A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms

A.K.M Sazzadur Rahman, F. M. Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, Syed Akhter Hossain

Abstract: Chronic Liver Disease is the leading cause of global death that impacts the massive quantity of humans around the world. This disease is caused by an assortment of elements that harm the liver. For example, obesity, an undiagnosed hepatitis infection, alcohol misuse. Which is responsible for abnormal nerve function, coughing up or vomiting blood, kidney failure, liver failure, jaundice, liver encephalopathy and there are many more. This disease diagnosis is very costly and complicated. Therefore, the goal of this work is to evaluate the performance of different Machine Learning algorithms in order to reduce the high cost of chronic liver disease diagnosis by prediction. In this work, we used six algorithms Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest. The performance of different classification techniques was evaluated on different measurement techniques such as accuracy, precision, recall, f-1 score, and specificity. We found the accuracy 75%, 74%, 69%, 64%, 62% and 53% for LR, RF, DT, SVM, KNN and NB. The analysis result shown the LR achieved the highest accuracy. Moreover, our present study mainly focused on the use of clinical data for liver disease prediction and explore different ways of representing such data through our analysis.

Keywords: Machine Learning, Liver Disease, Classification, Supervised learning, Computational Intelligence, Regression, Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes.

1. INTRODUCTION

THE liver is the largest organ of the body and it is essential for digesting food and releasing the toxic element of the body. The viruses and alcohol use lead the liver towards liver damage and lead a human to a life-threatening condition. There are many types of liver diseases whereas hepatitis, cirrhosis, liver tumors, liver cancer, and many more. Among them liver diseases and cirrhosis as the main cause of death [1]. Therefore, liver disease is one of the major health problems in the world. Every year, around 2 million people died worldwide because of liver disease [2]. According to the Global Burden of Disease (GBD) project, published in BMC Medicine, one million peoples are died in 2010 because of cirrhosis and million are suffering from liver cancer [3]. Machine learning has made a significant impact on the biomedical field for liver disease prediction and diagnosis [4-6]. Machine learning offers a guarantee for improving the detection and prediction of disease that has been made an interest in the biomedical field and they also increase the objectivity of the decision-making process [16]. By using machine learning techniques medical problems can be easily solved and the cost of diagnosis will be reduced. In this study, the main aspect is to predict the results more efficiently and reduce the cost of diagnosis in the medical sector. Therefore, we used different classification techniques for the classification of patients have liver disease or not. Six machine learning techniques have been applied including LR, KNN, DT, SVM, NB, RF and the performance of these techniques were estimated on various perspectives such as accuracy, precision, recall, f-1 score. Moreover, the performance was compared using the receiver operative characteristic (ROC).

The remains of the paper are arranged as follows, chapter 2 presents the dataset details, data preprocessing and methodology. Chapter 3 describes the classification algorithms. Chapter 4 describes results and discussion including Measurement of Classification Techniques, Analysis of the Results and Performance Evolution. Finally, chapter 5 presents the conclusion section.

- *A. k. M. Sazzadur Rahman Rahman is currently pursuing master's degree program in Computer Science and Engineering at Daffodil International University, Bangladesh. E-mail: sohag933@gmail.com*
- *F. M. Javed Mehedi Shamrat is currently pursuing Bachelor's degree program in Software Engineering at Daffodil International University, Bangladesh. E-mail: javedmehedicom@gmail.com*
- *Zarrin Tasnim is currently pursuing Bachelor's degree program in Software Engineering at Daffodil International University, Bangladesh. E-mail: zarrint25@gmail.com*
- *Joy Roy is currently Studying in Software Engineering at Daffodil International University, Bangladesh. E-mail: joy35-1706@diu.edu.bd*
- *Syed Akhter Hossain, Professor and Head of Department of Computer Science and Engineering at Daffodil International University, Bangladesh. E-mail: aktarhossain@daffodilvarsity.edu.bd*

2 MATERIALS AND METHODOLOGY

2.1. Data Collection

In this experiment, we collect a dataset from the UCI Machine Learning Repository. In addition, the original dataset was collected from the northeast of Andhra Pradesh, India [7]. This dataset consists of 583 liver patient's data whereas 75.64% male patients and 24.36% are female patients. This dataset has contained 11 particular parameters whereas we choose 10 parameters for our further analysis and 1 parameter as a target class. Such as,

