IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-7, July 2019 E-ISSN: 2347-2693

Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms

Binish Khan[1*], Piyush Kumar Shukla [2], Manish Kumar Ahirwar [3]

[1,2,3]Department of Computer Science, University Institute of Technology, RGPV, Bhopal, India

[*]*Corresponding Author: binish.khan190@gmail.com, Tel.: +91-7974277396*

DOI: https://doi.org/10.26438/ijcse/v7i7.7176 | Available online at: www.ijcseonline.org

Accepted: 15/Jul/2019, Published: 31/Jul/2019

Abstract — Liver diseases averts the normal function of the liver. Mainly due to the large amount of alcohol consumption liver disease arises. Early prediction of liver disease using classification algorithms is an efficacious task that can help the doctors to diagnose the disease within a short duration of time. Discovering the existence of liver disease at an early stage is a complex task for the doctors. The main objective of this paper is to analyse the parameters of various classification algorithms and compare their predictive accuracies so as to find out the best classifier for determining the liver disease. This paper focuses on the related works of various authors on liver disease such that algorithms were implemented using Weka tool that is a machine learning software written in Java. Various attributes that are essential in the prediction of liver disease were examined and the dataset of liver patients were also evaluated. This paper compares various classification algorithms such as Random Forest, Logistic Regression and Separation Algorithm with an aim to identify the best technique. Based on this study, Random Forest with the highest accuracy outperformed the other algorithms and can be further utilised in the prediction of liver disease.

Keywords — Healthcare, Prediction, Liver Disease, Classification Algorithms, Random Forest, Logistic Regression and Separation Algorithm.

I. INTRODUCTION

Healthcare is an efficacious part of a country's economy. It makes a provision to improve the health by taking the certain essential measures into consideration. It is mainly composed of the Primary Care, Secondary Care and Tertiary Care [5]. Thus, "The main aim behind the Healthcare System is to deliver the best quality of services and to predict the diseases at an early stage."

Liver is an essential organ of our body. There is a great need for an early detection of liver disease so as to prevent complete liver failure, which can result in patient's death. For the proper diagnosis, it is necessary to evaluate some of the main attributes of liver patient's dataset [12]. Some of the main attributes of liver disease include, "Total_bilirubin, direct_bilirubin, alkaline_phosphotas, total_protein, albumin and globulin_ratio." Below, figure 1 shows the various functions that are performed by the liver.

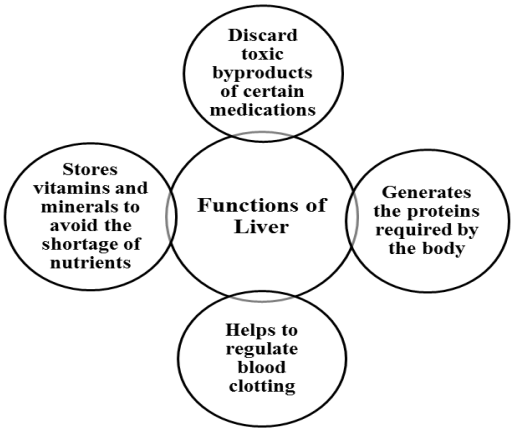

Figure 1. Functions performed by the liver

It is the challenging task for doctors to accurately predict the liver disease. Various classification techniques are used to classify the data and predict the liver disease through the datasets of liver patients [9]. "Having access to classification algorithms with large amount of data will help clinicians make better decisions and ultimately improve patient outcomes with an accurate prediction of liver disease." This paper shows a survey about the classification techniques that can be used for the prediction of liver disease and gives an idea for future work, that which classification technique can be utilised further for diagnosis of the liver disease.

