# LIVER CIRRHOSIS PREDICTION USING MACHINE LEARNING ALGORITHMS

M.Harishwar Reddy E-mail:harishwar.momula@gmail.com P.Pavan Kumar E-mail: Pavanbabbu01@gmail.com

R.Suryaprakash E-mail: Suryaprakashreddy0412@gmail.com T.Pushpa(Associate Professor) Email: pushpa.404@gmail.com

Electronics And Communication Engineering,Vbit,Hyderabad,India.

***Abstract*— Lifestyle diseases have become common these days and a sedentary way of life has paved the way for a range of syndromes and unknown diseases. Identification or diagnosis of the disease at an early stage is most crucial. This greatly helps in the prevention of the disease at an early stage with minimal medications. Less common causes of cirrhosis include autoimmune hepatitis, primary biliary cholangitis, and primary sclerosing cholangitis that disrupts bile duct function, genetic disorders such as Wilson's disease and hereditary hemochromatosis, and chronic heart failure with liver congestion Traditional methods involve physical examination and lab results. Identification of the Liver disease at an early stage is very difficult as the symptoms of the diseases are visible only at a later stage of the disease.The Application of Machine learning models would help in the early diagnosis of the disease and hence facilitates in identifying crucial factors that lead to liver damage.In this paper, the risk of liver disease was predicted using two different machine learning algorithms. The final output was predicted based on the most accurate machine learning algorithm i.e., XGBoost. Based on the accuracy we designed a system which asks a person to enter the details of his/her medical report. Then the system uses the most accurate model which is trained to predict, whether a person has risk of liver disease or not.**

**Keywords— Machine Learning,Logistic Regression, XGBoost, Gradient Boost**

## I . INTRODUCTION

Liver diseases have become the twelfth leading cause of death around the globe. Liver diseases are of many types. Hepatitis is a condition where the liver has inflammation and is caused by the Hepatitis virus. Hepatitis A, B, C, D and E are five types of strains that cause liver damage. Hepatitis B and Hepatitis C strain of virus results in chronic disease. 325 million people around the globe suffer from hepatitis B or C. A study conducted by WHO found that the prevention of 4.5 million deaths would be possible through powerful vaccination and efficient diagnostic test. WHO stated that in the year 2018, India had 2,64,193 deaths due to liver diseases which are approximately 3% of the total deaths across the globe. Liver Cirrhosis occurs due to the formation of fibrosis or distortion of the liver. In 2020, it has become one of the leading causes of death across the world. Consumption of alcohol has accounted for 3.8% of liver related-deaths. Obesity has also led to Fatty-Liver diseases. Some of the metabolic liver diseases are Hereditary hemochromatosis, Wilson's disease, Alpha-1 antitrypsin deficiency and Hepatocellular carscinoma.25% of the population around the globe suffers from Non- Alcoholic Fatty Liver disease (NAFLD).

Figure 1. Normal liver and Liver Cirrhosis

Childhood obesity and diabetes are factors that lead to NAFLD.Chronic Liver Diseases are on the rise across the globe for the past few decades. The liver becomes inflamed due to hepatitis caused by virus forms like A, B or C. Hepatitis A lasts for six months and is caused by ingesting substances that are contaminated. Hepatitis B spreads through bodily fluids and sometimes may lead to chronic liver infection. Hepatitis C is transmitted with infected blood and is more serious in nature. This liver disease shows no symptoms and may stay in the liver for many years. Fatty liver disease results due to high-fat deposits in the liver and amounts to 5 to 10% of the entire weight of the liver. People with excess weight and diabetes are more prone to this disease. This disease gradually progresses to liver failure. Alcohol and smoking habits increase the severity of thedisease.

Sclerosing Cholangitis (PSC) and Primary Biliary Cirrhosis (PBC). Hepatitis can be prevented by vaccination,regular exercise, low sugar and high fiber intake. In India, the average age of people affected due to liver diseases is between 30 to 40 and in the west, it is between 45 to 55.The dataset taken for the study was the Indian Liver disease dataset drawn from the UCI repository. The dataset was scaled for reducing the high variance in the data values.Highly colinear features were eliminated using RFE (Recursive Feature Elimination) and the impact of its application was tested on the Gradient Boosting Algorithms.

## II DATASET AND FEATURES

The data contains the information collected from the Mayo

Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984.A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo- controlled trial of the drug D- penicillamine.The first 312 cases in the dataset participated in the randomized trial and contain largely complete data.The additional 112 cases did not participate in the clinical trial but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants.

