Publisher homepage: www.universepg.com, ISSN: 2663-7804 (Online) & 2663-7790 (Print)
https://doi.org/10.34104/ajeit.020.085090
Australian Journal of Engineering and Innovative Technology
Journal homepage: www.universepg.com/journal/ajeit

Prediction of Liver Diseases by Using Few Machine Learning Based Approaches

Md. Shafiul Azam[1], Aishe Rahman[1], S. M. Hasan Sazzad Iqbal[1], and Md. Toukir Ahmed[1]*

[1]Department of Computer Science and Engineering, Pabna University of Science and Technology (PUST), Pabna, Bangladesh.

*Correspondence: toukirahmedreal@gmail.com (Md. Toukir Ahmed, Lecturer, Department of Computer Science and Engineering, PUST, Pabna, Bangladesh).

ABSTRACT

Advancement in medical science has always been one of the most vital aspects of the human race. With the progress in technology, the use of modern techniques and equipment is always imposed on treatment purposes. Nowadays, machine learning techniques have widely been used in medical science for assuring accuracy. In this work, we have constructed computational model building techniques for liver disease prediction accurately. We used some efficient classification algorithms: Random Forest, Perceptron, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) for predicting liver diseases. Our works provide the implementation of hybrid model construction and comparative analysis for improving prediction performance. At first, classification algorithms are applied to the original liver patient datasets collected from the UCI repository. Then we analyzed features and tweaked to improve the performance of our predictor and made a comparative analysis among the classifiers. We examined that, KNN algorithm outperformed all other techniques with feature selection.

Keywords: Classification, Feature selection, Liver disease, Machine learning, and Performance metrics.

INTRODUCTION

Researchers faces more challenging task in health-care sectors to predict the diseases from the voluminous medical databases. Nowadays data mining techniques are more essential in healthcare. Data mining tools and techniques including classification, clustering, association rule mining for assessing frequent patterns are applied to medical data for disease prediction. In data mining, classification techniques are much appreciated in medical diagnosis and predicting diseases (Ramana *et al*., 2011). Chronic Liver Disease is the leading cause of death worldwide which affects a large number of people worldwide. This disease it is caused by a combination of certain substances that damage the liver (Rahman *et al*., 2019). Liver is the largest internal organ in the human body, playing a major role in metabolism and serving several vital functions. It weighs about 3 lb (1.36 kg). The liver supports almost every organ in the body and is vital for our survival. Liver disease may not show any symptoms at earlier stage or the symptoms may seem low, like minor sickness and enervation. Symptoms somewhat rely on the type and the severity of liver disease. Liver diseases are diagnosed based on the liver functional test (Karthik *et al*., 2011). Classification techniques are widely applied in various automatic medical diagnoses. Problems with liver are not easily understood in primary stage as it will be functioning normally even when it is harmed (Liu and Huang, 2008). An early diagnosis of liver problems will accelerate patient's survival rate. Liver disease is often diagnosed by analyzing the enzyme levels in the blood (Schiff *et al*., 2007).

Review of Literature

Different researchers have worked on liver disease diagnosis previously and found accuracy of different machine learning algorithms using different tools. Rajeswari and Reena, (2010) in this paper, Authors perform data classification which is based on liver disorders this paper deals with the results in the field of data classification obtained with Naive Bayes algorithm, FT Tree algorithm and KStar algorithm. Dhamodharan, (2014) has predicted three major liver diseases such as Liver cancer, Cirrhosis and Hepatitis with the help of distinct symptoms. They used Naïve Bayes and FT Tree algorithms for disease prediction. Comparison of these two algorithms was assessed purely based on their classification accuracy measure. From the experimental results they concluded the Naïve bayes as the better algorithm which predicted diseases with maximum accuracy in classification than the other algorithm.

Ramana *et al*. (2011) the classification algorithms considered here are Naïve Bayes classifier, C4.5, Back propagation Neural Network algorithm, and Support Vector Machines. These algorithms are evaluated based on four criteria: Accuracy, Precision, Sensitivity and Specificity. Karthik *et al*. (2011) in first phase, ANN is used for classifying the liver disease. In second phase rough set rule induction using LEM (Learn by Example) algorithm is applied to generate classification rules. In third phase fuzzy rules are applied to identify the types of the liver disease.

