aydin, 2015).

In this paper, we used boosting technique to increase the prediction accuracy. To evaluate the results of this paper, we compare our results with other works in this subject by using the same dataset which can be seen in Table 8:

Table 8. Comparing the results of this paper with other methods in literature on liver disease

	Best used algorithm	The most important factors	Accuracy (%)
Ramana et al. (2011)	Back Propagation	Age, Sex, SGOT, SGPT and ALP	98.00
Ramana et al. (2012)	kNN	Alkphos, SGPT and SGOT	-
Tiwari et al. (2013)	SVM	TB, DB,TP, ALB and A/G Ratio for Men data TB, DB,TP, and A/G Ratio for Woman data	99.76 (for woman data)
Alfisahrin & Mantoro, (2013)	NBTree	TB, DB, Alkphos, SGOT, SGPT and ALB	67.01
Jin et al. (2014)	Decision Tree	-	69.40
Montazeri et al. (2014)	C5.0 by using INN classifier	DB, Age, Alkphos and SGOT	71.00
Abdar, (2015)	C5.0	-	87.91
Nagaraj & Sridhar, (2015)	NeuroSVM	SGPT, SGOT, TB, DB, Alkphos and Age	98.83
Weng, Huang & Han, (2016)	Artificial neural network (ANN) when K = 5	-	79.38
Our Study	Boosted C5.0	Alkphos, DB, SGOT, Age, TB, ALB, TP, A/G, Sex and SGPT	93.75

In Ramana et al. (2012), three major factors including Alkphos, SGPT and SGOT are identified. By comparing the results of our study and in their paper, it can be seen that Alkphos, SGOT and SGPT factors have been identified as the effective factors in boosted C5.0 and SGPT in CHAID. However in our study, in addition to identifying these factors, other major factors including DB, Age, TB, ALB, TP, A/G, Sex and SGPT by boosted C5.0 were also identified. Tiwari, Sharma and Krishna (2013) in their study, have reported the accuracy for two groups, that one of them, for the men data and another group for the women data. The best accuracy by using SVM, have been 97.76 and 97.70 percent for men and women data respectively. But unlike them, in our study in addition to the good accuracy, we have proposed simple rules and the most important factors in liver disease and as

well as in our study all data at the same time have been used.

Alfisahrin & Mantoro, (2013) have stated the importance of 10 factors in the liver disease patients from India dataset. These factors, named in the order of their importance, are TB, DB, Alkphos, SGOT, SGPT, A/G Ratio, Age, ALB, Gender (Sex) and TP. Comparison of their results and the results of Fig. 6 in our study shows that both studies emphasize the importance of five main factors which include DB, Alkphos, SGOT, Age and TB. But the accuracy of our algorithm is significantly higher than their paper.

In (Abdar, 2015) accuracy with C5.0 algorithm has been 87.91 percent but in this study, we have used the Boosting technique for improving the accuracy. According to the our results in Tables 7 and 8, it can be seen that accuracy in our study, is 93.75 percent, as well as, we have found a simple rules and effective solution to identify the most important factors in liver disease while in paper by Abdar (2015) has not pointed out to this details. In (Nagaraj & Sridhar, 2015), their results show that SGPT, SGOT, TB, DB, Alkphos, and Age are important in the prediction of liver disease. In our study we realized the importance of these factors as follows Alkphos, DB, SGOT, Age, TB, ALB, TP, A/G, Sex and SGPT by boosted C5.0 and DB, ALB, SGPT, TB and A/G through CHAID algorithm. In (Ramana, Babu, & Venkateswarlu, 2011; Jin, Kim, & Kim, 2014), authors obtained the best precision in Naïve Bayes with 95.10 percent but Random forest and Logistics algorithms had best recall and sensitivity while they had the worst specificity. The average in reports of the first paper was 96.552 and for the second paper that the decision tree have been used the best accuracy was 69.40 percent while according to Table 7, best specificity, sensitivity, precision and accuracy (for testing) in our study were 91.42, 94.40, 97.52 and 93.75 percent respectively. In (Montazeri, Montazeri, Beygzadeh and javad Zahedi, 2014) the best accuracy is 71 percent when they used the INN classifier technique in C5.0 algorithm. The present study has used the boosting technique to increase the accuracy of C5.0 algorithm. As can be seen, our proposed technique has better performance than their technique by having 93.75 percent for testing accuracy compared to 71 percent for accuracy in their study. In another research that has been done by Weng, Huang & Han, (2016) the best accuracy reported on ILPD dataset has been 79.38 percent while the best accuracy in our paper with boosted C5.0 is 93.75%.

