A Comparative Study of Various Data Mining Algorithms for effective Liver Disease Diagnosis –
A decade review from 2010 to 2019.

IfraAltaf
Research Scholar, Department of Computer Science,
University of Kashmir, Srinagar, J&K, India.

Muheet Ahmed Butt
Scientist, Department of Computer Science,
University of Kashmir, Srinagar, J&K, India.

Majid Zaman Baba
Scientist, Directorate of IT&SS
University of Kashmir. Srinagar, J&K, India

S Jahangeer Sidiq
School of Computer Science and Engineering
Lovely Professional University. Phagwara, Punjab, India

Abstract. Clinical databases contain information about patients and their health conditions and this information can be used to mine the hidden relationships and patterns present inside the data that helps in providing up-to-the-minute medicinal awareness with assistance in diagnoses and treatment of diseases. Classification and prediction models support medical diagnosis which helps to arrive at the accurate decision-making and actual prediction of different diseases. Many researchers have used various data mining techniques to accurately detect and predict the liver disease and they have diverse views and estimations of confirming their proposed algorithms as better algorithms in predicting and classifying the disease. To find out how data mining and machine learning techniques have developed during the past decade, this paper assesses certain classification algorithms for predicting various liver diseases through a survey of literature from 2010 to 2019.

Keywords: liver diseases, hepatitis, classification techniques, medical diagnosis, literature review

1. Introduction

The liver is an important internal organ that has multiple functions in the body. It is the largest gland that sits just under the rib cage on the right side of the abdomen and weighs about 1.36 kg. The liver plays an essential role in digesting food, detoxifying substances in the body, storing vitamins etc. Various infections can occur in the liver due to damage by medications, alcohol consumption, and obesity. Liver disease can be inherited. The perceptible signs and indications are not noticed in liver disease in the early stages. The most common liver disorders are fatty liver, hepatitis, cirrhosis and liver cancer [1, 2]. Microorganisms such as virus, bacteria damage the liver, thereby preventing it from functioning as it should be [3, 4]. Liver diseases are detected on the basis of the liver functional test [5] and by analyzing the levels of enzymes in the blood [6]. The usual symptoms of liver disease generally include nausea, vomiting, right upper quadrant abdominal pain, fatigue, eating disorder and weakness [7]. The other symptoms also include jaundice, abdominal swelling, and fluid in abnormal cavity, pale stool, weight loss, bloated spleen, itching and gallbladder [8].
Healthcare is an effective part of a country's economy. By taking certain crucial steps into consideration, there is a provision for improving health overall [9]. In health sphere, data mining is a process of finding and coming across with the concealed pattern of information from large medical datasets. It is one of the most significant phases for automatically predicting a disease by analyzing the medical data and converting a large collection of data into knowledge with the help of numerous distinct algorithms and elucidated statistical procedures which can describe new health related facts. Mining useful information can help to make a correct decision at right time [10]. The patterns found can be analyzed and used for disease prediction. The derived information can support strong clinical and administrative decision making from large clinical databases. In medical diagnosis, classification techniques play a very significant role in predicting the diseases automatically [11].

One of the major reasons behind human death is liver disorder diseases [12]. It is essential to detect and predict the liver disease earlier with better accuracy that can enhance the correct medication and treatment to the patients well in good time. Even though the liver tissue damage partially, it becomes very difficult to detect it in the early stages [13]. Often many medics do not succeed to become aware of the disease and this failure can mislead to inappropriate medical management [14]. The patient’s survival rate can increase with an early diagnosis of liver problems [6]. Data mining and machine learning approaches can be very effective in predicting the liver disease. Besides clinic and diagnostic tests various data mining algorithms pattern have been widely used for early diagnosis of liver diseases in medicine by experts. The potential mistakes by the medical professional can be reduced while diagnosing a disease. The medical data can be examined in short period and can be more detailed [15]. A general practitioner can understand the difficult diagnostic tests, combine the information from various sources and provide patient-specific prognosis [16].

In this paper we will focus on the prediction of liver disease i.e. we review on the applications of data mining in the prediction of liver disease. By reviewing, our study will help to provide the critical analysis of the latest data mining practices and algorithms used in predicting the liver disease. Also, this will contribute the researchers in analyzing the latest work done on liver disease prediction using various data mining algorithms. Furthermore, this study will provide the future directions and comparisons of the various machine learning algorithms implemented in liver disease prediction. In this paper, a brief review is provided to summarize the recent studies on the health care mostly –liver disease prediction, using various data mining techniques including supervised, unsupervised and semi-supervised learning algorithms, their pros & cons. Most importantly in this paper, we will also discuss the various challenges which are yet to be faced. To improve the precision in predicting the liver disease, researchers have been working by using various approaches which include Artificial Neural Networks, Data Mining and Machine Learning Algorithms, Fuzzy Inference systems, Support Vector Machines and many other techniques.