A. Tool Used

Weka Tool

Weka is an efficient machine learning software which is written in java and developed at Waikato. It is used to classify various parameters when different algorithmic approaches are applied based on the datasets. It is an accumulation of visualization tools and algorithms that can be utilised for analysing the data and predictive modelling. Classification algorithms are implemented using Weka that are utilised for the prediction of liver disease at an early stage [3]. The experimental results of classification algorithms provide ease to the doctors through the provision of accurate patient's pathological status. As shown in Figure 2, Weka can be utilised to perform various data mining tasks.

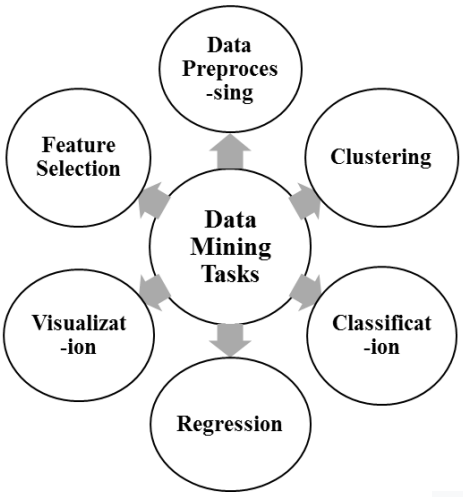

Figure 2. Data Mining Tasks performed by Weka [3]

B. Classification Algorithms Used

Random Forest

Random forests is a supervised machine learning algorithm that supports both classification and regression tasks. It mainly deals with the construction of multiple decision trees. It follows a basic approach where a dataset is divided into a batch of random datasets such that a decision tree is built for each random datasets [8]. Thus, "The forest is an ensemble of decision trees that are trained and all of them come up with a decision such that a majority vote is considered which results in a final single decision." It can operate on large data set and maintains the accuracy for missing data.

Logistic Regression

Logistic Regression is a supervised machine learning algorithm that is used for classification. It is utilised for predictive modelling and helps to calculate the possibility of a particular event taking place. It mainly deals with the prediction of binary outcomes for a given set of independent variables and determine the discrete values. It performs the binary classification and predict the future outcomes based on training from the previous output [11].

Separation Algorithm

Separation algorithm is a novel approach that mainly deals with an idea to first treat every data point as a fundamental entity. It mainly signifies that, "Every data point is separated which may be d-dimensional from every other point by hyper planes such that no two points are unseparated by at least one hyper plane." This algorithm is non-iterative and demonstrate an approach that once the data points are separated by planes then these planes are utilised to classify the data points [6].

C. Methodology

In this section, a flow diagram is demonstrated which consist of various classification algorithms that are evaluated and compared on the basis of accuracy parameter. The algorithms are analysed so as to find the best classifier, which can be further used to predict the liver disease [4]. On the basis of accuracy, Random Forest algorithm with 100% accuracy outperformed other algorithms. Below, Figure 3 depicts the work flow diagram that is used to determine the best classification technique.

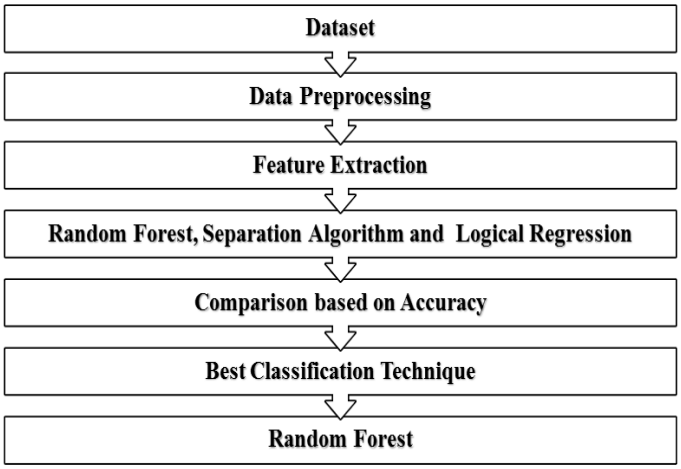

Figure 3. Flow Diagram [4]

The remaining portion of the paper is organized as follows, "Section I consists of Introduction that describes about the tool used and various classifiers that are required for an accurate prediction along with the methodology that contains a flowchart, Section II consists of Literature Survey that provides a description of the work performed by several authors and also include a table comparing the work of several authors, Section III consists of a Conclusion that describes about the accomplishment of this study by providing the best classification technique that can be utilised for an accurate prediction of liver disease and followed by References."