The paper were responded to the addressed questions. Firstly, the Indian liver patient dataset (ILPD) with 11 features, were considered for detection of liver disease. Second, through utilizing proposed algorithms (boosted C5.0 and CHAID), simple and comprehensible rules for liver disease were obtained (see Tables 2 and 3). Third, cut-off points in the five most important features (common between two algorithms) were presented (see Table 5). Fourth, seven metrics have been evaluated for comparison performance of the boosted C5.0 and CHAID algorithms (see Table 7). Based on our experimental results, C5.0 optimization algorithm through using boosting technique for detection/diagnosis of liver disease indicated meaningful guidance with a partial set of examples in literature, therefore providing insightful implications for physicians, hospitals, health organizations and even governments.

4 Conclusion and discussion

According to the statistics published by the relevant agencies, liver disease is among the most fatal disease which puts human life at risk. Decision trees are one of the most important and most well-known algorithms in data mining algorithms and therefore in this paper we used two algorithms named C5.0 and CHAID which are based on decision trees. One of the important features about C5.0 algorithm is the possibility to apply boosting techniques in it. Boosting techniques in C5.0 algorithm leads to increased accuracy and speed in creating rules. As there is no study in which Indian Liver Patient Dataset (ILPD) is classified with boosted C5.0. Therefore, to compare the performance of boosted C5.0 and CHAID, this dataset was also classified. According to the obtained results, we conclude that boosted C5.0 algorithm has higher accuracy and more rules than CHAID algorithm and also more important factors were found by using these algorithms particularly boosted C5.0 compared to previous studies on this disease and this increase in accuracy can be considerably important especially in medicine. In this paper the boosted C5.0 and CHAID were studied on data from the UCI repository of the liver disease. The main reasons for simple and understandable rules established by boosted C5.0 and CHAID algorithms are:

1. In general, as stated in (Das, 2010), the tree can be displayed as IF-THEN rules which are understandable to use.
2. Trees provide the possibility of Conjunction (AND) and Disjunction (OR) hypothesis.

Path from the root to the leaf determines the conjunction (AND) of the features and the tree itself makes the disjunction (OR) of these compounds. After creating the first tree, weights are determined and then the construction of the weighted tree which creates the set of rules is continued. After that, trees which define the set of rules are limited to be about the same size as the primary model. Finally, a simple average of the class of produced probabilities from each tree or set of rules will be the final prediction (i.e. no of stage weights). Considering the above description as well as description in sections 4.1 and 4.2 it becomes clear that by using the decision tree structure, C5.0 and CHAID generate tree or set of rules step by step and therefore at the last stage they consider the best answers. Also in both algorithms used in this paper, “AND” and “OR” primitives have been used for production of rules. These advantages make the generated rules in this paper, simple and understandable. It should be noted that boosting technique in C5.0 is similar to Adaboost algorithm and boosted C5.0 accuracy is much more than CHAID.

Our results show that the depth of the tree generated (depth of boosted C5.0 tree is 14) by the boosted C5.0 algorithm has higher accuracy compared to CHAID in the creation of existing rules among liver disease dataset. The accuracy of boosted C5.0 algorithm is 93.75 percent compared to the CHAID algorithm which has the accuracy of 65.00 percent. Three much more important factors in boosted C5.0 that should be considered are as follows: Alkphos, SGOT and Age. Also in both algorithms, DB, ALB, SGPT, TB and A/G factors have a significant impact in prediction of liver disease and it is recommended that these factors are considered more seriously by physicians. The results of our study show that the use of boosting technique in C5.0 algorithm will improve the accuracy of prediction as well as the production of rules on liver disease dataset.In addition, gender was a missing factor in the majority of previous studies and also according to Fig. 6, it can be seen that this factor has less importance than other factors. However the rules established by boosted C5.0 in Table 3 clear the