2. Related Work

Considered as one of the most fatal diseases in the world, liver disorders are having an exceptional increase in the recent years. Liver disease patients need dynamic medical dealing to trim down the mortality rate. A vast magnitude of research work has been performed by various researchers in this field. Since a lot of literature is available for prediction of liver disease, we have chosen some of the selected studies which are mentioned below:

P.Rajeswari, et al. (2010) has used the data collected from UCI repository consisting of 345 instances with 7 different attributes to create the training dataset. In this paper classification technique algorithms such as FT tree Naïve Bayes and KStar have been used to predict the liver disease disorder using the 10-fold cross validation tested on liver disease datasets. Weka Tool has been used to implement the algorithms. 10-fold cross validation has been used to evaluate the data. The outcome in terms of precision, accuracy and time taken to run the data when compared reveal that FT tree plays a fine role to increase the accuracy of the dataset in classification algorithms for early diagnosis the disease. The accuracy comes out to be 97.10% for FT Tree algorithm [17].
A.H. Roslina, et al. (2010) proposed to use the Support Vector Machines and Wrapper method to predict the hepatitis disease. In order to get the noise free features, the classification was preceded by the implementation of wrapper methods so as to minimize the noise or irrelevant data. The accuracy according to the authors was showed by the Support Vector Machine in enforcing feature selection first. Wrapper methods were implemented in Weka and the classification was done using LibSVM. The dataset was acquired from UCI machine learning repository and consisted of 155 instances distinguished into two classes. The original dataset was preprocessed and attributes were analyzed. The data is then split into training and test datasets. The dataset is subjected to classification algorithms and the output is evaluated. The authors further presented that by combining Wrappers Method and SVM techniques, there was a reduction in the cost and time of clinical lab test, reduction of attributes from 20 to 10 and along with that the accuracy also increased up to 74.55% [18].

Huda Yasin, et al. (2011) analyzed the risk factors of hepatitis-C virus. The dataset acquired from the University of California machine learning store consists of 20 features including the target class and 155 records. Data normalization techniques are employed to deal with the missing value problems. 15 binary attributes, 5 continuous attributes and a class attribute with binary categorization are included in the dataset. At the outset, the classification algorithms were applied to the full dataset and the accuracy was computed as 90.6%. Since there is a correlation between the attributes of the dataset, as a consequence, the authors found it impractical to use all the 19 attributes. Therefore, the dimension of the dataset is reduced using the PCM (principal component analysis) and the resulting subset dataset consists of only 37% of the total fields in the original dataset. Next binary logistic regression is applied to classify the reduced dataset and the calculable classification accuracy comes out to be more than 89% accurate. Categorically the classification accuracy is gone low but the authors argue about the fact that there is a low feature complexity with a good classification rate in their proposed approach. The authors concluded saying that the research can be extended by using different techniques like outlier analysis, association rule mining and fuzzy approach for estimating Hepatitis C virus [19].

Tahseen A. Jilani, et al. (2011) used the Neural Network and proposed an automatic diagnosis system PCA-ANN (Principal Component Analysis – Artificial Neural Network) for predicting hepatitis disease. They combined the feature extraction with classification in order to form a hybrid system to analyze the hepatitis dataset obtained from UCI machine learning data repository. The model works in two stages. Firstly, based on the importance there is a reduction of features in the original dataset and after reduction from 19 attributes to 6 imperative attributes, the reduced dataset is subjected to the classification. The missing values were imputed with the local mean method and then the reduced dataset was subjected to the Artificial Neural Network Classifier and the accuracy was calculated around 99.1% for the training data. The accuracy was obtained as 100% for the testing data [20].

Javed Salimi Sartakhti, et al. (2011) examined Simulated Annealing and Support Vector Machine techniques for diagnosis of hepatitis disease. A hybridized combination of support vector machine (SVM) and simulated annealing (SA) was used and thereby a novel machine learning method was developed to diagnose the Hepatitis disease. In this study the was collected from the UCI machine learning database and consisted of 13 nominal and 6 numeric features. The model was fundamentally parallel to a stochastic method for difficult optimization problem and the classification accuracy was made out by 10-fold cross validation and the classification accuracy was calculated as 96.25% [21].

Duygu Calisir, et al. (2011) used the concept of Least Square Support Vector Machine Classifier Principle Component Analysis (PCA–LSSVM) and thus presented an intelligent hepatitis diagnosis system. The authors used this system for feature extraction and reduction in order to perform an efficient classification. Firstly using PCA the hepatitis dataset acquired from UCI machine learning repository which had a total of 19 features was reduced to 10 features. Secondly the reduced features were fed as inputs to the LSSVM classifier. The classification accuracy obtained was 96.12% in their system. Two parameters were important in LSSVM classifier and those were the Gaussian kernels and the regularization factor. Total of 10 combinations of Gaussian kernels and the regularization factor were utilized to evaluate the performance of SVM. The range of
Gaussian kernels was adjusted between 0.1 and 25 and that of the regularization factor was adjusted between 1 and 100000 and the best accuracy was obtained from Gaussian kernels at 0.8 and regularization factor at 100 [4].

A.S. Aneeshkumar, et al. (2012) formulated a method to effectively classify liver and non-liver disease dataset. In order to achieve the effective classification the authors employed pre-processing method. The dataset had 2453 real medical records and was collected from a Public Charitable Hospital in Chennai. The dataset consisted of the 15 attributes of real medical data that were actually the symptoms recorded for liver and non-liver patients. The features included frequent alcoholic consumption (FAC), obesity, age, gender, fever, vomiting, abdomen pain (AP), yellowish urinary discharge (YUD), loss of appetite (LA), disturbance in abdomen (DA), pale stools, chill, rigor, head ache, acting differently (AD) and a target class for predicting the liver disease. The dataset was submitted to classification techniques C4.5 and Naive Bayes - the two algorithms used in the study. To evaluate the accuracy, the dataset was distributed into three different types of ratio on the basis of average and standard deviation of both class. The results revealed that C4.5 gives more accuracy of 99.20% and takes minimum time 0.16 seconds for its execution. So, C4.5 outperforms Naive Bayes classification algorithm [22].