II. LITERATURE SURVEY

Hoon Jin et al. [1], described the concept of various classification techniques that assist the doctors to determine

the disease quickly and efficiently. Various classifiers such as Naïve Bayes, Multi-Layer Perceptron, Decision Tree and k-NN were compared and analysed based on several parameters like specificity, sensitivity and so on. The algorithms were implemented using the Weka tool and dataset was collected from UCI Repository. The experimental results showed that in terms of precision, Naïve Bayes gave the better classification results whereas Logistic Regression and Random Forest gave better results in terms of recall and sensitivity.

Ayesha Pathan et al. [2], proposed the concept of diagnoses of liver disease. Various classification algorithms were used such as Naïve Bayes, Ada Boost, J48, Bagging and Random Forest. These algorithms were further compared based on the parameters such as Accuracy, Error rate and so on. Also, Pre-processing technique was utilised to divide the data into two groups- liver patients and non-liver patient that was accomplished using K means clustering algorithm. Further, the clustered dataset was applied to the various classification algorithms. The implementation of the different classification algorithms was performed using the Weka Tool. The overall comparison was done between Naïve Bayes, Ada Boost, J48, Bagging and Random Forest algorithms. After the comparison was performed, the comparative study showed that the Random Forest gave the better results as compared to the other algorithms.

Tapas Ranjan Baitharu et al. [3], presented an approach of diagnoses of liver disorder through an analysis of liver disorder datasets. The main focus of the research was to help the physicians with the medical decision making process. Several algorithms were compared on various parameters such as Naive Bayes, ANN, ZeroR, IBK, VFI, J48 and Multilayer Perceptron. The algorithms were implemented using the Weka tool and dataset was collected from UCI Repository. The experimental results showed that Multilayer Perceptron gave the better classification results as compared to other algorithms. Thus, Multilayer Perceptron can be further used to diagnose the liver disorder efficiently.

Dr. S. Vijayarani1 et al. [4], demonstrated the predictive analysis of liver disorder using various classification algorithms. In this approach, Naïve Bayes and Support Vector Machine classification algorithms were used. These two algorithms were compared on the basis of performance parameters that include classification accuracy measures and execution time measures. The proposed system was implemented using Matlab 2013 tool and evaluated the dataset that had been collected from UCI Repository. After the experimental results, it had been observed that Support Vector Machine outperformed Naïve Bayes Algorithm due to the highest classification accuracy and can be used further in the prediction of liver disease.

Rakhi Ray et al. [5], proposed the concept of healthcare management system using a data mining technique. Owing to the great advantages various organizations are using data mining technology. Healthcare is a vital part for everyone. Different new technologies are inventing to examine physical conditions and finding symptoms of the different disease. There is a huge amount of data involved with it including a patient's past medical records, examination history, and even the personal details. If the symptoms are known, anyone can take enough precaution, and the sudden risk of severe effect can be minimized. Since there are large amount of data related to the medical systems, an efficient method to find the appropriate data from the database is required. Thus, Data mining is one of the best solution for an appropriate prediction of diseases.

B.Saritha et al. [6], presented a novel approach to detect the liver disorders of patients at an early stage. Separation of points by planes algorithm was used to distinguish healthy patients from the unhealthy patients and assisted in the diagnoses of liver disorder. The Separation algorithm was deployed to classify the functional data of liver. The data was collected from a hospital in Hyderabad. Thus, "Separation algorithm could diagnose the liver disorder with the accuracy of 85.1% and the total time taken for completion of training is 1 second and testing is 1 second".

