RESEARCH ARTICLE OPEN ACCESS

# A Survey on Diagnosis of Liver Disease Classification

Harsha Pakhale[1],Deepak Kumar Xaxa[2]
Department of CSE, MATS University,Raipur, Chhattisgarh, India

**Abstract:**

Data mining techniques play very important role in health care industry. Liver disease is one of the growing diseases these days due to the changed life style of people. Various authors have worked in the field of classification of data and they have used various classification techniques like Decision Tree, Support Vector Machine, Naïve Bayes, Artificial Neural Network (ANN) etc. These techniques can be very useful in timely and accurate classification and prediction of diseases and better care of patients. The main focus of this work is to analyze the use of data mining techniques by different authors for the prediction and classification of liver patient.

*Keywords*— **Classification, Data Mining, Decision Tree.**

## I. INTRODUCTION

Liver is the largest internal organ in the human body, playing a major role in metabolism and serving several vital functions. The liver is the largest glandular organ of the body. It weighs about 3lb (1.36 kg). The liver supports almost every organ in the body and is vital for our survival [6]. Liver disease is any disturbance of liver function that causes illness. The liver is responsible for many critical functions within the body and should it become diseased or injured, the loss of those functions can cause significant damage to the body. Liver disease is also referred to as hepatic disease. Usually nausea, vomiting, right upper quadrant abdominal pain, fatigue and weakness are classic symptoms of liver disease. An early diagnosis of liver disease can lead to better treatment and increased chances of survival.

## II. DATA MINING AND ITS TECHNIQUES

The term data mining [1] refers to the findings of relevant and useful information from databases. Data mining is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency network, analyzing changes and detecting anomalies. Data mining and knowledge discovery is the new interdisciplinary field, merging ideas from statistics, machine learning, database and parallel computing.

### *2.1 KDD Process*

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data or KDD [2] which is a process of extracting the hidden knowledge from data warehouse. The knowledge discovery process is alterative sequence of several steps. The very first step is data cleaning which includes removal of noise and inconsistent data. The second step is data integration in which multiple data sources may be combined. The next step is data selection where data relevant to the analysis task are retrieved from the database. In data transformation data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations. Next step is data mining where intelligent methods are applied to extract data patterns. Next step is pattern evaluation where the identification of truly interesting patterns representing knowledge based on interestingness are performed. The last step is knowledge presentation where visualization and knowledge representation techniques are used to present mined knowledge to users. The following figure1 shows the KDD process:

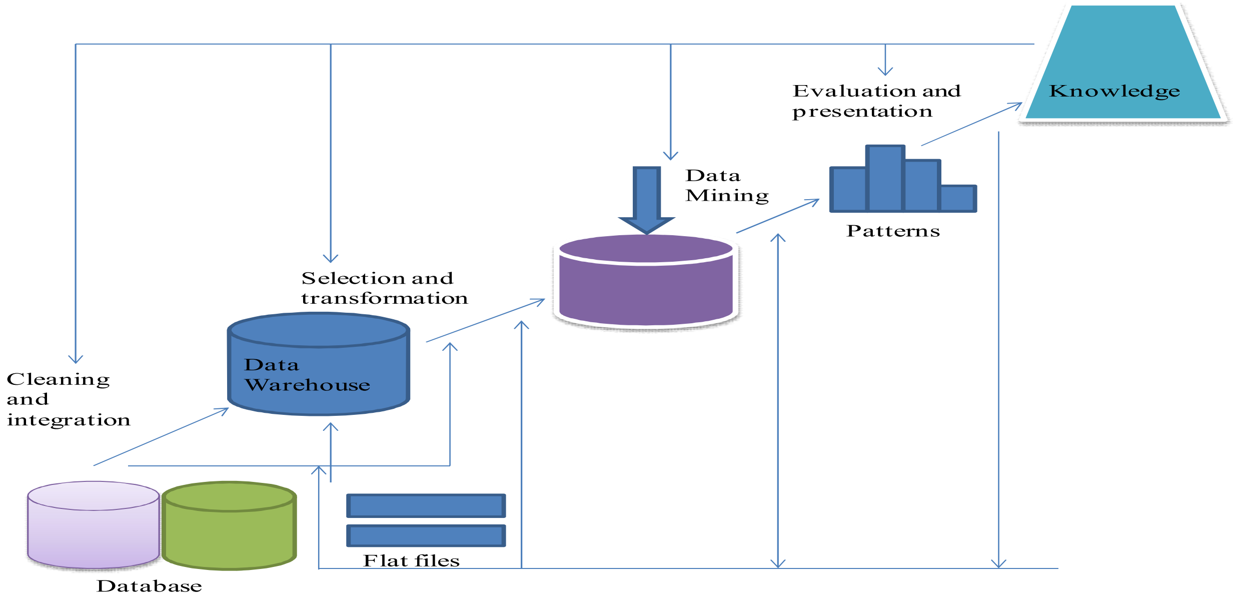

Figure 1: KDD process

There are various data mining applications like classification, clustering, association rule mining and prediction. In this paper,we have focus on classification techniques to classify the liver patient data. The figure 2 shows that general process of classification ofliver patientdata.In this process we have applied the ILPD data set on the classifier that is capable of classifying the data as liver and not liver data.

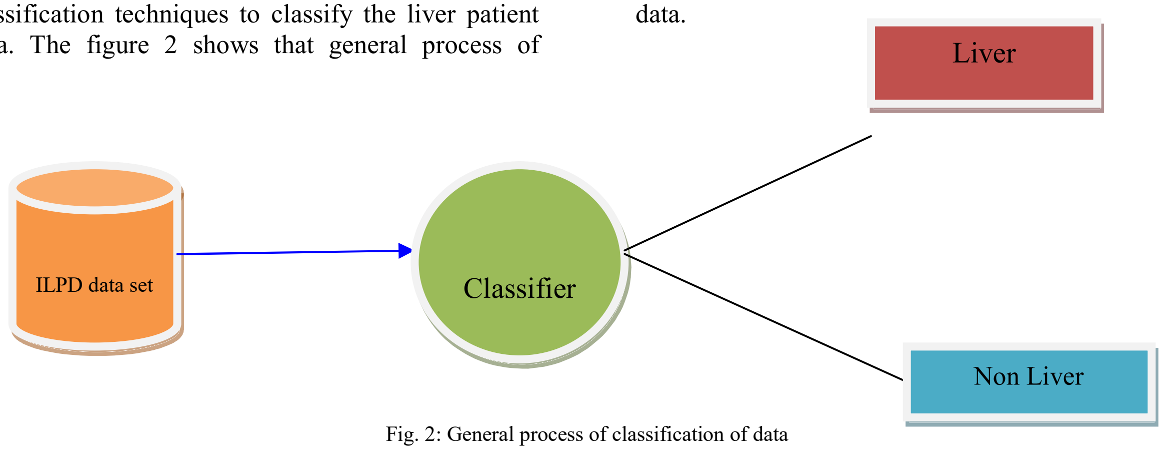

Fig. 2: General process of classification of data

*A. Classification*

Classification (Han, J. et al., 2006) [2] is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). Classification predicts categorical (discrete, unordered) labels.Classification models are tested by comparing the predicted values to known target values. Classification is composed of two steps – training and testing.

*A.1Decision Tree*

A decision tree[1]is classification scheme which generates a tree and a set of rules, representing the model of different classes, from a given data set. The set of records available for developing classification methods is generally divided into two disjoint subsets-a training set and a test set. The former is used for driving the classifier, while the latter is used to measure the accuracy of the classifier. The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified.

*A.1.1 C4.5*

C4.5 [1] is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rules derivation. C4.5 is classification algorithm that can classify records that have unknown attribute values by estimating the probability of various possible results unlike CART, which generates a binary tree.