BendiVenkata, et al. (2012) used a Modified Rotation Forest algorithm for classifying the Liver Disease patients. The authors calculated the accuracy of the classification techniques by utilizing the combination of feature selection technique and classification algorithm and results were in the form of improved accuracy. The authors in this paper focused about the challenges of limited medical recruits that have led to the involvement of intelligent medical diagnostic methods. The classification algorithms used in the study were Tree based (J48 and simple CART classification algorithms), Statistical based (Naive Bayes and Bayes Net classification), Neural Networks based (MLP and SMO classification algorithms), Rule based (PART and Zero classification algorithms) and Lazy learners (IBK and KStar). For evaluating the best combination of the algorithms, Principal component analysis (PCA), Correlation based feature selection (CFS), Random projection and Random subset feature selection methods are employed. There were two datasets used in the study. One was the BUPA Liver Disorders dataset acquired from UCI machine learning repository and another one was from Andhra Pradesh state of India. Using the 10-fold cross-validation, the experimental results reveal that 74.7826% accuracy is given by the Modified Rotation Forest (multi layer perception classification algorithm with random subset) for the UCI liver data set and Modified Rotation Forest (nearest neighbor with CFS) gives highest accuracy that is 73.0703% for the Andhra Pradesh liver data set [23].

Varun Kumar, et al. (2012) put forward a “Hepatitis Prediction Model” by utilizing the optimal feature selection in order to raise the analytical precision of the data mining algorithms. The dataset for the study was collected from UC Irvine machine learning repository containing 155 records and having 19 parameters with one target class. Firstly the authors identified the attributes and subjected to data cleaning. The missing values are filled with zeros in order to avoid reducing of the dataset. The most significant features are selected and assessed using Chi-Square attribute evaluation. The proposed HPM method undergoes data preprocessing method, imputation of missing values and choosing significant attributes by feature selection method improves the accuracy in predicting the disease. The proposed methodology gave the accuracy result as 79.33% without feature selection. The proposed HPM method reduces the attributes to 8 and results in the improved accuracy of 83.12% with feature selection [24].

FadlMutaher, et al. (2013) used the Rough Set Technique (RST) and presented the comparative analysis for diagnosing hepatitis disease. According to their experiments the performance of Naive Bayes algorithm was fast and enhanced the classification accuracy of hepatitis dataset. The result for accuracy of Naive Bayes algorithm was obtained as 96.52% and while analyzing the data the prediction of the result was more specific and precise using rough set technique. The dataset was acquired from UCI machine learning repository and consisted of 155 instances distinguished into two classes with 32 die instances and 123 live instances. There were a total of 20 attributes, together with the target class attribute and 20 missing values. The authors compared and analyzed the seven different classification algorithms namely, Naive Bayes, Naive Bayes updatable, FT Tree, KStar, J48, LMT, and Neural network for analyzing Hepatitis prognostic data. Based on
the results Naïve Bayes algorithm took less time to run and showed the highest figures in accuracy for hepatitis dataset while fully classifying the attributes. The authors were of the opinion that in order to boost the health index for making useful diagnosis, use of ensemble methods in medical datasets should be encouraged to add to the performance accuracy. The authors’ also suggested using fuzzy learning models for the improved Hepatitis diagnosis [25].

**Mahdieh Adeli, et al. (2013)** proposed a novel hybrid method for diagnosing the hepatitis disease. The authors tried selecting the significant features from the original dataset and then the reduced subset of features is subjected to the classification algorithm. Two intelligent methods have been proposed by the authors for feature selection and classification. Genetic Algorithm (GA) has been employed for the feature selection stage while Adaptive Network Fuzzy Inference System (ANFIS) has been used for classification purpose. That is how they have been able to create to a new fusion algorithm GA-ANFIS. The hepatitis dataset has been acquired from the UCI machine learning repository consisting of 155 instances and 19 attributes. The proposed algorithm is implemented in Matlab and predicts whether the patient lives or dies. Since both the algorithms of the hybrid algorithm are based on randomness, therefore the authors made the program run ten times. When evaluated, the proposed model acquired the best accuracy of 97.44% and took 2228.45 seconds. The sensitivity was recorded as 92.31% and the specificity was recorded as 100% for the proposed algorithm [26].

**Yilmaz Kaya, et al. (2013)** proposed a new hybrid decision support system known as RS-ELM based on rough set (RS) and extreme learning machine (ELM) for predicting the hepatitis disease. The hybrid system worked in two phases. Firstly, using the rough set approach, the redundant features were removed so as to avoid the data loss by removing missing value records. Then the classification was applied to the remaining features that is fatigue, malaise, prottime, and histology features of the dataset using extreme learning machine. The dataset has been acquired from the UCI machine learning repository. Using RS-ELM model the authors were able to classify with an accuracy of about 96.49%. The experimental results show the accuracy of 93.18%, 96.30% and 100.00% for training-test rates of 50-50%, 70-30% and 80-20% respectively through RS-ELM model. Thus, the average accuracy reached through the proposed model was 96.49% [27].

**Sa’diyah Noor Novita, et al. (2013)** worked upon the 10 important attributes of liver disease i.e. age, gender, total bilirubin (TB), direct bilirubin (DB), alkaline phosphotase (Alkphos), alamine aminotransferase (Sgpt), aspartate aminotransferase (Sgot), total proteins (TP), albumin (ALB), ratio albumin and globulin (A/G Ratio) and proposed to identify the liver disease patients. The dataset has been collected from (UCI) repository, California, which contained records of 583 patients and 11 attributes. A total of 416 patients suffered from liver disease and 167 patients didn’t. The authors used Decision Tree, Naïve Bayes, and NB Tree algorithms to classify the patients and the results calculated NB Tree algorithm (combination of Decision Tree and Naïve Bayes) has the highest decent accuracy score of 67.01%. The Naïve Bayes algorithm had the fastest execution time however. Authors put forwarded the future scope of finding the most considerable factor in categorizing liver disease patients in order to improve the performance of NB Tree algorithm [14].