importance of this factor more than before. Even though the significance of this factor in the prediction of liver disease is low, but the rules generated in this study show that this factor plays a special role in the identification of this disease. We emphasize that the results of this paper is visible in medical field associated with this disease and on large datasets. According to the rules generated by boosted C5.0 algorithm, we observed that females are more likely to develop liver disease with 53.125 percent of all rules by boosted C5.0 compared to male with 46.875 percent of all rules. With regard to the rules generated by boosted C5.0 and CHAID, we propose the use of boosted C5.0 and CHAID algorithms as a complementary algorithms for early diagnosis of liver diseases. Also, we introduced five ranges for five important common factors as follows: DB = [10.900 – 1.200], ALB [4.00 – 4.300], SGPT = [34 – 37], TB = [0.600 – 1.200] (by boosted C5.0), A/G = [1.180 – 1.390].

According to our experimental results and as well as other existing studies on liver disease, the boosted C5.0 had acceptable performance. As mentioned earlier, proposed method as novel method generated quite simple rules. Through using several metrics, we also observed that the accuracy of boosted C5.0 in this study is one of the high accuracies in literature. Moreover, the importance of features and cut-off points in the five most important factors have been presented. Overall, we would argue that this study considers all aspects about liver disease whereas previous papers only focused on some aspects such as good accuracy for diagnosis of liver disease. In conclusion, the presented study provided the use of rules generated by boosted C5.0 and CHAID algorithms for extraction of useful knowledge from them. Because physicians have different opinions about the importance of features in various diseases like liver disease, so this study has concentrated on the application of computational intelligence, especially the production of rules by mining-based classifiers to identify important factors behind this disease. This study has been performed on one sample of liver disease data set. Thus, as future work, we propose further research on this disease through using the combination of neural network and C5.0 algorithms both with booting approach. Feature Selection and n-fold cross validation approaches are important to improve the accuracy of methods. In this regard, we will use both approaches with various methods on liver disease. Other future work will be about a deep neural network (DNN) and the standard backpropagation algorithm with weight decay (l_2-regularization) or sparsity (l_1-regularization). Furthermore, weighted-Naïve Bayesian (W-NB) algorithm will be applied with neural network and genetic algorithms.

References

Abdar, M. (2015). A Survey and Compare the Performance of IBM SPSS Modeler and Rapid Miner Software for Predicting Liver Disease by Using Various Data Mining Algorithms. *Cumhuriyet Science Journal*, *36*, 3230-3241.

Alemayehu D & Berger M. L. (2016). Big Data: transforming drug development and health policy decision making. *Health Services and Outcomes Research Methodology,* Springer.

Alfisahrin, S. D. N. N., & Mantoro, T. (2013). Data Mining Techniques for Optimization of Liver Disease Classification. In *Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference on* (pp. 379-384). IEEE.

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, *111*, 52-61.

Alizadeh, S., Ghazanfari, M., & Teimorpour, B. (2011). Data Mining and Knowledge Discovery. *Publication of Iran University of Science and Technology*.

Althuwaynee, O. F., Pradhan, B., & Lee, S. (2016). A novel integrated model for assessing landslide susceptibility mapping using

CHAID and AHP pair-wise comparison. *International Journal of Remote Sensing*, *37*, 1190-1209.

Anisha, P. R., Reddy, C., & Prasad, L. V. (2015). A pragmatic approach for detecting liver cancer using image processing and data mining techniques. In *Signal Processing and Communication Engineering Systems (SPACES), 2015 International Conference on* (pp. 352-357). IEEE.

Baker, S., & Cousins, R. D. (1984). Clarification of the use of chi-square and likelihood functions in fits to histograms. *Nuclear Instruments and Methods in Physics Research*, *221*, 437-442.

Bahramirad, S., Mustapha, A., & Eshraghi, M. (2013). Classification of liver disease diagnosis: A comparative study. In *Informatics and Applications (ICIA), 2013 Second International Conference on* (pp. 42-46). IEEE.

Bittel, S., Kaiser, V., Teichmann, M., & Thoma, M. (2015). Pixel-wise Segmentation of Street with Neural Networks. *arXiv preprint arXiv:1511.00513*.

Boubekeur, M. B., Luo, S., & Labidi, H. (2015). A background subtraction algorithm for indoor monitoring surveillance systems. In *Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2015 IEEE International Conference on* (pp. 1-5). IEEE.