*A.1.2 CART*

CART (Classification and Regression Tree)[1] builds a binary decision tree by splitting the record at each node, according to a function of single attribute. The initial split produces two nodes, each of which we now attempt to split in the same manner as the root node. Once again we examine all the input fields to find the candidate splitters. If no split can be found that significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only one leaf node remains and we have grown the full decision tree.

*A.1.3 CHAID*

CHAID [1], proposed by Kass in 1980, is a derivative of AID(Automatic Interaction Detection), proposed by Hartigan in 1975. CHAID attempts to stop growing the tree before over fitting occurs, whereas CART, ID3 and C4.5 generate a fully grown tree and then carry out pruning as post processing step. In this sense CHAID avoids the pruning phase.

*A.1.4 Random Forest*

Random Forest (or RF) [3] is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of classes output by individual trees. Random forests are often used when we have very large training datasets and a very large number of input variables.

***B.*** *Support Vector Machine*

In formal definition, a support vector machine design [18] a hyper lane or a set of hyper planes in a high or infinite dimensional space, which can be used for classification regression or other tasks. A SVM is a promising new method for classification of both linear and nonlinear data. SVMs are based on the concept of design planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships (V.N. Vapnik, 1998). Support Vector Machine algorithms divide the n dimensional space representation of the data into two regions using a hyper plane. This hyper plane always maximizes the margin between the two regions or classes. The margin is defined by the longest distance between the examples of the two regions or classes and is computed based on the distance between closest instances of both classes to the margin, which are called supporting vectors(V. Vapnik, 1998).

***C.*** *Bayesian Net*

Bayesian Net [2] is statistical classifiers which can predict class membership probabilities, such as the probabilities that a given tuple belong to a particular class. Let X is a data sample whose class label is unknown. Let H be some hypothesis such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P(H|X), the probability that the hypothesis H holds given the observed data sample X. P(H|X) is the posterior probability, or a posteriori probability, of H conditioned on X.

***D.*** *Multilayer Perceptron (MLP)*

MLP [1] is a development from the simple perceptron in which extra hidden layers (layers additional to the input and output layers, not connected externally) are added. More than one hidden layer can be used. The network topology is constrained to be feed forward, i.e., loop-free. Generally, connections are allowed from the input layer to the first (and possible only) hidden layer, from the first hidden layers to the second and so on, until the last hidden layer to the output layer. The presence of these layers allows an ANN to approximate a variety of non-linear functions. The actual construction of network, as well as the determination of the number of hidden layers and determination of the overall number of units, is sometimes of a trial-and-error process, determined by the nature of the problem at hand. The transfer function generally a sigmoid function. Multilayer perceptron is a neural network that trains using back propagation.