**Anju Gulia, et al. (2014)** thought up a hybrid model construction for predicting liver disease in patients. This was done in three phases. The dataset is taken from UCI repository. Various classification algorithms were used on the original datasets of liver patients and then the dataset is subjected to the feature selection algorithm. The new reduced dataset comprising only important attributes is now rendered to the classification algorithms. The comparative analysis on the lookout for improved prediction accuracy discloses SVM algorithm as the better performance algorithm. Before applying the feature selection property on original dataset, the greater accuracy is given by SVM calculated as 71.3551%. After applying the feature selection property on original dataset to get the reduced dataset, the greater accuracy in that very case is given by SVM Random Forest algorithm. Next in the third and the last stage the authors compared all the results of classification algorithms obtained by means of and devoid of feature selection. As a result Random Forest algorithm surpassed other classification algorithms with the aid of feature selection. The accuracy came out to be 71.8696% as calculated [28].
Manjeevan Seera, et al. (2014) proposed to use the Fuzzy Min–Max neural network (FMM), Classification and Regression Tree (CART) and Random Forest model (RF), thus a hybrid intelligent system FMM-CART-RF was designed. The model used the benefits of the component models and could incrementally discover from the data sample. The model achieved high classification performances as the discovered results could predict outputs. The authors used various standard datasets which includes Breast Cancer Wisconsin, Pima Indians Diabetes and Liver Disorders from the UCI Warehouse of Machine Learning to assess the efficiency of their hybrid system. BUPA Medical Research Company obtained the features from blood tests and formed the liver disorder dataset. The results were analyzed and compared with the methods already published in the literature. The results enlightened that the proposed hybrid intelligent system was efficient for medicinal information classification and constructed excellent results and it also made known its knowledge base with a decision tree. The accuracy given by the model was calculated to be 95.01% with 9:1 train-to-test-ratio under 30 runs of the program. The prediction specified by the hybrid intelligent system could be used for medicinal decision [29].

Omar et al. (2014) studied that the Hepatitis C Virus affects millions of people every year and could be mortal and end up by taking lives. The authors used Modified Particle Swarm Optimization algorithm and Least Squares Support Vector Machine and suggested a hybrid classification system named Least Squares-Support Vector Machine (LS-SVM) classifier for HCV diagnosis. The feature vector was extracted using the Principle Component Analysis algorithm. Keeping the number of iterations low, Modified-PSO (Particle Swarm Optimization) Algorithm was applied to look for the optimal values of LS-SVM parameters. This model was implemented on the HCV dataset acquired from UCI repository of machine learning repository which contained 154 records and had 22 attributes involving 12 binary and 10 discrete attributes. The results thus calculated were compared and evaluated with the other classification system suitable for PCA and LS-SVM. The classification accuracy achieved by the proposed model was 98.86% while using 10-fold Cross Validation method [30].

S.Dhamodharan (2014) presented a paper and said that Naïve Bayes is at times enhanced than FT growth algorithm for predicting the class types from various classes which include liver cancer, cirrhosis, hepatitis and ‘no diseases’. A comparison has been made between FT growth and Naïve Bayes algorithms. A total of 29 datasets with 12 different attributes were used for the training the algorithm. The results ended with revealing Naïve Bayes is better than FT growth algorithm as the former gives the accuracy of 75.54% while the latter gives accuracy of 72.66% when compared using WEKA Tool using the 29 datasets [31].

Divya Tomar, et al. (2015) proposed a model for diagnosing the diseases efficiently that used weighted least squares twin support vector machine (WLSTSVM) for classification purpose. The method uses a hybrid feature selection (HFS) which combines filter and feature selection approaches. To cope up with the correlated features and imbalanced data, the method used sequential forward selection (SFS) method and correlation feature selection (CFS). The features were analyzed for their importance. The dataset consisting of 155 records for Hepatitis disease is taken from UCI machine learning warehouse and subjected to the proposed algorithm. On the basis of accuracy, sensitivity, specificity, and geometric mean, the method is evaluated. The accuracy comes out to be 86.67% with original features and 87.50% with reduced features using the proposed WLSTSVM model for diagnosing hepatitis disease [32].

Heba Ayeledeen, et al. (2015) proposed a model based on decision tree classifier for medicinal decision of liver fibrosis using classification strategies. The sample population dataset consisted of the records of 100 chronic HCV patients who were between 19-60 years old. The patients recorded in the dataset were analyzed at Department of Biochemistry and Molecular Biology of KasrAlainy Hospital of Cairo University. J48 was used as a classifier to classify the data using Weka data mining tool to predict the presence of fibrosis and Cirrhosis. The results demonstrated stated that the decision tree classifier accuracy was calculated in higher range as 93.7% and was superior to other classification calculations [33].
A.S. Anneshkumar, et al. (2015) handled the classification of liver disorder by using Fuzzy based arrangement to accomplish better precision for liver disease prediction. The authors used feature selection and fuzzy K-means classification that facilitated to set up a procedure for predicting liver disease. The dataset was collected from a hospital situated in the southern region of Tamil Nadu and consisted of 6078 instances with 48 attributes. The significant attributes were selected and were used to train the novel Fuzzy K – means model. When the results were compared, Fuzzy based classification gives better performance in categorizing liver disorder type appropriately with fundamental attributes. For each type of liver disorder, the accuracy achieved by Fuzzy K – means model based classification came out to be 94% [34].

