FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 5 No. 3, September, 2021, pp 1 - 8
DOI: https://doi.org/10.33003/fjs-2021-0503-515

HEPATITIS DISEASES PREDICTION USING MACHINE-LEARNING TECHNIQUES

YUSUF Aminat Bolatito*, AKANDE Oyelola

Department of Information and Communication Technology, Usmanu Danfodiyo University, Sokoto, Nigeria
*Corresponding Authors Email: aminbolly@gmail.com

ABSTRACT
The importance of research that contributes to the early diagnosis and management of lethal diseases is critical to society, and hepatitis is one of these killer diseases. Hepatitis is a life-threatening condition that develops when the liver becomes enlarged and injured. As a result, the primary goal of this article is to analyze the hepatitis dataset in order to accurately forecast outcome accuracy and dependability. Six machine learning classification methods: Support Vector Machines, Gaussian Naive Bayes, Logistic Regression, Decision Tree, K Nearest Neighbors, and Multiplayer Perceptron were tested on hepatitis dataset and a confusion matrix was plotted for each of the classification models. The accuracy, precision, and recall criteria were used to make the comparison. For each model, the accuracy was assessed using the root mean square value and mean absolute error. The selected algorithms, particularly the Multiplayer Perceptron (87%) and Logistic Regression (87%) algorithms, showed high accuracy rates. Furthermore, with a minimal root mean error of 0.35 and a minimal mean absolute error of 0.12 and 0.13, the two algorithms are the most dependable of all the methods.

Keywords: Hepatitis Disease, classification models, forecast.

INTRODUCTION

Hepatitis is a potentially fatal disease that occurs when the liver becomes inflamed and injured. It is a viral disease that has resulted in a high death rate worldwide (Nilashi, 2019). Hepatitis is transmitted by sewage pollution or direct contact with contaminated bodily fluids (Al-Thaqafy et.al, 2013). Viruses, bacteria, medicines, or drugs can also cause this condition(Trishna et al., 2019).Tattoos and piercing, drug abuse, sexual contact with an infected person, hemodialysis, blood transfusions are also methods by which an infected person can transmit this disease (Metwally et al., 2018) hepatitis may be acute or chronic(Metwally et al., 2018). Acute hepatitis causes intense and painful symptoms at the start of the disease, making it more painful for patients, but it only lasts a month or two (Trishna et al., 2019).
Consequently, there is only minor liver cell disruption and no effect on immune system function. Chronic hepatitis is a form of hepatitis that lasts more than six months and leads to cirrhosis, a condition in which the parenchymal cells of the liver are damaged (Metwally et al., 2018). Hepatitis A, B, C, D and E are 5 distinct forms of hepatitis(Ahmad et al., 2019). Hepatitis A and E are acute hepatitis, while Hepatitis B, C, and D are chronic hepatitis. Despite continuing studies into a treatment for hepatitis C, there is currently no available vaccine for the disease(Bhargav & Kumari, 2018). Early detection, as well as proper diagnosis and treatment, can cure the disease (Yarasuri et al., 2019). Also, health workers are most at risk with hepatitis disease (Polat & Günes, 2006). The cause for this is that the diagnosis of hepatitis disease is mostly by routine blood tests, which exposes medical personnel to associated risks during diagnosis. Hepatitis medical diagnosis is difficult since a specialist must weigh several aspects before performing the disease diagnosis process (Nilashi, 2019). As a result, this condition necessitates the creation of automated and reliable diagnostic systems that can aid in the identification of hepatitis for physician decision-making.
Machine learning is a valuable technique that clinicians can use in this instance. Machine Learning (ML) is a technique for teaching a system to learn by finding patterns and associations in captured data using various algorithms (AtifKhan et.al, 2012). As a result, ML allows for predicting and diagnosing any illness, taking into account two essential factors: parameter collection and the method used to analyze these parameters. This study compares six related ML algorithms that are beneficial to diagnose hepatitis. Support Vector Machines (SVM), Gaussian Naive Bayes, Logistic Regression, Decision Tree, K Nearest Neighbors (KNN), and Multiplayer Perceptron (MLP) are the algorithms considered. The main objective of this paper is to analyze hepatitis dataset data and correctly predict the outcome in each dataset using the six ML methods. The study makes substantial contributions in the following areas:

a) To improve the classification accuracy and reliability for predicting hepatitis diseases.
b) To make a comparison of six classification algorithms for ML on the data set for hepatitis.
c) Determine the most effective ML algorithm for predicting hepatitis.