## III. LITERATURE REVIEW

There are various authors have worked in the field of classification of liver patient data. **H. Jin et.al, (2014)** proposed “Decision Factors on Effective Liver Patient Data Prediction” and evaluated on Indian Liver Patient data (ILPD) dataset [4]. The WEKA data mining tool which is most widely known as a useful tool for data mining is used for validating the model. This research work analyses the classification algorithms such as Naïve Bayes, Decision Tree, Multilayer Perceptron, k-NN, Random Forest and Logistic. It compares the performance of different classification algorithms. These algorithms were compared in several kinds of evaluation criteria like precision, recall, sensitivity, specificity. For comparison of selected classification algorithms, cross validation method was used.Experimental results show that Logistic and Random Forest gives highest and second highest precision and recall value respectively as compared to other previous four algorithms which are Naïve Bayes, Decision Tree, Multilayer Perceptron and K-Nearest Neighbor. In case of sensitivity and specificity Naïve Bayes shows the highest value and Random Forest and Logistic show relatively higher than others.**P. Saxena et.al, (2013)** proposed “Analysis of Various Clustering Algorithms of Data Mining on Health Informatics” [5]. They worked on clustering technique of data mining. The evaluation took place on ILPD dataset. They used WEKA tools to predict the result. In this research work they used different clustering algorithms such as COBWEB clustering algorithm, DBSCAN clustering algorithm, Hierarchical clustering algorithm and K-means clustering algorithm on ILPD dataset. Experimental results show that k-means clustering algorithm is simplest and fastest algorithm as compared to other clustering algorithms which are COBWEB, DBSCAN and Hierarchical clustering algorithm.**A. Guliaet.al, (2014)** proposed “Liver Patient Classification using Intelligent Techniques” [6]. They used J-48 classifier, Multilayer Perceptron classifier, Random Forest classifier, Support Vector Machine classifier and Bayesian Network classifier for classification of liver patient data. ILPD dataset available on UCI Repository is used for evaluation. Performance analysis is done using WEKA tool. Feature selection strategy also known as Variable selection or attribute selection is used to reduce the number of features with maintaining good result. This research work compares the result of classification algorithms with and without the application of feature selection.The results obtained show that Support Vector Machine algorithm gives better performance with an accuracy of 71.3551% as compared to other algorithms when evaluated without feature selection and Random Forest algorithm gives better performance with an accuracy of 71.8696% as compared to other algorithms when evaluated after feature selection.They stated that the future work will include the construction of hybrid model for classification of health care data. They also proposed the use of new algorithms to gain improved performance than the techniques used.**J. Pahareeya et.al, (2014)** proposed “Liver Patient Classification using Intelligence Techniques” [8]. They used J-48, Multilayer Perceptron, Random Forest, Multilayer Regression, Support Vector Machine and Genetic Programming (GP) for classification of liver patient data. The ILPD dataset available on UCI Repository is used for evaluation. They employed under sampling and over sampling approach. They also used 10-fold cross validation in this research work.Experimental results show that Random Forest over sampling model proved to be better technique than all the other techniques. It also shows that Genetic Programming for the original data and Genetic Programming for the under sampling stood second with an accuracy of 84.75%.**S. Bahramiradet.al, (2013)** proposed “Classification of Liver Disease Diagnosis: A Comparative Study” [12]. They used two liver patient dataset. Both datasets are taken from UCI Repository. The first dataset is AP dataset and the Second dataset is BUPA dataset. They worked with different classification algorithms such as Logistic, Bayesian Logistic Regression, Logistic Model Trees (LMT), Multilayer Perceptron, K-star, RIPPER, Neural Net, Rule Induction, Support Vector Machine (SVM) and CART. Accuracy, Precision and Recall parameters are used to evaluate the performance of the proposed method. They also used Brute Force Optimization and Bayesian Boosting for improving the result.The result obtained shows that AP proved to be slightly better than the BUPA dataset in terms of accuracy while BUPA dataset proved to be better than AP dataset in terms of precision and recall. **E. M. Hashemet.al, (2013)** proposed “A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis” [10]. They used Support Vector Machine (SVM) classification technique for classification of liver patient data to achieve improved performance. Two dataset are used for performance evaluation. The first dataset is BUPA dataset and the second dataset is ILPD dataset. Both dataset are obtained from UCI Repository. Error Rate, Sensitivity, Prevalence, Specificity and Accuracy are used to evaluate the performance of Support Vector Machine (SVM). Feature Ranking is also used to reduce the number of features. MATLAB is used to write and implement the Support Vector Machine (SVM) algorithm. This research work also used cross validation.The result obtained shows that the Specificity at first 6 ordered features are best for BUPA dataset compared to ILPD dataset while the Sensitivity, Error Rate, Accuracy and Prevalence at first 6 ordered features are best for ILPD dataset as compared to BUPA dataset.