

Liver Disease Diagnosis Using Machine Learning Algorithm

Prof. Sayalee Deshmukh¹, Pratiksha Kawale², Manasi Khopade³, Anushka Sawant⁴, Yashika Palan³

¹Professor, Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune

^{2,3,4,5}Student, Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune

Abstract - Liver disease is one of the most terrifying diseases. This disease is caused by a combination of factors that harm the liver. Obesity, an undiagnosed hepatitis infection, alcohol abuse, which causes abnormal nerve function, coughing up or vomiting blood, kidney failure, liver failure, jaundice, liver encephalopathy, and many other conditions are examples. Early detection of a liver infection is critical for effective treatment. Because of the subtle symptoms, medical researchers face a difficult task in predicting the disease in its early stages. Symptoms frequently appear when it is too late. To address this issue, this project will use machine learning approaches to improve liver disease diagnosis. The primary goal of this study is to use a classification algorithm to distinguish between liver patients and healthy people. Based on chemical compounds (bilirubin, albumin, proteins, alkaline phosphatase) found in the human body and tests such as SGOT and SGPT, the outcome indicates whether a person is a patient who requires diagnosis or not. Patients with liver disease have been steadily increasing as a result of excessive alcohol consumption, inhalation of harmful gases, and consumption of contaminated food, pickles, and drugs. This project's goal is to analyze prediction algorithms in order to reduce doctors' workload.

Key Words: Liver disease, Liver Function Test (LFT), ML, Random Forest, python, etc.

1. INTRODUCTION

The liver part of the body is bulky and brawny, located on the right side of the abdomen, and is reddish-brown in colour and rubbery to the touch. The liver is normally inaccessible because it is protected by the rib cage. The liver is divided into two large sections known as the right and left lobes. The gallbladder is located beneath the liver and contains elements of the pancreas and intestines. The liver organ and its supporting organs work together to digest and process food. The liver gland is a very important organ that plays an important role in the body and weighs about 3 pounds. It performs numerous tasks that aid in the digestion of food, the conversion of food into energy, and the storage of energy. It also aids in the removal of noxious substances from the bloodstream. Hepatitis, fatty liver, bleeding, fatigue, and jaundice are all diseases that can affect the liver. Fatty liver infection occurs when fat accumulates around the liver.

The liver is difficult to detect early because it will continue to function normally even if it is partially destroyed. If a patient is diagnosed early, their chances of surviving liver disease improve. Indians are more likely to suffer from liver

failure. By 2025, India is expected to be the World Capital for Liver Diseases. A deskbound lifestyle, increased alcohol consumption, and smoking are all factors that contribute to the prevalence of liver infection in India. Over a hundred different types of liver infections exist. As a result, developing a machine to aid in disease identification will be extremely beneficial in the medical field. These technologies will help physicians make accurate patient decisions, and the use of automatic classification tools for liver illnesses (likely mobile or web enabled) will reduce patient wait times at liver experts such as endocrinologists.

Medical diagnosis and disease prediction rely heavily on classification techniques. According to Paul R Harper, there is no single best classification tool; instead, the best-performing method is determined by the characteristics of the dataset. The primary goal of this research is to use classification algorithms to differentiate between liver patients and healthy people. The Random Forest (RF) algorithm is implemented as a user-friendly Graphical User Interface (GUI) in Python using Flask in this study. The GUI is simple for doctors and medical practitioners to use as a screening tool for liver disease. The Indian Liver Patient Dataset (ILPD) was chosen from the UCI Machine Learning repository for this work. It is a representative sample of the Indian population as a whole.

2. LITERATURE SURVEY

Every day, the health-care and pharmaceutical industries deal with massive amounts of data. This information includes patient records, prognosis reviews, and clinical photographs. This fact must be used to decipher a decision-assistance system. To accomplish this, the knowledge domain must be discovered and extracted from the raw data. Knowledge discovery and data mining (KDD) [1] is used to do this. In the biological area, the use of machine learning techniques is widespread. Liver problems have vastly improved in recent years, and liver disease is now one of the most dangerous conditions in a number of countries. In this study, liver patient records are examined with the purpose of developing classification models to predict liver disease. For enhancing prediction accuracy of Indian liver patients, several function model development and comparison analysis are carried out. For the classification of liver illnesses, various research have been conducted. One of the most broad and relevant statistics mining approaches used in sickness prediction is the classification algorithm. In a variety of automatic medical health diagnostics, the classification algorithm is the most popular. Many of them have a high degree of categorization accuracy.

