PREDICTION OF LIVER DISEASE USING MACHINE LEARNING

Dr. M. Venkateshwara Rao, Mr. Rajesh Saturi
B.Sushanth[1], C. H. Vaishnavi[2], Ch.Dinesh[3]

Article History: Received: 23.03.2023 **Revised:** 05.05.2023 **Accepted:** 19.06.2023

Abstract

The liver is the most fundamental organ in humans and is responsible for many bodily processes, such as protein and carbohydrate metabolism, enzyme activation, glycogen storage, and vitamin and mineral absorption. Almost 1 million people in India have liver illnesses, which are primarily underdiagnosed and have resulted in a number of fatalities. This project aids in the early detection of liver disorders, reducing the number of deaths and preventing India from becoming the "Global capital of liver diseases" by 2025. The sharp rise in obesity and unhealthy lifestyle rates ultimately reflects the propensity and frequency of liver-related disorders. The patient data sets are studied in this study to determine if it is likely that the subject will develop liver disease, which is predicted using a variety of machine learning methods such as Naïve Bayes, Logistic Regression, MLP (Multi-Layer Perceptron) and Random Forest. The most precise machine learning algorithm was used to forecast the outcome. The method employs the most precise model that has been trained to determine whether or not a person is at risk for liver disease.

Keywords – Naïve Bayes, Logistic Regression, MLP (Multi-Layer Perceptron), Random Forest.

[1,2,3]Vignana Bharathi Institute of Technology, Aushapur, Ghatkesar

DOI: 10.31838/ecb/2023.12.si6.549

1. INTRODUCTION

Around the world, liver disease (LD) is a frequent illness. The liver's ability to function properly has a significant impact on several bodily processes, including protein synthesis, iron and sugar metabolism, and blood coagulation.

257 million people worldwide are afflicted by the Chronic Hepatitis B virus, which causes liver infections. Chronic liver disease claims the lives of about a million people who have chronic infections like HBV. As a result, there is a clear need for an efficient, precise, and useful framework to predict the outcome of such an infection. It will be useful for preventing problems and providing the right care.

The study of liver disease prediction and prevention using data mining and artificial intelligence principles is crucial in the current decade. Machine learning (ML) models are used by many researchers to forecast diseases. In this study, we offer the empirical statistical analysis to prevent liver diseases and use effective machine learning models for early, low-cost forecasts of liver disorders.

By leveraging historical information about the amounts of enzymes in patients' bodies, this approach aims to boost the survival rate of liver patients. Due to their lifestyle choices and the state of the environment today, individuals are susceptible to many diseases. The identification and prediction of such diseases at their earlier stages are much important, so as to prevent the extremity of it.

To prevent such diseases from becoming severe, early detection and prediction of these disorders are crucial. Even when liver tissue has been damaged sufficiently in the early stages of the disease, it is very difficult to diagnose, leading many medical professionals to repeatedly overlook the illness. Early detection is crucial and critical to protect the patient because this could lead to the inappropriate medication and treatment. The diagnosis of this illness is very expensive and difficult. In order to lower the high cost of chronic liver disease diagnosis by prediction, the objective of this work is to analyse the performance of several Machine Learning methods.

2. LITERATURE SURVEY

Six machine learning algorithms, including LR, KNN, DT, SVM, NB, and RF, have been used in [1] this research. The performance of these techniques was estimated from a variety of views, including accuracy, precision, recall, and f-1 score. Furthermore, the receiver operative feature was used to compare the performance (ROC). They obtained a dataset for this experiment from the UCI Machine Learning Repository. The original information was also gathered in Andhra Pradesh, India, in the northeast. There are 75.64% male patients and 24.36% female patients in this dataset of 583 liver patients. The main goal of this research is to create a system that can accurately diagnose patients with chronic liver infections using six different supervised machine learning classifiers. They studied how the classifiers performed when given patient information parameters, and found that the LR classifier provided the highest order exactness (75% based on the F1 measure to predict liver disease), while the NB classifier provided the lowest precision (53%).

Support vector machines are employed as classification algorithms in [2] this work to assess their effectiveness with several feature combinations, including aspartate aminotransferase (SGOT), glutamic pyruvic transaminase (SGPT), and alkaline phosphatase (Alkphos), utilising two datasets. The first is BUPA Liver Diseases datasets retrieved from the University of California at Irvine (UCI) Machine Learning Repository, and the second is from ILPD (Indian Liver Patient Dataset), which was gathered from Andhra Pradesh, India's northeast. In comparison to the BUPA dataset, the results demonstrate that the ILPD dataset has better accuracy, error rate, sensitivity, and prevalence at the top six ordered characteristics. This can be explained by the fact that the ILPD liver dataset, as opposed to the BUPA dataset, has a variety of relevant variables that can aid in the identification of liver illness, such as Total bilirubin, direct bilirubin, albumin, gender, age, and total proteins.