Aneeshkumar and Venkateswaran, (2012) in this paper authors are using classification. The overall performance of C4.5decision tree is better than Naive Bayesian. Pahariyavohra *et al*. (2014) this work mainly represents computational intelligence techniques and measures for Liver Patient Classification. The efficacy of the techniques viz. Multiple Linear Regression, Support Vector Machine, Multilayer Feed Forward Neural Network, J-48, Random Forest and Genetic Programming has been tested on the ILPD Data Set. Authors employed under sampling and over sampling for balancing it. The results obtained from experiments indicate that Random Forest over sampling with 200% outperformed all the other techniques.

A study on intelligent techniques to classify the liver patients is performed by the Gulia *et al*. (2014). Among algorithms, J48 presents 70.669% accuracy, 70.8405% exactness is gathered by the MLP algorithm, SVM provides 71.3551% accuracy, 71.8696% accuracy is displayed by Random forest and Bayes Net shows 69.1252% accuracy. Rajeswari and Reena, (2010) used Naive Bayes, K star and FT tree to analyze the liver disease. Data set is taken from UCI consisting 345 instances and 7 attributes. 10 fold cross validation test are imposed by using WEKA tool. Naïve Bayes shows 96.52% correctness in 0 sec. 97.10% accuracy is gathered by using FT tree in 0.2 sec. Paul R Harper reported that, there does not exist necessarily only one best classification tool but instead the best performing algorithm will rely on the features of the dataset to be examined. Ramana *et al*. (2012) modified rotation forest algorithm was proposed with multi layer perception classification algorithm and random subset feature selection method for UCI liver data set.

Rosalina and Noraziah, (2010) deduced prediction on a hepatitis prognosis disease assisted by SVMand Wrapper Method. From the experimental outcome, they observed the ongoing accuracy rate in the clinical lab test cost with lower execution time. They have fulfilled the goal by the combination of Wrappers Method and SVM techniques. Among the most influential work in Micro-Array Analysis can be attributed to Rifkin *et al*. (2003) Their work is attributed to a Support Vector Machine to accurately (80%) predict the origin of tumors collected from samples obtained at Massachusetts General and other medical institutions.

METHODOLOGY:

The research work about this paper can provide the solutions of liver disease features which include process of feature selection applied on dataset and the performance of model construction. Comparative analysis of classification algorithms is performed for ameliorating accuracy in prediction of liver patients with or without feature selection. This paper finds answers to these questions which can help to know the various aspects about classification of liver patients. By performing this work, it is shown that feature selection has a great significance as the process of choosing a subset of most relevant features for their usage in the construction of model. By using feature selection on ILPD (Indian Liver Patient Dataset) before a classification algorithm can

be applied, performance of classification algorithm increases. This also provides that the splitting size also differ the accuracy of different algorithms. In this paper, five Classification algorithms Decision Tree, Perceptron, Support Vector Machine, Random Forest and K-Nearest Neighbors algorithms have been considered for comparing their performance based on the ILPD.

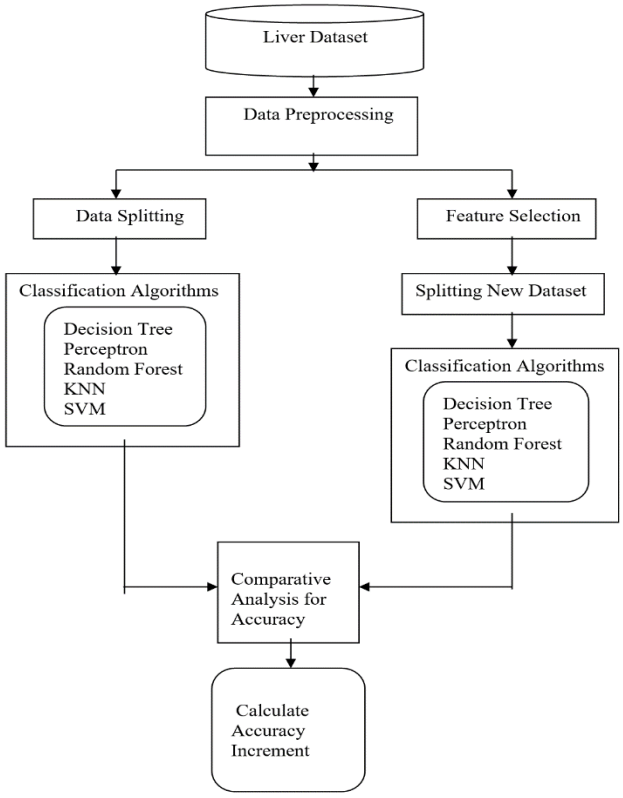

Fig 1: Hybrid model construction and comparative analysis for improving accuracy.