Below, Table 1 shows the comparison of various parameters that are utilized for an analysis of liver disease. Weka tool was utilized by the several authors in order to evaluate the results of various classification algorithms. On the basis of their performance, best algorithm was evaluated and analyzed for the prediction of liver disease.

Table 2 shows the comparison of various parameters utilized in different decision trees and classification algorithms by several authors. Weka tool was utilized in order to evaluate the results of various algorithms on the liver disease dataset. On the basis of performance, best algorithm was analyzed and evaluated for the detection of liver disorder.

Table 1. Comparison Table of parameters used in different Classification Algorithms for Liver Disease Prediction.

Features	***Hoon Jin et al. (2014)***				***Ayesha Pathan et al. (2018)***				***Tapas Ranjan Baitharu et al. (2016)***			
Objective	To evaluate the results of classification algorithms for better prediction of liver disease.				To implement different classification algorithms using Weka in order to predict the liver disorder.				To forecast liver disease from Liver Function Test dataset using various classification algorithms.			
Dataset	UCI Repository **(Liver Disease Dataset).**				UCI Repository **(Liver Disease Dataset).**				UCI Repository **(Liver Disease Dataset).**			
Concerned Disease	Liver Disease				Liver Disease				Liver Disease			
Environment Used	Weka				Weka				Weka			
Attributes Used	11				11				7			
Algorithms Used	**Naïve Bayes**	**Decision Tree**	**Multilayer Perceptron**	**k-NN**	**Naïve Bayes**	**Ada Boost**	**J48**	**Random Forest**	**J48**	**ZeroR**	**Naïve Bayes**	**Multilayer Perceptron**
Specificity	0.952	0.352	0.303	0.467	-	-	-	-	-	-	-	-
Sensitivity	0.374	0.831	0.829	0.727	-	-	-	-	-	-	-	-
TP Rate	-	-	-	-	-	-	-	-	-	-	-	-
Precision	95.1	76.3	74.9	77.4	0.796%	0.508	0.872	1	-	-	-	-
F Measure	-	-	-	-	0.56	0.594	0.872	1	-	-	-	-
Accuracy	53.9	69.4	67.9	65.3	55.84%	71.31%	87.46%	100%	68.97	57.971	62.8986	60.2899
Error Rate	-	-	-	-	44.16%	28.69%	12.54%	0.00%	-	-	-	-
Recall	-	-	-	-	0.558	0.713	0.875	1	-	-	-	-
Kappa Statistics	-	-	-	-	-	-	-	-	0.3401	0	0.153	0.4023
Mean Absolute Error	-	-	-	-	-	-	-	-	0.3673	0.4874	0.4597	0.3543
Root Mean Squared Error	-	-	-	-	-	-	-	-	0.5025	0.4936	0.5083	0.4523
Relative Absolute Error	-	-	-	-	-	-	-	-	75.3511	100	102.9673	72.68
Best Algorithm	Naïve Bayes				Random Forest				Multilayer perceptron			
Result	In terms of precision, Naïve Bayes gave better classification results. Also, appropriate algorithms were evaluated and analysed for prediction of the liver disease.				Random Forest Algorithm gave better performance results as compared to other algorithms.				In terms of accuracy, multilayer perceptron gave best classification results as compared to other classifiers.			

Table 2. Comparison Table of parameters used in different Classification Algorithms and Decision Trees for Liver Disease Prediction

