Prediction and Detection of Liver Diseases using Machine Learning

R.T. Umbare
Information Technology
JSPM's RSCOE
Pune, India
rtumbare_it@jspmrscoe.edu.in

Omkar Ashtekar
Information Technology
JSPM's RSCOE
Pune, India
odashtekar@gmail.com

Aishwarya Nikhal
Information Technology
JSPM's RSCOE
Pune, India
aishwaryanikhal12@gmail.com

Bhagyashri Pagar
Information Technology
JSPM's RSCOE
Pune, India
bhagyashripagar08@gmail.com

Omkar Zare
Information Technology
JSPM's RSCOE
Pune, India
omkarzare123@gmail.com

2023 IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET) | 979-8-3503-9781-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/TEMSMET56707.2023.10150135

***Abstract*—Liver is a very essential organ in the human body. It is essential to recognize or diagnose the disease early. This considerably aids in the early avoidance of disease with minimal medication. Conventional methods include Liver Function Tests and test results. Early detection of liver disease is very difficult. This is because symptoms of the disease become apparent only in the later stages of the disease. This system of Machine learning facilitates early disease detection, Identifying elements that lead to fatal liver impairment. Predicting the disease in its early stages is a difficult task for doctors and scientists due to the apparently sensitive signs. Effects will become apparent only when it is too late. The initiative seeks to use machine learning techniques to address this issue and improve the victims of the disease. Because there are few signs of liver disease, it is difficult to diagnose and symptoms usually do not appear until it is too late. The aim is to study and use a classification approach to distinguish between liver disease and healthy individuals, if diseased then further classified into the level of disease and its type. Also, precautions are provided for any symptoms. Consequently, ML techniques have identified liver disease in individuals.**

***Index Terms*—Disease, Naïve Bayes, KNN, Liver Disease, Logistic Regression, Machine Learning, Support Vector Machine.**

I. Introduction

Liver diseases are becoming more common than ever due to the increasingly sedentary lifestyle trend and lack of physical activity. The intensity is still controllable in rural areas, but liver disease is increasingly very prevalent in urban areas, particularly metropolitan areas. Prediction of Liver Diseases is necessary for Human beings' livelihood. The liver is the largest internal Organ of the human body and when it is injured, the human being's life is automatic gets troubled. The different types of liver diseases identified in Human beings are fatty liver, cirrhosis, liver cancer, hepatitis, liver tumor, etc. The percentage of diseases is increasing due to the increase in the consumption of alcohol, Drugs, pickles, and food. Each year Millions of people die due to liver diseases. The liver can be affected by several disease states. When alcohol is used for long periods of time, it changes the liver's metabolism, which may have overall bad impacts. hemochromatosis causes liver problems. In this system it identifies significant features and predicts whether a person may suffer from liver disease based on those features. Genetic algorithms will be used to identify significant features, and then those features will be used to train different classification models such as Naïve Bayes, KNN, Support Vector Machine (SVM), and Logistic Regression which will indicate whether a person has a chance of liver disease and if so then which disease a person has i.e. major or minor and which type of liver disease and what type of precautions need to take.

The following are the main results that can be expected from this project:

1) liver disease classification.
2) Precautions for particular symptoms
3) Reduction in liver disease-related deaths
4) More accurate medical diagnosis of liver disease

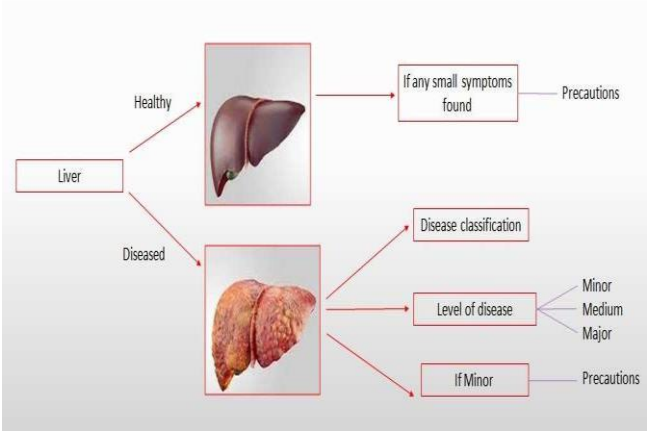

Fig. 1. Proposed System.

979-8-3503-9781-9/23/$31.00 ©2023 IEEE

II. Motivation

The motivation behind this study is that liver disease is among the most common diseases in the world today. As a result of the disease, the death rate is increasing alarmingly. Early detection of the disease may reduce the complication of the disease misfortune on patients. Using different algorithms, it predict a patient who suffers from which liver disease. Also, data mining tools can be used to assist physicians in predicting and diagnosing the disease to enhance necessary treatment. One more significant drive behind this study is to advance on the works of previous researchers who make their own contributions in this particular field of study.