A. Description of Dataset

Databases of 583 records/entries are taken from the ILPD (Indian Liver Patient Dataset) Data set for the purpose of solving problem of this paper. This dataset is downloaded from UCI machine Learning Repository (http://archive.ics.uci.edu/ml/). Entire ILPD dataset is comprised of data about 583 Indian liver patients where in total 416 are liver patient records and 167 numbered non liver patient records. The dataset was gathered from north east of Andhra Pradesh, India.

This dataset contains ten features and one output. The features are Age, Gender, Total Bilirubin, Direct Billirubin, Alkaline Phosphotase, Alamine Amino-transferase, Aspartate Aminotransferase, Total Protiens, Albumin, Albumin and Globulin_Ratio. Result is a class label used to divide into groups (liver patient or not). In our dataset, only Albumin and Globulin Ratio contains 579 data, rest are 583 in count. Only the feature Gender contains two unique values in which most patients are Male. All of the patient in average 45 years old and features contain average result. There minimum and maximum rate for all numerical attributes and they are counted in 25%, 50% and 75% measurement. We see that maximum and minimum values for "Result" are 2 or 1 which represents disease or no disease. Maximum minimum and average value for LFT in this study are also given.

B. Proposed Algorithms

Decision Tree Classifier - For our first algorithm we will be using Decision Tree classifier. It is vastly used machine learning algorithms to this date. They are applied for both classification and regression problems. Now a question might arise why we are willing to use Decision tree classifier over other classifiers. To answer that question we can have two reasons. One being, Decision trees often tries to mimic the same way human brain thinks so it is quite simple to understand the data and come to some good conclusions or interpretations. To start a decision tree is a tree where there are a bunch of nodes and each node represent a feature (attribute), each link (branch) represent a decision otherwise known as rule and each leaf of the tree represent an outcome otherwise known as categorical or continues value. The idea is to create a tree for the entire data and get an outcome at every leaf (Russell, 2002).

Perceptron - Perceptron is a single layer neural network and a linear classifier (binary). It is used in supervised learning. The perceptron consists of 4 parts, Input values or One input layer, Weights and Bias, Net sum, Activation Function(Russell, 2002).

Random Forest Classifier - Random forests are one of the ensemble learning methods for classification (and regression) that works by generating multiple decision trees at training time and showing outcome the class result by individual trees(Russell, 2002). It is an excellent algorithm in terms of accuracy among mentioned algorithms. It runs effectively on large data base. It can handle thousands of input variables without variable shrinking. It gives estimates of variables having importance in the classification. Random Forests generated number of classification trees. The forests then have a choice of the classi-fication having the most votes (Gulia *et al.*, 2014).

K-Nearest Neighbor Algorithm - K-Nearest neighbor algorithm (KNN) is one of the most widely used supervised learning algorithms that have been applied in many applications in data mining. It follows a way for classifying relied on closest training samples in the feature space. An object is distinguished by a majority of its neighbors. The neighbors are chosen from an array of objects for which the correct classification is observed (Russell, 2002).

Support Vector Machines (SVM) - SVM is a learning system that uses a hypothesis space of linear functions in a high dimensional space, trained with a learning algorithm from optimization theory denoting a learning bias got from statistical learning theorem (Russell, 2002). SVM incorporates with a linear model to impose non-linear class boundaries by mapping input vectors non-linearly into a high dimensional feature space using kernels. The training examples that remain closest to the maximum margin hyper plane are known as support vectors. All other training examples are not relevant for deducing the binary class boundaries. Support vector machines are supervised learning models with incorporated learning algorithms that analyze data and deduce patterns, used for classification and regression (Gulia *et al*., 2014).

C. Performance Metrics

To assess the result of the study accurately, rather than accuracy alone, some of the other performance metrices were introduced in the result sections too. By observing these metrices, a clear indication of better result was noticed among different folding and splits of the dataset.

Performance parameters are the most important factor to compare among classifier methods to get the best classifier. Applied performance metrices includes Accuracy, precision, Recall and F-Score. These parameters calculated from a confusion matrix which situated in every step of classification (Russell, 2002). About confusion matrix and detailed information about these proposed parameters are as follows:

Table1: Confusion Matrix

	Predicted YES	**Predicted NO**
Actual YES	TP	FN
Actual NO	FP	TN

TP represents the number of correctly classified positive instances.
FP represents the number of misclassified positive instances.
FN represents the number of misclassified negative instances.
TN represents the number of correctly classified negative instances.

