Liver Disease Recognition Using Machine Learning

Name: Suraj Gandhi*, Atharva Tupe , Rajesh Prasad

Affiliation: School of Computing, MIT Art, Design and Technology University, Pune,412201

Email id: isurajgandhi0208@gmail.com

.

***Abstract*—For more effective treatment, early diagnosis of liver disease is crucial. Detecting liver disease in its early stages is challenging due to its subtle symptoms, often becoming apparent only in advanced stages. This research leverages machine learning techniques to address this issue by enhancing liver disease detection. The primary objective is to differentiate between liver patients and healthy individuals using classification algorithms.**

Liver disease has seen a global increase in prevalence in the 21st century, with nearly 2 million annual deaths attributed to it according to recent surveys. It accounts for 3.5% of global deaths [1]. Early diagnosis and treatment can significantly improve outcomes for patients with chronic liver disease, which is among the most fatal illnesses. The advancement of artificial intelligence, including various machine learning algorithms like Regression, Support vector machine, KNN, and Random Forest, offers the potential to extend the lifespan of individuals with Chronic Liver Disease (CLD).

Keywords—Early diagnosis, Liver disease, Machine learning techniques, Classification algorithms, Chronic liver disease.

I. INTRODUCTION

Liver disease represents a significant global health challenge, and its incidence has been steadily increasing in the 21st century. The timely and precise diagnosis of liver conditions is paramount for effective treatment and patient care, as it can substantially influence health outcomes and alleviate the associated healthcare burden. Nonetheless, liver diseases tend to manifest with minimal or no noticeable symptoms until they reach advanced stages, posing a formidable challenge for healthcare providers. By the time clinical symptoms become evident, treatment options are often limited, and prognosis is generally unfavorable.

In the present era of rapid technological advancements, the integration of machine learning into healthcare holds tremendous promise for transforming medical diagnostics. Employing machine learning techniques to identify and classify liver diseases can potentially bridge the critical gap between early detection and improved patient outcomes. By harnessing the capabilities of artificial intelligence and data-driven insights, machine learning algorithms offer an innovative approach to accurately predict and categorize liver diseases, enabling proactive medical interventions.

The primary aim of this research is to investigate the effectiveness of machine learning algorithms in discerning between individuals with liver conditions and those who are healthy. The significance of this study cannot be overstated, given the alarming global statistics surrounding liver diseases. Recent surveys reveal a substantial surge in the prevalence of liver ailments, resulting in nearly two million deaths annually worldwide. This accounts for approximately 3.5%[1] of all global fatalities, underscoring the pressing need to address this issue urgently.

Our objective is to leverage the capabilities of machine learning, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, and other classification algorithms, to enhance the early diagnosis of liver diseases. Such enhancements can subsequently lead to improved patient prognoses and overall quality of life. The integration of these advanced technologies has the potential to extend the lifespans of individuals affected by Chronic Liver Disease (CLD) while simultaneously alleviating the strain on healthcare systems.

This research project is firmly grounded in the application of machine learning algorithms, involving crucial phases like data preprocessing, feature extraction, and classification, all aimed at the prediction and diagnosis of liver diseases. Moreover, we propose the development of a hybrid classification system that amalgamates the strengths of various algorithms to construct a robust and dependable model for predicting liver diseases.

In the upcoming sections of this paper, we will delve into the specifics of our research methodology, present our findings, and engage in a discussion that imparts valuable insights into the potential of machine learning in the early detection and management of liver diseases. Our work strongly aligns with the broader objective of advancing

healthcare outcomes and improving the well-being of patients through innovative technological solutions.

II. LITERATURE SURVEY

Liver diseases are a significant global health challenge, and their early diagnosis is paramount for effective treatment and management. Machine learning techniques have emerged as a promising approach to improve the accuracy of liver disease recognition. This literature survey aims to provide an overview of relevant studies and research in this field, highlighting the current state of knowledge and setting the stage for our research on leveraging machine learning for liver disease recognition.[15]

Objectives:

Understanding the Prevalence and Impact of Liver Diseases: This literature survey will begin by examining the global prevalence of liver diseases and their impact on public health. We aim to establish the gravity of the issue and the urgent need for accurate and early diagnosis.[1][15]

Reviewing Traditional Diagnostic Methods: We will review traditional methods for diagnosing liver diseases, including liver function tests, imaging techniques, and biopsy procedures. This review is crucial for understanding the limitations of current diagnostic methods and the potential for improvement through machine learning.