I. Age: Age of the patient
II. Gender: Gender of the Patients
III. TB: Total Bilirubin
IV. DB: Direct Bilirubin
V. Alkphos: Alkaline Phosphotase
VI. Sgpt: Alamine Aminotransferase
VII. Sgot: Asparatate Aminotransferase
VIII. TP: Total Proteins
IX. ALB: Albumin
X. AG Ratio: Albumin and Globulin Ratio
XI. Selector field used to split the data into two sets (labeled by the experts)

2.2. Data Preprocessing

In this study, we analyzed 583 liver patient's data whereas 416 samples are liver patient and 167 samples are non-liver patients. The ratio of total liver patients is presented in Fig. 1. Moreover, from the liver patient's dataset, (Fig. 2) 441 are male samples and 112 are female samples were taken for analysis.

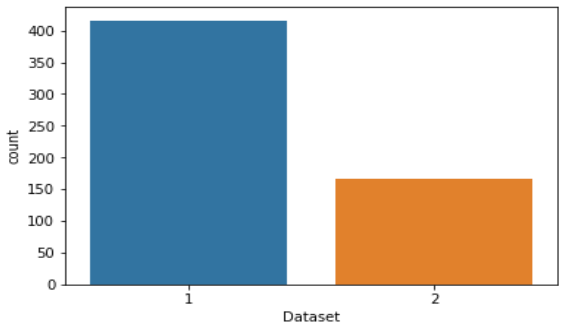


Fig. 1: *Count Plot shows the ratio of liver patients.*

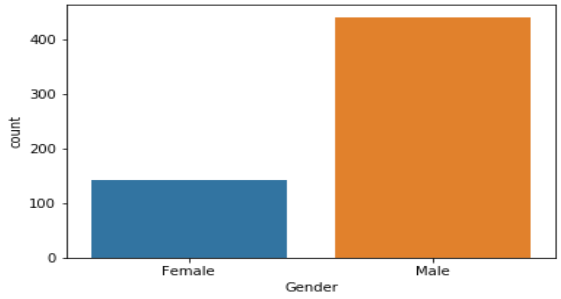


Fig. 2: *Count Plot shows the ratio of gender of liver patients.*

The heatmap is shown in Fig. 3 appear to have some correlated parameters. Some of these columns have a low correlation. Therefore, we omitted some of the features for better prediction of liver disease.

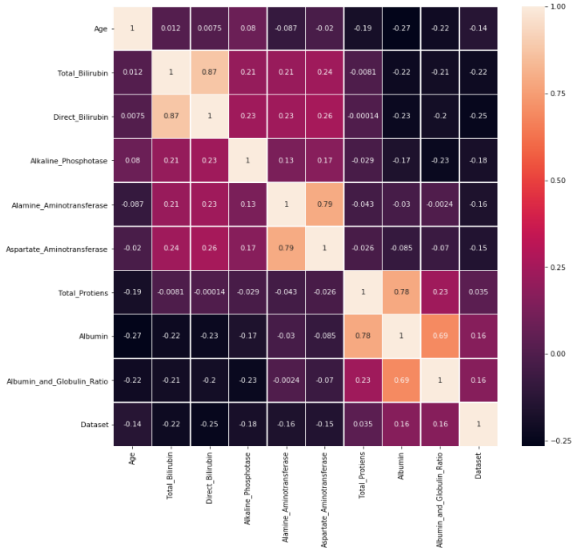


Fig. 3*. Heat map for checking correlated columns for the liver dataset.*

2.2. Tool and Language

In this study, we used the jupyter notebook as a tool and python 3.7 as a programming language.

3 DESCRIPTION OF THE CLASSIFICATION ALGORITHMS

3.1. Logistics Regression (LR)

Calculated Regression was for the most part utilized in natural research and applications in the mid-20th century [8]. Logistic regression can deal with any number of numerical as well as absolute factors. In addition, it introduces a discrete parallel item somewhere in the range of 0 and 1. Strategic Regression processes the connection between the element factors by surveying probabilities (p) utilizing an underlying logistic function. Regression equation given as,

$$p = \frac{1}{1+e^{-(b_0+b_1x_1+\cdots+b_nx_n)}} \quad (1)$$

3.2. Random Forest (RF)

Random forests or random decision forests are an ensemble learning technique for classification, regression and different assignments that works by developing a huge number of decision trees at training time and yielding the class that is the method of the classes (classification) or mean forecast (regression) of the individual trees. Random decision forests right for decision trees' propensity for overfitting to their training set. In the forest of trees has been the immediate connection between the combine trees and the outcome it can get. To get increasingly effective and precise predictions, random forest inserts an additional layer of irregularity to stowing [9].