REVIEW OF RELATED LITERATURE

ML methods have been used in a variety of experiments to diagnose and predict hepatitis disease. Neshat M. et al. 2012 used two approaches to diagnosing hepatitis: Particle Swarm Optimization (PSO) and Computer-Based Reasoning (CBR). The proposed (CBR-PSO) method results were compared to

five other classifications, and the proposed method (CBR-PSO) showed better results with an accuracy of 93.25% for diagnosing hepatitis disease.

Karthikeyan & Thangaraju 2013 deal evaluate classification techniques, namely, "Bayes.NaiveBayes, Bayes.BayesNet, Bayes. NaiveBayesUpdatable, J48, Random Forest, and Multi-Layer Perceptron." The model's findings are precise and timely. Finally, the conclusion was that Naive Bayes outperforms other classification strategies for hepatitis patients, with an accuracy rate of 84%. To diagnose hepatitis B virus disease, Mahesh et al. (2014) suggested a Generalized Regression Neural Network-based expert system. The system divides each patient into infected and non-infected categories. If infected, what is the severity of the infection in terms of intensity rate? Metwally et al. (2018) propose an Artificial Neural Network (ANN)-based technique for diagnosing hepatitis virus; they concluded by demonstrating the capacity of the ANN-based for diagnosing hepatitis patients by understanding the patient's chance of survival or death. To find a reliable predictor for hepatitis, Yarasuri et al. 2019 compared SVM, ANN, and KNN. ANN was the most accurate, with a 96% forecast precision and the lowest Mean Square Error (MSE). Bhargav et al. 2018 tried to classify whether a person with hepatitis C will survive or die using ML algorithms. Decision Trees, Naive Bayes Classifier, Logistic Regression, and Linear Support Vector Machine were the four data algorithms that were compared and tabulated. The Decision Tree Algorithm has the highest accuracy of 82.05 %, while the Logistic Regression algorithm has an accuracy of 87.17 %. These algorithms can be used to establish whether a person living with hepatitis C lives or die. Nilashi et al. 2019 proposed a practical approach for diagnosing hepatitis using ensemble learning. For data dimensionality reduction, they used "Nonlinear Iterative Partial Least Squares, Self-Organizing Map technique for clustering, and Neuro-Fuzzy Inference System ensembles" for predicting hepatitis disease. Decision trees were used to choose the essential features in the experimental dataset. The method was tested with real-world evidence, and the findings compared to previous research. According to the researchers, the approach outperforms the Neural Network, ANFIS, KNN, and SVM on the dataset.

MATERIAL AND METHODS

Data collection

The hepatitis dataset was retrieved from the University of California, Irvine (UCI) Repository. There are 155 samples in the database, and it has 20 attributes, together with the class label attribute. To diagnose and identify hepatitis, Machine Learning Algorithms were applied to this dataset. The specifics can be found in table 1 below. The dataset was trained and tested using six machine learning algorithms.

Table 1: Attributes in Dataset

ATTRIBUTES	VALUE
Class	Integer
Age	Integer
Sex	Integer
Steroid	Integer
Antivirals	Integer
Fatigue	Integer
Malaise	Integer
Anorexia	Integer
Liver_big	Integer
Liver_firm	Integer
Spleen_palable	Integer
Spiders	Integer
Ascites	Integer
Varices	Integer
Bilirubin	Float
Alk_phosphate	Float
Sgot	Float
Albumin	Float
Protime	Float
Histology	Integer

A comparison was made based on the tools' accuracy, precision, and recall. Loading data, attributing and preprocessing data, data classification, implementation of ML methodology, and disease prediction are the critical processes involved in this study. Figure 2 depicts a model method for hepatitis diagnosis, with phases of the procedure explained in the sections below:

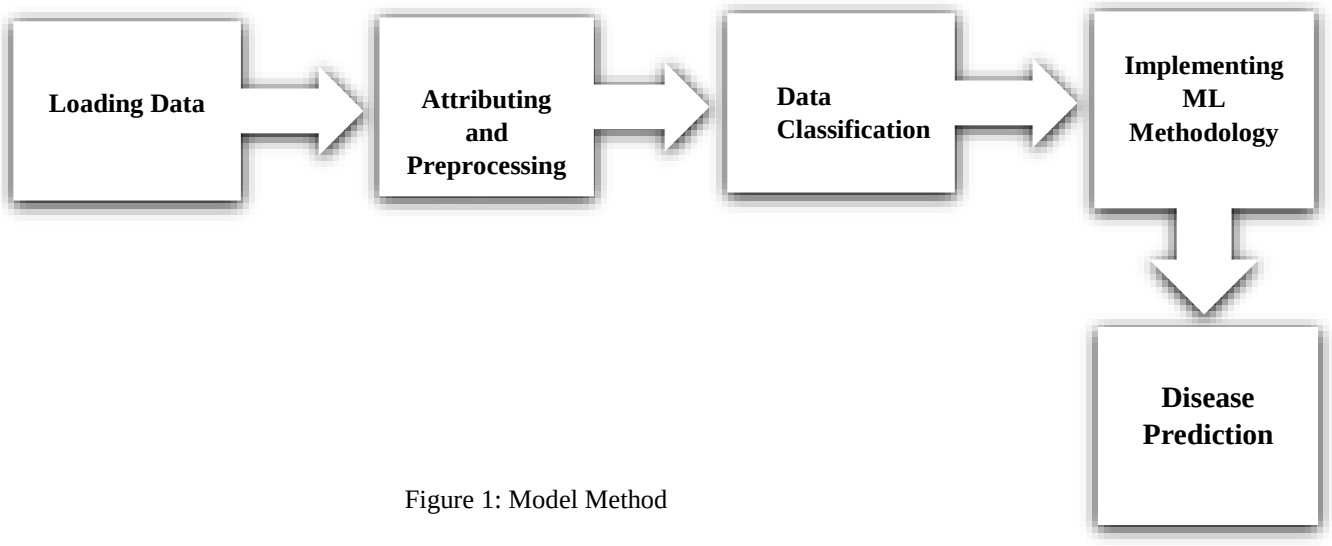

Figure 1: Model Method

Loading Data
The data came from the UCI register, which has 155 instances with 20 different attributes. Because machine learning learns from samples, the model requires smoothing large amount of data to produce results. Data imputation is performed on the available dataset to obtain a satisfactory amount of data.

Attributing and Preprocessing of Data
Missing data was resolved in order to obtain adequate data for preparation, validation, and testing by imputing the omitted data and substituting an individual global constant for each of the lost values. In this hepatitis data set, 75 of the 155 instances have missing values. If missing data in any field is not properly treated, it can result in error prediction and degrade the performance quality.

Classifying the Data.

The data for this analysis were divided using stratified splitting. Data were segmented into 10-fold cross-validation training and testing data sets before modeling. The scores collected for each fold are averaged out and utilized as a single score after a 10-fold cross-validation repeat. This means that the model is trained with 90% of the data for each fold and evaluated against the remaining 10%. This style's cross-validation avoids the bias of training the model primarily on negative or positive data.

Using ML tool to diagnose the disease
Training, forecasting, and testing are the basic three steps of the machine learning implementation. The classifier algorithm creates the model based on the training dataset during the training phase. The trained model then use to predict the hepatitis disease. The testing data set was use to validate the forecast's performance by determining the accuracy, precision, and recall of the prediction. The techniques used in this analysis were SVM, Gaussian Naive Bayes, Logistic Regression, Decision Tree, KNN, and MLP classifiers:

SVM is a widely used and practical method for dealing with data classification, interpretation, and prediction issues (Saangyong et.al, 2009). SVM is used to map the input variable to n-dimensional function space. For classified training outcomes, it generates a hyperplane that divides the function space by their class, preventing overfitting (Xiao & Leedham, 2002).

