Liver Disease Severity Detection Using Machine Learning Algorithms

M. Prameela
Department of CSE,
Amrita Sai Institute of Science and Technology
Paritala, Andhra Pradesh, India
prameelamotha@gmail.com

M. Shiva Rama Krishna
Department of CSE
Amrita Sai Institute of Science and Technology
Paritala, Andhra Pradesh, India
shivachowdarymadala@gmail.com

G. Vijay Kumar
Department of CSE
Amrita Sai Institute of Science and Technology
Paritala, Andhra Pradesh, India
gvk.vijay73@gmail.com

Abstract*—Machine learning is a powerful technique used to uncover meaningful patterns within vast amounts of data, enabling machines to learn and make informed decisions. This paper focuses on the application of supervised learning, specifically using a Liver Patient dataset sourced from the UCI Repository. The dataset comprises comprehensive information on patients undergoing medical examinations, particularly those with liver conditions. The collected data serves as a valuable resource for enhancing future patient care. The study employed a range of sophisticated algorithms including SVM, Naive bayes and random forest to forecast liver patient outcomes. Through meticulous analysis and result computations, it was determined that these algorithms demonstrated impressive levels of accuracy. This analysis provided valuable insights into the severity and progression of liver disease, enriching our understanding of the patient's overall health status of liver.

Keywords: ***Machine Learning, supervised learning, dataset, random forest, accuracy.***

I. INTRODUCTION

The liver is situated in the upper right quadrant of the abdominal cavity, commands an eminent presence as one of the largest organs in the human body. Its anatomical structure takes on a distinctive wedge shape, underscoring its prominence. Functionally, this remarkable organ reigns as the preeminent gland, orchestrating the secretion of a multitude of chemical substances known as hormones. The liver is a powerhouse organ, it undertakes a staggering repertoire of over 500 functions crucial to human physiology, cementing its indispensability for existence [1]. Remarkably, it serves as a linchpin, providing vital support to the majority of essential organs that are imperative for our survival.

In the realm of adult anatomy, the liver manifests its presence by accounting for approximately 2% of the total body weight. Within the male populace, this organ exhibits a weight ranging between 1.4 to 1.8 kilograms, while in females, it registers a slightly lower mass of 1.2 to 1.4 kilograms. In the delicate stage of newborns, the liver assumes a dainty stature, weighing a mere 150 grams. Functionally, this remarkable organ fulfills an array of crucial roles:

Firstly, its regal authority by skillfully secreting vital substances such as bile and glycogen;

Secondly, it takes charge of the intricate synthesis of serum protein lipids;
Thirdly, it assumes the formidable responsibility of purging the blood from both endogenous and exogenous toxins, drugs, and alcohol, thereby safeguarding our internal equilibrium;

Finally, it stands as a vigilant guardian, storing essential vitamins including D, A, K, E, and B1, ensuring a well-stocked reserve for our physiological well-being.

Liver disease, characterized by hepatic inflammation due to the presence of harmful substances, bacterial infections or hereditary conditions, disrupts the vital functions of digestion and bacterial elimination, which hinge upon a properly functioning liver [2].

Prevalently affecting individuals within the age bracket of 40 to 60 years, liver diseases disproportionately afflict the male population. Astonishingly, an alarming number of 1 million cases of liver disease are diagnosed annually, resulting in a staggering death toll of 1,40,000 per year in India. Machine Learning, an integral facet of Artificial Intelligence (AI), emulates human intelligence by endowing machines with the ability to learn and emulate human-like actions. In essence, Machine Learning imparts knowledge to systems without the need for explicit programming, facilitating autonomous learning and decision-making [3].