P.Thangarajul, et al. (2015) used evolutionary algorithm particle swarm optimization algorithm (PSO) with KStar classification. The liver disease dataset is acquired from the UCI machine learning repository having 345 records and 7 attributes including the target attribute. Using a java application, the proposed algorithm is implemented. The target attribute categorized the patients into two classes – either having a disease or not. The authors are of the opinion that the accuracy indubitably improved with PSO-KStar algorithm and hence considered it as the most appropriate algorithm for the categorization of liver ailments. The given framework is understandable, transformable and the accuracy calculated is 100% [35].

Mohammad Reza Moshkani, et al. (2016) presented a research paper on comparison of decision trees, neural networks, and SVM algorithms and tried to evaluate the algorithms on the basis of their accuracy. The dataset was acquired from the data base of University of California accessible through the university website “http://archive.ics.uci.edu”. The dataset consists of 155 data samples classified into two classes – ‘Live’ and ‘Die’. The total number of ‘Live’ records was 123 and total number of ‘Die’ records was 32. There were 17 attributes and 1 target attribute in the dataset. According to the comparative study the authors stated that network algorithm had the higher accuracy in comparison with other algorithms and it could accurately predict 89.74% of the hepatitis related cases. The model found logic of 109 cases out of 116 cases for training the classifier and was tested on the 39 cases to predict the classes to which they belong. The model could correctly classify 29 samples and thus achieved the accuracy of 89.74% using the neural networks. The authors are of the opinion that if more health related data becomes accessible, then better results can be anticipated [36].

B. Tapas Ranjan, et al. (2016) attempted to propose an intelligent medical decision support system that could learn from the patterns inferred from the collected Liver disorder data and help physicians with proper medical diagnosis. A comparative analysis of effectiveness and correction rate of various decision trees J48, Naïve Bayes, ANN, ZeroR, 1 BK and VFI algorithm for detecting Liver disease using case-based reasoning (under different scenarios) is presented. The results reveal that the Multilayer perceptron gives the overall best classification than other classifiers with an accuracy of 71.59%. The authors are of the opinion that early detection of liver disease is very important as it can lead to early treatment of the disease and also can lead to perform betterin terms of their vivid correctness [37].

Saba Bashir, et al. (2016) exhibited a most advantageous arrangement of classifiers presented with multilayer classification. The ensemble method namely HMV (Hierarchical Majority Voting) uses bagging with multi-objective optimized weights. The model prevails over the limitations of usual performance bottlenecks by making the most of an ensemble of seven heterogeneous classifiers and suggests liver illness diagnosis. The liver dataset i.e., BUPA liver disease dataset and Indian Liver Patient Dataset (ILPD) are taken from the UCI data repository. The model has been experimentally evaluated with a number of well known classifiers and the experimental assessment shows that the proposed model can accelerate the training speed of the network and achieved 96.67% accuracy. The HMV ensemble framework gave a classification accuracy of 71.53% for ILPD dataset and 67.54% accuracy for Bupa liver disease dataset. When compared to the other state of the art techniques, HMV achieves the highest accuracy of 86.45% [38].

Shapla Rani Ghosh, et al. (2017) considered some of the classification algorithms such as Naïve Bayes, Bagging, KStar, Logistic Model Tree (combination of logistic regression (LR) and decision tree learning) and REP (reduced-error pruning) Tree and two liver patient datasets for predicting liver diseases. One was
collected from Andhra Pradesh state of India which consisted of 583 liver patient records with 10 attributes and another was acquired from University of California at Irvine (UCI) machine learning repository which consisted of 345 records with 6 attributes. The selected algorithms were applied on the two datasets in Weka and were evaluated on the basis of accuracy, precision, sensitivity and specificity. 100% accuracy, precision, sensitivity and specificity was given by KStar in both the datasets while Naïve Bayes gave the minimum accuracy [39].

Maruf Pasha, et al. (2017) used the meta-learning algorithms to work with the Indian liver patient dataset acquired from the UCI machine learning repository. The dataset contained 583 records of the liver patients where a total of 416 patients suffered from liver disease. The dataset consisted of 11 attributes as age, gender, TB (total Bilirubin), DB (direct Bilirubin), alkhphos (Alkaline Phosphotase), SGPT Alamine Aminotransferase, SGOT Aspartate Aminotransferase, TP (total proteins), ALB (albumin), A/G ratio (Albumin and Globulin Ratio) and target attribute. The algorithms Adaboost, logitboost (logistic regression to AdaBoost method), Bagging (bootstrap aggregating) and Grading were used by the authors in Weka for classification purpose. The highest accuracy in terms of Correct Classification rate was given by Grading meta-learning algorithm with 71.3551% accuracy. The lowest Correct Classification rate was given by Adaboost with 70.3259%. 70.4974% accuracy in terms of Correct Classification Rate was provided by Logitboost and Bagging. Grading meta-learning algorithm also resulted in the less execution time. The incorrect classification rate given by Adaboost, Logitboost, Bagging and Grading were calculated as 29.6741%, 29.5026%, 29.5026% and 28.6449% respectively [40].