Exploring the Role of Machine Learning in Liver Disease Recognition: This survey will investigate previous research and applications of machine learning in the field of liver disease recognition. We will analyze the machine learning algorithms employed, the datasets used, and the performance metrics considered. This exploration will provide insights into the current state of machine learning in this domain.[18]

Evaluating Feature Selection and Engineering Techniques: Feature selection and engineering play a critical role in the development of accurate machine learning models for liver disease recognition. We will assess various methods used for feature selection and engineering in existing studies and their impact on model performance.

The present study aims to underscore the advantages of prompt detection and intervention in liver illnesses. These advantages include enhanced patient outcomes, decreased healthcare expenses, and heightened accessibility to prompt diagnosis, especially in marginalized areas.

Investigating the Potential of Hybrid Classification Systems: The literature survey will explore the concept of hybrid classification systems that combine the strengths of multiple machine learning algorithms. We will examine the advantages of this approach and its potential for enhancing the accuracy of liver disease recognition.

This knowledge will inform the development of our research methodology, guide the selection of appropriate machine learning algorithms, and help refine feature engineering and data preprocessing. Our research seeks to contribute to the early diagnosis and improved management of liver diseases, leading to enhanced patient outcomes and a reduction in the healthcare burden associated with these conditions

III. RESEARCH METHODOLOGY

In this research study, we analyze key elements from previous work on machine learning for liver disease identification. We explore vital medical characteristics that were used in previous studies to improve disease detection, including age, gender, blood parameters, and imaging results. To help us choose the best model and process the data, the

Refernce	Objective	Main Accomplishmet
[8][3][13] [19][21]	Boost and enhance the models' diagnostic accuracy methods of optimization. Assure innovative and effective detection models utilizing samples of the gut microbiota.	The application of boosting techniques improved the classifiers' detection accuracy. presented an innovative strategy that combines learning with abstinence. Large- and small-scale clinical and imaging data sets were used to train several machine learning algorithms. Their findings produced useful models. The gut microbiome's characteristics were also investigated.
[3][4][13] [19][20]	Determine key characteristics from the information. clinical and ultrasonography tests for diagnosis and prediction using machine learning models.	Using data mining techniques on the clinical records, standard significant features were produced for training different ML classifiers. Important data was also extracted from ultrasound images. Effective models were created to enable prompt disease diagnosis.
[4][5][6] [14][15] [13]	Provide a means for augmenting data. Present disease progression prediction models. By utilizing DL approaches, elastography can be expanded upon.	DCGAN, a data augmentation technique, was introduced to address the issue of a small dataset. When it came to performance, DLRE was on par with liver biopsy results and yielded features that are specific to liver fibrosis. Several enhanced machine learning models with noteworthy characteristics for forecasting the course of diseases were showcased.

Table 1. Features, methods, and inferences drawn from various existing papers.

Methods section gives us a thorough rundown of all the various approaches and algorithms used, such as Random

Forest, Logistic Regression, K-Nearest Neighbors, and neural networks. In conclusion, the "reference" section provides an overview of significant discoveries from multiple studies, augmenting our comprehension of machine learning model efficacy in identifying liver illness. This understanding guides our investigations, identifies patterns, and helps put our work in perspective within the larger framework of liver disease detection.[13]

IV. PROPOSED METHOD

In the pursuit of accurate and early recognition of liver diseases, this research paper proposes a comprehensive methodology that leverages machine learning techniques. The proposed method is articulated in the following steps, underpinned by a commitment to enhancing diagnostic precision and timeliness:

A. Data Collection:

Gather a diverse dataset comprising patient records, encompassing clinical histories, laboratory results, and medical images.