In the domain of supervised algorithms, the training process relies on utilizing user inputs and corresponding outputs to facilitate accurate predictions. The remarkable field of machine learning has not only expanded its horizons but has also permeated the realm of healthcare. One of the significant challenges confronted by the healthcare sector is the escalating number of patients seeking medical attention. Leveraging the potential of machine learning, applications within the healthcare domain have the capability to significantly enhance the precision of treatment. By employing sophisticated classification techniques, various automated medical diagnostic methods can effectively contribute to the early detection of liver disease, a condition

often challenging to identify in its initial stages due to the organ's resilient functioning despite partial destruction. Early diagnosis of liver problems plays a pivotal role in elevating patients' survival rates, underlining the crucial importance of leveraging machine learning for timely intervention and improved healthcare outcomes. [4]

This study focuses on utilizing the presence of enzymes in the bloodstream as a means to identify liver disease. The liver patient dataset is employed to predict whether individuals have liver disease or not. Within this paper, an assortment of Naive bayes, Random Forest, SVM have been meticulously compared for their efficacy in liver disease prediction. Notably, the study addresses and rectifies the limitations overlooked by prior researchers, leading to enhanced prediction accuracy [5].

To accurately predict liver patients from the dataset, a systematic approach was followed, encompassing Exploratory Data Analysis (EDA), comprehensive data pre-processing, outlier removal, Synthetic Minority Over-sampling Technique (SMOTE) application, as well as the utilization of various base and advanced classifiers and algorithms [6]. To tackle the problem at hand, a total of 583 records/entries were meticulously collected from the Indian Liver Patient Dataset (ILPD) [citation/reference].

The dataset utilized in this study was procured from the esteemed UCI Machine Learning Repository, accessible at the web address http://archive.ics.uci.edu/ml/. Within this comprehensive dataset, an assemblage of 583 Indian individuals with liver-related conditions is meticulously documented. Among these entries, a total of 416 instances pertain to patients afflicted with liver ailments, while 167 records correspond to individuals without liver-related issues. The data collection process was conducted in Andhra Pradesh, a region situated in the north-eastern part of India.
In the context of this study, the term "selector" assumes the role of a discerning class label, serving as a means to categorize and distinguish objects into distinct groups, namely those classified as liver patients and those classified as non-liver patients.

II. LITERATURE SURVEY

The comprehensive research conducted by Hartatik, Mohammad Badri Tamam, and Arief Setyanto focused on exploring the efficacy of Python applications in solving prediction challenges for patients afflicted with liver illness. The study employed the Indian Liver Patient Dataset procured from the esteemed UCI Machine Learning Repository (ILPD) . Notably, by employing a prediction model incorporating six variables, the Naive Bayes algorithm showcased superior performance compared to the KNN algorithm, yielding enhanced accuracy when compared to previous studies [7][8]. Similarly, Thirunavukkarasu K, Ajay S. Singh, Md Irfan, and Abhishek Chowdhur undertook a comprehensive study employing various classification techniques, including Logistic Regression, Support Vector Machine, and K-Nearest Neighbor, for liver illness prediction. These algorithms were meticulously compared based on their classification accuracy, as determined by the confusion matrix. While both Logistic Regression and K-Nearest Neighbor demonstrated high accuracy, logistic regression emerged as the method with the highest sensitivity in the experimental setting. Hence, it can be concluded that Logistic Regression stands as a reliable approach for predicting liver illness [9].

III. PROPOSED METHODOLOGY

The elucidation of each component is presented below:

Data Acquisition:
The initial step involves acquiring the requisite dataset for the study. In this case, the Indian Liver Patient Dataset from the UCI Machine Learning Repository (ILPD) serves as the primary source.

Data Preprocessing:
The process of data preprocessing involves various techniques to ensure the quality and completeness of the dataset. One such technique is the imputation of missing values, which plays a crucial role in handling incomplete data. In the context of the Indian liver disease patient's dataset, there were instances where the Albumin and Globulin ratio had missing values. To address this, a robust approach was employed to restore these missing values using the median values [11]. By imputing the missing values with the median, the dataset achieves enhanced completeness and maintains the integrity of the analysis, paving the way for more reliable and accurate predictions.

Feature selection:
Feature selection is a meticulous process of curating the optimal subset of input variables, strategically tailored to enhance the efficiency and effectiveness of the machine learning algorithm. By carefully handpicking and retaining the most relevant and informative features, this process not only expedites the model training but also mitigates computational complexities, resulting in streamlined operations. Moreover, feature selection enables a clearer and more interpretable representation of the underlying patterns and relationships within the data, facilitating a deeper understanding of the predictive factors influencing the outcome. Through judicious selection, feature selection acts as a discriminating filter, spotlighting the essential attributes and empowering the model to focus on the most influential variables, ultimately culminating in refined and expedited machine learning algorithms.

