**Analysis and Prediction of Liver Disease for the Patients in India using various Machine Learning Algorithms**

U.Sinthuja1, 2 [0000-0001-5873-3459], Vaishali Hatti3, Dr.S.Thavamani 4

1, 4 Sri Ramakrishna College of Arts and Science, Coimbatore, Tamilnadu, India.

2 Hindusthan College of Arts and Science, Coimbatore, Tamilnadu, India.

3 Vivek Vardhini Public School, Aland, Karnataka, India.

sint@techie.com

**Abstract.** Predicting diseases in humans used to be an extremely time-consuming and complex procedure. It is now easier to save information and photographs due to the availability of multiple workstations and computers. Machine learning is vital in the healthcare sector because the number of liver patients is expanding on a big scale, so predicting liver illness at an early stage is essential to keep the patient from suffering more. The liver is a complicated organ located on the right side of our stomach that serves several key tasks in the human body. Many Machine Learning techniques are used in this study to categorize liver patient datasets, including Logistic regression, K-nearest neighbor, Decision tree, Random forest, AdaBoost, LightGBM, XGBoost, and Multilayer Perceptron. These techniques are employed in the frame model to cleanse the collected dataset by using Data pre-processing methodology, and Data visualization was used to visualize the null values and substitute duplicates.

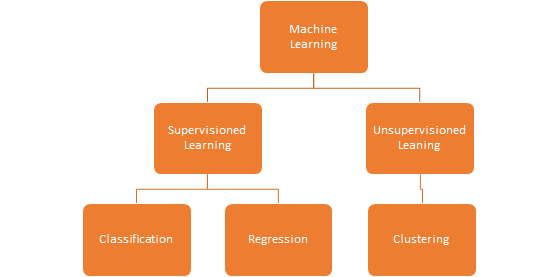
**Keywords:** Liver disease, Machine Learning Algorithms, Data pre-processing Methodology, Data Visualization Technique.

1. **Introduction**

Machine learning algorithms have advanced in recent years significantly in the healthcare industry, this is extremely significant. Diseases judgment based on even a clinical system. Several Machine learning is being used by both scholars and businesses, to help with clinical diagnosis. The Carried work concentrated liver disease identification, as the initial stage taken sample data has been preprocessed from selected features followed by the classification process done via some of the following Machine Learning Methods has been processed using python. The literature review also carried out to show up the ideas of various researchers. Finally, the findings demonstrated the accuracy of several machine learning algorithms in predicting liver illness.

* 1. **The Role Machine Learning in Liver Disease Perception**

Machine learning is a subset of Artificial Intelligence that enables computers to behave like humans and make decisions with no need for social interaction. Machine learning has made significant advances in the detection of various diseases like liver disease as a result of recent advances in Artificial Intelligence. Furthermore, machine learning technology allows us to make more correct estimates and improve our efficiency. Machine learning can be classified into several types, as indicated in the diagram (Fig 1) below.



**Fig 1** Machine Learning Algorithm Types

In this study, some of the algorithms were chosen and processed to predict liver disease. This has happened in both supervised and unsupervised learning categories of then machine learning.

### **1.2 COVID-19 and Acute Liver Disease**

Liver disease is also regarded as one of the world's most serious and deadly diseases. [1]Liver fibrosis, poor diet, cirrhosis of the liver, and HCV virus are some of the symptoms of liver disease, extreme alcohol consumption, drug usage, and hazardous and hereditary factors anomalies. There is no way to save a liver that is completely failing. There is only one method to regain it, is through liver transplant. It is extremely difficult to detect liver illness in its early stages, even when liver tissue has been moderately damaged; in these cases, many medical expert systems struggle to detect the sickness. It is critical to provide correct treatment in order to avoid this early prognosis and save the patient's life.

According to the Centers for Disease Control and Prevention, some COVID-19 patients exhibited high levels of liver enzymes. This indicates that a person's liver has been harmed, at least briefly, as a result of their sickness. Furthermore, patients with pre-existing liver illness who have been diagnosed with COVID-19 have a greater mortality rate than people who do not have pre-existing liver disease. It is vital to successfully this research right now in order to avoid future problems of the Patient with COVID-19.

1. **Literature Survey**

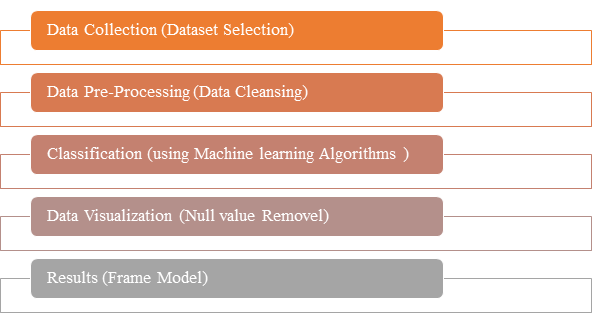
The review has been done for the past one decade by the various researchers which is clearly depicted in the table below.