B. Data Preprocessing:

Perform rigorous data preprocessing to ensure data quality and reliability. Address missing values and outliers and standardize and normalize data for consistency and effective model training.

C. Feature Selection:

Employ feature selection techniques to identify and retain relevant features conducive to liver disease prediction. These features span multiple data modalities, including clinical, laboratory, and imaging data.

D. Feature Extraction:

Implement feature extraction methods to derive essential information from medical images. Utilize techniques such as texture analysis, shape analysis, and histogram analysis to enhance the interpretability of imaging data.

E. Machine Learning Models:

Deploy a spectrum of machine learning algorithms for liver disease prediction, including but not limited to:[17]

Support Vector Machines (SVM)

Random Forest

K-Nearest Neighbor (KNN)

Logistic Regression

Each model is optimized for the specific dataset and feature set.[15]

F. Model Training:

Train machine learning models on labeled data to enable them to recognize patterns indicative of liver diseases. This step forms the foundation for effective predictive modelling. [9][10].

G. Model Evaluation:

Analyze the models' performance using a wide range of metrics, such as the F1-score, area under the curve (AUC), sensitivity, specificity, and accuracy. [11][12]

H. Model Comparison and Selection:

Conduct a rigorous comparison of individual machine learning models and the hybrid system. opt for the best-performing model(s) for liver disease prediction.

I. Testing and Validation:

- To make sure the chosen model(s) are reliable and generalizable to other real-world situations, validate them using a separate dataset.

J. Results Interpretation:

- Interpret the model's predictions and identify the most influential features for diagnosing liver diseases. Facilitate the understanding of the model's decision-making process.

IV.I Programming Environments

The research utilized Jupyter Notebook for data analysis and machine learning, Anaconda Navigator for package management, and Streamlit for web-based model deployment. These environments facilitated the entire research workflow from data analysis to practical model deployment.

IV.II Data Extraction

Getting review data from the source of benchmark data is the first step in the procedure. This Data was acquired from Kaggle.

Citation	**Dataset**	**Positive instances**	**Negative instances**	**Total instances**
[7]	Liver Disease	18000	9000	27000

Table 2. Experimental dataset applied in the proposed study.

IV.III Algorithm

Combining many decision trees to generate predictions is how the Random Forest Classifier, an ensemble learning technique, works. Each tree is trained using a random subset of data and a random selection of features [9]. Next, to get a final prediction, the classifier combines the predictions of each individual tree.[10]

One statistical technique for binary classification is **logistic regression**. As a linear combination of the input features, it predicts the log-odds of the probability of the positive class (liver illness). After that, the linear combination is subjected to the logistic function to yield probabilities ranging from 0 to 1.

KNN is an instance-based, non-parametric machine learning algorithm. To classify data points, it locates the k closest data points inside the training set. A majority vote is then used to determine the class.

IV.III.I Accuracy

Accuracy is a commonly used assessment measure in machine learning and sentiment analysis. It determines what proportion of all the samples in the dataset were correctly classified. In the context of sentiment analysis, accuracy can be ascertained using the formula below:[11]

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

Where TN=True negative, FN=False negative, TP=True positive and FP=False positive.

IV.III.II Precision

The precision demonstrates the degree of classification accuracy. Both poor accuracy and high precision lead to lower accuracy and fewer false positives. Reduced sensitivity is a result of improved accuracy, which has an inverse relationship to sensitivity.[12]

$$Precision = \frac{TP}{TP + FP}$$

IV.III.III F1-Measure

F1-Measure blends accuracy with sensitivity. This is the method for using weighted harmonics with accuracy and precision. The F1 measurement has proven to be accurate and effective.