Nancy.P, et al. (2017) performed a comparative study of about fifteen data mining classification algorithms viz. Rnd Tree, Quinlan decision tree algorithm (C4.5), K-Nearest Neighbor algorithm etc. These algorithms were used on a large Hepatitis dataset acquired from the UCI Machine Learning Repository. The dataset comprised of the 20 attributes including the target class attribute and a total of 155 instances. The authors discussed the importance of feature selection in their work and also used Fisher filtering, Relief filtering, Step Disc on the dataset and subsequently classified the dataset using fifteen most common supervised algorithms. In their paper they have mentioned that the classification algorithms such as BVM, CVM and Rnd Tree produced 100 percent accuracy for classification of bivalued classes of training data. The classification algorithms were later applied to test data. Their results point out the classification accuracy in terms of error rate and importance of all the instances in identifying the survival of a liver disease patient in future [41].

NazmunNahar, et al. (2018) used various decision tree techniques and investigated the early prediction of liver disease. The liver disease dataset used for carrying out the experimental analysis was collected from the UCI repository and consisted of attributes like total bilirubin, age, gender, total proteins, albumin and globulin ratio. The authors calculated and compared the performance of seven decision tree techniques, specifically J48, LMT, Random Forest, Random Tree, REPTree, Decision Stump and Hoeffding Tree on the basis of Accuracy, Mean Absolute Error, Precision, Recall, Kappa statistics and Runtime. At the end claimed that according to their analysis Decision Stump provided the highest accuracy of 70.67%. The authors alleged of using this study to develop other advanced decision trees in future [7].

M. Priya, et al. (2018) used the feature selection methodology and put forward the construction of a feature model that improved prediction accuracy of Indian liver patients. The dataset has been collected from the UCI repository. The accuracy improves in three phases. Firstly, min max normalization algorithm is applied on the original liver patient datasets. Secondly, by using the PSO feature selection, whole normalized liver patient datasets gets reduced to a dataset comprising the significant attributes only. Then the data set is submitted to the classifiers and lastly, the accuracy is calculated using root mean square value and root mean error value. Based on the performance of the algorithms after applying PSO feature selection, J48 algorithm’s performance is considered the better than the rest. The accuracy of J48 other classification algorithm with the help of feature selection was calculated as 95.04% [42].
Naiping Li, et al. (2018) made use of the Logistic regression to create the non-invasive assessment model. The research intends to establish their model with serum markers using cloud computing and internet of things for the analysis of liver disease. The dataset consists of the records of patients who went through the liver biopsy in the Second Xiangya Hospital, Central South University. Using logistic regression the accuracy to recognize significant liver fibrosis got to more than 70%. With 12 selected indexes from the 24 indexes when used, the general accuracy moved ahead from 70% to 72% [43].

Mehrbakhsh Nilashi, et al. (2019) proposed a precise method of neuro-fuzzy technique in diagnosing Hepatitis disease to support the ensemble learning of the data. In order to perform the data dimensionality reduction, the authors used a Non-linear Iterative Partial Least Squares method. For clustering task Self Organizing Map technique was used and for prognosis of the hepatitis disease ensembles of Neuro-Fuzzy Inference System was used. The selection of the most important features in the experimental dataset was done by using decision trees. The authors in addition tested their proposed method on real-world dataset obtained from UCI repositories the accuracy of the proposed method measured by ROC was 93.06% on this dataset. The results were contrasted with the results of previous studies. The analysis done by the authors demonstrate that the performance of the proposed method is greater than that of the Neural Network (NN), ANFIS, K-Nearest Neighbors and (SVM) Support Vector Machine and hence their method has the ability to be used as an Intelligent Learning System for the diagnosis of hepatitis. The authors have suggested that whenever some new information becomes available, a method should be developed which can automatically update the trained models and can be more competent in terms of the memory requirement [44].

Sivakumar D, et al. (2019) utilized the K-means clustering calculation and the C4.5 decision tree approaches to predict the chronic liver disease. The research focused on the life quality factors along with their ranges at different level of risk. Classification algorithms k-means clustering algorithm and the C4.5 decision tree have been used for predicting the chronic liver disease and the results have been compared in terms of accuracy, recall, precision and the F-measures to show the results with the distinct error measures RMSE, MAE and Kappa measurement. The dataset acquired from the online repository consists of a number of attributes such as family history of liver disease, alcohol consumption, smoking, contaminated food intake, diabetes and obesity etc. The percentage of 80:20 is considered for training and testing liver disease predicting structure. The calculable accuracy for C4.5 is 94.36% and that of K-means is 93.7% and undeniably shows that C4.5 provides a better precision. The research study provides establishment of an outline formation for early diagnosis for liver diseases [45].

N Komal Kumar, et al. (2019) compared the performance of the classification algorithms such as Logistic regression, random forest, decision tree, C4.5 and Multilayer perceptron classifiers for examining and forecasting hepatitis disease. The assessment of performance is based on the precision values and the highest accuracy is achieved by the Random forest classifier with a highest accuracy of 90.3226% among five classifiers and suitably categorized the records in the dataset. The execution time for the Random forest classifier was recorded as 0.14 sec. The authors have presented the comparison of the five classifying algorithms for classifying and predicting infectious hepatitis disease. The authors have suggested utilizing hybrid algorithms with a genetic algorithm for predicting hepatitis disease in future [46].