In this paper [2], different machine learning algorithms are used such as the methods of Random Forest (RF), Support Vector Machines (SVM), Decision Tree (DT) are proposed to predict liver disease with better precision, accuracy and reliability. [3] The important goal of this paper is to predict liver ailments the usage of different classification algorithms. These classification algorithms are compared primarily based on the overall performance.

In [4] this work to build the machine-learning model, Indian Liver Patient Dataset is used, which is based totally on Indian patient and Random Forest (RF) algorithm is used to predict the sickness with different pre-processing techniques. Data set is checked for skewness, outliers and imbalance the usage of univariate and bivariate evaluation and then appropriate algorithms used to remove outliers and various oversampling and under sampling strategies are used to stability the data. Further refinement of model is completed through hyper parameter tuning the use of grid search and function selection. The closing model presents 100% accuracy and additionally right score across different metrics.

[5] The lab test findings of patients who have had a Liver Function Test are used by Dr. Vijayalakshmi M.N. The classification techniques SVM, Logistic Regression, and Decision Tree are utilised in MATLAB2016 to construct a model. The accuracy of the Logistic Regression was 95.8%. The existence of a problem in the liver is determined by drawing a graph that tests various predictions.

To forecast liver disease, Nazmun Nahar and Ferdous Ara et al. [6] used decision tree algorithms J48, LMT, Random Forest, REPTree, Decision Stump, and Hoeffding Tree. A comparative examination of these algorithms has also been carried out. The system evaluates all algorithms' performance by calculating their accuracy, F-measure, precision, recall, and recommended absolute error. [7] With feature selection, this research examines various classification models and visualisation techniques used to predict liver disease. The prediction of liver disease has been studied and analysed. The greatest qualities necessary for liver disease prediction are fetched using a genetic algorithm paired with XGBoost. [8] This paper proposes a deep-learning-based system for segmenting vacuoles in liver pictures, as well as a study of the automated quantification's association with expert pathologist's manual evaluation.

3. METHODOLOGY

This research work comprises of following phases:

Phase I: Collection and Pre-processing of Data

Patients who are predicted of infected liver or liver disorder through the initial clinical tests are considered for Liver Function Test (LFT). Various parameters that are focused in LFT are Total-Bilirubin, Direct-Bilirubin, Alkaline Phosphatase Alanine Aminotransferase (ALT), Aspartate Aminotransferase (AST), Total, Proteins, Albumin, Globulin, Albumin and Globulin Ratio. These parameters help in prediction of liver disease.

The liver patients are classified from non-liver patient by using above different parameters. There are 10 parameters used to predict the liver disease. The outline/description of attributes are as follows:

Age: Age (in years) of patient

Gender: Gender (Male/Female) of patient.

Total Bilirubin: It is a term that any form of a yellowish pigment made in the liver when crimson blood cells are broken down and normally excreted with the bile. Bilirubin can be classified as direct bilirubin and indirect bilirubin.

Direct Bilirubin: Indirect bilirubin is shaped by way of the breakdown of hemoglobin in the red blood cells. The liver converts this bilirubin into direct bilirubin, which can then be launched into the gut with the aid of the gallbladder for elimination.

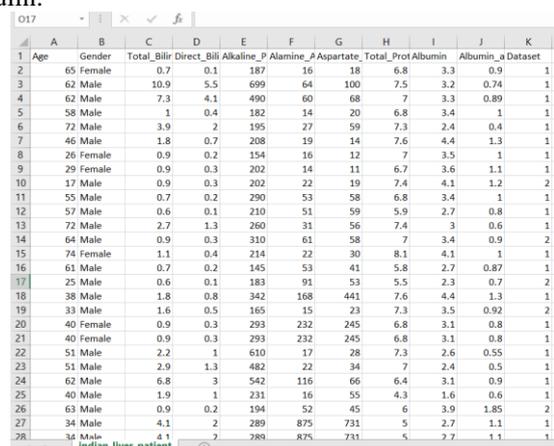
Alkaline phosphatase: It is also referred to as Alk Phos. It is an enzyme discovered in bones and liver. An enzyme is some thing that helps velocity up a chemical reaction in body. This blood check measures the amount of alkaline phosphatase in the blood.