**Table 1.** List of ML Algorithms and Its Accuracy value for past decade to predict liver disease

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Year & Ref No** | **Disease** | **Name of the ML Algorithm** | **Accuracy (%)** |
| 1 | 2011 [2] | Liver | C4.5, NB,  KNN, Backward  propagation,  & SVM | NB- 95.07, C4.5-96.27,  KNN-96.93, Backward  Propagation-97.47 & SVM-97.07 |
| 2 | 2012 [3] | Liver | Modified Rotation  Forest | MLP- 74.78  NN +CFS-73.07 |
| 3 | 2013 [4] | Liver | DT, NB,SVM &  ANN | DT-98.46 (Which has given higher Accuracy) |
| 4 | 2014 [5] | Liver cancer,  Hepatitis and  Cirrhosis | FT Tree , NB | FT Tree-72.66 NB-75.54 |
| 5 | 2015 [5] | Liver  Fibrosis | DT | DT-93.7 |
| 6 | 2016 [6] | Liver disease  Disorder | C4.5, BPNN, Regression ,NB,  SVM &DT | C4.5 Given Higher Accuracy |
| 7 | 2017 [7] | Liver disease | Back propagation & SVM | Back propagation-73.2 & SVM-71 |
| 8 | 2018 [8] | Liver Disease | KNN, ANN, Logistic  Regression &  SVM | Logistic regression- 73.23, KNN- 72.05, SVM-75.04, & ANN-  92.8 |
| 9 | 2019 [9] | Liver Disease | C4.5 & K-Means | C4.5- 94.36 (Highest Accuracy) |
| 10 | 2020 [10] | Liver Disease | SVM, J48 & NB | J48- 95.04 (Highest Accuracy) |

1. **Materials and Methods**

This study was conducted using the block diagram shown in Figure 2



**Fig 2** Block Diagram of the Study Model

### **3.1 Data Collection**

The Liver Patient Dataset was used to create the dataset (ILPD). This is from the UCL Machine Learning Repository, which you can find here. There are 567 instances and 10 attributes in this dataset. Age, gender, DB, TB, ALB, SGOT, SGPT, TP, ALP, and A/G ratio are all attributes.

**Table 2.**  List of Parameters & Its Data types

|  |  |  |
| --- | --- | --- |
| **S.No** | **Parameter name** | **Data type** |
| 1 | Age | Integer |
| 2 | Gender | String |
| 3 | tot\_bilirubin | Real |
| 4 | direct\_bilirubin | Real |
| 5 | tot\_proteins | Integer |
| 6 | Albumin | Integer |
| 7 | ag\_ratio | Integer |
| 8 | sgpt | Integer |
| 9 | Sgot | Real |
| 10 | Alkphos | Real |
| 11 | is\_patient | Integer |

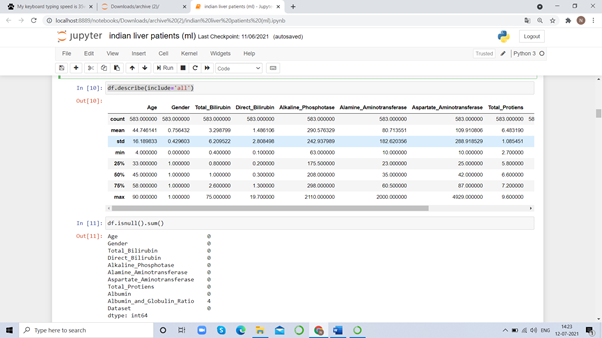
### **3.2 Data Pre-Processing**

**Imputation of missing values:** By viewing the data in seaborn or matplotlib, we can identify missing values and replace them with the mean (average) value, resulting in a high level of accuracy. If the number of null values (missing values) exceeds the number of null values, the contained column can be removed or feature engineering used.

**Label Encoding:** The data now contains some string values in the gender column that should be changed to integers to improve the analysis. Label encoding, a phase in data preprocessing that focuses on transforming string values to integers, is now used (Machine readable form).

**Elimination of duplicate values:** To increase the quality and efficiency of data, this technique involves the removal of unnecessary variables.