**C. Liang et.al, (2013)** proposed “An Automated Diagnosis System of Liver Disease using Artificial Immune and Genetic Algorithms” [11]. They used a combination of two methods to diagnose the liver disease which are artificial immune and genetic algorithm. The experiments and assessments of the proposed method were performed with ILPD dataset and Liver Disorder dataset both taken from UCI Repository. F-measure, Accuracy, Sensitivity, Specificity and Precision are parameters that are used for performance evaluation. This research work also performed 20-fold cross validation on datasets.According to experimental results, the predicting accuracy obtained by this system is highest as compared to other classification methods such as Support Vector Machine (SVM), Naïve Bayes classifier, k-nearest neighbor, C4.5 and Backpropogation.**Suryakantet.al, (2015)** proposed “An Improved K-means Clustering with Atkinson Index to Classify Liver Patient Dataset” [9]. They worked on clustering technique of data mining to classify liver patient dataset. The dataset being used is ILPD dataset taken from UCI Repository. Two evaluation matrices are used. The first matrix is Gold Standard and the second matrix is F-score. They used k-means clustering algorithm with Atkinson Index which is the technique for measuring the inequality.The result obtained shows that K-means clustering algorithm with Atkinson index gives better result as compared to K-means clustering algorithm for both evaluation matrices.They stated that the future work can be carried out by applying k-nearest neighbor (K-NN) to the data received by Atkinson Index which will improve the result further. Further work can be carried out by using Information Gain technique instead of Atkinson Index for selecting initial centroids.**Dr. S. Vijayaraniet. al, (2015)**have proposed “Liver Disease Prediction Using SVM and Naïve Bayes Algorithm” [7]. They have used Support Vector Machine (SVM) and Naïve Bayes algorithms for classification of liver patients. Indian Liver Patient Dataset(ILPD) available on UCI repositoryhas been used for performance evaluation. This work is implemented using Matlab 2013 tool. These classification algorithms are evaluated on the basis of accuracy and execution time.Experimental result shows that Support Vector Machine (SVM) is better than Naïve Bayes algorithm in classifying liver patient dataset.**B. V. Ramanaet. al, (2012)** have proposed “Liver Classification Using Modified Rotation Forest” [17]. They have used 10 different classification algorithms from 5 different categories. They are J48 and simple cart classification algorithm from tree based algorithm, Naïve Bayes and Bayes Net classification algorithms from statistical classification algorithm, MLP and SMO classification algorithms from multilayer perceptron based algorithm, IBK and K Star classification algorithm from lazy learner and PART and zero classification algorithm from rule based algorithm. Two different datasets are used for evaluation. One is Indian Liver Dataset (ILPD) and second one is BUPA liver disorder dataset.They have used feature selection strategy on dataset to reduce the number of features. The four feature selection techniques used are PCA, CFS, Random Projections and Random subset.Results obtained show that multilayer perceptron classification algorithm with Random subset gives highest accuracy of 74.7826% for BUPA liver disorder dataset and nearest neighbor with CFS gives highest accuracy of 73.0703% for ILPD dataset.**Reetuet. al, (2015)** have proposed “Medical Diagnosis for Liver Cancer Using Classification Techniques” [13]. They have used decision tree induction J48 algorithm to classify liver patient dataset. The dataset is taken from Pt. B. D. Sharma Postgraduate Institute of Medical Science, Rohtak. WEKA data mining tool has been used for the processing of data.The result obtained shows that the performance of J48 algorithm is better than other classification algorithms.**S. Dhamodharan (2014)** has proposed “Liver Prediction Using Bayesian Classification” [14]. He has used Naïve Bayes and FT tree classification techniques to classify the liver patient dataset into four classes i.e. liver cancer, cirrhosis, hepatitis and no disease. He has used WEKA data mining tools for the processing of data.The experimental results show that Naïve Bayes gives better performance than FT tree with an accuracy of 75.54% .**H. R. Kirubaet. al, (2014)** have proposed “An Intelligent Agent Based Framework for Liver Disorder Diagnosis Using Artificial Intelligence Techniques” [15]. They have used C4.5 decision tree algorithms and Random Tree algorithm for classifying the liver patient dataset. Two different liver dataset are used. One is Indian Liver Dataset (ILPD) and another one is BUPA Liver Disorder Dataset.Experimental results show that 100% accurate classification is recorded for C4.5 and RT algorithms for both datasets.**A.S. Aneeshkumaret.al, (2012)** have proposed "Estimating theSurvallience of Liver Disorder Using Classification Algorithms" [16]. They have worked on Naïve Bayesian and C4.5 decision tree algorithms for classification of liver patient. For this research work, real medical data with 15 attributes were collected from a public charitable hospital in Chennai. This work gives us a binary classification i.e. either a person is a liver patient or not. They have also preprocessed the data before evaluation in order to remove any inconsistencies.The result obtained shows that C4.5 decision tree gives better performance compared to Naïve Bayesian in classification of liver patient dataset.

