Machine Learning for Liver Cirrhosis Disease Detection

Sapna Kharche, Akash Rathod, Himanshu Rai

Department of Data Science, G. H. Raisoni College of Engineering, Nagpur-440016, India

Abstract

Patients with liver diseases have been steadily increasing over the last few years as a result of excessive alcohol consumption, inhaling hazardous chemicals, and consuming contaminated food, drinks, pickles, and drugs. This dataset of 417 patient blood reports is used to assess the accuracy of various machine learning algorithms in order to reduce the burden on doctors and diagnose early liver cirrhosis disease.

Keywords: Liver Cirrhosis Disease, Machine learning algorithm.

1. INTRODUCTION

The liver is one of the largest organs in the human body, and it is also the second largest organ after the skin. It is located in the upper right part of the abdominal cavity. It is shaped like a wedge. It is also the body's largest gland, secreting chemical substances known as hormones. The liver performs over 500 functions in the human body and supports the majority of organs that are essential for our survival [1]. Adults have a liver weight of about 2% of their body weight, males have a liver weight of about 1.4 - 1.8 kgs, females have a liver weight of about 1.2 - 1.4 kgs, and newborns have a liver weight of 150 g. The liver is responsible for various functions some are listed below: 1. It produces bile and glycogen.2 It produces serum protein lipids.3. It cleanses the bloodstream of endogenous and exogenous substances such as toxins, drugs, and alcohol. It contains the vitamins D, A, K, E, and B125.It has the ability to regenerate (the remaining liver tissue can regenerate back to its previous size in 5-7 days if two-thirds of the liver is removed) Liver disease is defined as liver swelling caused by toxic substances, bacteria, or inherited disease, which causes the liver to fail to function properly as it is essential for digestion and bacteria removal [2 Liver disease is most common in men between the ages of 40 and 60. Every year, approximately 10 lakh people are diagnosed with liver disease, with a total of 1.4 lakh deaths occurring in India. Machine Learning is a subset of Artificial Intelligence (AI) that simulates human intelligence in machines that can be programmed to think like humans and perform actions like theirs. In other words, ML assists the system in learning without the need for specific knowledge [3]. The user inputs and outputs are used in the Supervised algorithm for training and accurate prediction. Machine learning has also made inroads into health care. One of the issues confronting health care is the growing number of patients. Machine Learning applications have the potential to improve treatment accuracy [4]. Classification techniques are used in a variety of automatic medical diagnostic methods. Because the organ functions normally despite being partially destroyed, the symptoms of liver disease are difficult to detect early on. Early detection of a liver problem improves patient survival. Enzyme levels in the blood can be used to diagnose liver disease. In this paper, we use a dataset of 417 patient blood reports to predict

whether or not the patient has liver disease. For predicting a patient's liver disease, several ML models were compared in this paper, including Logistic Regression, Random Forest, XGBoost, KNN, Decision Tree, Gradient Boosting, and Neural Networks. All issues overlooked by previous researchers are taken into account, and prediction accuracy is improved [5].The following steps were taken to predict liver patients from the dataset: EDA, data pre-processing, outlier removal, and various classifiers, base, and advanced algorithms [6]. This data set includes 417 liver patient records and 167 non-liver patient records from Andhra Pradesh's North East region. The "Dataset" column is a class label that is used to categorize groups as liver patients (liver disease) or not (no disease). There are 441 male patient records and 142 female patient records in this data set.

2. LITERATURE REVIEW

Varun Vats and colleagues (2018) investigated three different ML (Machine Learning) algorithms. A comparison of these algorithms was performed to assess their forecasting accuracy and computing complexity [6]. AP (Affinity Propagation), K means, and DBSCAN were among the algorithms used. This work focused on a medical dataset based on lever disorders. The Silhouette coefficient was used in this study to compare the efficiency of the algorithmic approaches under consideration.