Alanine Aminotransferase (ALT) and Aspartate Aminotransferase (AST): These enzymes, formerly known as SGOT and SGPT, are normally found in liver cells that leak out of cells and mixes in blood when liver cells gets injured. The ALT is a more specific indicator of liver inflammation as AST can also be found in other organs like heart and skeletal muscles.

Total proteins: The total protein test measures the total amount of two classes of proteins found in the fluid portion of blood. These are albumin and globulin. Proteins are important parts of all cells and tissues.

Albumin: It is a major protein which is formed by the liver. It helps to keep fluid in bloodstream so it doesn't leak into other tissues. It also carries various substances throughout the body, including hormones, vitamins, and enzymes. Low albumin levels can indicate a problem with our liver or kidneys.

Albumin and Globulin Ratio: It is a ratio of albumin and globulin.



Age	Gender	Total_Bilir	Direct_Bilir	Alkaline_P	Alanine_A	Aspartate_A	Total_Prot	Albumin	Albumin_a_Dataset
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74
62	Male	7.3	4.1	490	60	68	7	3.3	0.89
58	Male	1	0.4	182	14	20	6.8	3.4	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3
26	Female	0.9	0.2	154	16	12	7	3.5	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2
55	Male	0.7	0.2	290	53	58	6.8	3.4	1
57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8
72	Male	2.7	1.3	260	31	56	7.4	3	0.6
64	Male	0.9	0.3	310	61	58	7	3.4	0.9
74	Female	1.1	0.4	214	22	30	8.1	4.1	1
61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87
25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7
38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3
33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92
40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8
40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8
51	Male	2.2	1	610	17	28	7.3	2.6	0.55
51	Male	2.9	1.3	482	22	34	7	2.4	0.5
62	Male	6.8	3	542	116	66	6.4	3.1	0.9
40	Male	1.9	1	231	16	55	4.3	1.6	0.6
63	Male	0.9	0.2	194	52	45	6	3.9	1.85
34	Male	4.1	2	289	875	731	5	2.7	1.1
34	Male	4.1	2	289	875	731	5	2.7	1.1

Fig 1: ILPD Dataset

The Indian Liver Patient Dataset (ILPD) contains records of 583 patients with 416 liver patient records and 167 non-liver patient records. The patients were described as either 1 or 2 on the basis of liver disease. The last column, Dataset, is the label (with '1' representing presence of disease and '2' representing absence of disease). All the features except gender are real valued integers. The feature gender is converted to numeric value (0 and 1) in the data pre-processing step.

The pre-processing is carried out in following steps:

- **Cleansing:** The size of the dataset is reduced by refining unwanted and repeated data.
- **Integrating:** Structured and unstructured data formats are unified.
- **Redundancy Elimination:** Unnecessary data is removed and only necessary attributes are extracted which results in reduction of data size.
- **Transforming:** The data is transformed into scaled values so that it can fit within minimal range.
- **The columns which contain null values are replaced with mean values of the column.**

Following figure shows a sample image of Liver Function Test (LFT) report of a person which user can upload to the system and system takes the values of these parameters automatically using Optical Character Recognition (OCR) Tool that is pytesseract. Pytesseract (Python Tesseract) is a OCR tool for python. It will recognize and read the text embedded in images and writes text to a file.

<u>TESTS</u>	<u>RESULTS</u>
LIVER FUNCTION TEST	
Bilirubin- Total	: 5.90
Bilirubin- Direct	: 1.13
Bilirubin- Indirect	: 4.77
SGPT	: 54
SGOT	: 62
Alkaline Phosphatase	: 124
Total Protein	: 6.5
Albumin	: 3.7
Globulin	: 2.8

Fig 2: Sample image of LFT report

Phase II: Dataset Preparation and Classification

Data extracted from source is organized as collection of data or individual analytical data. Each attribute of data serves as variable and every instance has individual characterization. Liver disease is predicted using ILPD dataset. The ILPD dataset is created by adapting the methods of data collection and pre-processing.