KNN is one of the most fundamental classification algorithms. This algorithm prioritizes the best k nearest neighbors [It is a common machine learning algorithm for datasets due to its ability to select neighbors. We will not get the right results if we choose the lower and upper values of k. As a result, in order to obtain a particular result, we select an optimal k value for the algorithm.

Gaussian Naive Bayes implies the presence of one function in a class has no effect on the existence of any other feature. [The idea behind the term "naive" is that it reduces the difficulty of computation to a general probability multiplication. The primary advantage of GNB is its speed, as it is a simple algorithm in comparison to other classification algorithms. Due to its simplicity, this GNB algorithm is capable of efficiently processing datasets with a large number of dimensions

MLP is a form of feed-forward artificial neural network that maps input data datasets to a set of suitable outputs. A MLP is made up of multiple layers of nodes in a directed graph, each layer being completely connected to the previous one. Excluding the input nodes, a unit node represents a processing unit with a nonlinear activation function. In the MLP classification dataset, back propagation is a supervised learning approach that was employed to train the network. MLP is a version of the normal linear perceptron which can classify data in datasets that are not linearly separable.

Logistic Regression is a computational method for evaluating a data set in which the result is calculated by one or more independent variables. The aim of logistic regression is to determine the optimal model that describes the relationship

between a collection of predictor variables and an observed dichotomous feature.

Decision Tree is the most frequently used classification algorithms are decision tree algorithms (Karthikeyan & Thangaraju, 2013)(Twa eta.al,2005). A decision tree is a straightforward modeling technique that employs tree structure to construct classification or regression models. It generates a related decision tree incrementally as a data set is subdivided into smaller categories. Consequently, a tree with leaf and decision nodes is formed. A decision node with more than two branches is referred to as a leaf node, and the upmost decision node in a tree is referred to as the root node, which represents the best predictor (Soofi, & Awan, 2017).

Classification performance measures

The following are the metrics used to evaluate the classification mentioned above.

a) Precision is measured as the ratio of the number of correctly identified positive samples to the overall number of positive samples. The precision of the model is measured as shown in Equation 1:

$$Precision = \frac{True_{\ positive}}{True_{\ positive} + False_{positive}} \qquad (Equation\ 1)$$

b) Accuracy is a metric that summarizes a model's performance across all classes. It is advantageous when all classes are equally relevant. It is calculated as the ratio of correct predictions to total predictions as shown in Equation 2.

$$Accuracy = \frac{True_{\ positive} + True_{\ negative}}{True_{\ positive} + True_{\ negative} + False_{positive} + False_{negative}} \qquad (Equation\ 2)$$

c) Recall is determined by dividing the total number of positive samples by the number of positive samples correctly identified as positive. The recall is a metric that tests how well a model can detect positive samples. The higher the recall, the greater the number of positive samples found. This is shown in Equation 3

$$Recall\ = \frac{True_{\ positive}}{(True_{\ positive} + False_{negative})} \qquad (Equation\ 3)$$

d) The Mean Absolute Error (MAE) is a statistic that assesses the average size of mistakes in a set of forecasts without taking into account their direction. It is the average of the absolute differences between forecast and actual observation over the test sample, where all individual deviations are given equal weight as shown in Equation 4.

$$MAE = {}^{1}\!/_{n} \sum_{j=1}^{n} |y_j - \hat{y}_j| \qquad (Equation\ 4)$$

It is the measure of the difference between the two continuous variables. The MAE is the average vertical distance between each actual value and the line that best matches the data. MAE is also the average horizontal distance between each data point and the best matching line.

e) Root Mean Square Error (RMSE) is defined as the square root of the average squared distance between the actual score and the predicted score as shown in Equation 5:

$$RMSE = \sqrt{{}^{1}\!/_{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (Equation\ 5)$$