L. Alice Auxilia et al. (2018), the use of medical datasets has piqued the interest of medical experts worldwide [8]. The application of ML (Machine Learning) algorithms was quite common as a branch of creating expressively helpful networks for disease prediction by arranging therapy-based datasets. In general, grouping schemes have been used as a segment of the curative domain to extract order more efficiently than a signal classification model. Liver malady disorders can be described as liver damage or sickness. A liver disorder can be classified into several types. This study used standard Indian liver disease patient records as a database to assist the researcher.

Vyshali J Gogi et al. (2018), the healthcare sector had a lot of data, but it was useless [7]. This massive amount of data necessitated the use of a cutting-edge analytic tool in order to uncover the hidden relationship and valuable knowledge. A medical condition affecting the human liver was referred to as liver disease. The liver diseases caused abrupt changes in health conditions that governed the liver's functioning, affecting other internal body organs. Several data mining-based classification algorithms were used in this work. DT (Decision Tree), LD (Linear Discriminant), SVM Fine Gaussian, and LR were among the algorithms used. This study made use of patient lab-based metrics in the form of a liver dataset.

Sanjay Kumar et al. (2018) described various classification approaches by applying them to a dataset of patients with liver diseases [12]. The main goal here was to accurately predict liver disease using a variety of data mining algorithms. This study used a dataset of real-time patients to build classification paradigms for the prediction of liver diseases. On the used dataset, this work implemented five classification algorithms. This study examined various metrics such as precision, recall, and accuracy to determine the effectiveness of the implemented classification models.

Nazmun Nahar and Ferdous Ara (2018), their research investigates the early prediction of liver disease using various decision tree techniques. The liver disease dataset chosen for this study includes total bilirubin, direct bilirubin, age, gender, total proteins, albumin, and globulin ratio[4]. The main goal of this work is to compute and compare the performance of various decision tree techniques. J48, LMT, Random Forest, Random Tree, REPTree, Decision Stump, and Hoeffding Tree are the decision tree techniques used in this study. The analysis shows that Decision Stump outperforms all other techniques in terms of accuracy.

3. METHODOLOGY

Logistic Regression, K-nearest Neighbor, and Support Vector Machine classification accuracy are compared using the proposed methods. The data must first be cleaned. Filling in the blanks, then converting the nominal attribute to a binary attribute. The next step is feature selection, which involves determining the best attribute for a subset of features. A pivot table was used in this work to visualize the relationship between the attributes and the predictor variable. Attributes were chosen based on the findings. The third step is to transform the data. Data is standardized in this technique so that it has a Gaussian distribution with a mean of 0 and a standard deviation of 1. The classification model is trained in the fourth step.

Table 1 shows the processing steps for determining the best algorithm for the given dataset.

Table 1: Data Pre-processing steps

Algorithm steps:

Step 1: First, read the dataset.

Step 2: To balance the data set, random sampling is used.

Step 3: Separate the dataset into two parts: Train dataset and Test dataset.

Step 4: The proposed models are subjected to feature selection.

Step 5: Accuracy and performance metrics were calculated to determine the efficiency of various algorithms.

Step 6: Next, find the best algorithm for the given dataset based on efficiency.

