IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 8, Issue.12, December 2020 E-ISSN: 2347-2693

Condition Based Disease Detection Using Machine-Learning Algorithms Based Prediction

Surender Singh[1*], Jyoti[2]

[1]Department of Information Technology, MSIT, Delhi, India
[2]IGNOU, Delhi, India

**Corresponding Author: surenderbhanwala@msit.in*

DOI: https://doi.org/10.26438/ijcse/v8i12.9497 | Available online at: www.ijcseonline.org

Received: 20/Dec/2020, Accepted: 22/Dec/2020, Published: 31/Dec/2020

***Abstract*-** Diseases are increasing rapidly now a days due to number of reasons. It will be very helpful to cure that disease if we predict occurrences of diseases in the early stages. Even though doctors and health centers collect data daily but most of them are not using machine learning and pattern matching techniques to extract the knowledge that can be very useful in prediction. We have chosen dataset of liver diseases to evaluate prediction algorithms in an effort to reduce burden on doctors. In our work, we have trained eight models Logistic Regression, Random Forest, XGBoost, KNN, Decision Trees, SVC, Gradient Boosting and Neural Network. The analysis compare all these models and choose the best model.

***Keywords*-** Data Mining, Classification, Decision Tree, Liver Disease

I. INTRODUCTION

The liver plays a vital role in many bodily functions from protein production and blood clotting to cholesterol, glucose (sugar), and iron metabolism [1]. It has a range of functions, including removing toxins from the body, and is crucial to survival. The loss of those functions can cause major damage to the body. The definition of acute liver disease is based on duration, with the history of the disease does not exceed six months. Acute viral hepatitis and drug reactions account for the majority of cases of acute liver disease. Symptoms of liver disease can vary, but they often include swelling of the abdomen and

legs, bruising easily, changes in the color of your stool and urine, and jaundice, or yellowing of the skin and eyes [2]. Sometimes there are no symptoms. Diagnosis of liver disease at a preliminary stage is important for better treatment. It is a very challenging task for medical researchers to predict the disease in the early stages owing to subtle symptoms. Often the symptoms become apparent when it is too late. To overcome this issue, we aims to improve liver disease diagnosis using machine learning approaches and compare the classification algorithms based on their performance factors. Eight algorithms such as Logistic Regression, Random Forest, XGBoost, KNN, Decision Trees, SVC, Gradient Boosting and Neural Network are taken to analyze the liver disease dataset and compared their performance.

II. RELATED WORK

Machine learning models is very useful in various areas especially in medical field to solve diagnosis of medical problems in early stages. These models can be used to reduce the burdens on medical practitioner [3]. Aneesh kumar [4] used an approach to effective classification of liver and non-liver disease dataset. Some pre-processing method is used to clean the data and used two algorithms (C4.5 and Naive Bayes) in his research. In research of Gunasundari [5], neural networks and Genetic algorithm gives very good result for liver disease disorder diagnosis. CART and C4.5 both algorithms also gives satisfactory result [6]. Feature selection plays a vital role in improving the accuracy. Various methods have been developed, each having unique properties and selecting different features. Bendi [7], proposed a Modified Rotation Forest algorithm to calculate the accuracy of the liver classification techniques in UCI liver dataset using the combo of feature selection technique and selected classification technique algorithm. Liver abscess is the commonest factor of hepatomegaly and it is due to amoebiasis which is followed by fatty liver, congestive cardiac failure, hepatocellular carcinoma, and viral hepatitis [8]. Malathi [9] proposed hybrid classifier algorithm which is better in predicting liver diseases.

III. MACHINE LEARNING MODELS

Eight machine learning models: Logistic Regression, Random Forest, XGBoost, KNN, Decision Trees, SVC, Gradient Boosting and Neural Network is used for analyzing the liver disease dataset. (a) Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. (b) Random forests or random decision forests is an ensemble learning method or classification, regression and other tasks that operates by constructing a multitude of decision

trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. (c) XGBoost stands for extreme gradient boosting. The implementation of the algorithm was engineered for efficiency of compute time and memory resources. A design goal was to make the best use of available resources to train the model. (d) k-NN is a type of instance-based learning where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. (e) Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. (f) Support Vector Machine (SVM) is a type of supervised machine learning model used for regression, classification, and outlier detection. (g) Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. (h) Artificial Neural Network (ANN) is computational algorithm used to learn the behavior of a given system and subsequently to simulate and predict this behavior. They have great capacity in predictive modeling and provide a computationally efficient way of determining an empirical possibly nonlinear relationship between a number of inputs and outputs. ANNs are constructed in interconnected layers to one or more hidden layers where the realistic processing is the performance through weighted connections. Each neuron in the hidden layer joins to all neurons in the output layer. The result of the processing are acquired from the output layer.

IV. METHODOLOGY

4.1 Dataset and Pre-processing

The liver data set (https://archive.ics.uci.edu/ml/datasets/ILPD) contains 416 liver patient records and 167 non liver patient records. This contains 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos (Fig. 1).