3.3. Decision Tree (DT)

Decision Tree calculation has a place with the supervised learning algorithms [10]. In contrast to other supervised learning algorithms, a decision tree algorithm can be utilized for taking care of regression and classification issues as well. The general thought process of utilizing Decision Tree is to make a training model that can use to predict class or estimation of objective factors by taking in choice standards derived from earlier data (training data). In Fig. 4 we have shown a sample picture of decision trees.

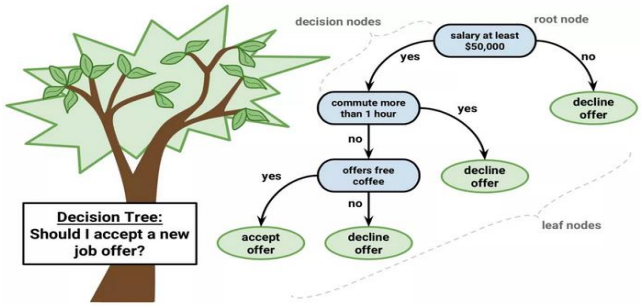


Fig. 4: *Sample of the process of Decision Trees.*

3.4. Support Vector Machine (SVM)

SVM is a supervised learning calculation. It can utilize for both grouping or relapse issues however generally it is utilized in characterization issues. SVM function admirably for some, human services issues and can comprehend both linear and non-linear issues. SVM grouping strategy which is an endeavor to pass a linearly separable hyperplane to order the dataset into two classes [11-12]. At long last, the model can without a doubt gauge the objective groups (labels) for new cases. For Classification type 1 of SVM, training involves the

minimization of the error function:

$$\frac{1}{2}w^T w + c\sum_{i=1}^{N}\zeta_i \quad (2)$$

In contrast to Classification SVM Type 1, the Classification SVM Type 2 model minimizes the error function:

$$\frac{1}{2}w^T w - vp + \frac{1}{N}\sum_{i=1}^{N}\zeta_i \quad (3)$$

3.5. K-Nearest Neighbors (KNN)

KNN is one of the most fundamental occasion-based classification algorithms in Machine Learning. In any case, the KNN takes a shot at the idea that examples are near fit in similar examples class [13]. A KNN sorts an example to the class that is most decided among K neighboring. K is a limitation for adjusting the classification algorithms [14].

3.6. Naive Bayes (NB)

Naive Bayes is one of the basic, best and ordinarily utilized, AI techniques. It is a probabilistic classifier that classifies utilizing the speculation of restrictive freedom with the pre-trained datasets [15]. From this time forward, Naive Bayes classifiers are procedures for finding the conventional arrangement of grouping issues, for example, spam identification, and furthermore all-around fit for medical issues. Bayes' Theorem finds the probability of an occasion occurring given the probability of another occasion that has just happened. Bayes' theorem is expressed mathematically as the following equation:

$$P(A/B) = \frac{P(B/A)*P(A)}{P(B)} \quad (4)$$

4 RESULT AND DISCUSSIONS

4.1. Measurement of Classification Techniques

In the work, we utilized some factual estimations that measure the test execution of various classification algorithms. The performance of the classification methods was assessed by various evaluation procedures, for example, accuracy, sensitivity, specificity, and precision and f1 measure. Consequently, the exhibition evaluation variables are determined by the confusion matrix. Here, True Positive (TP): The result of prediction correctly identifies that a patient has liver disease. False Positive (FP): The result of prediction incorrectly identifies that a patient has liver disease. True Negative (TN): The result of prediction correctly rejects that a patient has liver disease. False Negative (FN): The result of prediction incorrectly rejects that a patient has liver disease. The precision gives the contrast between sound and patient capacity ratio utilizing the prediction model. To discover the precision of classification is determined by the true positive, true negative, false positive and false negative.