The LFT dataset helps us in diagnosing the disorder based on its parameters. In this proposed work ILPD dataset with 11 attributes are considered for obtaining accurate results. Each attribute should lie between the specified threshold ranges. If the attribute values deviated from the specified threshold, the liver disorder is predicted.

Classification is a process that comprises of identification of problem statement. The characteristics of the liver disease among the patients are diagnosed through the random forest algorithm. Attributes considered in the dataset are: Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase Alanine Aminotransferase, Aspartate Aminotransferase, Total, Proteins, Albumin, Albumin and Globulin Ratio, Dataset (Liver Disease prediction).

Table 1: Liver Function Test; U/L – units per litre; g/dL-grams per deciliter.

Liver Function Test	Reference Range
Bilirubin- Total	0.1-0.3 mg%
Bilirubin- Direct	0-0.3 mg%
Alkaline Phosphatase	64-306 U/L
Alanine Aminotransferase (SGPT)	8-40 IU/L
Aspartate Aminotransferase (SGOT)	8-35 IU/L
Total Protein	5.5-7.5 gm/dl
Albumin	3.0-5.2 gm/dl
Globulin	2.3-3.5 gm/dl
Albumin and Globulin Ratio	0.8-2.0

Phase III: Classification Algorithm: Random Forest

A supervised classification technique, the Random Forest algorithm. We can tell from the name that the goal is to produce a forest in a way that is random. The amount of trees in the forest has a direct link with the accuracy of the results: the more trees, the more exact the result. However, establishing the forest is not the same as building the decision using the information gain or gain index technique. They're decision trees built by classifiers for training input. For dividing the tree, a random value is assigned to feature space as range. The modal value of a separate tree is anticipated based on the training ensemble class value. It creates a forest out of an ensemble of decision trees, which are normally trained using the bagging approach. The bagging method's fundamental concept is to use a variety of learning models to improve the overall output. By estimating out-of-bag error, the approach is used to rank the features. The relevant score for each characteristic is then calculated.

```

from sklearn.ensemble import RandomForestClassifier
import sklearn.metrics
random_forest = RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, y_train)
# Predict Output
rf_predicted = random_forest.predict(X_test)

random_forest_score = round(random_forest.score(X_test, y_test) * 100, 2)
print("Random Forest accuracy: ", random_forest_score)
confusion_matrix=sklearn.metrics.confusion_matrix(y_test,rf_predicted)
classification_report=sklearn.metrics.classification_report(y_test,rf_predicted)
print("\nconfusion matrix\n",confusion_matrix)
print("\nclassification_report\n\n",classification_report)

Random Forest accuracy: 69.5

confusion matrix
[[ 9 31]
 [12 89]]

classification_report

              precision    recall  f1-score   support

0               0.43        0.23        0.30         40
1               0.74        0.88        0.81        101

 accuracy avg         0.59        0.55        0.55        141
 macro avg          0.59        0.55        0.55        141
 weighted avg       0.65        0.70        0.66        141
    
```

Fig 3. Random Forest Implementation

Fig. 4(b) After uploading image user have to check all the values of all parameters and if any value is missed user have to enter that and submit it. And if user have not uploaded the image then user have to enter all the values and then submit it.

Liver Function Test		Reference Range
Bilirubin Total		0.1-0.3 mg/dl
Bilirubin-Direct		0.0-0.2 mg/dl
Alkaline Phosphatase		66-168 U/L
Alanine Aminotransferase (SGPT)		1-40 IU/L
Aspartate Aminotransferase (SGOT)		1-37 IU/L
Total Protein		5.5-7.7 g/dl
Albumin		3.8-5.2 g/dl
Globulin		2.3-3.3 g/dl
Albumin and Globulin Ratio		0.8-2.0

Fig 4(c)

Fig 4: Development of GUI

Fig. 4(c) On this page user can see the result and normal rang of all the parameters.

4. DEVELOPMENT OF GUI

Graphical User Interface (GUI) is a visual way of interacting with a computer. A GUI extant items that deliver information, and signify actions which can be taken by the user. GUI is created using Flask and html and implemented using Random Forest. Three GUIs are created, one for uploading image of LFT report, one for predicting and the other for training new data. The GUI contains input fields for all attributes in the dataset. The system will predict whether the patient’s liver is infected or not based on the trained model. The GUI will be useful tool for medical staff in the early diagnosis of liver disease in patients. Pictures of the developed GUI are shown below.

