

# Machine Learning-Based Approach to Predict Liver Disease Using an Imbalanced Dataset

Abid Bin Ahosan  
Dept. of Computer Science and  
Engineering  
Dhaka International University  
Dhaka, Bangladesh  
abidbinahosan@gmail.com

Khandaker Mohammad Mohi  
Uddin  
Dept. of Computer Science and  
Engineering  
Southeast University  
Dhaka, Bangladesh  
jilanicsejnu@gmail.com

Forhadul Islam  
Dept. of Computer Science and  
Engineering  
Dhaka International University  
Dhaka, Bangladesh  
forhad92149@gmail.com

Fatema Tuj Janin  
Dept. of Computer Science and  
Engineering  
Southeast University  
Dhaka, Bangladesh  
jeninahmed20@gmail.com

Md Abdul Based  
Dept. of Computer Science and  
Engineering  
Dhaka International University  
Dhaka, Bangladesh  
based@diu.ac

**Abstract**— Liver Disease describes a condition that damages the liver and can cause issues with its ability to operate. Liver disease affects an enormous percentage of people globally and it is currently one of the most common causes of death. It is important to diagnose liver disease as soon as possible in order to save human life and implement the necessary control measures. In this study, the emphasis lies on data pre-processing and feature extraction leading to improved performance for any machine learning classifier to predict liver disease. An array of Machine Learning models and ensemble approaches were assessed in the proposed system. Based on the outcomes from the study, it can be concluded that the Light Gradient Boosting Machine (LGBM) classifier perform better than the remaining models. It achieved 83.74% accuracy, 78% precision, 85% recall, 81% F1-score, and 90.3% AUC after utilizing the Synthetic Minority Oversampling Technique (SMOTE).

**Keywords**—Liver Disease, SMOTE, Feature Selection, LGBM

## I. INTRODUCTION

Liver is the largest and powerful organ that performs hundreds of essential functions in our body. It helps in the removal of toxins, energy storage, and the digestion of food. It is a major multitasker. Situated close to the diaphragm in the right upper quadrant of the abdominal cavity, this triangular organ has a reddened-brown appearance. It weighs around 1.6 kg (3.5 pounds) on average and has 500 processes, both on its own and in relation to other bodily systems. Positioned on the right side of the body, the bigger mass of the liver, referred to as the right lobe, descends towards the right kidney. In the meantime, the abdomen is close to the location of the narrower left lobe. The liver portal vein and the liver artery are the main sources of the kidney's plentiful supply of blood. The hepatic veins known as portals supply nutrient-rich blood from the spleen and the digestive tract, which includes the stomach, intestine, and colon. The hepatic artery receives blood that has been enhanced with

oxygen by the heart as it beats. Strengthening the body's immune system, metabolism, digestion, and nutrition preservation are all made possible by the function of kidneys. The liver is a vital component of the body because without it, the body's tissues run the risk of dying from a lack of nutrients and energy. Hepatic or liver illnesses encompass a wide range of conditions damaging the liver. These disorders are referred to as chronic liver disease (LD) because they typically continue for a long time. It's intriguing to note that liver illnesses frequently have similar symptoms notwithstanding their specific variances.

Prior to being transferred to other areas of the human being, blood that has come from the digestive system must first pass through the liver. It is astonishing because at any one time, it uses around 13% of the blood that is going around the tissues of the body. The liver is made up of two lobes, the left and right lobes, each containing eight distinct phases. Thousands of lobules inside each segment are essential to the liver's amazing capabilities.

LD has the ability to disrupt these processes (metabolism, detoxification, Bile Production, blood regulation, storage, immune function, synthesis) and, in severe circumstances, even completely halt them. Liver has two unique qualities that impact the progression of disease. Initially, a large portion of the harm inflicted on the liver can be repaired, and regeneration occurs naturally. Secondly, the liver lacks nerve endings that would indicate the presence of illness. Consequently, the majority of individuals are ignorant about the onset and progression of disease.

Damage to the liver might become more irreversible as it starts to outweigh its natural ability to repair itself. A common misunderstanding is that excessive alcohol usage is the only cause of liver disease. While it is a prevalent felon, alcohol is just one of several. Other causes include poor diet and obesity, as well as viruses such as hepatitis A, B, C, D, or E [1]. An

individual may be considered at-risk because of having a family history of LD, engage in excessive use of prescription and over-the-counter medications, misuse illicit drugs, or have an autoimmune condition that triggers the immune system to attack the liver. Notable examples of inherited disorders include Wilson disease and hemochromatosis [2].