**Fig 3** Sample Screen for Dataset Process



The above Figure (Fig 3) has shown that the dataset processing technique for Feature selection (also known as variable selection, attribute selection, or variable subset selection) is a method of selecting the features in your data that contribute the most to the prediction (output). It simply means limiting the amount of input variables that do not contribute to the model. If the appropriate subset is chosen, it allows machine learning algorithms to train faster and enhances model accuracy. The data is now divided into two categories: independent (X) and dependent (y) features. In [11] NB has done research with researchers.

### **3.3 Classification using Machine learning Algorithms**

In [14] AI based work done by the authors, the data is divided into two categories: training data and testing data, which are used to train and test the model, respectively.[12] [13] The classification is done using a variety of machine learning algorithms which are discussed below.

**Logistic Regression:** The supervised learning classification technique logistic regression, also known as logit regression or logit model, is used to predict the likelihood of a categorical dependent variable. There are several types of logistic regression (for example, binary logistic regression, multinomial logistic regression, and ordinal logistic regression), but because the dependent variable in the Indian liver patient dataset is a binary variable with data coded as 1 or 0, binary logistic regression is used to train the model[15].

**K-Nearest neighbor (KNN):** KNN is a supervised Machine learning algorithm that can perform both classifier and regression tasks using numbers (K) of neighbors (instances). Before the implementation of KNN the given data must be preprocessed (i.e... Imputation, Label Encoding, and Elimination of duplicate values, Resampling, Outlier detection and Feature selection) which means data should be balanced. Next the data is divided as Independent and Dependent features.

# Formulae to measure the distance between two points

# Euclidean Distance Formula

# d(x,y) = √(x1 – y1) + (x2 – y2)

But other measures can be more suitable for a given setting that include the Manhattan, Chebyshev , and Hamming distance. KNN is used to train a model to predict which class or feature a new data point belongs to base on a value of K (it is preferable to consider the value of k as an odd number; by default, it is 5). The distance between the new data point and the rest of the data points. K closest numbers are chosen, and the new data point is assigned to the class or feature based on the chosen points.

**Decision Tree classifier: Decision** trees are a type of supervised Machine learning in which data is continuously separated based on a parameter. Two entities, nodes and leaves, can be used to explain the tree. The decision nodes are where the data is split, and the leaves represent decisions or ultimate outcomes. Two mathematical equations, Gini Impurity and Entropy, are used to apply decision trees for training any model [16]

**Random Forest classifier:** Random forests, also known as random decision forests, are an ensemble learning method for classification and regression that uses random samples from training data to create numerous decision trees. Random forest combines the results of numerous decision trees (each of which trains a different observation) to provide a more accurate and reliable forecast [17].

**AdaBoost Ensemble Technique:** AdaBoost, short for Adaptive Boosting, is a boosting approach used in machine learning as an ensemble method. It works on the premise of converting weak learners into strong ones. During the training of data, it creates n-number of decision trees. The record that was mistakenly classified during the initial model is given additional priority as the first decision tree is constructed. Only these records are sent to the second model as input. The procedure will continue until we designate how many base learners we wish to produce.

**LightGBM:** LightGBM is a high-performance gradient boosting framework based on decision tree techniques that may be used for ranking, classification, and a variety of other Machine Learning applications. It divides the tree leaf by leaf (instead of depth wise). It may result in over fitting, which can be reduced by setting the splitting depth.