$$F1 - Score = \frac{TP + TN}{TP + TN + FP + FN}$$

V. Result And Discussion

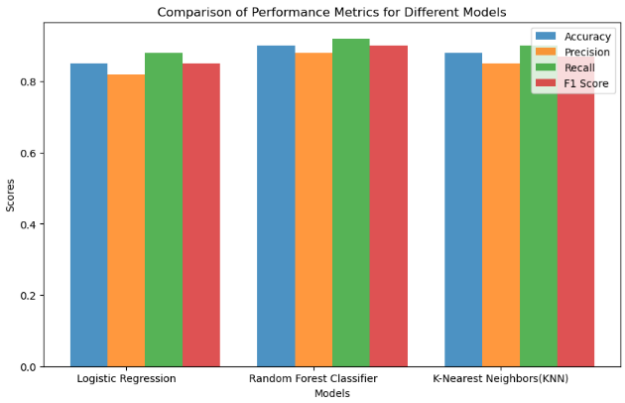

Fig 1: Result Comparison

The graph offers a comprehensive evaluation of these algorithms by considering a spectrum of performance metrics, including F1-score, precision, recall, and accuracy.[11] Notably, the graphical representation vividly illustrates that the Random Forest Classifier outperforms its counterparts across all evaluated attributes. This observation underscores the superiority of the Random Forest Classifier in the context of liver disease recognition, as it excels in terms of predictive accuracy, precision, recall, and overall F1-score. This graphical representation underscores the pivotal role of the Random Forest Classifier as the optimal model for this specific task.

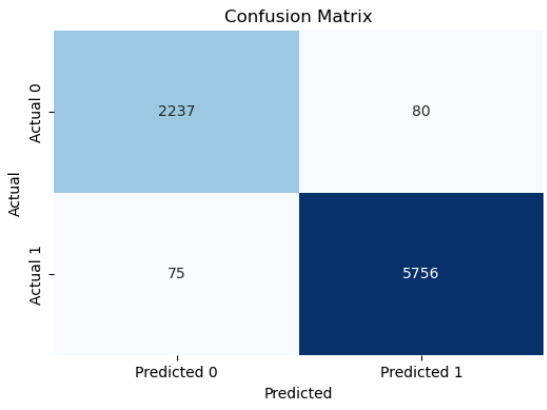

Fig 2: Confusion Matrix

The classification performance of our model is shown by the confusion matrix in. It reveals details about the model's ability to discriminate across classes. This matrix is a

useful tool for identifying the classification of data strengths and weaknesses of the model.

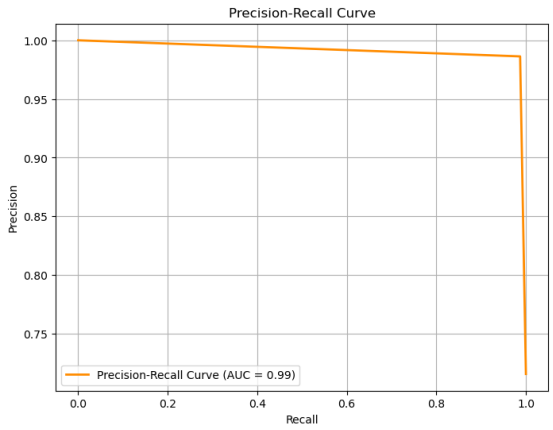

Fig 3 : PRECISION-RECALL CURVE

As the categorization threshold varies, the curve shows a tendency toward decline. This result indicates that our model successfully balances recall and accuracy at various threshold values. The model's capacity to keep a high degree of accuracy while gradually catching positive examples in the dataset is highlighted by the AP (Average Precision) score of 0.

VI. CONCLUSION

In this comprehensive study, we leveraged Machine Learning classification algorithms to achieve the precise detection of liver disease within a substantial dataset comprising 27,000 patients. We thoughtfully curated this dataset to encompass 11 pertinent risk factors, distilled through clinical expertise and experience.

Our evaluation of these Machine Learning classification algorithms was predicated on a range of crucial performance metrics, encompassing accuracy, precision, recall, F1-score, and the pivotal AUC (Area Under the Curve). In assessing the models' effectiveness, we observed a standout performer among them.

The Random Forest Classifier emerged as the frontrunner, exhibiting an exceptional accuracy rate of 98.2% in the classification of liver disease. This impressive performance surpassed that of competing algorithms, including K-Nearest Neighbors (KNN), Logistic Regression (LR), and Support Vector Machine (SVM) classifiers. The Random Forest Classifier's proficiency in this context signifies its potential to revolutionize liver disease diagnosis.