Lipid buildup in the liver is a hallmark of non-alcoholic fatty liver disease (NAFLD). 'Non-alcoholic steatohepatitis' is the term used to describe liver cell damage and inflammation. Another dangerous liver condition called cirrhosis develops when good liver tissue is replaced by scar tissue, causing irreparable damage and malfunction. Alcohol use, non-alcoholic fatty liver disease (NAFLD), chronic hepatitis B, and chronic hepatitis C are the main causes of liver cirrhosis [3].

Hepatitis or alcoholism are typically linked to LD, but obesity and diabetes are also increasingly linked to possibly catastrophic liver damage [4]. A person with advanced fatty LD has a nearly seven-fold increased risk of dying. Fatty liver disease is often a silent 'killer,' and matters can deteriorate rapidly once signs of its presence emerge [5,6].

Although the signs and symptoms of liver illness might vary, they frequently include jaundice (yellowing of the skin and eyes), bruising easily, changes in the color of the urine and stool, and swelling in the legs and abdomen. Not every instance has symptoms. Liver function tests and imaging studies are among the diagnostic procedures for liver injury. Patients should refrain from strong alcohol usage to prevent liver damage.

It is recommended that those with hepatitis B, hepatitis C, or alcoholic hepatitis absolutely abstain from alcohol [7]. Using condoms during sexual activity, not sharing catheters or needles, receiving a hepatitis A and B vaccination, and adopting protective steps to shield the skin from potentially harmful chemicals are further preventive measures. In the end, eating a balanced diet, exercising frequently, and keeping a healthy weight all support ideal liver function [8].

Around 2 million peoples worldwide pass away from LD each year. About a million of these deaths are caused by problems associated with cirrhosis, while another million are the consequence of hepatocellular cancer and viral hepatitis. In terms of total mortality, liver cancer ranks 16th globally, accounting for 3.5% of all deaths, whereas cirrhosis ranks 11th. In addition, cirrhosis is among the top 20 causes of years lived without a disability and years of life adjusted for disability, contributing 1.6% and 2.1% of the worldwide burden, respectively. Approximately 2 billion peoples use alcohol worldwide, and up to 75 million of them people have been identified with alcohol-related disorders and LD susceptibility. In addition, more than 400 million adults have diabetes and over 2 billion individuals are overweight or obese. These illnesses increase the chance of non-alcoholic fatty liver disease (NAFLD) and hepatocellular carcinoma [9]. Acute hepatitis is mostly caused by drug-induced liver damage, however viral hepatitis is also common around the world. Even though liver replacement is the second most frequent solid organ transplant surgery, fewer than 10% of the world's transplant demands are being met by current rates. Since most liver disorders are avoidable, these numbers also highlight a tremendous opportunity to improve public health [10].

The incidence of LD will get special consideration in the context of this investigation. The following noteworthy contributions have been produced by the employed methodology:

- The initial steps in data preprocessing were to remove duplicate instances, handle null values, and SMOTE. Because the instances in the dataset are distributed in this manner effective classification models can be developed to predict the earliest signs of liver illness.
- The Pearson Correlation matrix feature selection approach has been used to determine the importance of features.
- In the study, common measures including precision, accuracy, recall, F1 score, and AUC were used to compare and assess different machine learning (ML) models. The experimental results indicate that the ensemble ML approach, and more especially the LGBM classifier, performed better than the other models suggested in the study.
- This study's main emphasis, the LGBM classifier, showed more accuracy than earlier published research using the same dataset.

## II. MATERIALS AND METHOD

Here, the outline includes the dataset used, along with the essential elements of the process employed to estimate the risk of liver disease: class balancing and feature ranking in the balanced data. Lastly, the ML models used to interpret the experimental data will be mentioned.

### 2.1. Description of the dataset

The dataset of Indian Liver Patients' Records served as the foundation for our study [11]. There are 583 participants in this particular dataset, of whom 441 (75.64%) are men and 142 (24.36%) are women. The target class indicates whether an individual has received a diagnosis of LD or not. There are 416 participants (71.35%) who have been diagnosed with LD. [Table 1](#) provides a quick summary of the features of the dataset.