The true score for the i[th] data point is denoted by y_i, and the predicted value is denoted by $\hat{y}_i$.

f) Area Under Curve (AUC) is the likelihood that the classifier would score a randomly chosen positive example higher than a randomly chosen negative example. The AUC is based on a plot of the false positive rate against the true positive rate and ranges between 0 and 1 which are defined as shown as shown in Equation 6 and 7.

$$TPR\ (sensitivity\ or\ recall) = \frac{True_{\ positive}}{(True_{\ positive} + False_{negative})} \qquad (Equation\ 6)$$

$$FPR\ (1 - specificity) = \frac{False_{positive}}{(True_{\ negative} + False_{positive})} \qquad (Equation\ 7)$$

RESULTS AND DISCUSSION

The classification techniques was implemented with python. A number of health-related attributes are included in the dataset, as well as the class label, which corresponds to a patient's hepatitis status. The data was separated into two categories: training data and validating data. Using the training data given, we trained the six models; SVM, Gaussian Naive Bayes, Logistic Regression, Decision Tree, KNN, and MLP. The models were tested using validating data, and a confusion matrix was plotted for each of the models. The Table 2 depict the confusion matrix of the SVM, Gaussian Naive Bayes, Logistic Regression, Decision Tree, KNN, and MLP on the hepatitis dataset respectively. The confusion matrix for all the models are as shown below:

Table 2: Confusion Matrix for all the Models

		Predicted = Yes	**Predicted = No**
Support Vector Machine classifier	Does not have	1	28
	Has	3	7
Gaussian Naive Bayes classifier	Does not have	10	0
	Has	11	18
Logistic Regression classifier	Does not have	5	5
	Has	0	29
Decision Tree classifier	Does not have	8	2
	Has	4	25
K Nearest Neighbors classifier	Does not have	3	7
	Has	0	29
Multiplayer Perceptron classifier	Does not have	6	4
	Has	0	29

The classifier's accuracy in making correct predictions is measured using the uncertainty matrix. The count value of the uncertainty matrix represents the number of accurate and inaccurate classifier predictions. The upper row of the uncertainty matrix lists predicted positive events with true positives, while the lower row lists no events with true negatives. The diagonal elements denote the number of projected target classes that are equal to the actual target class. The misclassified or wrongly predicted targets class belongs to the off-diagonal elements.

From the matrix, the true positives, true negatives, false positives, false negatives along with the true positive rate and false positive rate were utilized to calculate the recall, precision, accuracy and AUC were calculated by implementing specified modules. The recall, precision and accuracy give the performance of the various classification algorithms when applied on the Hepatitis dataset are display in the chart figure 2 and 3 of the ROC graph as shown.