There are some interesting insights if we observe closely. Take the case at Ascites, we observe that the risk of disease is higher with increase in Ascites. also presence of spiders has a positive relation with disease risk.

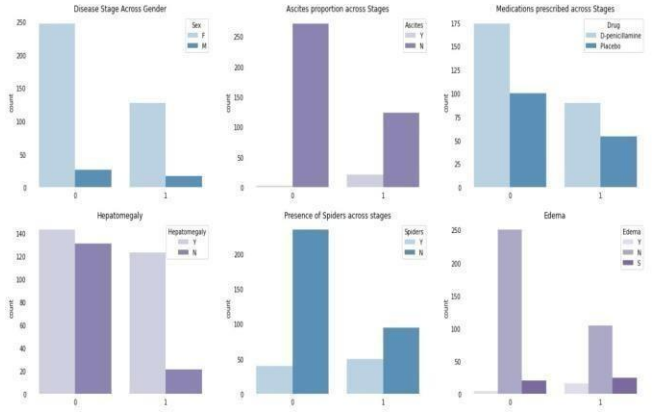


Figure 2: Some Features

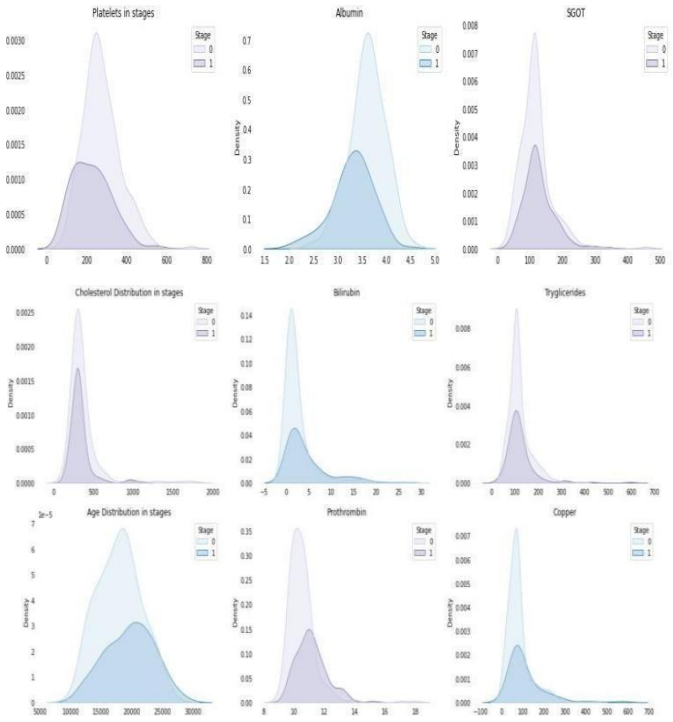


Figure 3: Feature distribution

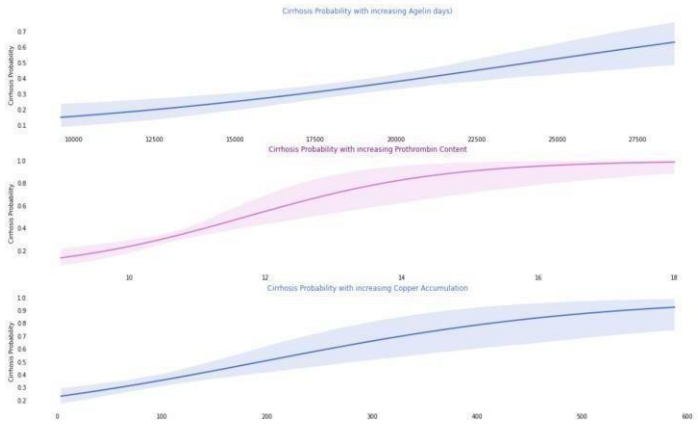


Figure 4: some features (platelets etc)

We can also observe some features such as Platelets, Albumin, Cholesterol where the probability of disease decrease with increase in feature value. Lets tally that with some more regression plots.