### 2.2 Risk Prediction of The Liver Disease

In order to predict disease using medical databases, ML techniques have become increasingly significant in the healthcare industry these days. To enhance medical diagnosis, a lot of scientists and businesses are using ML. Classification algorithms are heavily utilized in the field of disease prediction among other ML methodologies. The long-term risk prediction of LD is expressed here as a classification issue using two alternative classes,  $c = \text{"Liver-Disease" (LD)}$  or  $c = \text{"Non-Liver-Disease" (Non-LD)}$ . The trained ML models will be able to predict the class of a newly unclassified occurrence as either LD or Non-LD based on the values of the input data, and therefore the likelihood of LD developing.

#### 2.2.1 Preprocessing of Data

Firstly, the presence of duplicate rows were checked. After inspection, 13 duplicate rows were identified and

TABLE I. DESCRIPTION OF THE DATASET

Feature	Type	Description
Gender	Nominal	This characteristic displays the gender of the user.
Age (years)	Numeric	Participants range in age from 4 to 90 years old.
Total Protein—TP (g/L)	Numeric	The total protein of the individual is recorded in this feature.
Albumin—ALB (g/dL)	Numeric	The albumin of the participant is recorded by this function.
Albumin and Globulin Ratio—AGR	Numeric	The participant's albumin and globulin ratio are recorded in this feature.
Total Bilirubin—TB (mg/dL)	Numeric	The total bilirubin of the subject is recorded by this feature.
Direct Bilirubin—DB (mg/dL)	Numeric	This characteristic records the individual's direct bilirubin.
Aspartate Aminotransferase—SGOT (U/L)	Numeric	The participant's aspartate aminotransferase is captured by this characteristic.
Alkaline Phosphatase—ALP (IU/L)	Numeric	This characteristic records the alkaline phosphatase level of the individual.
Alanine Aminotransferase—SGPT (U/L)	Numeric	This characteristic records the alanine aminotransferase level of the individual.
Liver Disease	Nominal	This characteristic indicates if the individual has received a LD diagnosis or not.

subsequently removed. Next, an examination for any NULL cells revealed a total of 4. The missing data were handled using the fillna approach. Label Encoder technique was applied to convert the gender string type to numerical values. The uneven distribution throughout the dataset may have an effect on the precise identification of LD and Non-LD cases. That is why an oversampling technique called SMOTE [12] is used to produce synthetic data [13] on the minority class. “Fig. 1” depicts the approach of the suggested LD prediction strategy after the dataset has been scaled using the MinMaxScaler. Due to an oversampling of instances in the Non-LD class, the subjects covered in the two classes are equally split. Following the application of SMOTE, there are 812 participants’ total; 258 (31.77%) are women and 554 (68.23%) are males. With 406 occurrences of LD and Non-LD in the objective class, the information set is now harmonized. 25% of the data was used for testing and rest 75% for training without k fold validation.

### 2.2.2 Features Selection

“Fig. 2” displays the Pearson correlation of all features for this study. The magnitude and direction of the correlation between two different features, or between a feature and the class of concern, are reflected in this parameter. The values of the coefficients span from -1 to 1. Upon concentrating on this parameter, we find that TB and DB have a strong connection of 0.98. We can thus remove one of these functionalities. DB was dropped for this study. Subsequently, the chosen characteristics are saved in the data frame (df), including  $df = \{\text{Age, Gender, TB, ALP, SGPT, SGOT, TP, ALB, AGR, Dataset}\}$ .