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

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 Score} = \frac{2*precision*recall}{precision+recall}$$

D. Model Evaluation

To evaluate and carry out the analysis, we first preprocessed our dataset by removing null values and converting textual features to numerical values. Then, we found out the relation between different features in our dataset in visually. After processing, dataset was split into Training (75%) and Testing (25%) set for algorithmic model construction.

After feeding training data to build model, testing data was applied to find out performance of result. Then, we tried tweaking in features for having even better result. For this, we tried to find the correlation of each two features and omit one of them if there is linearity. Then we have to split again and feed in algorithms. Lastly, we measured performance of each algorithm if they are increased or not.

RESULTS AND DISCUSSION:

After processing dataset, we fed it into above mentioned machine learning classifiers one by one. Firstly, we run the classifiers on raw dataset after processing, assessed results and run a comparative analysis among the classifiers. Then we ran classifiers again after selecting useful features to improve performance of our existing classifiers. We have done an elaborate experiment on all the classifiers mentioned above and found KNN as the best performing classifier. Performance Comparison is made among these classification algorithms before and after applying feature selection. From the analysis of each algorithm we can say that, at first we get most accuracy in Support Vector Machine

(71.22%). The experimental results are shown **Table 2** and also shown in **Fig 4**.

Table 2: Performance measure of various machine learning approaches.

	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
DTWOFST	60	61	60	60
DTWFST	72	72	72	72
PWOFST	39	80	39	31
PWFST	66	55	66	58
RFWOFST	64	65	64	64
RFWFST	73	74	73	73
KNNWOFST	66	64	66	65
KNNWFST	74	72	74	72
SVMWOFST	71	80	71	60
SVMWFST	72	80	72	61

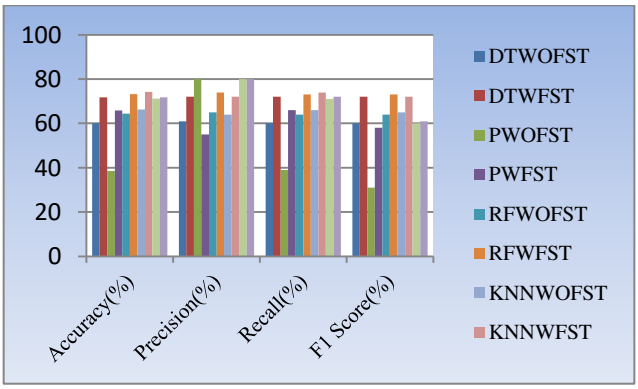

Fig 4: Performance measure of various machine learning approaches.

Abbreviations used in Table 2 and Fig 4:

DTWOFST: Decision Tree without Feature Selection Techniques
DTWFST: Decision Tree with Feature Selection Techniques
PWOFST: Perceptron without Feature Selection Techniques
PWFST: Perceptron with Feature Selection Techniques
RFWOFST: Random Forest without Feature Selection Techniques
RFWFST: Random Forest with Feature Selection Techniques
KNNWOFST: K-Nearest Neighbour without Feature Selection Techniques
KNNWFST: K-Nearest Neighbour with Feature Selection Techniques
SVMWOFST: Support Vector machine without Feature Selection Techniques
SVMWFST: Support Vector machine with Feature Selection Techniques

Rest algorithms are below 70 which should be increased for better observation for this dataset. To increase this we did feature selection techniques. After that all algorithms reached up to 70 without Perceptron. But this also increased in a large rate (38.54 to 65.85).

So our feature selection technique is efficient for this purpose of Liver Disease Prediction. We find out that, K-Nearest Neighbors algorithm outperformed all other techniques with 74.15% accuracy after applying Feature Selection.

CONCLUSION AND FUTURE DIRECTION:

This work presents an approach that will be used for hybrid model construction of community health services. These classification algorithms can be implemented for other dominant diseases also like cardiac and diabetes prediction and classification. More than one dataset may be used for better approach and comparison (Ahmed *et al.*, 2020). Another scope is to see whether by applying new algorithms will result any improvements over techniques which are used in this work in future. More techniques for accuracy increment may be applied. Wrapper method may be applied for removing noise in the dataset.

Classification rules and disease identifying techniques may also be generated by using different efficient algorithms. More than one database for comparative analysis may also be used. Our works has certain limitations as the model has underperformed having less accuracy than expectations. So, in future, inclusion of deep learning methods may improve our results further.