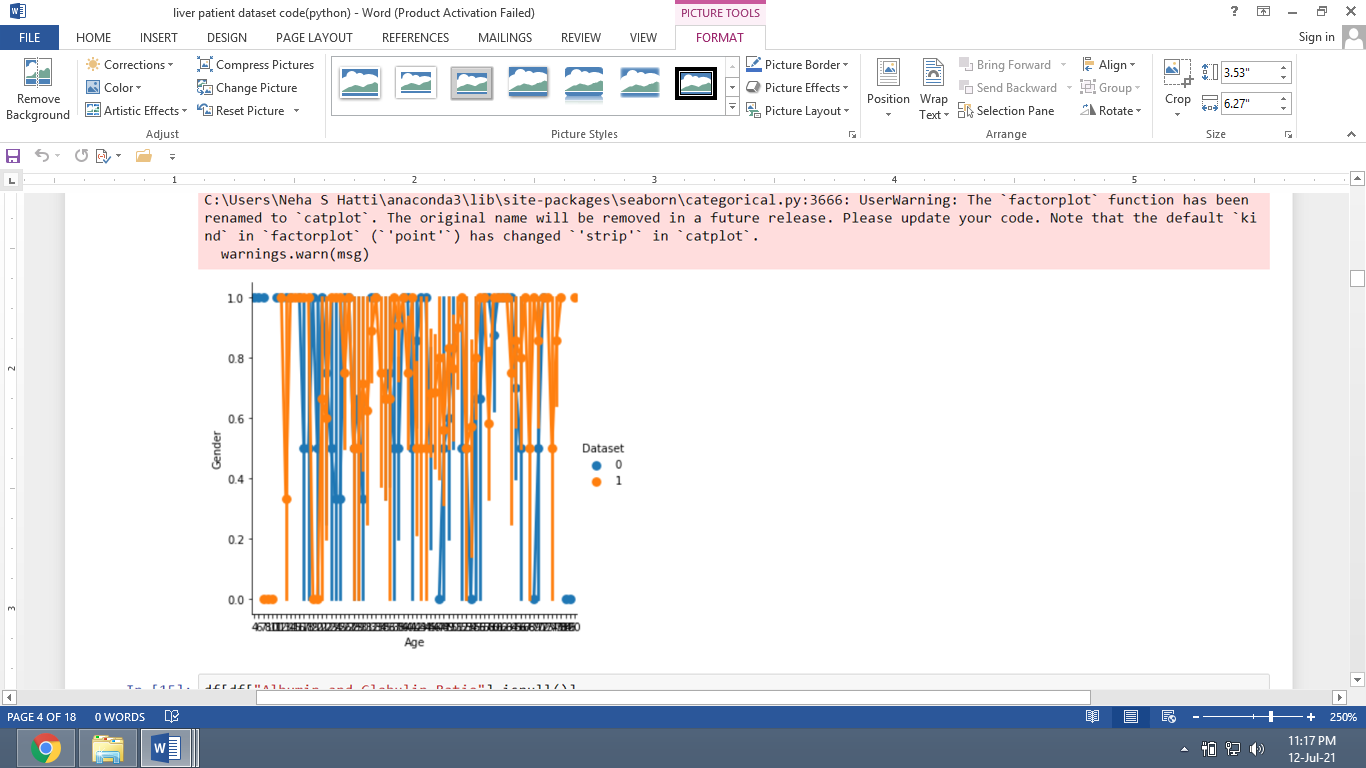
**XGBoost:** XGBoost (eXtreme Gradient Boosting) is a gradient boosting framework-based decision tree-based ensemble ML method. To improve the algorithm, the processing is done in parallel. In [18] [19] given importance to security in communication technologies.

**Multilayer Perceptron:** It is a feed forward neural network augmentation. It is made up of three different layers. –

* The input signal to be processed is received by the input layer.
* The required task, such as prediction and classification, is executed in the output layer.
* An arbitrary number of hidden layers are inserted between the input and output layers in a hidden layer.

### **3.4 Data Visualization**

After the process of the dataset python will be displaying the processed data as shown in the below figure. Here age and gender data has been visualized as an example.



**Fig** **4** Sample Screen for **Age** and Gender Visualization of Dataset 0&1

### **3.5 Results and Discussion**

The Accuracy of each classification Machine Learning Algorithm result has been given below after processing the dataset of the liver disease patient which has been taken as an example.

**Table 3:** The Results of Proposed Classification Algorithms

|  |  |  |
| --- | --- | --- |
| **S.No** | **Proposed Machine Learning Algorithms** | **Accuracy  (Out of 1.00)** |
| 1 | 1. Logistic Regression | 0.70 |
| 2 | 1. K-NN | 0.78 |
| 3 | 1. Decision Tree | 1.00 |
| 4 | 1. Random Forest | 1.00 |
| 5 | 1. AdaBoost | 0.88 |
| 6 | 1. Light GBM | 1.00 |
| 7 | 1. Multilayer Perceptron | 0.70 |
| 8 | 1. XGBoost | 1.00 |

The table above shows that out of 11 classification-based machine learning algorithms studied, Decision Tree, Random Forest, Light GBM, and XGBoost generated the best results. According to these comparative studies, the best result supplied algorithms can be applied to improve prediction outcomes.

1. **Conclusion**

This research summarises previous research on the identification and diagnosis of liver disease using various machine learning algorithms. This survey and study definitely found and noticed that several machine learning algorithms, used to provide superior accuracy in detecting and predicting liver illness. Different algorithms perform differently in different scenarios, but the dataset and feature selection are also crucial in obtaining superior prediction outcomes.Various procedures, such as imputation of missing values with the mean value and conversion of categorical data to numerical data, are used to clean the data. To forecast the existence or absence of liver disease, a variety of algorithms are used. In the future based on the prediction results of the liver disease can lead to finding the way for treatment by the clinician, further assistance can also be done in future.