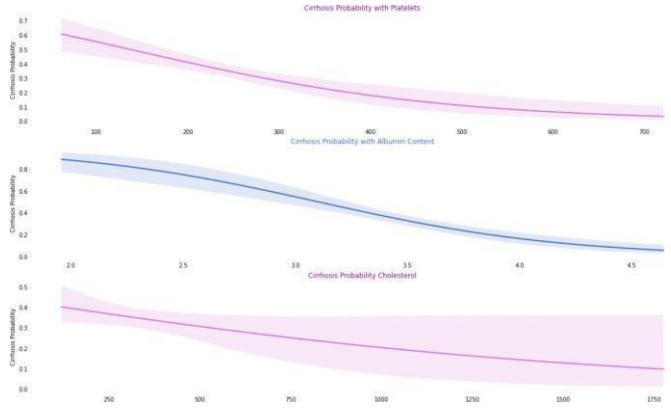


Figure 5: Features (cholesterol e.t.c)

Platelets, Albumin checks with our logic the findings about Cholesterol seems interesting! Looks like people with high Cholesterol have lower risk of Cirrhosis, this might not sound correct but our data certainly shows so.

This should help our model predict the target. We will be looking at what features contribute the most in later part of the project.

Looking at the feature distribution we can observe that in features such as Age, Prothrombin, Copper the risk of the disease increase with increase in feature value, thus having a positive co-relation on with the disease probability. Lets fit a regression line to check.

## III LITERATURE SURVEY

The proposed methods used are to compare classification accuracy of Logistic Regression, K-nearest neighbour and Support Vector Machine. The first step is to clean the data. Filling the missing values followed by transforming nominal attribute to binary attribute. Only catogorial values are taken.[1]

In this they have compared several ML methods, such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Extra Trees (ET) for the prediction of liver disorders.At the preprocessing step, categorical values are encoded through label encoding.Accuracy was less[2]

## IV. METHODOLOGY PROPOSED

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.In prediction problems involving unstructured data (images, text, etc. ) artificial neural networks tend to outper form all other algorithms or frameworks.

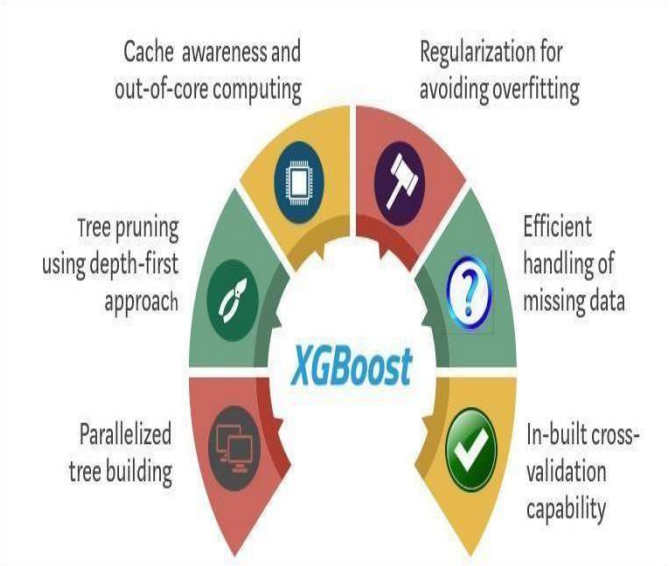


Figure 6 How XGBoost optimizes standard GBM algorithm.

## V. MACHINE LEARNING

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to **Parallelization**: XGBoost approaches the process of sequential tree building using parallelized implementation. This is possible due to the interchangeable nature of loops used for building base learners; the outer loop that enumerates the leaf nodes of a tree, and the second inner loop that calculates the features.

**Tree Pruning:** The stopping criterion for tree splitting within GBM framework is greedy in nature and depends on the negative loss criterion at the point of split. XGBoost uses 'max_depth' parameter as specified instead of criterion first, and starts pruning trees backward. This 'depth-first' approach improves computational performance significantly.

**Hardware Optimization**: This algorithm has been designed to make efficient use of hardware resources. This is accomplished by cache awareness by allocating internal buffers in each thread to store gradient statistics. Further enhancements such as 'out-of-core' computing optimize available disk space while handling big data-frames that do not fit intomemory.

do so. Machine learning algorithms use historical data as input to predict new output values.

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.There are four basic approaches:supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict.

**Supervised learning:** In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

**Unsupervised learning:**This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.

Unsupervised machine learning algorithms do not require datato be labeled. They sift through unlabeled data to look for patterns that can be used to group data points into subsets.