ACKNOWLEDGEMENT:

Many thanks to the co-author supported with proper assistance and help for analysis and writing to conduct successful research study.

CONFLICTS OF INTEREST:

The authors declare they have no competing interests with respect to the research.

REFERENCES:

1. Ahmed MT, Imtiaz MN, and Mitu NS. (2020). Impact of weather on crops in few northern parts of Bangladesh: HCI and machine learning based approach, *Aust. J. Eng. Innov. Technol*., **2**(1), 7-15. https://doi.org/10.34104/ajeit.020.07015
2. Aneeshkumar A. S. and Venkateswaran C. J. (2012). Estimating the surveillance of liver disorder using classification algorithms. *International Journal of Computer Applications*, **57**(6), Pp. 0975-8887. https://www.ijcaonline.org/archives/volume57/number6/9121-3281
3. Dhamodharan, S. (2014), Liver disease prediction using bayesian classification. In *4th National Conference on Advanced com-puting, applications & Technologies*, pp. 1-3.
4. Gulia, A, Vohra R, and Rani P. (2014). Liver Patient Classification Using Intelligent Techniques. *International J. of Computer Science and Information Techno-logies,* **5***(4)*, 5110-5115. http://www.ijcsit.com/docs/Volume%205/vol5issue04/ijcsit2014050462.pdf
5. Karthik, S., Priyadarishini, A., Anuradha, J. and Tripathy, B.K. (2011). Classification and rule extraction using rough set for diagnosis of liver disease and its types. *Adv Appl Sci Res*, **2**(3), pp.334-345. https://www.researchgate.net/publication/318645553
6. Liu, K.H. and Huang, D.S., 2008. Cancer classification using rotation forest. *Com-puters in biology and medicine*, **38**(5), pp.601-610. https://doi.org/10.1016/j.compbiomed.2008.02.007
7. Pahariyavohra J, Makhijani J. and Patsariya S. (2014). Liver patient classification using intelligence techniques. International journal of advanced research in computer science and software engineering. **4**(2): 295-299.
8. Rahman, A. S., Shamrat, F. J. M., Tasnim, Z., Roy, J. and Hossain, S. A. (2019). A Comparative Study on Liver Disease Prediction Using Supervised Machine Learning Algorithms. *International J. of Scientific & Technology Research*, *8*(11), pp.419-422.
9. Ramana BV, Babu MSP, Venkateswarlu NB. (2012). Liver Classification Using Modified Rotation Forest. *International Journal of Engineering Research and Development,* **1**(6), 17-24.
10. Ramana B. V., Babu M. S. P. *Venkateswarlu* N. B. (2011). A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *International Journal of Database Management Systems (IJDMS) ,* **3**(2), 101-114 . https://pdfs.semanticscholar.org/c92d/38a7a76c20a317de63fb9278bb10102c758b.pdf
11. Rajeswari P., Reena G. S. (2010). Analysis Of Liver Disorder Using Data Mining Algorithm. *Global Journal Of Computer Science And Technology* , **10**(14). Pp. 48-52. https://core.ac.uk/download/pdf/231162636.pdf
12. Rifkin R, Mukherjee S, Pablo, Tamayo, P, and Mesirov JP. (2003). An Analytical Method For Multi-Class Molecular Cancer Classification, *SIAM Review* **45**(4): 706-723. https://doi.org/10.1137/S0036144502411986
13. Rosalina. A.H, Noraziah. A. (2010). Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method, *IEEE*, Pp. 2209- 22. https://doi.org/10.1109/FSKD.2010.5569542
14. Russell, S. and Norvig, P. (2002). Artificial intelligence: a modern approach. Prentice Hall, Englewood Cliffs, New Jersey 07632. https://www.cin.ufpe.br/~tfl2/artificial-intelligence-modern-approach.9780131038059.25368.pdf
15. Schiff, E. R., Sorrell, M. F., & Maddrey, W. C. (2007). *Schiff's Diseases of the Liver* (10 ed.). https://miami.pure.elsevier.com/en/publications/schiffs-diseases-of-the-liver

Citation: Azam MS, Rahman A, Iqbal SMHS, and Ahmed MT. (2020). Prediction of liver diseases by using few machine learning based approaches, *Aust. J. Eng. Innov. Technol.*, **2**(5), 85-90. https://doi.org/10.34104/ajeit.020.085090