Features	***B.Saritha et al. (2017)***	***Han Ma et al. (2018)***				***Nazmun Nahar et al. (2018)***		
Objective	Rapid Initial Diagnosis of the liver disease by classifying liver function data using the separation algorithm.	To predict the Non-alcoholic Fatty Liver Disease using various machine learning techniques.				To predict the liver disease using various decision tree techniques.		
Dataset	Medics Path Labs India private Limited, Hyderabad **(Liver Disease Dataset).**	College of Medicine, Zhejiang University, China **(Liver Disease Dataset).**				UCI Repository **(Liver Disease Dataset).**		
Concerned Disease	Liver Disease	Liver Disease				Liver Disease		
Environment Used	Weka	Weka and Keel				Weka		
Attributes Used	11	18				11		
Algorithms Used	**Separation Algorithm**	**Logical Regression**	**Support Vector Machine**	**Naïve Bayes**	**Bayesian Network**	**Decision Stump**	**Hoeffding Tree**	**Logistic Model Tree**
Accuracy	85.1%	83.41%	82.73%	81.31%	82.92%	70.67%	69.75%	69.47%
Specificity	-	0.934	0.946	0.913	0.878	-	-	-
Precision	-	0.713	0.725	0.644	0.636	0.499	0.634	0.632
Recall	-	0.518	0.452	0.496	0.675	0.707	0.700	0.695
F-measure	-	0.600	0.557	0.560	0.655	0.585	0.619	0.628
Tree Size	-	-	-	-	-	Single Level	1	1
Mean Absolute Error	-	-	-	-	-	0.4392	0.4091	0.4116
Kappa Statistics	-	-	-	-	-	0.379	0.0501	0.065
Runtime	-	-	-	-	-	0.01	0.12	0.88
Best Algorithm	Separation Algorithm	Bayesian Network				Decision Stump		
Result	The Separation algorithm could diagnose the liver disorder with the accuracy of 85.1% and the total time taken for completion of training is 1 second and testing is 1 second.	In this study, F-measure was considered as an evaluative measure due to which Bayesian Network with the highest F-measure demonstrated the best performance and could be used as a major classifier for the earlier diagnoses of Non-alcoholic fatty liver disease (NAFLD).				In terms of accuracy, decision stump outperformed well as compared to other algorithms.		

III. CONCLUSION

The main aim of this study was to provide an overview of different classification algorithms popular in the field of data driven prediction of liver disease. In this survey, various classification algorithms were analysed that help doctors to predict the liver disease at an early stage. The purpose of this study was achieved by performing a comparative study of various papers. Thus, Random forest with 100% accuracy, Separation algorithm with 85.1% accuracy and Logistic Regression with 83.4% accuracy were compared and analysed as they have the better performance in terms of accuracy than other algorithms. From this evaluation study, the authors conclude that Random Forest with an accuracy of 100% outperformed the other algorithms and should be further explored for an accurate prediction of liver disease.

REFERENCES

[1] Hoon Jin, Seoungcheon Kim, Jinhong Kim, "Decision Factors on Effective Liver Patient Data Prediction", International Journal of Bio-Science and Bio-Technology, Vol. **6**, Issue.**4**, pp. **167-178**, **2014**.

[2] Ayesha Pathan, Diksha Mhaske2, Shrutika Jadhav, Rupali Bhondave, Dr.K.Rajeswari, "Comparative Study of Different Classification Algorithms on ILPD Dataset to Predict Liver Disorder", International Journal for Research in Applied Science & Engineering Technology (IJRASET**)**, Vol. **6**, Issue.**2,** pp. **388-394**, **2018**.

[3] Tapas Ranjan Baitharu, Subhendu Kumar Pani, "Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset", International Conference on Computational Modeling and Security, **India**, pp. **862-870**, **2016**.

[4] 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**.

[5] Rakhi Ray, "Advances in Data Mining: Healthcare Applications", International Research Journal of Engineering and Technology (IRJET), Vol. **5**, Issue.**3,** pp. **3738-3742**, **2018**.

[6] B.Saritha, S.V. Ramana, Narra Manaswini, RamaPriyanka, D.Hiranmayi, K.Eswaran, "Classification of liver data using a new algorithm", 4th International Conference on New Frontiers of Engineering, Science, Management and Humanities, **Hyderabad**, **2017**.