5. CONCLUSION AND FUTURE WORK

Diseases of the liver and heart are becoming more common over time. With continuous technological advancements, these will only grow in the future. Although people are becoming more health conscious and enrolling in yoga and dance classes, the sedentary lifestyle and luxuries that are constantly being introduced and enhanced will continue to be a problem. In this case, our project will be extremely beneficial to society. In this paper, we present a system that predicts liver disease using the Indian Liver Patient Dataset and images of Liver Function Test reports.

Today, almost everyone over the age of 12 owns a smartphone, so we can incorporate these solutions into an Android or iOS app. It can also be integrated into a website with commercial hosting and storage space, and both the apps and the website will be extremely beneficial to a large segment of society.

REFERENCES

[1] Hastie T, Robert, T, Jerome F (2009). The Elements of Statistical Learning: Data mining, Inference and Prediction. Springer. 485–586.

[2] A.Sivasangari, Baddigam Jaya Krishna Reddy, Annamareddy Kiran, P.Ajitha, “Diagnosis of Liver Disease using Machine Learning Models,” IEEE Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020.

[3] M.Ardrha Meghana Simon, N. Kalyan Saradhi, P.Sahithi, R. Sai Kiran, Mrs. A. Aparna, “Liver Disease Prediction using Machine Learning”, A Journal Of Composition Theory, Volume XIV, Issue VI, JUNE 2021.

[4] Sateesh Ambesange, Vijayalaxmi A, Rashmi Uppin, Shruthi Patil, Vilaskumar Patil, “Optimizing Liver disease prediction with Random Forest by various Data balancing



Fig. 4(a)

Fig. 4(a) Here user can upload image of LFT report or can skip this page.

Liver Disease Prediction

Age:

Gender:

Total_Bilirubin:

Direct_Bilirubin:

Alkaline_Phosphotase:

Alamine_Aminotransferase:

Aspartate_Aminotransferase:

Total_Proteins:

Albumin:

Albumin_and_Globulin_Ratio:

Fig. 4(b)

Techniques”, IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2020.

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

[6] 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, March 2018.

[7] Maria Alex Kuzhippallil, Carolyn Joseph, and Kannan A, “Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients,” 6th International Conference on Advanced Computing & Communication Systems (ICACCS), 2020.

[8] Sanket Deshmukh, Avinash Lokhande, Ratul Wasnik, and Nitin Singhal, “Vacuole Segmentation and Quantification in Liver Images of Wistar Rat,” 978-1-7281-1990-8/20/\$31.00 IEEE, 2020.

[9] Hartatik, Mohammad Badri Tamam, Arief Setyanto, “Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms”, 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020.

[10] R. Kalaiselvi, K. Meena, V. Vanitha, “Liver Disease Prediction Using Machine Learning Algorithms”, International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Oct 2021.

[11] G. Shobana, K. Umamaheswari, “Prediction of Liver Disease using Gradient Boost Machine Learning Techniques with Feature Scaling”, 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021.

[12] Rahul Katarya, Polipireddy Srinivas, “Predicting Heart Disease at Early Stages using Machine Learning: A Survey”, International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.

[13] Abderrahmane Ed-daoudy, Khalil Maalmi, “Real-time machine learning for early detection of heart disease using big data approach”, 2019.

[14] Rahma Atallah, Amjed Al-Mousa, “Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method”, 978-1-7281-2882-5/19/\$31.00 IEEE, 2019.

[15] Gautam Chitnis, Vidhi Bhanushali, Aayush Ranade, Tejasvini Khadase, Vaishnavi Pelagade, Jitendra Chavan, “A Review of Machine Learning Methodologies for Dental Disease Detection”, IEEE India Council International Subsections Conference (INDISCON), 2020.

[16] Yedilkhan Amirgaliyev, Shahriar Shamiluulu, Azamat Serek, “Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods”, 2018.

[17] Aida Brankovic, Ali Zamani, Amin Abbosh, “Electromagnetic Based Fatty Liver Detection Using Machine Learning”, 13th European Conference on Antennas and Propagation (EuCAP), 2019.