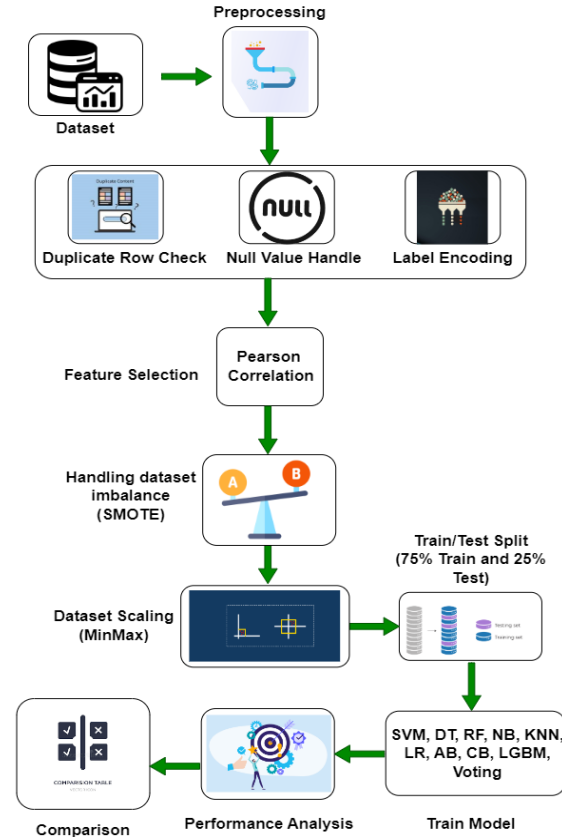


Fig. 1: The workflow of the proposed methodology.

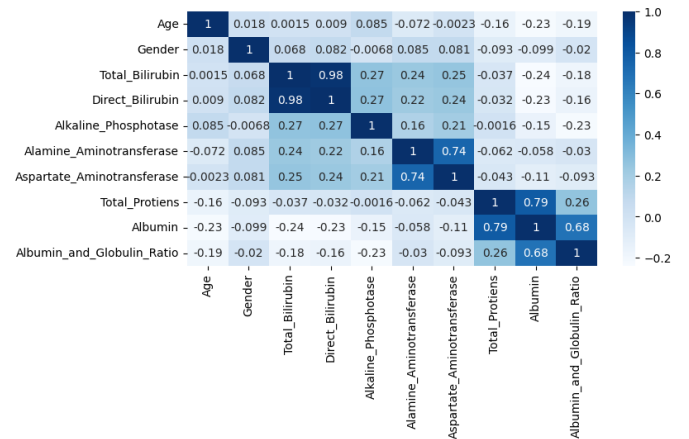


Fig. 2: Pearson correlation matrix of all features

### 2.3 Machine Learning Models

In this study, A number of ML models such as Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), AdaBoost (AB), CatBoost (CB), Logistic Regression (LR), LGBM, and VotingClassifier (VC) were experimented to determine which performs better by assessing each model's capacity for prediction.

### 2.4 Evaluation Metrics

The suggested system is tested and assessed using the performance metrics. We applied the most commonly utilized metrics found in the relevant literature, such as F-Measure, AUC, accuracy, precision, and recall [14,15], to evaluate the ML models' performance. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the elements that make up the confusion matrix. Equations 1 to 4 show the definition of the previously stated measures:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (1)$$

$$Precision = \frac{TP}{TP + FP} * 100\% \quad (2)$$

$$Recall = \frac{TP}{TN + TP + FN + FP} * 100\% \quad (3)$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} * 100\% \quad (4)$$

Utilizing the AUC allows for the assessment of a model's distinguishability. This measure has a range of [0, 1]. The ML model's ability to discriminate between LD and Non-LD cases improves with proximity to one.

## III. EXPERIMENTAL RESULTS

### 3.1. Experimental Setup

An extensive summary of the relevant experiment results is given in this section. After randomization, the dataset was split into training and testing sets. At first, several data preparation techniques were suited just to the training set. To reduce data leaking and overfitting, these strategies were then used for the testing and training sets. To get the results, this study used a range of scenarios. Table II highlights the necessary environment configuration for the investigation.

TABLE II. SPECIFICATIONS OF THE SYSTEM

CPU	1 × 11TH Gen Intel®Core™i5-1135G7 @ 2.40GHz
RAM	8GB
GPU Memory	8GB
GPU	Intel iRISXe
Disk Space	256GB
Cache	46MB
Session Limit	10h

### 3.2. Performance Evaluation

First, it was explained about how the SMOTE approach helps in the development of useful models for performance assessment. In Table III, Recall and Precision scores amply demonstrate the effect of the particular class balancing technique. These results pertain to the minority class (healthy people) and demonstrate how the SMOTE approach greatly enhanced the effectiveness of ML models. After that, SMOTE is applied to the models that are being examined, and Table IV shows these algorithms' typical results. Additionally, despite retaining greater overall performance levels of difficulty, the constant distribution of categories considerably improves the accuracy of identifying people who are healthy. This implies that the models under investigation are still useful for extremely accurately and precisely classifying diseases. The LGBM Ensemble approach performs better than the other models, with 83.74% accuracy, 78% precision, 85% recall, 81% F1-score, and 90.3% AUC. Furthermore, the voting classifier model performs well, with 83.25% accuracy, 83% precision, 86% recall, 85% F1-score, and 90.2% AUC.