The authors of [3] have used machine learning algorithms on a dataset of Indian liver patients to forecast patients at an early stage based on the enzyme level in their bodies. They preprocessed the data first, followed by some data visualisation work, and then they trained the model using various methods before choosing the one that produced the best results. They used various machine learning methods for classification, including Logistic Regression, SVC, and Random Forest. They also applied bagging to Random Forest and AdaBoost to Logistic Regression. Processing time for logistic regression is quick, and it provided accuracy of 73.5%. AdaBoost was employed to increase its accuracy, and the result was an accuracy of 74.36%.

Proposed System

The solution that is being suggested here makes use of the machine learning idea, and the models are trained before being tested. The final outcome will be predicted by the model that is most correct. The system initially requests that we provide information about our age, gender, total and direct bilirubin levels, total proteins, albumin, A/G ratio, SGPT, SGOT, and alkphos. The user's blood test report can be used to determine the values of the final eight parameters mentioned above. When the user provides these inputs, the system compares the supplied data to the training dataset of the most accurate model and then forecasts the outcome as risky or not. The system offers the following benefits:

No medical knowledge is necessary:
Using this application, we can anticipate liver disease without any prior medical science or understanding of liver problems. Entering the information that is already in the blood test report (some, like age and gender, are already known) will allow us to obtain the prediction results.

High degree of accuracy:
With the dataset we used to build this application, the algorithm predicts the outcomes with 94% accuracy. Even though the accuracy might vary in some instances, it will still be sufficient to be relied upon on a wide scale.

Immediate Results:
Results are projected in a matter of seconds after entering the information. In contrast to the conventional procedure, we don't need to wait for a doctor to arrive.

The tasks listed below make up the majority of the application:

Developing and training the system:
The end user has no input at all during this phase, which is entirely handled by the system's developer. At this phase, the dataset was divided into training and test sets, and the training sets were used to train the models.

Evaluating the models:
Using the test dataset created in the previous phase, we evaluated the models' correctness in this phase and determined which model is the most accurate.

Input information and making a prediction:
At this point, the end user enters the picture. He or she uses the application's GUI to enter the blood test report's details. The application then compares the information with the best accurate model's training set and predicts the outcome, displaying "Healthy Liver" or "Liver Disease" on the screen.

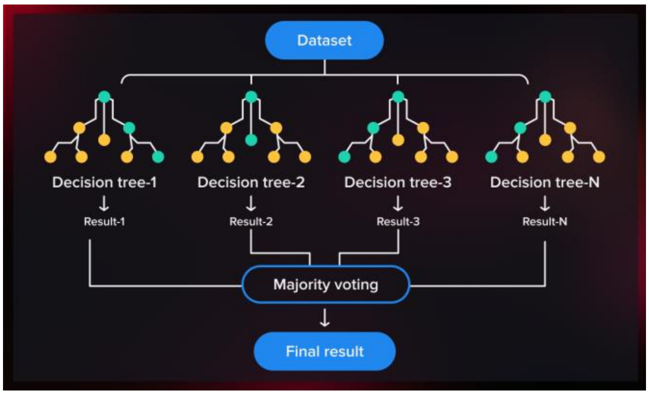
3. METHODOLOGY

It is insufficient to compare accuracy, or the proportion of accurate forecasts to all predictions, in problems of disease classification like this one. This is due to the fact that, depending on the context, such as the severity of the sickness, it is often more crucial that an algorithm does not incorrectly identify a disease as a non-disease, although doing so will result in a relatively less severe penalty. As a result, we will utilise the F-beta score, which is essentially the weighted harmonic mean of recall and precision, as our performance indicator in this instance. Recall is calculated as follows: Precision=TP/(TP+FP), where TP stands for True Positive and FP for False Positive. False Negative, or FN

Algorithms and Techniques:
For this issue, three supervised learning strategies are chosen. In order to include as many potential ways as possible, care is made to ensure that all of these techniques are essentially distinct from one another. For instance, though Random Forest and Ada Boost are to the same family of "ensemble" techniques, we won't choose them together.