[7] HanMa, Cheng-fu Xu, Zhe Shen, Chao-hui Yu, You-ming Li, "Application of Machine Learning Techniques for Clinical Predictive Modeling: A Cross-Sectional Study on Nonalcoholic Fatty Liver Disease in China", BioMed Research International, pp. **1-9**, **2018**.

[8] Nazmun Nahar, Ferdous Ara, "Liver Disease Prediction using different decision tree techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. **8**, Issue.**2,** pp. **1-9**, **2018**.

[9] Shapla Rani Ghosh and Sajjad Waheed, "Analysis of classification algorithms for liver disease diagnosis", Journal of Science,Technology&Environment Informatics, Vol. **5**, Issue.**1,** pp. **360-370**, **2017**.

[10] Insha Arshad, Chiranjit Dutta, Tanupriya Choudhury, Abha Thakra, "Liver Disease detection due to excessive alcoholism using Data Mining Techniques", International Conference on Advances in Computing and Communication Engineering, **Paris**, **France**, pp. **163-168**, **2018**.

[11] V.V. Ramalingam, A.Pandian, R. Ragavendran, "Machine Learning Techniques on Liver Disease – A Survey", International Journal of Engineering & Technology, Vol. **7**, Issue.**4,** pp. **493-495**, **2018**.

[12] Shambel Kefelegn, Pooja Kamat, "Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey, International Journal of Pure and Applied Mathematics, Vol. **118**, Issue.**9,** pp. **765-769**, **2018**.

[13] R.S. Walse, G.D. Kurundkar, P. U. Bhalchandra, "A Review: Design and Development of Novel Techniques for Clustering and Classification of Data**",** International Journal of Scientific Research in Computer Sciences and Engineering, Vol.**6**, Special Issue.**01**, pp. **19-22**, **2018**.

[14] Pawan S. Wasnik, S.D.Khamitkar, Parag Bhalchandra, S. N. Lokhande, Ajit S. Adte, "An Observation of Different Algorithmic Technique of Association Rule and Clustering**",** International Journal of Scientific Research in Computer Sciences and Engineering, Vol.**6**, Special Issue.**01**, pp. **28-30**, **2018**.

Authors Profile

Miss Binish Khan is currently pursuing a Dual degree PG Integrated course (B.E. and MTech.) in Computer Science from University Institute of Technology, RGPV, Bhopal. She is eager to work in the field of healthcare in order to improve the prevailing conditions of healthcare system. Her research interests include various Classification Algorithms, Web Development, Disease Prediction, Machine Learning and Data Mining. The main idea behind her research is to provide ease to the doctors, various classification techniques can be utilized to predict the liver disease at an early stage so that essential precautions could be undertaken to improve the overall experience of the patients.

Dr. Piyush Kumar Shukla received his Bachelor's degree in Electronics & Communication Engineering, LNCT, Bhopal in 2001, MTech. (Computer Science & Engineering) in 2005 from SATI, Vidisha and Ph.D. (Computer Science & Engineering) in 2013 from RGPV, Bhopal, India. He is a member of ISTE (Life Member), IEEE, ACM, IACSIT, IAENG. Currently he is working as an Assistant Professor in Department of Computer Science, University Institute of Technology, RGPV, Bhopal. He has published more than 60 research papers in various International and National Journals and Conferences, including 4 papers in SCIE Journals and more than 10 papers in Scopus Journals. He has also published an Indian patent.

Mr. Manish Kumar Ahirwar is currently working as an Assistant Professor in Department of Computer Science, University Institute of Technology, RGPV, Bhopal. He has work experience of several years in the field of teaching. His research interests include Data Mining Algorithms, Internet of Things (IoT), Machine Learning and Cyber Security. He is a member of IEEE, ACM, IACSIT, IAENG. He has published more than 30 research papers in various International and National Journals and Conferences, including 4 papers in SCIE Journals and more than 10 papers in Scopus Journals. He has also published two Indian patents and two copyrights.