TABLE III. BEFORE AND AFTER SMOTE INTO CLASS "NAME" THE PERFORMANCE OF ML MODELS

Name	Precision		Recall	
	NoSMOTE	SMOTE	NoSMOTE	SMOTE
CatBoost	59%	78%	33%	85%
LGBM	50%	78%	30%	85%
SVM	nan	67%	0%	79%
KNN	32%	69%	23%	79%
LR	59%	71%	25%	72%
GNB	38%	66%	97%	96%
DT	33%	75%	35%	75%
AdaBoost	42%	74%	33%	79%
RF	52%	79%	35%	83%
Voting	57%	83%	33%	86%

TABLE IV. EVALUATION OF ML MODELS' PERFORMANCE FOLLOWING SMOTE

Classifiers	Accuracy	Precision	Recall	F1	AUC
SVM	67.98%	67%	79%	73%	75.3%
LR	68.47%	71%	72%	71%	75.3%
KNN	68.97%	69%	79%	73%	71.6%
GNB	70.94%	66%	96%	78%	76.8%
DT	72.91%	75%	75%	75%	72.7%
AdaBoost	73.89%	74%	79%	77%	81.6%
CatBoost	78.82%	78%	85%	81%	88.0%
RF	81.28%	79%	83%	81%	87.7%
Soft Voting	83.25%	83%	86%	85%	90.2%
LGBM	83.74%	78%	85%	81%	90.3%

After applying SMOTE, “**Fig. 4**” shows how the ML model is evaluated using AUC ROC curves and the aforementioned metrics. Furthermore seen in the image, the LGBM, voting, CatBoost, RF, and AdaBoost models perform significantly better than the others. Numerical results further corroborate the above claim.

Furthermore, **Table V** presents models that have been proposed earlier by other studies that have used the same dataset [11]. Notably, with an accuracy of 83.74%, the suggested LGBM model performs better than these alternative methods.

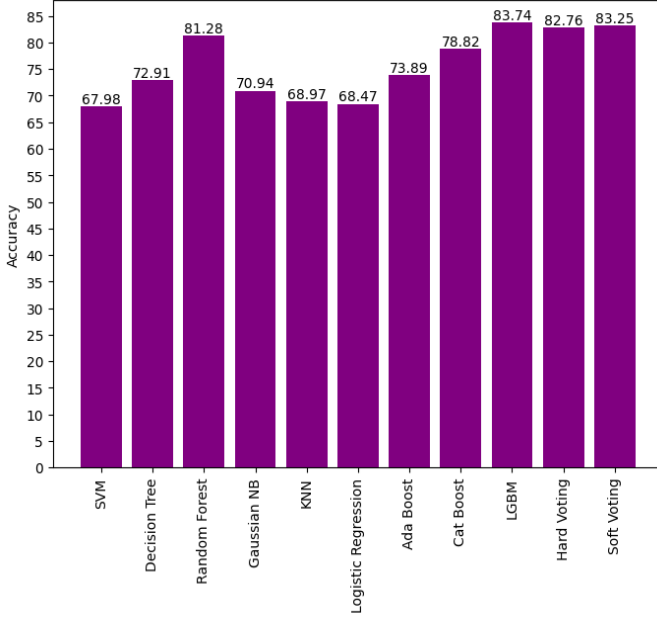


Fig. 3: Overall classification accuracy percentage for all used algorithms.

#### IV. DISCUSSION

**Table V** presents a comparison among the suggested system and relevant studies. Research based on the Indian Liver Patients' Records dataset [11] is presented in this section that utilize several ML models to forecast the occurrence of LD. In particular, With 72% accuracy, the Gradient Tree Boosting classifier for a balanced dataset performed best in [16]. Further research was conducted by Azam et al. [17] to see whether KNN using feature selection techniques (KNNWFST) performs better than any other comparable.

The authors of [18, 19] offered the LR model, which had a 75% accuracy rate. Geetha et al.[20] showed that the SVM achieves 75.04% accuracy. Gajendran et al. [21] suggested a hybrid ML model named Mathematical Approach on Multilayer Feedforward Neural Network with Backpropagation (MAMFFN). Dritsas et al. [22] proposed voting classifier and achieved accuracy of 80.10%.