Ain Najwa Arbain, et al. (2019) predicted the liver disease on imbalanced data through random sampling and explored the data mining algorithm on it. The results were analyzed and evaluated on the basis of accuracy and ROC index. The authors performed their work with 583 records of patients of Andhra Pradesh classified into liver patient and non-liver patient. The dataset suffered from class imbalance apart from missing values and high range. To ensure that the target class is balanced, random sampling is done before applying data mining algorithms. Upon calculating the performance they referred that K – Nearest Neighbour (k-NN) surpasses the other algorithm such as Logistic Regression, AutoNeural and Random Forest. The accuracy given by K-Nearest Neighbour came out to be 99.794% and they claimed of better performance on Andhra Pradesh liver disease dataset. The accuracy rate of AutoNeural according to the authors was 99.779% and that of Logistic Regression with backward selection method was 99.764%. The authors are of the opinion that early prediction of liver disease is necessary to increase the survival rate of liver disease patient. The authors
have concluded that oversampling method can be applied to the dataset to address this issue of overfitting due to which Random Forest algorithm was not regarded as the most suitable algorithm in their research [47].

Table 1: Summary of Literature Review

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Author(s)</th>
<th>Year</th>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Dataset Used</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P.Rajeswari, G.Sophia Reena</td>
<td>2010</td>
<td>FT Tree</td>
<td>97.10%</td>
<td>Liver Disease Dataset - UCI Machine Learning Data Repository</td>
<td>[17]</td>
</tr>
<tr>
<td>2</td>
<td>A.H.Roslin a, A.Noraziah</td>
<td>2010</td>
<td>Support Vector Machines and Wrapper method</td>
<td>74.55%</td>
<td>Hepatitis Dataset - UCI Machine Learning Data Repository</td>
<td>[18]</td>
</tr>
<tr>
<td>3</td>
<td>Huda Yasin, Tahseen A. Jilani, Madiha Danish</td>
<td>2011</td>
<td>Binary Logistic Regression</td>
<td>89%</td>
<td>Hepatitis Dataset - UCI Machine Learning Data Repository</td>
<td>[19]</td>
</tr>
<tr>
<td>4</td>
<td>Tahseen A. Jilani, Huda Yasin, Madiha Mohammad Yasin</td>
<td>2011</td>
<td>PCA-ANN</td>
<td>100%</td>
<td>Hepatitis Dataset - UCI Machine Learning Data Repository</td>
<td>[20]</td>
</tr>
<tr>
<td>5</td>
<td>Javad SalimiSartakht, Mohammad Hossein Zangoeei, Kourosh Mozafari</td>
<td>2011</td>
<td>SVM-SA</td>
<td>96.25%</td>
<td>Hepatitis Dataset - UCI Machine Learning Data Repository</td>
<td>[21]</td>
</tr>
<tr>
<td>7</td>
<td>A.S.Anees hkumar and C.JothiVen kateswaran</td>
<td>2012</td>
<td>C4.5</td>
<td>99.20%</td>
<td>Liver Disease Dataset - Public Charitable Hospital in Chennai</td>
<td>[22]</td>
</tr>
<tr>
<td>8</td>
<td>Bendi VenkataRamanal and, Prof. M. Surendra Prasad Babu</td>
<td>2012</td>
<td>Modified Rotation Forest</td>
<td>74.7826%</td>
<td>BUPA Liver Disorder Dataset</td>
<td>[23]</td>
</tr>
<tr>
<td>9</td>
<td>Varun</td>
<td>2012</td>
<td>HPM</td>
<td>83.12%</td>
<td>Liver Disease Dataset</td>
<td>[24]</td>
</tr>
<tr>
<td>No.</td>
<td>First Name</td>
<td>Last Name</td>
<td>Year</td>
<td>Algorithm</td>
<td>Accuracy</td>
<td>Dataset - UCI Machine Learning Data Repository</td>
</tr>
<tr>
<td>-----</td>
<td>------------</td>
<td>-----------</td>
<td>------</td>
<td>-----------</td>
<td>----------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>10</td>
<td>FadlMutaher Ba-Alwi, Houzifa M. Hintaya</td>
<td>2013</td>
<td>Naive Bayes</td>
<td>96.52%</td>
<td>Hepatitis Dataset - UCI Machine Learning Data Repository</td>
<td>[25]</td>
</tr>
<tr>
<td>11</td>
<td>Mahdieh Adeli, Nooshin Bigdeli, Karim Afroshar</td>
<td>2013</td>
<td>GA-ANFIS</td>
<td>97.44%</td>
<td>Hepatitis Dataset - UCI Machine Learning Data Repository</td>
<td>[26]</td>
</tr>
<tr>
<td>12</td>
<td>Yilmaz Kaya, Murat Uyar</td>
<td>2013</td>
<td>RS-ELM</td>
<td>96.49%</td>
<td>Hepatitis Dataset - UCI Machine Learning Data Repository</td>
<td>[27]</td>
</tr>
<tr>
<td>13</td>
<td>Sa’sdiyah Noor Novita Alfisahrin, Teddy Mantoro</td>
<td>2013</td>
<td>NB Tree</td>
<td>67.01%</td>
<td>Liver Disease Dataset - UCI Machine Learning Data Repository</td>
<td>[14]</td>
</tr>
<tr>
<td>14</td>
<td>Anju Gulia, Dr. Rajan Vohra, Praveen Rani</td>
<td>2014</td>
<td>Random Forest</td>
<td>71.8696%</td>
<td>Liver Disease Dataset - UCI Machine Learning Data Repository</td>
<td>[28]</td>
</tr>
<tr>
<td>15</td>
<td>Manjeevan Seera, Chee Peng Lim</td>
<td>2014</td>
<td>FMM-CART-RF</td>
<td>95.01%</td>
<td>BUPA Liver Disorder Dataset – UCI Machine Learning Data Repository</td>
<td>[29]</td>
</tr>
<tr>
<td>17</td>
<td>S. Dhamodharan</td>
<td>2014</td>
<td>Naïve Bayes</td>
<td>75.54%</td>
<td>Not mentioned</td>
<td>[31]</td>
</tr>
<tr>
<td>18</td>
<td>DivyaTomar, Sonali Agarwal</td>
<td>2015</td>
<td>WLSTSVM</td>
<td>87.50%</td>
<td>Hepatitis Dataset - UCI Machine Learning Data Repository</td>
<td>[32]</td>
</tr>
<tr>
<td>19</td>
<td>Heba Ayeldeen, Olfat Shaker, Ghada Ayeldeen, Khaled M</td>
<td>2015</td>
<td>Decision Tree Classifier</td>
<td>93.7%</td>
<td>Sample HCV Dataset - KasrAlainy Hospital of Cairo University</td>
<td>[33]</td>
</tr>
<tr>
<td></td>
<td>Authors</td>
<td>Year</td>
<td>Method</td>
<td>Accuracy</td>
<td>Dataset</td>
<td>Reference</td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>------</td>
<td>--------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>20</td>
<td>A.S.Anees hkumar, Dr. C.JothiVen kateswaran</td>
<td>2015</td>
<td>Fuzzy K-means</td>
<td>94%</td>
<td>Liver Disorder Dataset – Tamil Nadu Hospital</td>
<td>[34]</td>
</tr>
<tr>
<td>21</td>
<td>P. Thangaraju l, R.Mehala</td>
<td>2015</td>
<td>PSO-KStar</td>
<td>100%</td>
<td>Liver Patient Dataset - UCI Machine Learning Data Repository</td>
<td>[35]</td>
</tr>
<tr>
<td>22</td>
<td>Mohamma d Reza Moshkani, Mahdi Rousta, Dr.Yaghou bFajrjami</td>
<td>2016</td>
<td>Neural Network</td>
<td>89.74%</td>
<td>Hepatitis Dataset - Website of California University</td>
<td>[36]</td>
</tr>
<tr>
<td>23</td>
<td>B. Tapas Ranjan, Subhendu Kumar Pani</td>
<td>2016</td>
<td>Multilayer perceptron</td>
<td>71.59%</td>
<td>Liver Disorder Dataset</td>
<td>[37]</td>
</tr>
<tr>
<td>24</td>
<td>Saba Bashir, Usman Qamar, Farhan. Hassan Khan, Lubna Naseem.</td>
<td>2016</td>
<td>HMV</td>
<td>86.45%</td>
<td>BUPA Liver Disorder Dataset and ILPD (Indian Liver Patient Dataset) Dataset – UCI Machine Learning Data Repository</td>
<td>[38]</td>
</tr>
<tr>
<td>25</td>
<td>Shapla Rani Ghosh, SajjadWah eed</td>
<td>2017</td>
<td>KStar</td>
<td>100%</td>
<td>Andhra Pradesh Liver Patient dataset and BUPA Liver Disorder Dataset</td>
<td>[39]</td>
</tr>
<tr>
<td>26</td>
<td>Maruf Pasha, Meherwar Fatima</td>
<td>2017</td>
<td>Grading meta-learning</td>
<td>71.3551%</td>
<td>ILPD (Indian Liver Patient Dataset) Dataset - UCI Machine Learning Data Repository</td>
<td>[40]</td>
</tr>
<tr>
<td>28</td>
<td>NazmunNa har, Ferdous Ara</td>
<td>2018</td>
<td>Decision Stump</td>
<td>70.67%</td>
<td>Liver Patient Dataset - UCI Machine Learning Data Repository</td>
<td>[7]</td>
</tr>
<tr>
<td>29</td>
<td>M. BanuPriya,</td>
<td>2018</td>
<td>J48 with feature</td>
<td>95.04%</td>
<td>ILPD (Indian Liver Patient Dataset)</td>
<td>[42]</td>
</tr>
</tbody>
</table>
### 3. Conclusion

