IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact Factor: 6.078

(Volume 7, Issue 5 - V7I5-1200)

Available online at: https://www.ijariit.com

A study of Machine Learning algorithms to predict liver cirrhosis and its stage

Prakash Aryan
2019btechaiprakash7275@poornima.edu.in
Poornima University, Vidhani, Rajasthan

Abstract: ***In early stages, cirrhosis usually doesn't cause symptoms. Only through routine blood tests or liver biopsy does a doctor diagnose damage to the liver. Using Machine learning we develop a model that can assist doctors in diagnosing the early stages of liver cirrhosis before it gets fatal. In this study we use various machine learning algorithms to determine the liver cirrhosis stage. Data is collected from the Mayo Clinic trial, USA, in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. Performance of the algorithms were evaluated using ROC-AUC curve which is a very practical method of model evaluation for classification problems. Results showed that by applying Logistic Regression to predict the cirrhosis stage we get a ROC-AUC score of 0.74 which is considerable in view of the instance we have.***

Keywords: *Supervised Machine Learning, Disease Prediction, Python*

1. INTRODUCTION

Liver is the second largest organ in our body after our skin. It helps the body in removing toxins that are present in the blood supply, breaking down drugs, making the digestive fluid "Bile" (helps in digestion and removing wastes from the body), storage and release of glucose as required and various other processes and it is thus called as the "body's chemical processing plant" [1]. Fibrosis of the liver is the early stage when a healthy tissue is replaced with non-living scarred tissue (**Figure I**) and obstructs the functioning of the liver. Cirrhosis is the late or end stage of liver fibrosis that causes permanent and serious scarring of the liver. Liver cirrhosis makes the functioning of the liver very difficult. End-stage liver disease (ESLD) leads to a liver in which the functioning of the liver has drastically deteriorated such that it causes disruptions in flow of blood to the liver which leads to pressure build up in the portal vein. It is a very severe case of cirrhosis.

According to The Lancet (Gastroenterology & Hepatology) the age - standardized death rate globally, due to liver cirrhosis in 2017 was 16.5 per 100,000 population. In the sub-Saharan Africa super-region it was 32.2 per 100,000 population as compared to 10.1 per 100,000 population in the high-income super-region [2]. According to The Global Health Observatory of the World Health Organization, for countries such as Cambodia, Egypt, Nigeria and Sao Tome and Principe the age-standardized death due to liver cirrhosis is over 100 per 100,000 male population for the year 2016 [3].

In the initial stages around 40 percent of those affected with liver cirrhosis are asymptomatic [5]. The Most prevailing causes of liver cirrhosis are alcohol abuse, Hepatitis A, Hepatitis B and Nonalcoholic fatty liver disease (NAFLD) [6]. Global Alcohol-attributable fraction(AAF) for liver cirrhosis deaths(%) is 44 percent [7].Patients with liver cirrhosis have a number of complications for examples, ascites, spontaneous bacterial peritonitis, hepatic encephalopathy, portal hypertension, variceal bleeding, and hepatorenal syndrome [8]. Various important parameters such as drug, age, ascites, hepatomegaly, spiders, edema, bilirubin, cholesterol, albumin, copper, alkaline phosphatase, SGOT, triglycerides, platelets and prothrombin are collected for each individual patient.

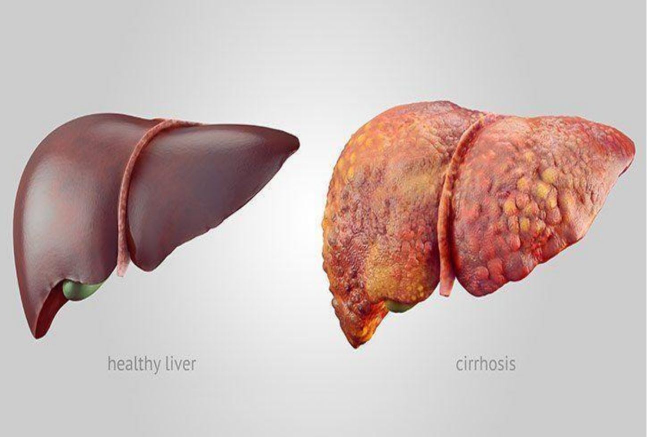

Figure I. Left: Healthy liver Right: Liver cirrhosis[4]

Artificial Intelligence in healthcare has helped in the reduction of the burden on doctors and their patients. An artificial intelligent system is a system that can perform tasks that would otherwise require human intelligence. In this study machine learning, which is a subdivision of artificial intelligence, is used to create a model that studies patterns from data that is very hard for humans to make sense of. Machine learning models can be built in all three areas of healthcare - diagnostic, predictive and prognostic.