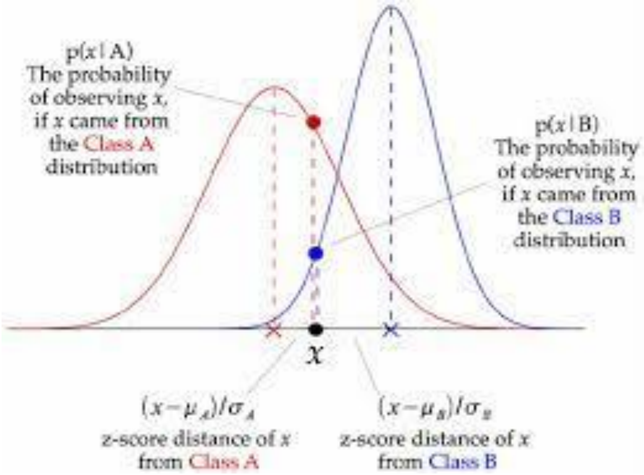
To find the optimal classifier for each method, we will experiment with a few hyperparameters at various values. The grid search cross validation method will be used to carry accomplish this. Here is a description of the algorithms:

1. Random Forest Classifier: The Random Forest supervised machine learning algorithm develops and merges various decision trees to produce a "forest." Random Forest is a strong and flexible classifier.



2. Gaussian Nave Bayes Classifier: The Nave Bayes technique is used to classify issues with binary (two-class) and multiple classes. When the method is explained using binary or category input values, it is simplest to understand. Probabilities serve as the Naive Bayes representation. A learnt Naive Bayes model keeps a file with a list of probabilities. This comprises:

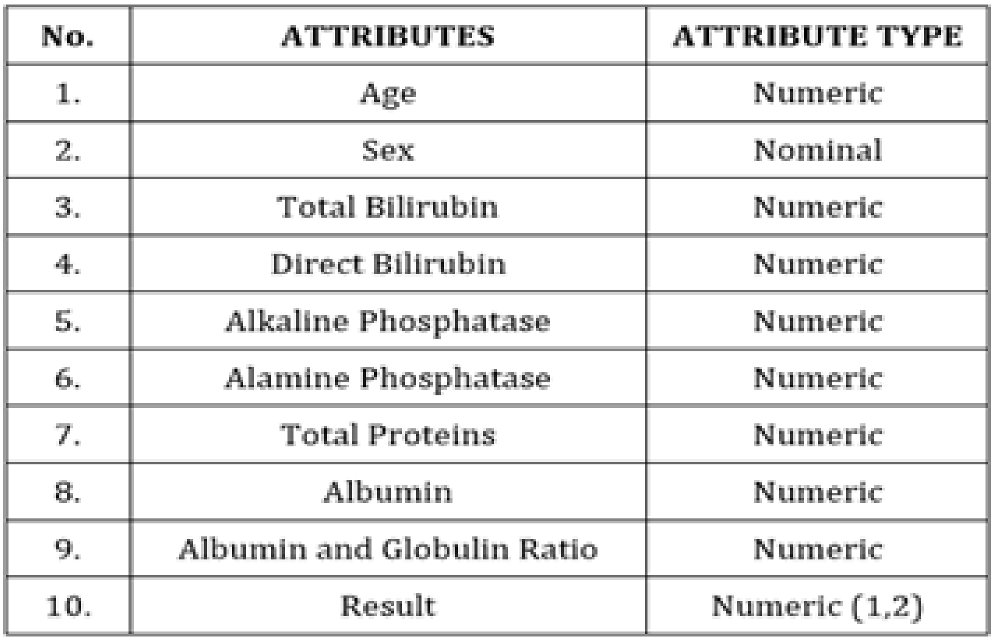
Class Probabilities: In the training dataset, these are the probabilities for each class. The conditional probabilities of each input value given each class value are known as conditional probabilities.


p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

3. Logistic Regression: This method seems appropriate because the result is binary and we have a sufficient number of samples compared to the number of features. A logistic or sigmoid function that measures the discrepancy between each forecast and its associated true value is the basis of this strategy. It gives varying weights to features depending on the number of inputs it receives (based on their relative importance).

Since it already knows the outcome for this data, it continuously modifies the weights to ensure that the outcomes are as accurate as possible when these weights and their features are entered into the logistic function. When given a test value, it again enters that value into our logistic function and outputs a number between 0 and 1, which denotes the chance of that test value
falling into a specific class.