III. Review of Literature

Prof.Thirunavukkarasu has predicted diseases in the liver using three various types of ML classification models algorithms like K Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression (LR). KNN and Logistic Regression algorithms had 73% accuracy, while SVM had 71.97%. They used the 70:30 ratio for training and testing their models. They Found that KNN and Logistic Regression had more prediction accuracy compared to others [1].

Maria Alex Kuzhippallil used multiple machine-learning techniques and compared them to the Indian Liver Patient dataset from Kaggle. It has 416 patient data with 11 features that were used for the project. Feature selection was done by using a Genetic algorithm. Machine learning models like Gradient Boost, Random Forest (RF), Multilayer Perceptron (MLP), K Nearest Neighbor (KNN), AdaBoost, Logistic Regression (LR), Decision Tree (DT), XGBoost. They found that it resulted in a higher prediction Rate using the feature selection method [2].

Vyshali J Gogi used machine learning algorithms like Support Vector Machine (SVM) Linear Discriminant, and LR, Decision Tree applied for the classification of the liver disease dataset. The parameters were collected through Liver Function Testing from the blood. Logistic Regression (LR) resulted in 95.8% accuracy among all other classification algorithms used [3].

Shivangi Gupta identified Liver diseases by using Machine Learning Algorithms They used different Machine Learning models like Random Forest (RF), Naïve Bayes (NB), AdaBoost, and Support Vector Machine (SVM). From the "https://archive.ics.uci.edu/ml" the Indian Liver Patient Dataset was accessed. 583 patients data with 11 features were taken for the project study and training and testing of the ML models. 70:30 ratio that is 70% of the data from the ILPD dataset was used for the training purpose and 30% of the ILPD datasets were used for the testing purpose of the models. Different ML Models K Nearest Neighbour, ANN, and DT together gained a prediction rate of 93%[4].

Dongyang Zhang studied the classification of human body liver data by comparing the performance of XGBoost and LightGBM. The dataset was taken from the Kaggle platform. There were 30 features and 8785 observations. XGBoost performed with an accuracy rate of 0.75%, where XGBoost - FA accuracy was 0.67%, and LightGBM secured an accuracy rate of 0.75% where LightGBM - FA performed with an accuracy rate of 0.672%[5].

Bendi et al. authors used two different input dataset and evaluate that datasets has better than UCLA dataset for all the different selected algorithms. Based on performance on their classification KNN, Backward propagation and SVM are giving better results. The AP data set is better than UCLA for the entire selected algorithm. And found out Naïve Bayes, C4.5, KNN, Backward propagation and SVM has 95.07, 96.27, 96.93, 97.47, and 97.07% accuracy respectively [6].

Joel Jacob et al. proposed a paper to diagnosis of liver disease by using three different algorithms, Logistic regression, K-Nearest Neighbor, SVM, and ANN and used Indian Liver Patient Dataset comprised of 10 different attributes of 583 patients. And concluded Logistic regression, KNN, SVM, and ANN has 73.23, 72.05, 75.04 and 92.8% accuracy respectively [7].

Bendi et al. proposed a paper based on Modified Rotation Forest, used two dataset as an input UCI liver dataset and Indian liver dataset. And results show that MLP algorithm with random subset gives better accuracy of 94.78% for UCI dataset than CFS achieved accuracy of 73.07% for Indian liver dataset.[8].

The prime purpose of this project is to diagnose Liver diseases and predict whether a patient is suspected of any particular liver disease and it will try to classify the specific type of disease of the Liver. Using multiple ML Classification Models, we will Then select the model that predicts the final result with the highest accuracy. At first, the system asks for your liver report features such as total bilirubin, total proteins, direct bilirubin, albumin, alkphos, gender, and age. When the system receives these inputs from the user, it compares them with the most accurate training dataset and then predicts the outcome accordingly. When a person is diagnosed with liver disease, the system determines if the disease is minor or major according to what precautions are required and what type of liver disease will occur.

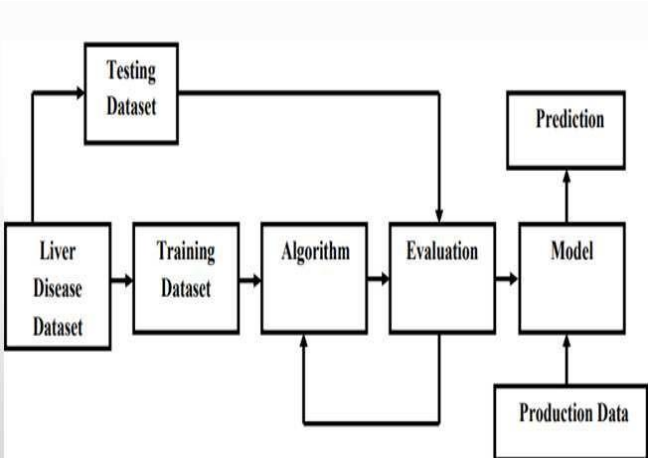

Fig. 2. Overview of the Data Flow Diagram.