SYSTEM ARCHITECTURE

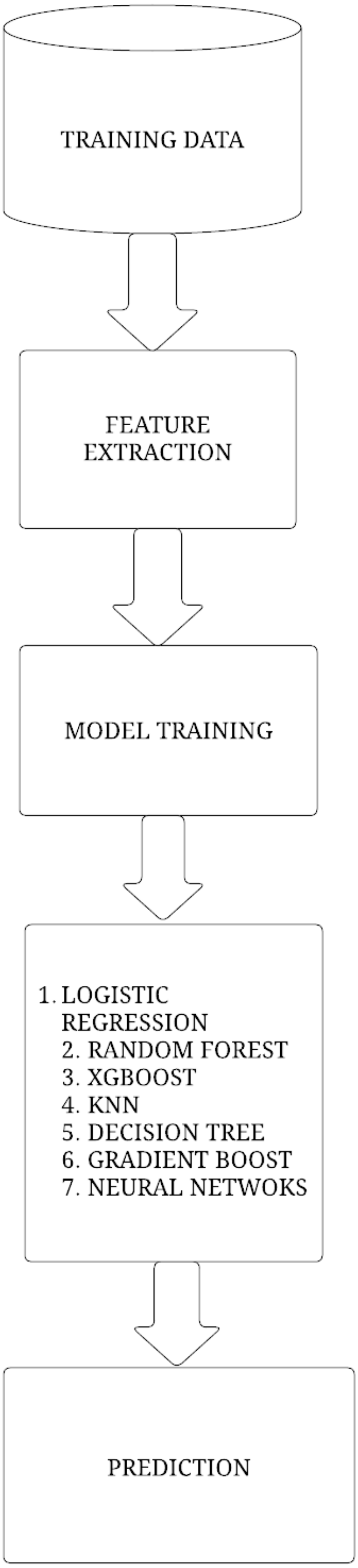

Figure 1: System Architecture

4. EXPERIMENT RESULT

In this section, the results of the machine learning algorithms Logistic Regression, Random Forest, XGBoost, KNN, Decision Tree, Gradient Boosting, and Neural Networks are analysed. The dataset is divided into two parts in the experiment: the training set and the testing set. The training set's ratio is 70% and 30%, respectively. The experiment is written in Python, and the libraries used are pandas and sci-kit learn.

It is a binary classification problem in which we must predict whether a given patient has liver disease or not based on the above set of features.
We use the following metrics for evaluation because this is a binary classification problem:
Confusion matrix - To gain a better understanding of the model's correct/incorrect predictions.
ROC-AUC - It considers the rank of the output probabilities and intuitively measures the model's ability to distinguish between a positive and a negative point. (It should be noted that ROC-AUC is typically used only for binary classification.) To choose the best model, we will use AUC.

A. Performance Measure

We will measure the performance accuracy of the model by two methods:

The AUC-ROC curve is a performance metric for classification problems with varying threshold settings. AUC represents the degree or measure of separability, while ROC is a probability curve. It indicates how well the model can distinguish between classes. The greater the AUC, the better the model predicts 0 classes as 0 and 1 classes as 1. Similarly, the higher the AUC, the better the model distinguishes between patients with and without the disease.

Confusion Matrix is the tabular representation of actual or predicted values. Accuracy is calculated by (true positive (TP) + true negative(TN))/ (true positive(TP) + true negative(TN) + false positive(FP) + false negative(FN)).

The sensitivity which is also called true positive rate state that how many positive values out of all positive values have been correctly predicted.
Sensitivity = TP/(TP+FN)

Specificity which is also known as true negative rate state that how many negative values out of all negative values have been correctly predicted.
Specificity = TN/ (TN + FP)

B. Results.

All the machine learning algorithm have been tested:

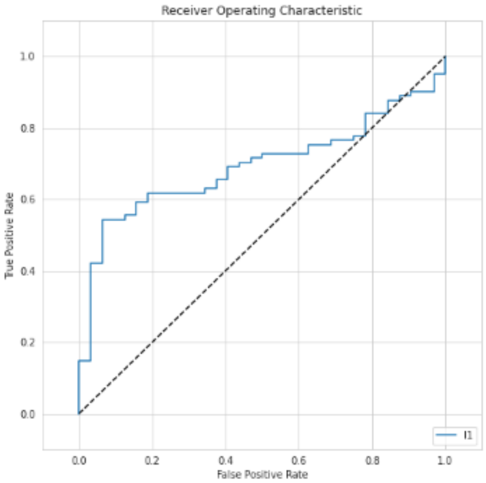

Figure 2: AUC curve of Logistic Regression

From the AUC curve of Logistic Regression model, accuracy has been calculated and it is coming out to be 53.8%.