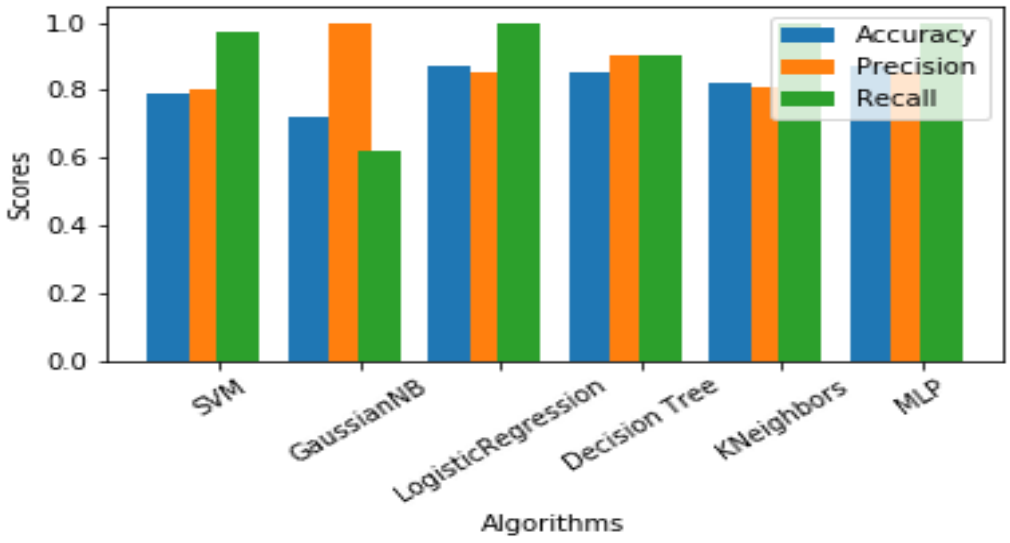

Figure 2: Performance measure comparison

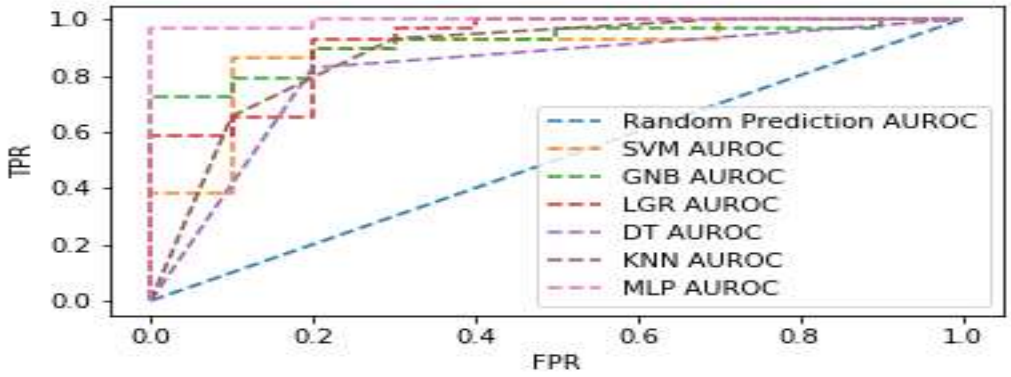

Figure 3: ROC Plot

The following conclusion may be drawn from the findings: the MLP and Logistic Regression algorithms have the highest accuracy of 87 percent, followed by the Decision Tree Algorithm with an accuracy of 85 percent. The KNN comes next, with an ideal accuracy of 82 percent, while the Gaussian Nave Bayes Algorithm comes in third, with an ideal accuracy of 72 percent. The ROC curve reveals that the AUC for MLP model beat all other models on the validation data set, with substantially higher and steady performance.

The following figure 4 and Figure 5 describe a Mean absolute error analysis and root mean square error for all the models.