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative})} \quad (5)$$

The affectability test gives the pace of effectively distinguishes the patient with their liver disease. It mainly demonstrates the positive instances of the test. It additionally is known as Recall and True Positive Rate (TPR).

$$\text{Sensitivity} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} \quad (6)$$

Particularity is showing the negative consequence of the disease. It gives the extent of the missing disease of the patients. It is otherwise called the True Negative Rate (TNR).

$$\text{Specificity} = \frac{\text{True Negative}}{(\text{False Positive} + \text{True Negative})}$$

Precision is otherwise called positive predictive value. It gives the proportion of an accurately predicted positive outcome by classifier algorithms.

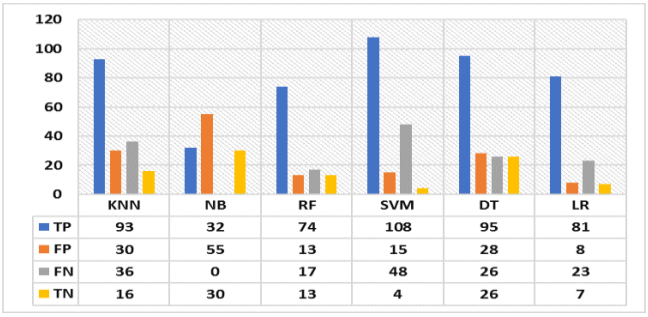
$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

F1 measures the precision of the model by a blend of accuracy and recall. It gives the proportion of both FP and FN of a model.

$$\text{F1} = \frac{2*(\text{Recall}*\text{Precisiion})}{(\text{Recall} + \text{Precision})} \quad (9)$$

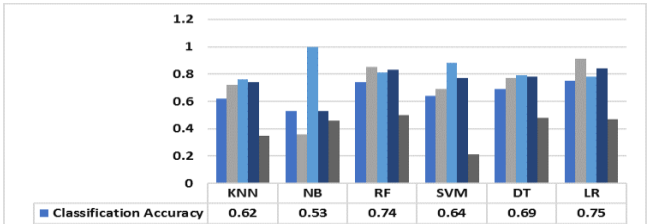
4.2. Analysis of the Result

In this experiment, we considered different analyses to examine the six-machine learning classifier for the classification of liver disease dataset. In terms of accuracy, LR achieved the highest accuracy of 75% and NB achieved the worst performance 53%. With respect to precision, LR achieved the highest score 91% and NB performs worst 36%. When considering the sensitivity, SVM achieved the highest value 88% and KNN obtained the worst 76%. Logistics Regression was also the best performer in terms of f1 measure 83% and NB obtained the worst performance 53%. When considering specificity DT achieved the highest value 48% and LR the lowest 47%. According to compare these measurement criteria LR classification technique is more effective than the other classifiers for predicting chronic liver disease. The confusion matrix of prediction results is shown in figure 5. The performance comparison of six supervised machine learning techniques is presented in figure 6.



	KNN	NB	RF	SVM	DT	LR
TP	93	32	74	108	95	81
FP	30	55	13	15	28	8
FN	36	0	17	48	26	23
TN	16	30	13	4	26	7

Fig. 5: *The Confusion Matrix of prediction results.*



	KNN	NB	RF	SVM	DT	LR
Classification Accuracy	0.62	0.53	0.74	0.64	0.69	0.75
Precision	0.72	0.36	0.85	0.69	0.77	0.91
Sensitivity	0.76	1	0.81	0.88	0.79	0.78
f-1	0.74	0.53	0.83	0.77	0.78	0.84
Specificity	0.35	0.46	0.5	0.21	0.48	0.47

Fig. 6*: The performance comparison of six supervised machine learning techniques.*

Figure 7 shows the Receiver Operating Characteristics (ROC). ROC is used to represent the performance of machine learning techniques which is based on the true positive rate (TPR) and false-positive rate (FPR) of these classification results. Moreover, SVM achieved the highest AUC (area under the curve) for ROC.