Chambers, M., & Dinsmore, T. W. (2014). *Advanced analytics methodologies: Driving business value with analytics*. Pearson Education.

Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus*, *5*, 1.

Choudhury, S., & Bhowal, A. (2015). Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. In *Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015 International Conference on* (pp. 89-95). IEEE.

Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*, *37*, 1568-1572.

Daş, R., Turkoglu, I, Sengur, A., (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert Systems with Applications*, 36, 7675-7680.

Das, R., Turkoglu, I, Sengur, A., (2009). Diagnosis of valvular heart disease through neural networks ensembles. *Computer Methods and Programs in Biomedicine*, 93, 185-191.

Das, R., Sengur, A., (2010). Evaluation of ensemble methods for diagnosing of valvular heart disease. *Expert Systems with Applications*, 37, 5110-5115.

Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using K-nearest neighbor and genetic algorithm. *Procedia Technology*, *10*, 85-94.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*, 861-874.

Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. (2005). An electric energy consumer characterization framework based on data mining techniques. *Power Systems, IEEE Transactions on*, *20*, 596-602.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*, 119-139.

Fürnkranz, J. (2002). Round robin classification. *The Journal of Machine Learning Research*, *2*, 721-747.

Fürnkranz, J. (2002). Pairwise classification as an ensemble technique. In *Machine Learning: ECML 2002* (pp. 97-110). Springer Berlin Heidelberg.

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques*. Springer Science & Business Media.

Grzymala-Busse, J. W., & Hu, M. (2000). A comparison of several approaches to missing attribute values in data mining. In *Rough sets and current trends in computing* (pp. 378-385). Springer Berlin Heidelberg.

Gullo, F. (2015). From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. *Physics Procedia*, *62*, 18-22.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

Han, J., & Kamber, M. (2001). Data mining: concepts and techniques. 2001. Morgan Kauffman.

Hou, S., Hou, R., Shi, X., Wang, J., & Yuan, C. (2014). Research on C5. 0 Algorithm Improvement and the Test in Lightning Disaster Statistics. *International Journal of Control and Automation*, *7*, 181-190.

Hunt, C. M., Yuen, N. A., Stirnadel-Farrant, H. A., & Suzuki, A. (2014). Age-related differences in reporting of drug-associated liver injury: data-mining of WHO Safety Report Database. *Regulatory Toxicology and Pharmacology*, *70*, 519-526.

Jiawei Han, M. K., & Pei, J. (2011). Data Mining: Concepts and Techniques: Concepts and Techniques, Elsevier.

Jin, H., Kim, S., & Kim, J. (2014). Decision factors on effective liver patient data prediction. *Int. J. BioSci. BioTechnol*, *6*, 167-178.

Kasabov, N., & Capecci, E. (2015). Spiking neural network methodology for modelling, classification and understanding of EEG spatio-temporal data measuring cognitive processes. *Information Sciences*, *294*, 565-575.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 119-127.

Khader, N., Lashier, A., & Yoon, S. W. (2016). Pharmacy robotic dispensing and planogram analysis using association rule mining with prescription data. *Expert Systems with Applications*.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd.

LDCR: Liver Disease in Canada Report, http://www.Liver.ca/support-Liver-foundation/advocate/Liver_Disease_in_Canada_Report.aspx**,** Accessed 06.04.15.

LDI: Liver Disease Information, American Liver Foundation, Liver Disease Information Center, http://www.Liverfoundation.org/abouttheLiver/info/, Accessed 06.04.15.

Lin, R. H. (2009). An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine*, *47*, 53-62

Marinakis, Y., Marinaki, M., & Matsatsinis, N. (2008). A stochastic nature inspired metaheuristic for clustering analysis. *International Journal of Business Intelligence and Data Mining*, *3*, 30-44.

Marateb, H. R., Mansourian, M., Faghihimani, E., Amini, M., & Farina, D. (2014). A hybrid intelligent system for diagnosing microalbuminuria in type 2 diabetes patients without having to measure urinary albumin. *Computers in biology and medicine*, *45*, 34-42.

Matusevich, D. S., Cabrera, W., & Ordonez, C. (2015). Accelerating a Gibbs sampler for variable selection on genomics data with summarization and variable pre-selection combining an array DBMS and R. *Machine Learning*, 1-22.