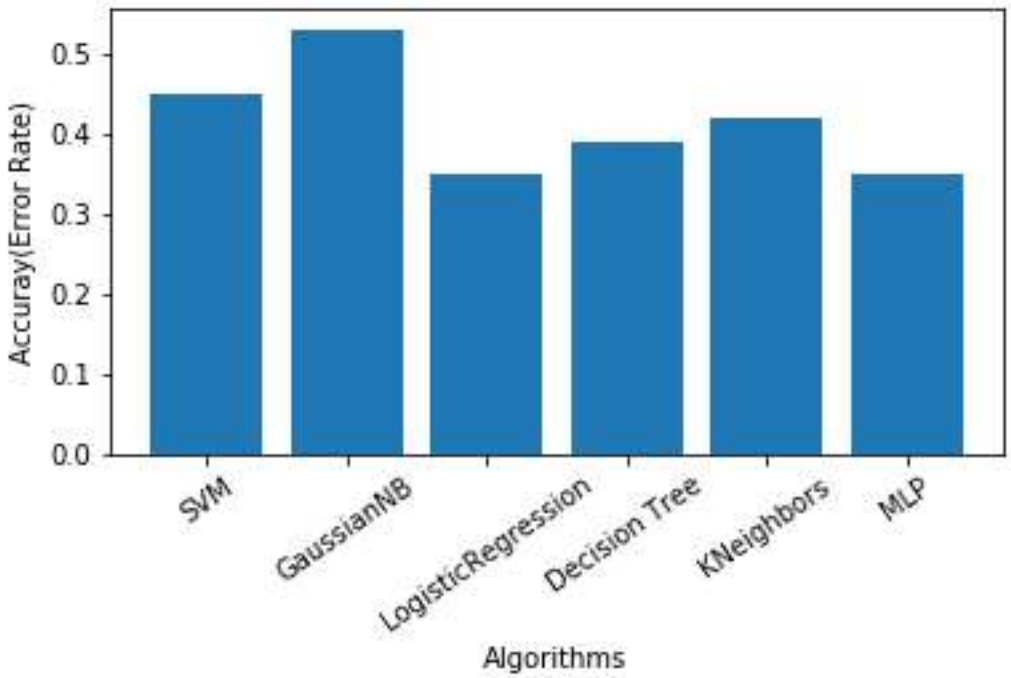

Figure 4: Root Mean Squared Error

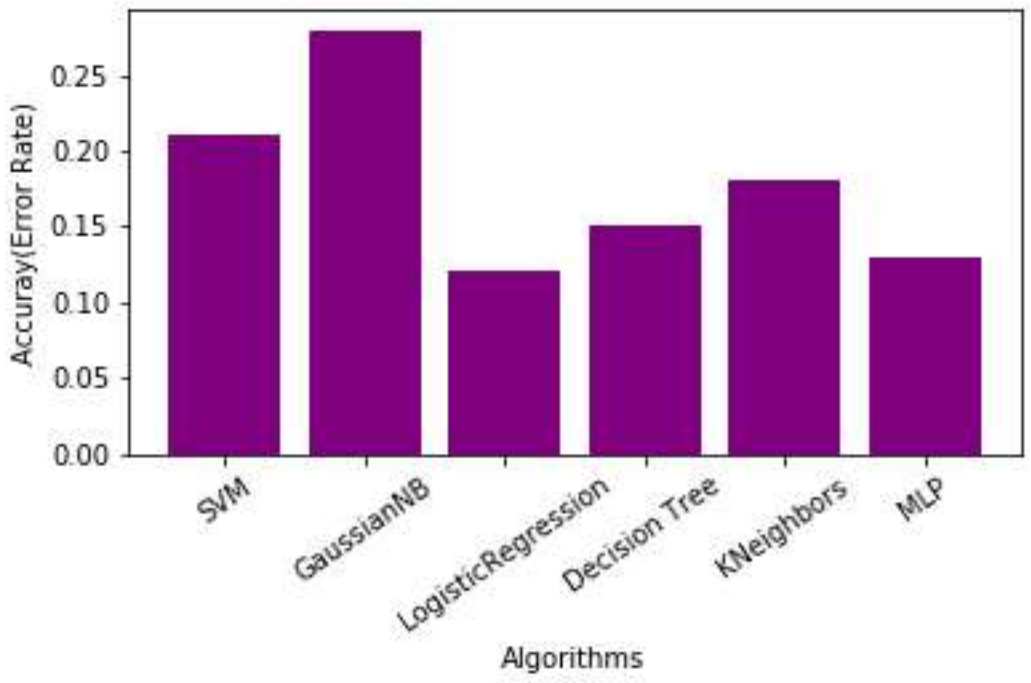

Figure 5: Mean Absolute Error

The lowest mean absolute error rate of 0.13 was achieved with MLP and Logistic regression, and the root mean square error rate of 0.35. The lower the MAE and RMSE for a given model, the more closely the model can predict the actual values.

Figure 4 below shows the comparison graph of the MSE for the ML tools. To further verify the accuracy of the models, the Mean Absolute Error (MAE) for each one of the models was determined. The MAE states the average difference between the actual data value and the value predicted by the models. The lower the MAE for a given model, the more closely the model can predict the actual values. The RMSE and MAE were also used to validate the algorithms' predictability.

CONCLUSION

In this paper, evaluation of performance using classification performance measures was carried out on selected Machine Learning (ML) algorithms. Accuracy, precision, and recall were used to determine if an individual has hepatitis or not from the various independent attributes. According to the results of this analysis, the selected algorithms demonstrated some good accuracy percentages, especially MLP (87%), Logistic Regression (87%), Decision Tree (85%) and KNN (82%) algorithm. These algorithms can be applied for determining whether or not hepatitis is present in a person. MLP, on the other hand, is the most dependable, with a Mean Absolute Error 0f 0.13 and a minimum Root Mean Square Error of 0.35.