Data Analytics in the healthcare system has improved the quality of health care services to patients [48]. To improve the quality of healthcare, data mining techniques have played a vital role in supporting medical diagnosis. Predicting diseases in order to reduce the mortality rate still remains a major challenge in the medical field. For medically diagnosing and predicting the diseases precisely, classification and prediction models have been adopted by healthcare sectors. This paper represents a review of the literature related to various data mining algorithms used for prediction and prognosis of liver diseases and a brief comparison between the accuracies attained by them in classifying the disease. This paper adds to the literature and uses a systematic approach beginning with the review process. The results of this review show that the type of dataset used and
the significant attributes (liver patient’s symptoms recorded in the disease dataset) considered for classification greatly influences the accuracy of the algorithm. It also revealed that in recent years there is a dramatic rise in adopting the data mining techniques for predicting the liver disease and there is a need to use the data mining algorithms in order to improve the method of predicting the disease, to decrease the time of computation, boost the precision of the system and apply it to other suitable datasets as well. Hybrid techniques as well as Meta learning algorithms can be successfully used to get enhanced performance precision for liver disorder diseases prediction. Accordingly, Random forest algorithm and advanced decision trees such as CART can be further explored for an exact prediction of liver diseases. Most recent and up to date data from different parts of the world can be collected for liver disease prognosis. Further, a method for automatic updating of trained models can be worked upon in classifying the disease.

References


[2] Liver Disease, Mayo Clinic, Available at: https://www.mayoclinic.org/diseases-conditions/liver-problems/symptoms-causes/syc-20374502