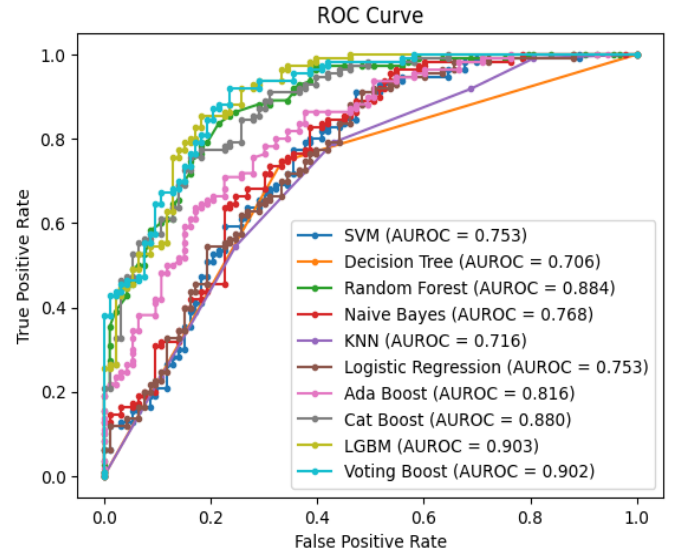


Fig. 4: ML Model's evaluation based on AUC ROC curves.

Using the same dataset as [16–22], the goal of this work was to employ ensemble learning to investigate improved alternatives. The biochemical parameters used in the training and evaluation processes were the same prior to and following class balance. Additionally, ensemble models that had not received significant consideration in earlier studies were now the focus. In addition, a feature evaluation strategy and an equitable data set based on SMOTE were taken into consideration. Notably, specific models fared worse than the classifiers trained and evaluated in the studies [16–22], especially in terms of accuracy. Best of all, the LGBM classifier—which was the main recommendation—performed better than any other approach, including the voting classifier. The importance of life quality makes it imperative to create a ML model based on biochemical data that appropriately reflects human health state, from both a therapeutic and scientific standpoint.

Finally, this section concludes that the current study will support researchers and clinicians in monitoring LD and designing high-performance personalized models with the flexibility to include quality-of-life features that demonstrate patients' well-being as well as the challenges related to this condition. Prior to concluding the examination of the representations that have been offered, it is important to recognize several limitations with this paper. The Indian Liver Patients' Records dataset [11] is one publicly accessible information that serves as the foundation for this present research undertaking. Nevertheless, the diagnostic requirements and precise terminology needed to identify patients with LD are missing from this dataset. Investigating data from a hospital or another institution could yield extra information for a more fruitful assessment of the ML models.

TABLE V. PROPOSED MODELS ARE ILLUSTRATED USING THE SAME DATASET AS THE SOURCE PUBLICATIONS

Research work	Proposed Model	Accuracy
A. Sokoliuk et al. [16]	Gradient Tree Boosting	72%
M. Azam et al. [17]	KNNWFST	74%
A. Srivastava et al. [18]	LR	75%
R. Choudhary et al. [19]	LR	75%
C. Geetha et al. [20]	SVM	75.04%
G. Gajendran et al. [21]	MAMFFN	75.30%
E. Dritsas [22]	Voting	80.10%
Proposed Method	LGBM	83.74%

## V. CONCLUSIONS

Due to the fact that LD has the potential to be fatal, it is a serious sickness that has to be treated right away. Pathological procedures are used by medical professionals to record a patient's status and provide an extensive medical report. The aim of this work was to use ML techniques to predict LD early. To be more precise, the precision, recall, F1-score, accuracy, and AUC of a variety of ML models, including GNB, KNN, DT, RF, SVM, AB, CB, LR, LGBM, and voting, were assessed in relation to their capacity to predict the likelihood of LD. After using SMOTE, the LGBM classification approach outperformed the other models and was the main suggestion in this study. It had an accuracy of 83.74%, precision of 78%, recall of 85%, F1-score of 81%, and an AUC of 90.3%. Notably, when compared to comparable published research efforts utilizing the same dataset, the suggested LGBM model exhibits greater accuracy. The transparencies of ML models can be demonstrated in the future using LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations).