**Semi-supervised learning:** This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.

Semi-supervised learning works by data scientists feeding a small amount of labeled training data to an algorithm.

**Reinforcement learning:** Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules.

Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task.

**Algorithms**

**XGBoost**

The leaf nodes are proportionately reduced and randomized parameters are used for computation. It has an efficient method to penalize the trees.

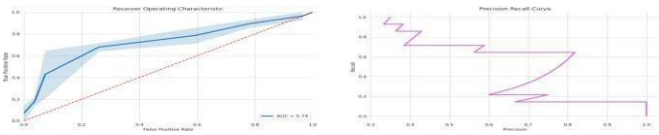

**FIGURE 7** Improvement over the logestic regression.

**Logistic Regression**

Sigmoid is a function, where the Logistic Regression Model selects the best-parameters and fits it to a non-linear function. Stochastic gradient ascent is one of the optimization algorithms used in Logistic Regression. Logistic Regression is one of the supervised machine learning models used for Regression.

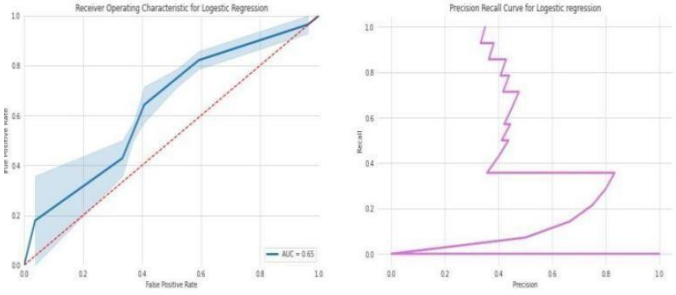


**FIGURE 8** Logistic regression

## VI RESULTS

This paper presents the best model to predict the liver disease than the existing model. This project helps to predict out whether a person is suffering from liver cirrhosis or not.

```
For Fold 1 the accuracy is 0.6904761904761905
For Fold 2 the accuracy is 0.7857142857142857
For Fold 3 the accuracy is 0.6190476190476191
For Fold 4 the accuracy is 0.6666666666666666
For Fold 5 the accuracy is 0.8095238095238095
For Fold 6 the accuracy is 0.6666666666666666
For Fold 7 the accuracy is 0.6904761904761905
For Fold 8 the accuracy is 0.7380952380952381
For Fold 9 the accuracy is 0.7317073170731707
For Fold 10 the accuracy is 0.6341463414634146

Logestic Regression Mean Accuracy =  0.7032520325203252
              precision    recall  f1-score   support

           0       0.70      0.78      0.74        27
           1       0.45      0.36      0.40        14

    accuracy                           0.63        41
   macro avg       0.58      0.57      0.57        41
weighted avg       0.62      0.63      0.62        41

AUC :  0.6507936507936508
```

**Figure 9** Accuracy of logistic regression

```
For Fold 10 the accuracy is 0.7804878048780488

XGboost model Mean Accuracy =  0.7344367015098723
              precision    recall  f1-score   support

           0       0.82      0.85      0.84        27
           1       0.69      0.64      0.67        14

    accuracy                           0.78        41
   macro avg       0.76      0.75      0.75        41
weighted avg       0.78      0.78      0.78        41

AUC :  0.738095238095238
```