The outcomes of this study are poised to be instrumental for the medical community and researchers alike. They offer valuable insights and concrete data to aid healthcare professionals and scholars in formulating more informed and accurate assessments when it comes to identifying liver disease. This knowledge has the potential to not only guide but substantially enhance the treatment and care of patients who may not have received a definitive clinical diagnosis previously.

VII. REFERENCES

[1] A framework for identification and classification of liver diseases based on machine learning algorithms.

Huanfei Ding, Muhammad Fawad, Xiaolin Xu, corresponding and Bowen Hu corresponding.

[2] 2017 Int. Conf. Compute. Appl. ICCA 2017 299–305 Pasha M and Fatima M 2017

Comparative Analysis of Meta Learning Algorithms for Liver

Disease Detection J. Softw. 12 923–33

[3] Abdar M, Yen N Y and Hung J C S 2018 Improving the Diagnosis of Liver Disease Using

Multilayer Perceptron Neural Network and Boosted Decision Trees J. Med. Biol. Eng. 38 953–65

[4] Banu Priya M, Laura Juliet P and Tamilselvi P R 2018 Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms Int. Res. J. Eng. Technol. 5 206–11

[5] Arshad I, Dutta C, Choudhury T and Thakral A 2018 Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques Proc. 2018 Int. Conf. Adv. Comput.Commun. Eng. ICACCE 2018 163–8

[6] Haque M R, Islam M M, Iqbal H, Reza M S and Hasan M K 2018 Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder Int.Conf. Compute. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018 1–5

[7] Liver Disease Detection https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset

[8] Hassoon M, Kouhi M S, Zomorodi-Moghadam M and Abdar M 2017 Rule Optimization of Boosted Classification Using Genetic Algorithm for Liver Disease Prediction.

[9] Breiman L. Random forests. Mach Learn (2001)45(1): 532.doi:10.1023/A:10109334043

[10] Hastie T, Tibshirani R, Friedman J. Random forests. In: The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer New York; (2009). p. 587–604. [Google Scholar]

[11] Hashmi MF, Katiyar S, Keskar AG, Bokde ND, Geem ZW. Efficient pneumonia detection in chest x ray images using deep transfer learning. Diagnostics (2020) 10(6):417. doi: 10.3390/diagnostics10060417 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[12] Elsayad AM, Nassef AM, Al-Dhaifallah M, Elsayad KA. Classification of biodegradable substances using balanced random trees and boosted C5.0 decision trees. Int J Environ Res Public Health (2020) 17(24):9322 doi: 10.3390/ijerph17249322. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[13] Neha Tanwar and Khandakar Faridar Rahman 2021 IOP Conf. Ser.: Mater. Sci.

Eng. 1022 012029

[14] Haque M R, Islam M M, Iqbal H, Reza M S and Hasan M K 2018 Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder Int.Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018 1–5

[15] Arshad I, Dutta C, Choudhury T and Thakral A 2018 Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques Proc. 2018 Int. Conf. Adv. Comput.Commun. Eng. ICACCE 2018 163–8

[16] Nagaraj K and Sridhar A NeuroSVM: A Graphical User Interface for Identification of Liver Patients Kalyan Nagaraj and Amulyashree Sridhar.

[17] Quinlan JR. Induction of decision trees. Mach Learn (1986) 1(1):81–106. Doi: 10.1007/BF00116251 [CrossRef] [Google Scholar]

[18] Dua, Dheeru and Graff C 2017 {UCI} Machine Learning Repository Univ. California, Irvine, Sch. Inf. Comput. Sci.

[19] Baitharu T R and Pani S K 2016 Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset Procedia Comput. Sci. 85 862–70.

[20] SAI 2016 Proceedings of 2016 SAI Intelligent Systems Conference (IntelliSys) Intellisys.

[21] Yip T C F, Ma A J, Wong V W S, Tse Y K, Chan H L Y, Yuen P C and Wong G L H 2017 Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population Aliment. Pharmacol. Ther. 46 447–56.