This study aims to build models and measure the accuracy of well-known machine learning algorithms - logistic regression, decision tree, random forest, SVM, KNN and gradient boosting algorithms to predict the liver cirrhosis stage using the dataset and compare the performance of these algorithms.

2. RELATED WORKS

Several machine learning models for instance Logistic Regression, J48, KNN, SVM, ANN, RF, GB, ANFIS, GANFIS and one proposed model were compared to predict diseases like cancer, diabetes and COVID-19.The results shows that the proposed model outperforms other models by 1.4765% and 1.2782 in accuracy and F-measure for COVID-19 dataset, 1.8274% and 1.7264 accuracy and F-measure for diabetes dataset and 1.7362% and 1.3821 accuracy and F-measure respectively, for heart disease dataset [9].

A convolutional neural networks (CNN) model is used to determine the presence of pneumonia from a collection of chest X-ray image samples. Results of this research give a training loss of 0.1288, training accuracy of 0.9531, validation loss of 0.1835 and validation accuracy of 0.9373 [10].

Three mathematical models, i.e., Logistic model, Bertalanffy model and Gompertz model is used to study the epidemic trends of SARS to predict the COVID-19 death toll in China and it is found that the Logistic model performed better than the other two models in Wuhan and non-Hubei areas [11].

Two machine learning models, artificial neural network(ANN) and support vector machine(SVM) are used to study individuals for prediabetes where an area under curve(AUC) of 0.768 is obtained for ANN and an AUC of 0.761 is obtained for SVM [12].

A deep learning model is used for classification of microarray cancer data which has a lot of complexity in processing. The model is a 7-layer deep neural network. The methods proposed in this study performs perfectly with an accuracy of 1.00 on four of the datasets- Leukemia, Lung-Michigan, Ovarian, and Prostate datasets. An accuracy of 0.99 is obtained on the Lung-Harvard dataset. Also an accuracy of 0.96 and 0.95 is obtained datasets-CNS and colon, and Breast cancer respectively [13].

A study to compare machine learning algorithms to predict liver disease shows that using random forest gives an accuracy of 83% whereas the precision, recall, and F1 scores of RF were 83.76%, 87%, 93.5%, and 90.1%, respectively [14].

A study employs classification and regression tree (CART) and case-based reasoning (CBR) techniques to build an intelligent system to raise the accuracy of liver disease prediction model. The results specify that the rate of accuracy for CART is 92.94%. The CBR diagnostic accuracy rate is 90.00%. The knowledge gained from CART is helpful to physicians in diagnosing liver diseases. CBR also proved to be helpful to physicians in identifying and solving a new liver disease [15].

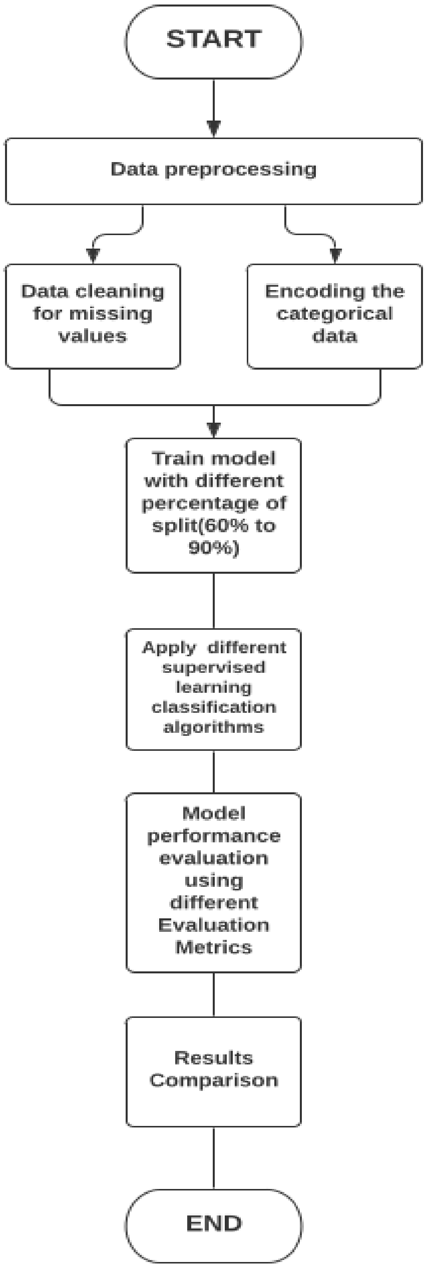
START
Data preprocessing
Data cleaning for missing values
Encoding the categorical data
Train model with different percentage of split(60% to 90%)
Apply different supervised learning classification algorithms
Model performance evaluation using different Evaluation Metrics
Results Comparison
END