**References**

[1]. D.A. Saleh F. Shebl M. Abdel-Hamid et al. "Incidence and risk factors for hepatitis C infection in a cohort of women in rural Egypt"Trans. R. Soc. Trop. Med. Hyg. vol. 102 pp. 921928 2008. <https://doi.org/10.1016/j.trstmh.2008.04.011>.

[2] Ramana, Bendi Venkata, M. Surendra Prasad Babu, and N. B. Venkateswarlu. "A critical study of selected classification algorithms for liver disease diagnosis." International Journal of Database Management Systems 3.2 (2011): 101-114.

[3] Ramana, Bendi Venkata, MS Prasad Babu, and N. B. Venkateswarlu. "Liver classification using modified rotation forest." International Journal of Engineering Research and Development 6.1 (2012): 17-24.

[4] Kumar, Yugal, and G. Sahoo. "Prediction of different types of liver diseases using rule based classification model." Technology and Health Care 21, no. 5 (2013): 417-432.

[5] Ayeldeen, Heba, Olfat Shaker, Ghada Ayeldeen, and Khaled M. Anwar. "Prediction of liver fibrosis stages by machine learning model: A decision tree approach." In 2015 Third World Conference on Complex Systems (WCCS), pp. 1-6. IEEE, 2015.

[6] Sindhuja, D., and R. Jemina Priyadarsini. "A survey on classification techniques in data mining for analyzing liver disease disorder." International Journal of Computer Science and Mobile Computing 5.5 (2016): 483-488.

[7] Sontakke, S., Lohokare, J., & Dani, R. (2017, February). Diagnosis of liver diseases using machine learning. In 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI) (pp. 129-133). IEEE.

[8] Jacob, Joel, Joseph Chakkalakal Mathew, J. Mathew, and E. Issac. "Diagnosis of liver disease using machine learning techniques." Int Res J Eng Technol 5, no. 04 (2018).

[9] Sivakumar D , Manjunath Varchagall , and Ambika L Gusha S “Chronic Liver Disease Prediction Analysis Based on the Impact of Life Quality Attributes.” (2019). International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume7, Issue-6S5, April 2019

[10] Durai, Vasan, Suyan Ramesh, and Dinesh Kalthireddy. "Liver disease prediction using machine learning." (2019).

[11] S. Gupta, M. Sameer and N. Mohan, "ROC Analysis of EEG Subbands for Epileptic Seizure Detection using Naïve Bayes Classifier" 2021 Vol 17, Issue 1-3,Journal of Mobile Multimedia, doi:<https://doi.org/10.13052/jmm1550-4646.171315>

[12] Sameer, M., Gupta, B. Detection of epileptical seizures based on alpha band statistical features. Wireless Pers Commun 115, 909–925 (2020). <https://doi.org/10.1007/s11277-020-07542-5>

[13] S. Gupta, M. Sameer and N. Mohan, "Detection of Epileptic Seizures using Convolutional Neural Network," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 786-790, doi: 10.1109/ESCI50559.2021.9396983

[14] A. Mahajan, K. Somaraj and M. Sameer, "Adopting Artificial Intelligence Powered ConvNet To Detect Epileptic Seizures," 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 427-432, doi: 10.1109/IECBES48179.2021.9398832.

[15 ]M. Sameer, A. K. Gupta, C. Chakraborty and B. Gupta, "Epileptical Seizure Detection: Performance analysis of gamma band in EEG signal Using Short-Time Fourier Transform," 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC), 2019, pp. 1-6, doi: 10.1109/WPMC48795.2019.9096119.

[16] M. Sameer and B. Gupta, "Beta Band as a Biomarker for Classification between Interictal and Ictal States of Epileptical Patients," 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), 2020, pp. 567-570, doi: 10.1109/SPIN48934.2020.9071343

[17] M. Sameer, A. K. Gupta, C. Chakraborty and B. Gupta, "ROC Analysis for detection of Epileptical Seizures using Haralick features of Gamma band," 2020 National Conference on Communications (NCC), 2020, pp. 1-5, doi: 10.1109/NCC48643.2020.9056027.

[18] M. et al., “EFFICIENT EMPLOYMENT OF THE MQTT AND S-MQTT PROTOCOL IN IOT NETWORK”, IJAST, vol. 28, no. 17, pp. 858 - 864, Dec. 2019.

[19] Sinthuja et al “Evaluating systems and tools for vulnerability study on multi-broker MQTT

Instances”, 2020 GEDRAG & ORGANISATIE REVIEW, VOLUME 33: ISSUE 04 – 2020, doi-<https://www.doi.org/10.37896/GOR33.04/070>