## REFERENCES

1. H. Razavi, "Global epidemiology of viral hepatitis," *Gastroenterol. Clin.*, 49, 179–189, 2020.
2. M. Ringehean, J.A. McKeating, U. Protzer, "Viral hepatitis and liver cancer," *Philos. Trans. R. Soc. B Biol. Sci.*, 372, 20160274, 201.
3. A. Smith, K. Baumgartner, C. Bositis, "Cirrhosis: Diagnosis and management," *Am. Fam. Physician* 100, 759–770, 2019.
4. G. Marchesini, S. Moscatiello, S. Di Domizio, G. Forlani, "Obesity-associated liver disease," *J. Clin. Endocrinol. Metab.*, 93, s74–s80, 2008.
5. H.K. Seitz, R. Bataller, H. Cortez-Pinto, B. Gao, B. A. Gual, C. Lackner, P. Mathurin, S. Mueller, G. Szabo, H. Tsukamoto, "Alcoholic liver disease," *Nat. Rev. Dis. Prim.*, 4, 1–22, 2018.
6. F. Åberg, M. Färkkilä, "Drinking and obesity: Alcoholic liver disease/nonalcoholic fatty liver disease interactions," In *Seminars in Liver Disease*; Thieme Medical Publishers: Volume 40, pp. 154–162, New York, NY, USA, 2020.
7. M. Bae, Y.K. Park, J.Y. Lee, "Food components with antifibrotic activity and implications in prevention of liver disease," *J. Nutr. Biochem.* 2018.
8. J. Cai, X.J. Zhang, H. Li, "Progress and challenges in the prevention and control of nonalcoholic fatty liver disease," *Med. Res. Rev.*, 39, 2019.
9. S.K. Asrani, H. Devarbhavi, J. Eaton, and P.S. Kamath, "Burden of liver diseases in the world," *Journal of hepatology*, 70(1), pp.151-171, 2019.
10. E.B. Keeffe, "Liver transplantation: current status and novel approaches to liver replacement," *Gastroenterology*, 120(3), pp.749-762, 2001.
11. Indian Liver Patient Records. Available online: [Indian Liver Patient Records \(kaggle.com\)](https://www.kaggle.com/datasets/indian-liver-patient-records) (accessed on 14 November 2022).
12. S. Maldonado, J. López, C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl. Soft Comput.*, 76, 380–389, 2019.
13. E. Dritsas, N. Fazakis, O. Kocsis, K. Moustakas, N. Fakotakis, "Optimal Team Pairing of Elder Office Employees with Machine Learning on Synthetic Data," In *Proceedings of the 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 12–14, pp. 1–4, Chania Crete, Greece, July 2021.
14. G.S. Handelman, H.K. Kok, R.V. Chandra, A.H. Razavi, S. Huang, M. Brooks, M.J. Lee, H. Asadi, "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods," *Am. J. Roentgenol.*, 212, 38–43, 2019.
15. J. Zhou, A.H. Gandomi, F. Chen, A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, 10, 593, 2021.
16. A. Sokoliuk, G. Kondratenko, L. Sidenko, Y. Kondratenko, A. Khomchenko, I. Atamanyuk, "Machine learning algorithms for binary classification of liver disease," In *Proceedings of the 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, pp. 417–421, Kharkiv, Ukraine, 6–9 October 2020.
17. M.S. Azam, A. Rahman, S.H.S. Iqbal, M.T. Ahmed, "Prediction of liver diseases by using few machine learning based approaches," *Aust. J. Eng. Innov. Technol.*, 2, 85–90, 2020.
18. A. Srivastava, V.V. Kumar, T. Mahesh, V. Vivek, "Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms," In *Proceedings of the 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1–4, 21–22, Bhilai, India, April 2022.
19. R. Choudhary, T. Gopalakrishnan, D. Ruby, A. Gayathri, V.S. Murthy, R. Shekhar, "An Efficient Model for Predicting Liver Disease Using Machine Learning," In *Data Analytics in Bioinformatics: A Machine Learning Perspective*; pp. 443–457, Wiley Online Library: Hoboken, NJ, USA, 2021.
20. C. Geetha, A. Arunachalam, "Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms," In *Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–4, Coimbatore, India, 27–29 January 2021;
21. G. Gajendran, R. Varadharajan, "Classification of Indian liver patient's data set using MAMFFN," In *Proceedings of the AIP Conference Proceedings*, Volume 2277, p. 120001, Coimbatore, India, 17–18 July 2020;
22. E. Dritsas, and M. Trigka, "Supervised machine learning models for liver disease risk prediction," *Computers*, 12(1), p.19, 2023.