Figure II. A flowchart to understand the approach of the study

3. MATERIALS AND METHODS

A detailed overview of how the project is carried out has been discussed in this section. Various important parts of building a machine learning model such as Data acquisition, Data preprocessing, feature selections and importance are also described. A flowchart of the study is shown in Figure II**.**

Software Requirements

- **Operating System -** Pop!_OS **, Release –** 21.04
- Python 3.8.3
- Jupyter Notebook **Version –** 6.0.3

Hardware Requirements

- **CPU –** Ryzen 5 3550H Quad core
- **RAM –** 8gb ddr4
- **Storage** - 20gb

3.1 Dataset

(The dataset has a total of 424 instances of PBC patients with 19 attributes and 1 outcome [16]. The data was collected in a ten-year interval from the Mayo Clinic. **Table I** describes the features with their mean value.

3.2 Data Preprocessing

The dataset that was collected had a lot of missing values in rows. Age was given in the number of days which was converted to years. Attributes such as Sex, Ascites, Hepatomegaly, Spiders and Edema had outputs in categorical variable that was encoded to Integer. A heatmap was visualized using seaborn and matplotlib to check the correlation of the independent variables with the "Stage" variable.

3.3 Data Splitting

Data was split into training and testing using the train_test_split function of the scikit learn library [17]. Data was split into 70:30 ratio for training: testing set.

Table 1: Information of the attributes and their properties

Attribute	Information
ID	unique identifier
N_Days	number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
Status	status of the patient C (censored), CL (censored due to liver tx), or D (death)
Drug	type of drug D-penicillamine or placebo
Age	age in [days]
Sex	M (male) or F (female)
Ascites	presence of ascites N (No) or Y (Yes)
Hepatomegaly	presence of hepatomegaly N (No) or Y (Yes)
Spiders	presence of hepatomegaly N (No) or Y (Yes)
Edema	presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)
Bilirubin	serum bilirubin in [mg/dl]
Cholesterol	serum cholesterol in [mg/dl]
Albumin	albumin in [gm/dl]
Copper	urine copper in [ug/day]
Alk_Phos	alkaline phosphatase in [U/liter]
SGOT	SGOT in [U/ml]
Triglycerides	triglycerides in [mg/dl]
Platelets	platelets per cubic [ml/1000]
Prothrombin	prothrombin time in seconds [s]
Stage	histologic stage of disease (1, 2, 3, or 4)

3.4 Algorithms

Random Forest Classifier

Random forest is based on the Decision tree algorithm. It learns through bootstrap aggregation, which is a machine learning ensemble meta-algorithm used to reduce variance in noisy data [18].

Logistic Regression

Logistic Regression is a classification algorithm that performs well for linearly separable classes. It is designed for two class classification but with the help of One-vs-Rest (OvR) classifier it can handle multi-class problem also. The output of logistic regression is binary dichotomous [19].

Decision Tree Classifier

Decision tree is a collection of decision nodes that originates from the root node. Each branch is connected to a node that is the possible outcome for a particular case [20].

K-Nearest Neighbor

K-Nearest Neighbor is a machine learning algorithm that is used for pattern recognition problems. It does a comparison between the available cases and the new case is allotted to the most similar one [21].

Artificial Neural Network

Artificial Neural Network is a deep learning model based on biological neurons that has weight and activation function instead of synapses and threshold respectively. It has three layers - input, hidden and output layer. The hidden layer further can be of multiple layers. It is where the processing takes place [22].