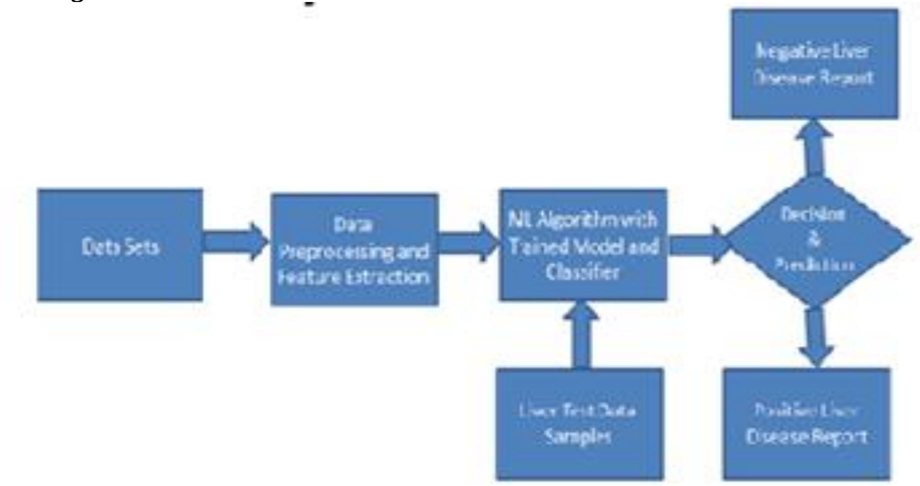
Patient Dataset:
There are ten attributes in the dataset, but gender has been removed because it affects our target variable. Currently, there are 30.000 records and 9 attributes in the collection



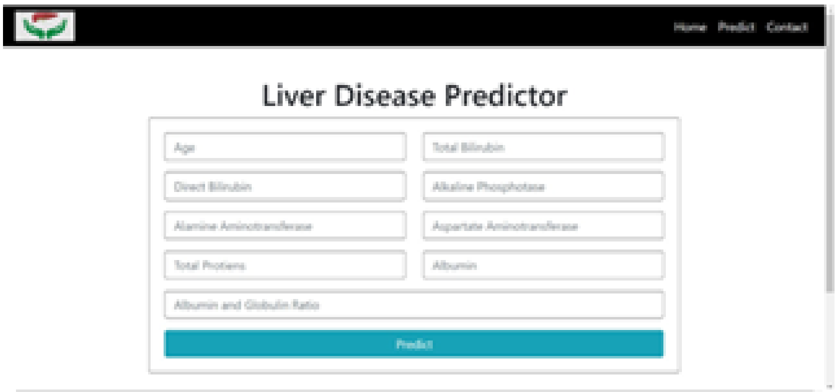
No.	ATTRIBUTES	ATTRIBUTE TYPE
1.	Age	Numeric
2.	Sex	Nominal
3.	Total Bilirubin	Numeric
4.	Direct Bilirubin	Numeric
5.	Alkaline Phosphatase	Numeric
6.	Alamine Phosphatase	Numeric
7.	Total Proteins	Numeric
8.	Albumin	Numeric
9.	Albumin and Globulin Ratio	Numeric
10.	Result	Numeric (1,2)

.

Architecture Diagram:

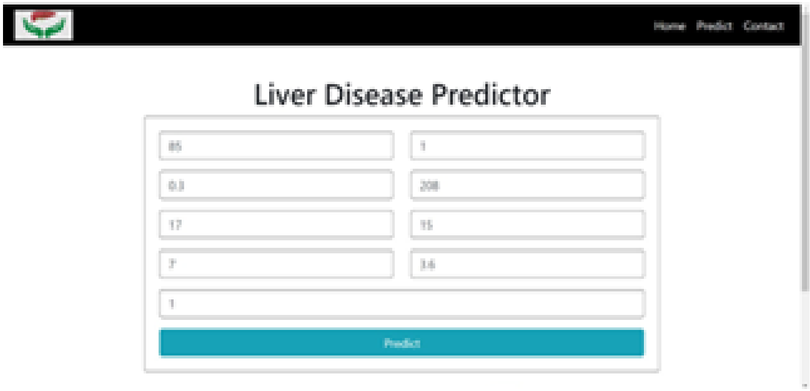

Deta Sets
Data Preprocessing and Feature Extraction
NL Algorithm with Trained Model and Classifier
Decision & Prediction
Negative Liver Disease Report
Positive Liver Disease Report
Liver Test Data Samples