The general data flow of the system is given in the above diagram. It has gathered the dataset, Training model, testing model, evaluation, and final model. The dataset can be gathered through various open platforms like Kaggle and UCI Repository for datasets. Then we will apply various Machine Learning Algorithms for predicting and deciding the result. The normal ratio for dividing the gathered dataset for training the model and testing the model is 80:20 where 80% of the dataset will be used for training purpose and 20% will be used for testing purpose. Then after that, we will evaluate our model with the 20% testing dataset and will predict the result. The final model will be then used for the prediction and detection of various liver diseases.

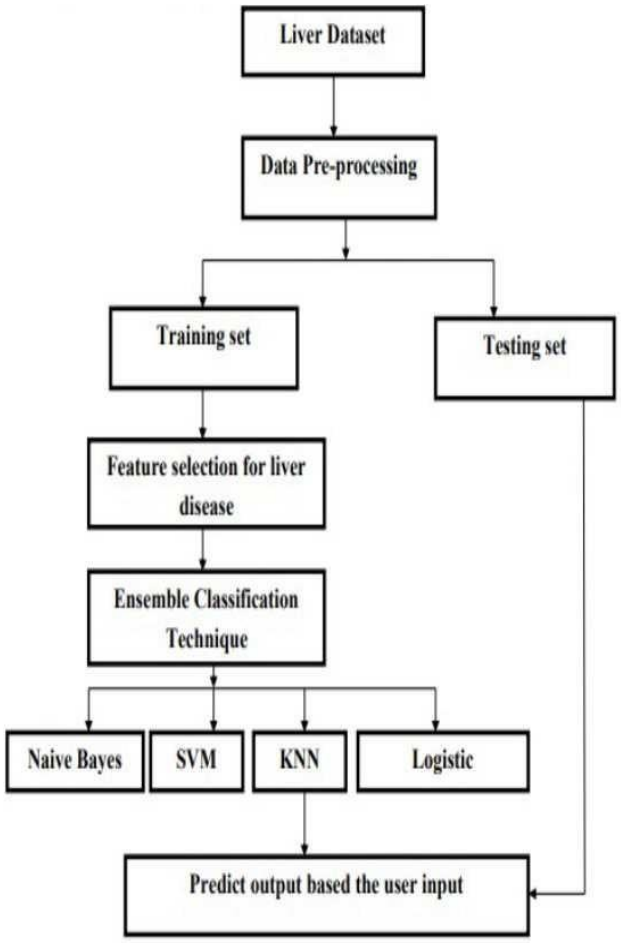

Fig. 3. Overview of the Workflow Diagram.

For every machine learning project, gathering data is critical. The data set can be acquired from a variety of sources, such as files, databases, and many other sources, depending on the type of project. After gathering our data from platforms like Kaggle and the UCI ML repository, preprocessing is done. Preprocessing of data is an important step in machine learning. To build more accurate machine learning models, this step is important and crucial. In ML, training and testing is done in 80/20 ratio. In order to train a model, dataset split it into two sections: first is training and the second is testing the dataset. An ML classifier model is trained and then it is evaluated and tested on an unseen 'test data set'. It is necessary to consider that only the training set is available during classifier training. The test set will be used during the testing of the model. In the case of liver disease, the dataset gets divided into 80/20. Next, then applied different classification techniques.

IV. Algorithms

1] Support Vector Machine(SVM)

"In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis." Support vector used to classify models.[9]

2] Logistic Regression(LR)

"It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1" Logistic Regression will produce the in the form of Yes or No.[10]

3] Naive Bayes(NB)

"In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features (see Bayes classifier). They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve high accuracy levels" it is class of probabilistic classifiers.[11]

4] K-Nearest Neighbor(KNN)

"k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically".[12]