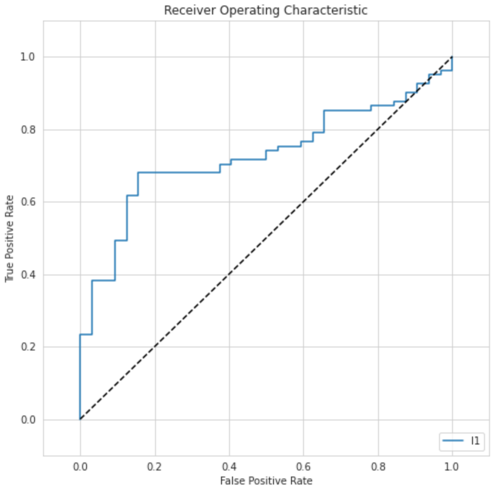

Figure 3: AUC curve of Random Forest

From the AUC curve of Random Forest model, accuracy has been calculated and it is coming out to be 87.7%.

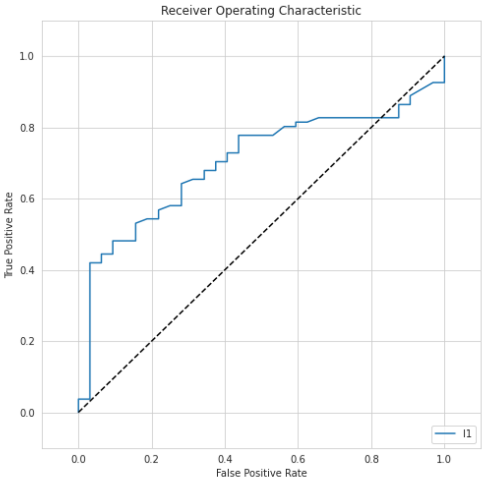

Figure 4: AUC curve of XGBoost

From the AUC curve of XGBoost model, accuracy has been calculated and it is coming out to be 64.7%.

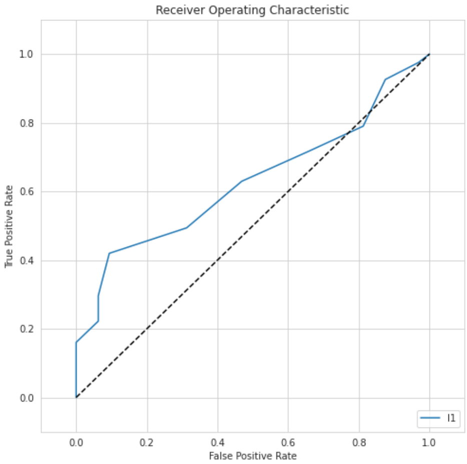

Figure 5: AUC curve of KNN

From the AUC curve of KNN model, accuracy has been calculated and it is coming out to be 63.7%.

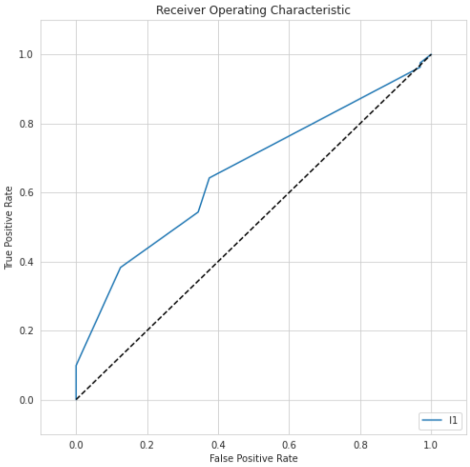

Figure 6: AUC curve of Decision Tree

From the AUC curve of Decision Tree model, accuracy has been calculated and it is coming out to be 59.6%

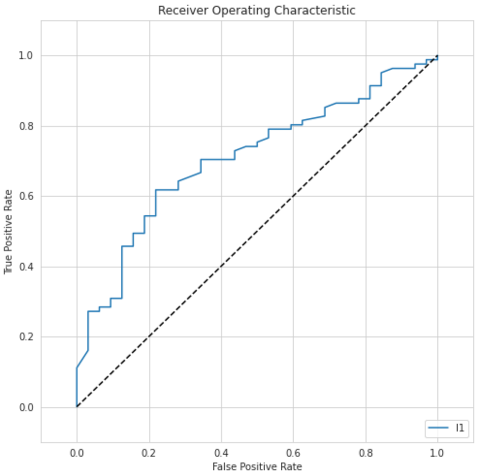

Figure 7: AUC curve of Gradient Boost

From the AUC curve of Gradient Boost model, accuracy has been calculated and it is coming out to be 86.1%