4. RESULT



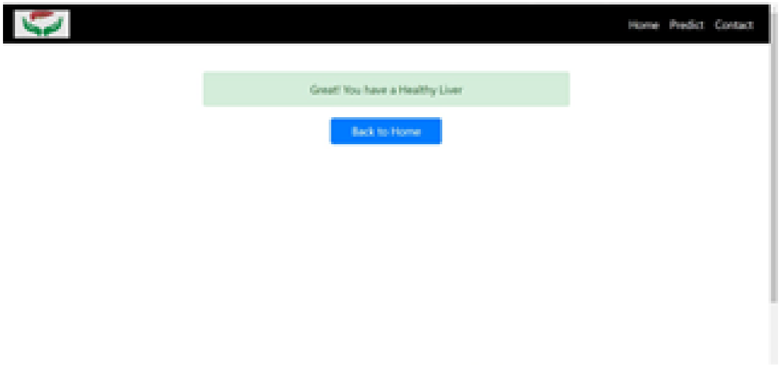
Liver Disease Predictor

Age | Total Bilirubin
Direct Bilirubin | Alkaline Phosphotase
Alamine Aminotransferase | Aspartate Aminotransferase
Total Protiens | Albumin
Albumin and Globulin Ratio

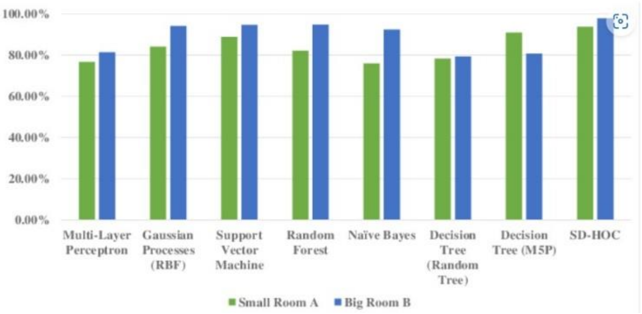


Liver Disease Predictor

85 | 1
0.3 | 208
17 | 15
7 | 3.6
1



Great! You have a Healthy Liver



5. CONCLUSION AND FUTURE WORK

In this study, we have suggested approaches for employing machine learning techniques to diagnose liver illness in patients. Naive Bayes, Logistic Regression, MLP, and Random Forest were the four machine learning methods employed. Different models were taken in reference and their performance metrics were evaluated. When compared with previous research studies, Random Forest had a highest accuracy when implemented. Hence, it was proven effective. Accordingly, it was used to develop a GUI that can be used as a medical tool.

The suggested system has a great deal of room for development in a number of different ways. They consist of:

1. Make the algorithms more accurate.
2. Modifying the algorithms to increase system performance and efficiency.
3. Developing some additional qualities to combat diabetes even more.

6. REFERENCES

1. **A. K. M. Sazzadur Rahman, F.M. Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy(2019).** "A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms".
2. **Arshad, C. Dutta, T. Choudhury, and A. Thakral. (2018),** "Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques". In IEEE International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 163-168
3. **Esraa M. Hashem, Mai S. Mabrouk (2014),** "A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis "
4. **Joel Jacob, Johns Mathew (2018),** "Diagnosis of Liver Disease Using Machine Learning Techniques. International Research Journal of Engineering and Technology", Vol.5 Issue 4
5. **Khan Idris, Sachin Bhoite (2019),** "Applications of Machine Learning for Prediction of Liver Disease"
6. **M. B. Priya, P. L. Juliet, P.R. Tamilselvi. (2018),** "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms", International Research Journal of Engineering and Technology, vol. 5 no. 1, pp. 206-211.
7. **Prof Christopher N(2010)**. "New Automatic Diagnosis of Liver Status Using Bayesian Classification".
8. **Panduranga Vital Terlapu (2021),** "Intelligent Liver Disease Prediction (ILDP) System Using Machine Learning Models".
9. **Joloudari, J. H., Saadatfar, H., Dehzangi, A., & Shamshirband, S. (2019)**, Computer-aided decision-making for predicting liver disease using PSO-based optimised SVM with feature selection.
10. **El-Shafeiy, E. A., El-Desouky, A. I., & Elghamrawy, S. M. (2018)**. Prediction of liver diseases based on machine learning technique for big data. International Conference on Advanced Machine Learning Technologies and Applications.
11. **Dutta, K., Chandra, S., & Gourisaria, M. K. (2022)**. Early-Stage detection of liver disease through machine learning algorithms.