In the future, the data set will be used to build the model will be increased, and this will result in more unique rules and better accuracy. Different weighing techniques are suggested to enhance the accuracy. Also, other classification methods can be employed to extend the research further.

REFERENCES

Al-Thaqafy MS, Balkhy HH, Memish Z, Makhdom YM, Ibrahim A, Al-Amri A,Al-Thaqafi A(2012). Improvement of the low knowledge, attitude and practice of hepatitis B virus infection among Saudi National Guard personnel after educational intervention. *BMC Res Notes*,5(1):597.

Al-Thaqafy MS, Balkhy HH, Memish Z, Makhdom YM, Ibrahim A, Al-Amri A,et al(2013). Hepatitis B virus among Saudi National guard personnel: seroprevalence and risk of exposure. *Journal of Infection and Public Health* ,6(4):237–45.

Atif Khan, John A. Doucette, Robin C.(2012), "Integrating Machine Learning into a Medical Decision Support System to Address the Problem of Missing Patient Data," *IEEE DOI* 10.1109/ICMLA.2012.82.

Bhargav, K. S., & Kumari, T. D. (2018). *Application of Machine Learning Classification Algorithms on Hepatitis Dataset*. *13*(16), 6.

Brownlee J. (2016). How to Implement Logistic Regression with Stochastic Gradient Descent from Scratch with Python - Machine Learning Mastery.

Karthikeyan T., Thangaraju P (2013). Analysis of classification algorithms applied to hepatitis patients. *International Journal of Computing*. 15,62.

Kotsiantis S. B.(2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica 31*, 249-268.

Mahesh, C., Kiruthika, K., Professors, M. D., & Rr, T.D (2014). Diagnosing Hepatitis B Using Artificial Neural Network Based Expert System. 7. ISBN No.978-1-4799-3834-6/14 IEEE

Metwally, N. F., AbuSharekh, E. K., & Abu-Naser, S. S. (2018). *Diagnosis of Hepatitis Virus Using Artificial Neural Network*. *2*(11), 7.

Nilashi, M. (2019). A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. *Journal of Infection and Public Health*, 8.

Polat K, Günes (2006), S. Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation. *Digit Signal Process*;16(6):889–901.

Saangyong U, Dong-Hoi Kim, Young-Woong Ko, Sungwon Cho, Jaeyoun Cheong, and Jin Kim (2009).A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis. *Expert Systems* 26, (1) 60-69.

Soofi, A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, *13*, 459–465. https://doi.org/10.6000/1927-5129.2017.13.76

Trishna, T. I., Emon, S. U., & Ema, R. R. (2019). *Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier*. 7.

Twa MD, Parthasarathy S, Roberts C, Mahmoud AM,Raasch TW, Bullimore MA(2005). Automated decision tree classification of corneal shape. *Optometry and vision science: official publication of the American Academy of Optometry*; 82: 1038.
https://doi.org/10.1097/01.opx.0000192350.01045.6f

Wu JC, Chen TZ, Huang YS, Yen FS, Ting LT, Sheng WY(1995). Natural history of hepatitis D viral superinfection: significance of viremia detected by polymerase chain reaction. Gastroenterology;108(3):796–802.

Xiao X, Leedham G(2002). Signature verification using a modified Bayesian network. Pattern Recognition; 35: 983-995.https://doi.org/10.1016/S0031-3203(01)00088-7

Yarasuri, V. K., Indukuri, G. K., & Nair, A. K. (2019). Prediction of Hepatitis Disease Using Machine Learning Technique. *2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 265–269. https://doi.org/10.1109/I-SMAC47947.2019.9032585

Zhang D, Nunamaker J.F. (2003). Powering e-learning in the new millennium: an overview of e-learning and enabling technology. *Information Systems Frontiers*; 5: 207-218. https://doi.org/10.1023/A:1022609809036