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

Sr.No	Authors	Year	Disease	Machine learning algorithm	Dataset input	Remarks	Conclusion
1	Prof.Thirunavukkarasu	2018	Prediction of liver disease	KNN, SVM, LR	ILPD	KNN, SVM, LR has 73% and 71.97% accuracy.	Found that KNN and logistic regression had more accuracy.
2	Maria Alex Kuzhippll	2020	Liver disease	Gradient Boost, RF, MLP, KNN, AdaBoost, LR	ILPD	Gradient Boost, RF, MLP, KNN, AdaBoost, LR has 69.43%	It gives higher prediction rate using the feature selection method.
3	Vyshali J Gogi	2018	Prognosis of Liver Disease	SVM, Linear Discriminant, LR, Decision Tree	Parameter collected through liver function testing from the blood.	SVM, Linear Discriminant, LR, Decision Tree has 95.8% accuracy.	Logistic regression gives high accuracy than other algorithm.
4	Shivangi Gupta	2020	Identify of liver disease	RF, NB, AdaBoost, SVM	ILPD	RF, NB, AdaBoost, SVM has 93% accuracy	KNN, ANN, DT together give a better prediction rate.
5	Dongyang Zhang	2020	Prediagnosis of liver failure	XGBoost, LightGBM	Acute liver failure	XGBoost, LightGBM has 0.75% and 0.67% accuracy	LightGBM gives better accuracy than XGBoost.
6	Bendi Venkata	2012	Diagnosis of Liver Disease	KNN, Backward Propagation and SVM	AP has better dataset result than UCLA	KNN, Backward Propagation and SVM has 95.07% and 96.27% accuracy	Used two dataset and evaluate that dataset better than UCLA dataset.
7	Joel Jacob	2018	Diagnosis of liver disease	Logistic regression, K-nearest Neighbor, SVM, ANN	Indian liver patient dataset	Logistic regression, K-nearest Neighbor, SVM, ANN has	More accuracy result in ANN Algorithm.

						73.23%,72.05%,75.04% and 92.8% accuracy	
8	Bendi Venkata	2012	Liver Disease	Modified Rotation Forest	UCI liver dataset and Indian dataset	MLP algorithm with random subset gives better accuracy 74.78% than NN with CFS of accuracy 73.07%	MLP algorithm with UCI liver dataset has better accuracy than NN with Indian liver dataset.

Table 1: Comparison table on existing machine learning technique

V. Conclusion

Using a hybrid classifier as a machine learning model, the proposed system to diagnose liver disease that will increase prediction and help users in identify the disease and prescribe further treatment and examinations with more consciousness

References

[1] k. Thirunavukkarasu, A. S. Singh, M. Irfan and A. Chowdhury, "Prediction of Liver Disease using Classification Algorithms," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India.

[2] M. A. Kuzhippallil, C. Joseph and A. Kannan, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020.

[3] V.J.Gogi and V.M.N., "Prognosis of Liver Disease: Using Machine Learning Algorithms," 2018 International Conference on Recent Innovations in Electrical, Electronics, and Communication Engineering (ICRIEECE), Bhubaneswar, India.

[4] S. Gupta, G. Karanth, N. Pentapati, and V. R. B. Prasad, "A Web- Based Framework for Liver Disease Diagnosis using Combined Machine Learning Models," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India.

[5] K. Prakash, S. Saradha: A Deep Learning Approach for Classification and Prediction of Cirrhosis Liver: Non-Alcoholic Fatty Liver Disease (NAFLD)(2022).

[6] 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.

[7] 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).

[8] 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.

[9] Support Vector Machine (2022) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Support_vector_machine (Accessed: January 24, 2023).

[10] Logistic regression (2023) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Logistic_regression (Accessed: January

979-8-3503-9781-9/23/$31.00 ©2023 IEEE

24, 2023).
[11] Naive Bayes classifier (2023) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Naive_Bayes_classifier (Accessed: January 24, 2023).
[12] K-nearest neighbors algorithm (2022) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm (Accessed: January 24, 2023).
[13] Gullo, F. From patterns in data to knowledge discovery: what data mining can do. Phys. Procedia 62, (2015).
[14] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac "Diagnosis of Liver Disease Using Machine Learning Techniques", International Research Journal of Engineering and Technology (IRJET)
[15] Sivakumar D, Manjunath Varchagall, and Ambika L Gusha S" Chronic Liver Disease Prediction Analysis Based on the Impact of Life Qual- ity Attributes."(2019), International Journal of Recent Technology and Engineering (IJRTE)
[16] H. Jin, S. Kimand J . Kim, Decision Factors on Effective Liver Patient Data Prediction, International Journal of Bio-Science and Bio-Technology.
[17] C. Geetha, Dr. AR. Arunachalam: Evaluation-based Approaches for Liver Disease Prediction using Machine Learning Algorithms (2021).
[18] Kandasamy Sellamuthu, Sylviya P, Pugazharasi K, Rajalakshmi S: Liver Disease Prediction using Logistic Regression (2022).
[19] Prof Christopher N. New Automatic Diagnosis of Liver Status Using Bayesian Classification.

979-8-3503-9781-9/23/$31.00 ©2023 IEEE