Attribute Name	Possible datatype / value
Age of the patient	Numeric
Gender of the patient	Numeric
Total Bilirubin	Numeric
Direct Bilirubin	Numeric
Alkaline Phosphatase	Numeric
Alanine Aminotransferase	Numeric
Aspartate Aminotransferase	Numeric
Total Proteins	Numeric
Albumin	Numeric
Albumin and Globulin Ratio	Numeric
Class	Nominal

Fig 1. Attribute Datatypes

4.2 Data Cleaning and Outliers Removal

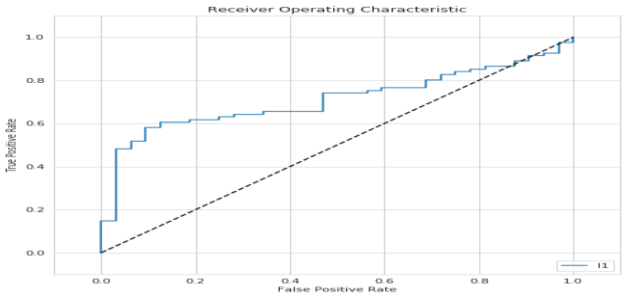
Data cleansing or data cleaning is the process of identifying incomplete, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty data. An outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses. We have used Weka (open tool) for data cleaning and outlier removal.

4.2 Accuracy Metrics

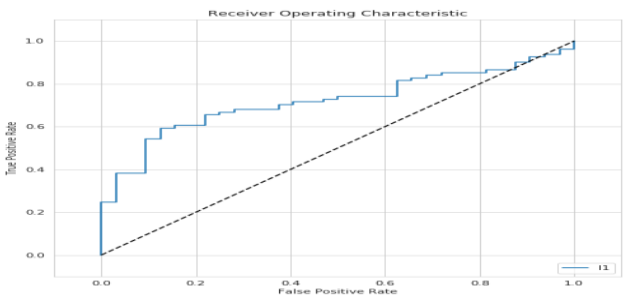
A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Sensitivity, accuracy, precision and specificity are some statistics of confusion metrics. Here, we used the receiver operating characteristic curve (ROC) for accuracy. ROC plots the false positive rate and true positive rate at different thresholds. ROC curves are judged visually by how close they are to the upper left-hand corner.

V. EXPERIMENTAL RESULTS

The study employed eight algorithms; Logistic Regression, Random Forest, XGBoost, KNN, Decision Trees, SVC, Gradient Boosting and Neural Network to predict the liver disease at an earlier stage. These techniques were evaluated and their performance was compared based on ROC (Fig. 3).



(i)



(ii)