## IV. CONCLUSION

In medical science, diagnosis of health condition is very challenging task. In this paper, we have surveyed many data mining techniques that have been used for classification of liver patients. Because of the rapidly increasing medical data, it has become extremely important to manage these data properly and use these data for accurate prediction of diseases and for providing better care to the patients. This paper has provided the summary of the use of data mining techniques for the classification of liver patient.

## References

[1] A.K. Pujari, Data mining techniques, 4th edition, University Press (India) Private Limited, 2001.

[2] J. Han. & K. Micheline, Data mining: Concepts and Techniques, Morgan Kaufmann Publisher,2006.

[3] R., Parimala et al.,A Study of Spam E-mail classification using Feature SelectionPakage".Global General of Computer Science and Technology, Vol. 11, ISSN 0975-4172,2011.

[4] H. Jin, S. Kim and J. Kim ,Decision Factors on Effective Liver Patient Data Prediction.International Journal of Bio-Science and Bio-Technology. VOL.6, No.4, pp. 167-178,2014.

[5] P. Saxena and S. Lahre,Analysis of Various Clustering Algorithms of Data Mining on Health Informatics.International Journal of Computer & Communication Technology, Vol.4, Issue-2, pp.108-112, 2013.

[6] A. Gulia, Dr. R. Vohra, P. Rani, Liver Patient Classification using Intelligent Techniques, International Journal of Computer Science and Information Technologies, Vol.5 (4), pp.5110-5115,2014.

[7] Dr. S. Vijayarani, Mr. S. Dhayanand , Liver Disease Prediction using SVM and Naïve Bayes Algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) Vol. 4, Issue 4, pp.816-820, 2015.

[8] J. Pahareeya, R. Vohra, J. Makhijani and S. Patsariya, Liver Patient Classification using Intelligence Techniques, International Journal of Advanced Research in Computer science and Software Engineering, Vol. 4, Issue 2, pp.295-299,2014.

[9] S. and I.A. Ansari , An Improved K-means Clustering with Atkinson Index to classify Liver Patient dataset, springer, 2015.

[10] Er. M. Hashem and M. S. Mabrouk, A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis, AmericanJournal of Intelligent Systems, pp.9-14,2014.

[11] C. Liang and L. Peng, An Automated Diagnosis System of Liver Disease using ArtificialImmuneand Genetic Algoriithms, Springer, 2013.

[12] S. Bahramirad, et al. , Classification of Liver Disease Diagnosis: A Comparative Study, IEEE,pp.42-46,2013.

[13] Reetu and Narendra Kumar , Medical Diagnosis for Liver Cancer using Classification Techniques, International journal of Recent Scientific Research, Vol 6, Issue 6,pp.4809-4813,2015.

[14] S. Dhamodharan, Liver Prediction using Bayesian Classification, An International Journal of Advanced ComputerTechnology,2014.

[15] H. R. Kiruba and Dr. G. T. Arasu, An Intelligent Agent Based Framework for Liver Disorder Diagnosisusing Artificial Intelligent Techniques, Journal of Theoretical andApplied Information Technology, pp.91-99, 2014.

[16] A.S. A. kumar and C. J. Venkateswaran, Estimating the Survallience of Liver Disorder using Classification Algorithms.International Journal of Computer Applcation, Vol 57, pp. 39-42, 2012.

[17] B. V. Ramana and Prof. M. S. Prasad Babu, Liver Classification using Modified Rotation Forest,International Journal of Engineering Research and Development, Vol 1,Issue6,pp. 17-24,2012.

[18] V. N. Vapnik, Statistical Learning Theory, New York: John Wiley and Sons, 1998.