Gradient Boosting Machines

A gradient boosting machines is another decision tree algorithm, just like random forest. The process by which weak learners are converted into strong learners is known as boosting. In boosting, each new tree is a fit on a modified version of the original data set [23].

Naive Bayes Classifier

The Naive Bayes Classifier is based on Bayes' Rule. It is a probabilistic classifier. Naive Bayes directly estimates parameters based on the probability [24].

Support Vector Machines

SVMs are classification algorithms that are based on determining vectors that in turn creates a hyperplane. The classes are divided based on this hyperplane [25].

3.5 Model Evaluation Techniques

AUC-ROC Score

Area under the curve-receiver operating characteristics curve is a performance measurement tool for classification problems. It gives an idea of the capability of a model to distinguish classes [26].

Accuracy

Accuracy of a model is one of the most used evaluation technique. It tells us how accurate our model is in making predictions. Accuracy is calculated by – Correct predictions/ Total predictions.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

3.6 Results and Comparison

In Table II we compare the results of the algorithms using ROC-AUC curve and Accuracy for evaluation metrics. In the study we see that algorithm such as Logistic Regression and Random Forest performed better than the other algorithms. ROC-AUC score for all the models is above 0.50. The best performing model i.e. logistic regression has an accuracy of 0.55 and roc-auc score of 0.74. The second-best performing model is random forest with an accuracy of 0.49 and roc-auc score of 0.72. Models such as gradient boosting machines and naïve Bayes classifier also performed well with an accuracy of 0.50 and 0.35 respectively and a roc-auc score of 0.68 for both.

The reason why logistic regression performs better than the other models is due to its probabilistic approach and low variance. ANN performed poorly despite increasing the number of hidden layers. Random forests are well known for their performance for input data that are correlated and in biology domain that is why they performed quite well here also. Decision tree performed Poorly because the model was overfitting i.e. it performed well on the training dataset but poorly on the test data set. SVM also faced the similar problem. GBM also did great as it can reduce variance that is high in decision trees.

Table 2: Metrics Evaluation of different Algorithm using ROC-AUC score and Accuracy

Model Name	ROC-AUC Score	Accuracy
Random Forest Classifier	0.72	0.49
Logistic Regression	0.74	0.55
Decision Tree	0.55	0.39
KNN	0.49	0.35
ANN	0.5	0.42
GBM	0.68	0.50
Naïve Bayes Classifier	0.68	0.35
SVM	0.58	0.45

4 CONCLUSION

The study was based on comparing machine learning model to predict liver cirrhosis and its stage. We worked with various classification machine learning algorithms on our dataset. Despite having a lot of missing values and less instances in the datasets our models performed well .The comparison of our model showed us which model works best for this case and the reason behind it. Further studies can be conducted based on this study providing we have data with less noise and more instances. By this study we deduce that machine learning can help lessen the burden on the doctors and their patients. Artificial Intelligence is a growing field and if we dedicate proper research and tool to the healthcare sector we can help save a lot of lives.

5. REFERENCES

[1] L. Packard, "Anatomy and Function of the Liver," 2013. http://www.lpch.org/DiseaseHealthInfo/HealthLibrary/transplant/liverant.html (accessed Sep. 08, 2021).

[2] S. G. Sepanlou *et al.*, "The global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017," *Lancet Gastroenterol. Hepatol.*, vol. 5, no. 3, pp. 245–266, Mar. 2020, doi: 10.1016/S2468-1253(19)30349-8.

[3] The Global Health Observatory, "Liver cirrhosis, age-standardized death rates (15+), per 100,000 population," 2016. https://www.who.int/data/gho/data/indicators/indicator-details/GHO/liver-cirrhosis-age-standardized-death-rates-(15-)-per-100-000-population (accessed Sep. 08, 2021).

[4] "Definition & Facts for Cirrhosis | NIDDK." https://www.niddk.nih.gov/health-information/liver-disease/cirrhosis/definition-facts (accessed Sep. 08, 2021).

[5] Center for Disease Control and Prevention, "FastStats - Chronic Liver Disease or Cirrhosis," *National Center for Health Statistics*, 2016. https://www.cdc.gov/nchs/fastats/liver-disease.htm (accessed Sep. 08, 2021).

[6] NIH-NIDDK, "Symptoms & Causes of Cirrhosis | NIDDK," 2018. https://www.niddk.nih.gov/health-information/liver-disease/cirrhosis/symptoms-causes (accessed Sep. 08, 2021).