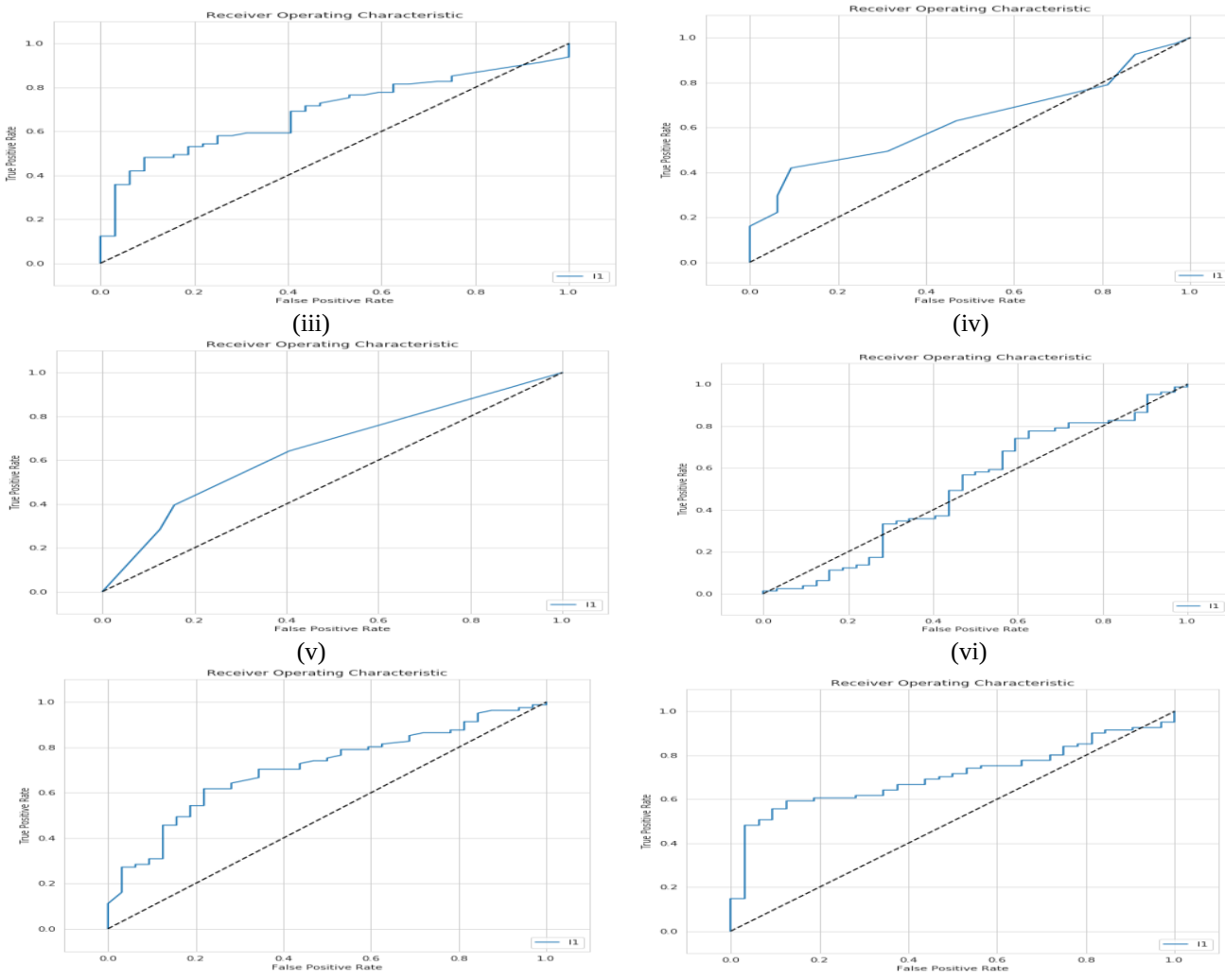


Fig. 2. ROC for (i) Logistic Regression, (ii) Random Forest, (iii) XGBoost, (iv) KNN, (v) Decision Trees, (vi) SVC, (vii) Gradient Boosting and (viii) Neural Network

From the analysis, XGBoost outperforms well than other algorithms and its achieved accuracy is 81% (Fig 3).

S.No.	Model	Cross-Validation	Accuracy %
1	Logistic Regression	2	71
2	Random Forest	3	72
3	XGBoost	2	81
4	k-Nearest Neighbors	4	62
5	Decision Tree	3	66
6	Support Vector Classifier	3	66
7	Gradient boosting	2	70
8	Neural Network	-	69

Fig. 3 Accuracy comparison for all models

VI. CONCLUSION AND FUTURE SCOPE

Eight machine learning models are evaluated and compared based on their performance. From the analysis, XGBoost outperforms well than other algorithms and its achieved accuracy is 81%. The application of algorithms in predicting liver disease will benefit in managing the health of individuals. However, in future, we will collect the very recent data from various regions across the world. We will also change the architecture and parameter setting as well as add more accuracy measures metrics.

REFERENCES

[1] Nahar, N. and Ara, F. "Liver disease prediction by using different decision tree techniques". *Int. J. Data Min. Knowl. Manag. Process*, ***8*(2), pp.01-09, 2018.**

[2] D. Sindhuja and R. J. Priyadarsini, "A survey on classification techniques in data mining for analyzing liver disease disorder", International Journal of Computer Science and Mobile Computing, **Vol.5, no.5, pp. 483-488, 2016.**

[3] B. V. Ramana, M. R. P. Babu and N.B. Venkaeswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", International Journal of Database Management Systems (IJDMS), **Vol.3, no.2, pp. 101-114, 2011.**

[4] A.S.Aneeshkumar and C.J. Venkateswaran, "Estimating the Surveillance of Liver Disorder using Classification Algorithms",

International Journal of Computer Applications (0975 –8887) , **Vol. 57, no. 6, pp. 39-42, 2012.**

[5] G. Selvara and S. Janakiraman, "A Study of Textural Analysis Methods for the Diagnosis of Liver Disease from Abdominal Computed Tomography", International Journal of Computer Applications (0975-8887), **Vol. 74, no.11, PP.7-13, 2013.**

[6] H. Sug, " Improving the Prediction Accuracy of Liver Disorder Disease with Oversampling", Applied Mathematics in Electrical and Computer Engineering, American-MATH 12/CEA12 proceedings of the 6th Applications and proceedings on the 2012 American Conference on Appied Mathematics, **PP. 331-335, 2012.**

[7] B. V. Ramanaland and M.S. P. Babu, "Liver Classification Using Modified Rotation Forest", International Journal of Engineering Research and Development ISSN: 2278-067X, **Vol. 1, no. 6, PP.17-24. 2012.**

[8] C.K. Ghosk, F. Islam, E. Ahmed, D.K. Ghosh, A. Haque and Q.K. Islam, "Etiological and clinical patterns of Isolated Hepatomegaly" Journal of Hepato-Gastroenterology, **vol.2, no. 1, PP. 1-4.**

[9] R. Malathi, S. Ravichandran, "Dual Threshold Based Classification Technique (DTBCT) For Assessing Liver Abnormalities from Medical Images," *International Journal of Computer Sciences and Engineering*, **Vol.7, Issue.5, pp.1436-1439, 2019.**