**FIGURE 10:** Accuracy of XGboost

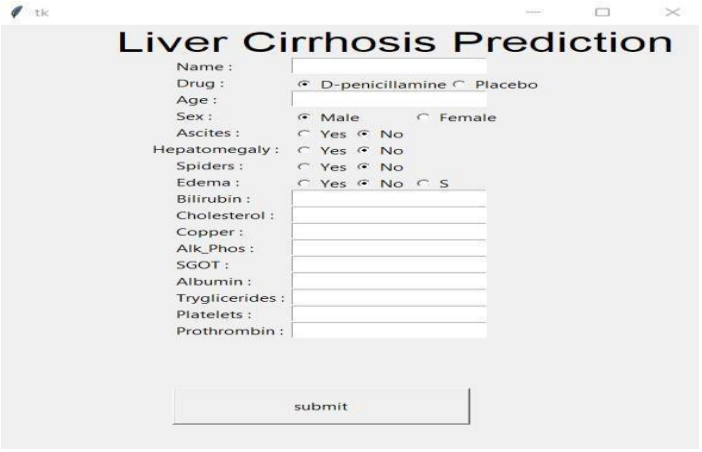


**Figure 11** Graphical user interface implementation

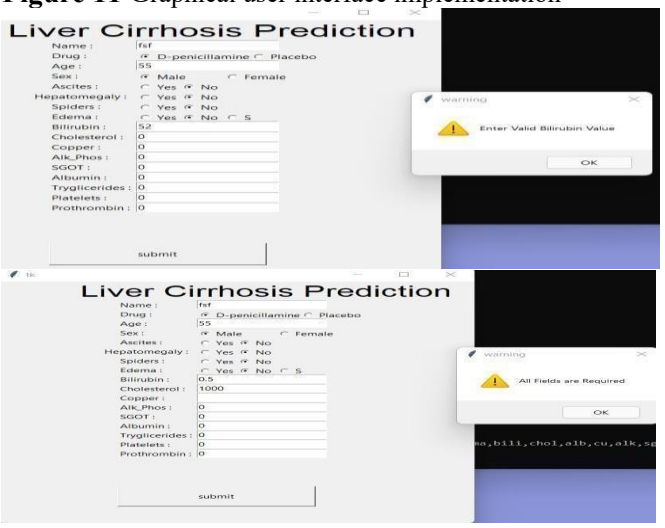


**Figure 12** Enter valid values(warning)

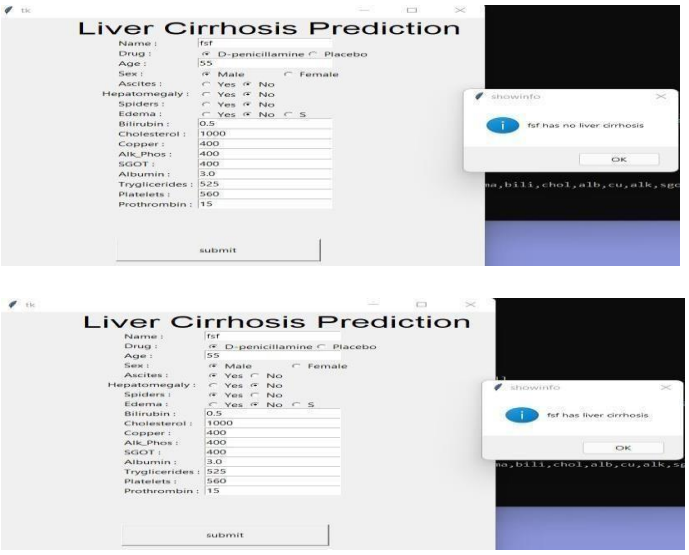


**Figure 13:** Prediction

## VII CONCLUSION

As we know the importance of liver in functioning of human digestive system. so, we need to protect it before it gets damaged completely. This project helps to predict out whether a person is suffering from liver cirrhosis or not. To reach it out to the public we have also come up with graphical user interface which helps people to access the model Easily.

## VIII REFERENCES


[1]M. Yamakawa, T. Shiina, N. Nishida and M. Kudo, "Computer aided diagnosis system developed for ultrasound diagnosis of liver lesions using deep learning," 2019 IEEE International Ultrasonics Symposium(IUS),Glasgow, United Kingdom,2019,pp.2330-2333, doi:10.1109/ULTSYM.2019.8925698.

[2] N. Li et al., "Machine Learning Assessment for Severity of Liver Fibrosis for Chronic HBV Based on Physical Layer With Serum Markers," in IEEE Access, vol. 7, pp. 124351-124365,2019, doi:10.1109/ACCESS.2019.2923688.

[3 ] M. R. Haque, M. M. Islam, H. Iqbal, M. S. Reza and M. Hasan, "Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder,"2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2),Rajshahi,2018,pp.1-5, doi: 10.1109/IC4ME2.2018.8465658.
.

[4 ]T. I. Trishna, S. U. Emon, R. R. Ema, G. I. H. Sajal, S. Kundu and T. Islam, "Detection of Hepatitis (A, B,C andE) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur,India, 2019,pp. 1-7, doi: .1109/ICCCNT45670.2019.8944455.

[5 ]S. Gupta, G. Karanth, N. Pentapati and V. R. B. Prasad, "A Web Based Framework for Liver Disease Diagnosis using Combined Machine Learning Models," 2020 International Conference on Smart Electronics and Communication (ICOSEC),Trichy,India,2020,pp.421-428, doi:10.1109/ICOSEC49089.2020.9215454