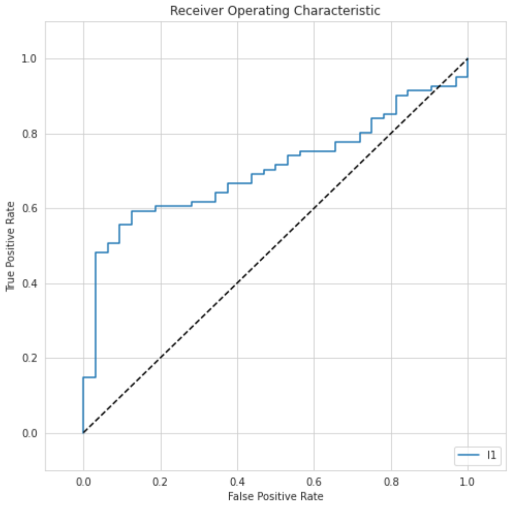

Figure 8: AUC curve of Neural Networks

From the AUC curve of Neural Networks model, accuracy has been calculated and it is coming out to be 69.9%

5. CONCLUSION

Logistic Regression, Random Forest, XGBoost, KNN, Decision Tree, Gradient Boosting, and Neural Networks were used in this paper to detect Liver Cirrhosis Disease. The proposed system's performance is evaluated using sensitivity, specificity, accuracy, and error rate. Logistic Regression, Random Forest, XGBoost, KNN, Decision Tree, Gradient Boosting, and Neural Networks have accuracy values of 53.8, 87.7, 64.7, 63.7, 59.6, 86.1, and 69.9% respectively. We discovered that the Random Forest classifier is the best machine learning model for this project, outperforming Logistic Regression, XGBoost, KNN, Decision Tree, Gradient Boosting, and Neural Networks.

6. REFERENCES

[1] A. Charleonnan,T. Fufaung, T. Niyomwong, W Chokchueypattanakit, S. Suwannawach, N. Ninchawee “Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques ” MITiCON2016.

[2] Anatomy and function of the liver. [Online]. Available: https://www.medicinenet.com/liver_anatomy_and_function/article.html

[3] Abhishek Chowdhury, Thirunavukkarasu K, Sidhyant Tejas(2017), Predicting whether song will be hit using Logistic Regression. Volume 6 Issue 9 September 2017.

[4] Nazmun Nahar and Ferdous Ara, “LIVER DISEASE PREDICTION BY USING DIFFERENT DECISION TREE TECHNIQUES”,International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, 01-09 (2018).

[5] P.Mazaheri, A. Narouziand A. Karimi (2015), Using Algorithms to Predict Liver Disease Classification,Electronics Information and Planning. 3:255-259.

[6] Varun Vats, Lining Zhang, Sreejit Chatterjee, Sabbir Ahmed, Elvin Enziama, Kemal Tepe, "A Comparative Analysis of Unsupervised Machine Techniques for Liver Disease Prediction", IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 303-307 (2018).

[7] Vyshali J Gogi, Vijayalakshmi M.N, "Prognosis of Liver Disease: Using Machine Learning Algorithms", International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), 875-879 (2018).

[8] L. Alice Auxilia, "Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease", 2nd International Conference on Trends in Electronics and Informatics (ICOEI), 45-50 (2018).

[9] David Diez, Christopher Barr and Mine Cetinkaya-Rundel. Open Intro Statistics. 3rd Edition. OpenIntro.org

[10] Reetu and N.Kumar (2015), Medical Diagnosis for Liver Cancer Using Classification Techniques, International Journal of Recent Scientific. Volume 6. Issue, 6, pp 4809-4813.

[11] Tina R. Patil, Mrs. S. S. Sherekar. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013

[12] Sanjay Kumar, Sarthak Katyal, "Effective Analysis and Diagnosis of Liver Disorder by Data Mining", International Conference on Inventive Research in Computing Applications (ICIRCA), 1047-1051 (2018)