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*, 93-99.

Montazeri, M., Montazeri, M., Beygzadeh, A., & javad Zahedi, M. (2014). Identifying efficient features in diagnose of liver disease by decision tree models. *HealthMED*, 8, 1115-1124.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, *40*, 1086-1093.

Nagaraj, K., & Sridhar, A. (2015). NeuroSVM: A Graphical User Interface for Identification of Liver Patients. *arXiv preprint arXiv:1502.05534*.

PANG, S. L., & GONG, J. Z. (2009). C5. 0 classification algorithm and application on individual credit evaluation of banks. *Systems Engineering-Theory & Practice*, *29*, 94-104.

Pashaei, E., Ozen, M., & Aydin, N. (2015). Improving medical diagnosis reliability using Boosted C5. 0 decision tree empowered by Particle Swarm Optimization. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (pp. 7230-7233). IEEE.

Patidar, S., Pachori, R. B., & Acharya, U. R. (2015). Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. *Knowledge-Based Systems*, *82*, 1-10.

Pazhanirajan, S., & Dhanalakshmi, P. (2016). MRI Classification of Parkinson's Disease Using SVM and Texture Features. In *Proceedings of the Second International Conference on Computer and Communication Technologies* (pp. 357-364). Springer India.

Quinlan, J. R. (1979). *Discovering rules by induction from large collections of examples* (pp. 168-201). Expert systems in the micro electronic age. Edinburgh University Press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81-106.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Quinlan, J. R. (1996). Bagging, boosting, and C4. 5. in *AAAI/IAAI*, *1*,725-730.

Rajeswari, P., & Reena, G. S. (2010). Analysis of liver disorder using data mining algorithm. *Global Journal of computer science and Technology*, 10, 48-52.

Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2012). A critical comparative study of liver patients from usa and India: An exploratory analysis. *International Journal of Computer Science Issues*, *9*, 506-516.

Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2011). A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, *3*, 101-114.

Riganello, F., Candelieri, A., Quintieri, M., Conforti, D., & Dolce, G. (2010). Heart rate variability: an index of brain processing in vegetative state? An artificial intelligence, data mining study. *Clinical Neurophysiology*, *121*, 2024-2034.

Ritschard, G. (2010). CHAID and earlier supervised tree methods.

Rodríguez-Jiménez, J. M., Cordero, P., Enciso, M., & Mora, A. (2016). Data mining algorithms to compute mixed concepts with negative attributes: an application to breast cancer data analysis. *Mathematical Methods in the Applied Sciences*.

Souillard-Mandar, W., Davis, R., Rudin, C., Au, R., Libon, D. J., Swenson, R., & Penney, D. L. (2015). Learning classification models of cognitive conditions from subtle behaviors in the digital Clock Drawing Test. *Machine Learning*, 1-49.

Tanha, J., van Someren, M., & Afsarmanesh, H. (2015). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*, 1-16.

Tiwari, A. K., Sharma, L. K., & Krishna, G. R. (2013). Comparative Study of Artificial Neural Network based Classification for Liver Patient. *Journal of Information Engineering and Applications*, *3*, 2225-0506.

Tomczak, J. M., & Zięba, M. (2015). Probabilistic combination of classification rules and its application to medical diagnosis. *Machine Learning*, *101*, 105-135.

UCI. ILPD (Indian Liver Patient Dataset) Data Set. (2012) https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29, Accessed 10.04.15.

Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. *Int. J. Sci. Eng. Technol. Res.(IJSETR)*, *4*, 816-820.

Weng, C. H., Huang, T. C. K., & Han, R. P. (2016). Disease prediction with different types of neural network classifiers. *Telematics and Informatics*, *33*, 277-292.

Witten, I. H., & Frank, E. (2005). *Data Mining:* Practical machine learning tools and techniques. Morgan Kaufmann.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*, 1-37.

Xu, S., Liu, Z., & Zhang, Y. (2015). Least squares support vector regression and interval type-2 fuzzy density weight for scene denoising. *Soft Computing*, 1-12.

Yadav, H., Bansal, P., & KumarSunkaria, R. (2015). Color dependent K-means clustering for color image segmentation of colored medical images. In *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on* (pp. 858-862). IEEE.