[7] The Global Health Observatory, "Alcohol-attributable fractions (15+), liver cirrhosis deaths (%)." https://www.who.int/data/gho/data/indicators/indicator-details/GHO/alcohol-attributable-fractions-(15-)-liver-cirrhosis-deaths-(-) (accessed Sep. 08, 2021).

[8] J. J. Heidelbaugh and M. Sherbondy, "Cirrhosis and Chronic Liver Failure: Part II. Complications and Treatment," *Am. Fam. Physician*, vol. 74, no. 5, pp. 767–776, Sep. 2006, Accessed: Sep. 08, 2021. [Online]. Available: www.aafp.org/afp.

[9] N. Kumar, N. Narayan Das, D. Gupta, K. Gupta, and J. Bindra, "Efficient Automated Disease Diagnosis Using Machine Learning Models," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/9983652.

[10] O. Stephen, M. Sain, U. J. Maduh, and D. U. Jeong, "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare," *J. Healthc. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/4180949.

[11] L. Jia, K. Li, Y. Jiang, X. Guo, and T. zhao, "Prediction and analysis of Coronavirus Disease 2019," *PLoS One*, vol. 15, no. 10 October, Mar. 2020, Accessed: Sep. 08, 2021. [Online]. Available: https://arxiv.org/abs/2003.05447v2.

[12] S. B. Choi *et al.*, "Screening for prediabetes using machine learning models," *Comput. Math. Methods Med.*, vol. 2014, 2014, doi: 10.1155/2014/618976.

[13] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," *CAAI Trans. Intell. Technol.*, vol. 5, no. 1, pp. 22–33, Mar. 2020, doi: 10.1049/TRIT.2019.0028.

[14] M. Ghosh *et al.*, "A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease," *Intell. Autom. Soft Comput.*, vol. 30, no. 3, pp. 917–928, 2021, doi: 10.32604/IASC.2021.017989.

[15] R. H. Lin, "An intelligent model for liver disease diagnosis," *Artif. Intell. Med.*, vol. 47, no. 1, pp. 53–62, Sep. 2009, doi: 10.1016/J.ARTMED.2009.05.005.

[16] "Cirrhosis Prediction Dataset | Kaggle." https://www.kaggle.com/fedesoriano/cirrhosis-prediction-dataset (accessed Sep. 08, 2021).

[17] "sklearn.model_selection.train_test_split — scikit-learn 0.24.2 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed Sep. 08, 2021).

[18] "Random forest classifier - Statistics for Machine Learning [Book]." https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/f35ce768-e55f-4a27-85cf-dfea0a7b92c0.xhtml (accessed Sep. 10, 2021).

[19] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. .

[20] "Chapter 8: Decision Trees - Discovering Knowledge in Data: An Introduction to Data Mining, 2nd Edition [Book]." https://www.oreilly.com/library/view/discovering-knowledge-in/9781118873571/c08.xhtml (accessed Sep. 10, 2021).

[21] O. Kramer, "K-Nearest Neighbors," pp. 13–23, 2013, doi: 10.1007/978-3-642-38652-7_2.

[22] K. Gurney and N. York, "An introduction to neural networks," 1997.

[23] "6. Gradient Boosting Machines - Practical Machine Learning with H2O [Book]." https://www.oreilly.com/library/view/practical-machine-learning/9781491964590/ch06.html (accessed Sep. 10, 2021).

[24] "CHAPTER 3 GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION Machine Learning 1 Learning Classifiers based on Bayes Rule," Accessed:

Sep. 10, 2021. [Online]. Available: www.cs.cmu.edu/~tom/mlbook.html.

[25] "1.4. Support Vector Machines — scikit-learn 0.24.2 documentation." https://scikit-learn.org/stable/modules/svm.html (accessed Sep. 10, 2021).

[26] S. Narkhede, "Understanding AUC - ROC Curve." https://www.48hours.ai/files/AUC.pdf (accessed Sep. 10, 2021).

BIOGRAPHY

Prakash Aryan, is an undergraduate student in the Department of Computer Science and Engineering pursuing his degree in Computer Engineering with specialization in Artificial Intelligence. His research interests are artificial intelligence, Machine learning, and Bioinformatics.