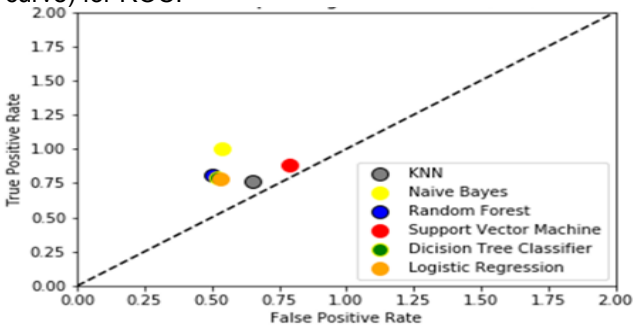


Fig. 7: *Receiver Operating Characteristics (ROC).*

5 CONCLUSION

The principal part of this work is to make an effective diagnosis system for chorionic liver infection patients utilizing six distinctive supervised machine learning classifiers. We researched all classifiers execution on patient's information parameters and the LR classifier gives the most elevated order exactness 75% dependent on F1 measure to predict the liver disease and NB gives the least precision 53%. From now on, the outperform classification procedure will give for the decision support system and diagnosis of chronic disease. The application will have the option to predict liver infection prior and advise the wellbeing condition. This application can be surprisingly gainful in low-salary nations where our absence of medicinal foundations and just as particular specialists. In our study, there are a few bearings for future work in this field. We just explored some popular supervised machine learning algorithms, more algorithms can be picked to assemble an increasingly precise model of liver disease prediction and performance can be progressively improved. Additionally, this work likewise ready to assume a significant role in health care research and just as restorative focuses to anticipate liver infection.

ACKNOWLEDGMENT

The authors are grateful and pleased to all the researchers in this research study.

REFERENCES

[1] K. Sumeet, J.J. Larson, B. Yawn, T.M. Therneau, W.R. Kim, Underestimation of liver-related mortality in the United States. Gastroenterology;(2013) 145:375–382, e371–372.

[2] A.A. Mokdad, A.D. Lopez, S. Shahraz, R. Lozano, A.H. Mokdad, J. Stanaway, et al, Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. BMC Med 2014; 12:145.

[3] Byass, Peter, The global burden of liver disease: a challenge for methods and for public health. BMC medicine 12.1 (2014); 159.

[4] L. A. Auxilia, Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease. 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE (2018).

[5] Hashem, M. Esraa, S. Mai, A study of support vector machine algorithm for liver disease diagnosis. American Journal of Intelligent Systems 4.1 (2014); 9-14.

[6] P. Sajda, "Machine learning for detection and diagnosis of disease." Annu. Rev. Biomed. Eng. 8 (2006); 537-565.

[7] UCI Machine Learning Repository. ILPD (Indian Liver Patient Dataset) Data Set. https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

[8] Logistic Regression, Retrieve from: HTTPS://WWW.SAEDSAYAD.COM/LOGISTIC_REGRESSION.HTM, LAST Accessed: 5 Octobor,2019

[9] L. Breiman, Random Forests. Machine Learning, 45(1), (2001); 5–32. https://doi.org/10.1023/A:1010933404324

[10] Decision Trees, Retrieve from: https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/, Last Accessed: 5 Octobor,2019

[11] Support vector machine, Retrieve from: http://www.statsoft.com/textbook/support-vector-machines, Last Accessed: 5 Octobor,2019

[12] V. Vapnik, I. Guyon, T. H.-M, Learn, and undefined 1995. Support vector machines. statweb.stanford.edu (1995).

[13] Zhang M, Zhou Z, "ML-KNN: A lazy learning approach to multi-label learning." Pattern recognition40.7:(2007); 2038-2048.

[14] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN Model-Based Approach in Classification (pp. 986–996). Springer, Berlin, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39964-3_62

[15] Naive Bayes, , Retrieve from: https://www.geeksforgeeks.org/naive-bayes-classifiers/, Last Accessed: 5 Octobor,2019

[16] S. M. Mahmud, et al. "Machine Learning Based Unified Framework for Diabetes Prediction." Proceedings of the 2018 International Conference on Big Data Engineering and Technology. ACM (2018).

[17] S. Safavian, D. Landgrebe, A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), (1991); 660-674.

[18] A. Chervonenkis, Early history of support vector machines. In Empirical Inference (pp. 13-20). Springer, Berlin, Heidelberg (2013).

[19] K.M. Leung, Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering (2007).