Random Forest Feature Selection:
Feature selection plays a pivotal role in model development, and the utilization of Random Forest algorithm for feature selection offers a plethora of advantages. By harnessing the power of Random Forest, feature selection attains unparalleled precision, mitigating the risks of overfitting and delivering reliable outcomes. Moreover, the interpretability of the selected features is greatly enhanced through the evaluation of their importance in the decision-making process of individual trees within the Random Forest ensemble. The utilization of Random Forest ensures the independence and decorrelation of the constituent decision

trees, effectively reducing the likelihood of overfitting. Each decision tree captures distinct conditions based on one or more attributes, thereby unraveling the intrinsic relationships and salient patterns within the dataset. This comprehensive approach to feature selection not only guarantees accurate and robust results but also provides an intuitive framework for understanding the underlying mechanisms governing the predictive capabilities of the chosen features.

Classification using Machine Learning Algorithms:
To train the machine for classification tasks, a diverse set of machine learning algorithms were employed, each offering unique approaches to model training and prediction. These algorithms encompass a wide spectrum of methodologies, enabling the machine to effectively learn and generalize from the provided data.

SVM:
Support Vector Machine (SVM) is a powerful supervised machine learning algorithm predominantly employed for classification tasks, although it can also be utilized for regression problems. The fundamental aim of the SVM algorithm is to identify an optimal hyperplane within an N-dimensional space that effectively segregates the data points into distinct classes. The dimensionality of the hyperplane is contingent upon the number of input features present in the dataset. In cases where there are two input features, the hyperplane takes the form of a line, while with three input features, it manifests as a two-dimensional plane. However,envisioning hyperplanes in spaces with more than threefeatures becomes progressively challenging, requiring advanced cognitive abilities.

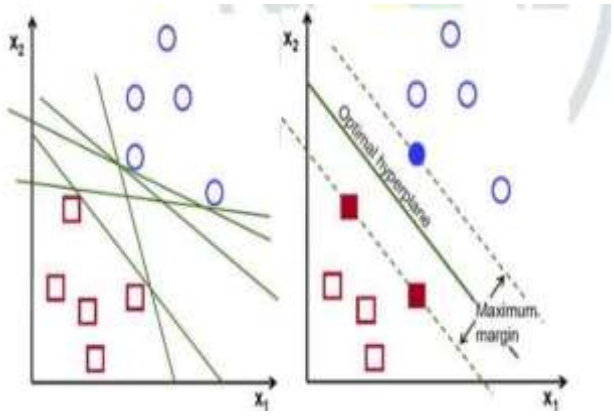

Fig.1 Support vector machine

Naïve Bayes Algorithm
The Naïve Bayes' classifier employs the principles of Bayes' theorem for statistical classification. It is characterized as an eager learning algorithm, swiftly classifying new instances without waiting for test data. Its classification capabilities are akin to those of Neural Networks and Decision Trees [15, 17]. The sklearn library provides a diverse range of Naive Bayes classifiers, including GaussianNB and Multinomial NB.

Assumptions of Naïve Bayes:

Naïve Bayes operates under two key assumptions:

1. Independence Assumption: It assumes that the predictors are mutually independent, meaning that each predictor provides unique and unrelated information for the classification task. This assumption enables efficient computation and streamlined modeling.

2. Equal Effect Assumption: Naïve Bayes assumes that all predictors exert an equal influence on the outcome. In other words, each predictor contributes equally to the decision-making process.

The Naïve Bayes classifier leverages Bayes' theorem to perform classification, employing the following formula:
$P(t|X) = P(t) * P(X|t) / P(X)$.....(1)
In this equation, $P(t|X)$ represents the **conditional probability of class 't'** given the predictor X, commonly referred to as the **posterior probability**.

$P(X|t)$ denotes the probability of observing predictor X given class 't', while $P(X)$ signifies the **probability of observing predictor X.**

By utilizing these principles, the Naïve Bayes algorithm offers a robust framework for statistical classification, enabling rapid and accurate classification of new instances based on the available data.

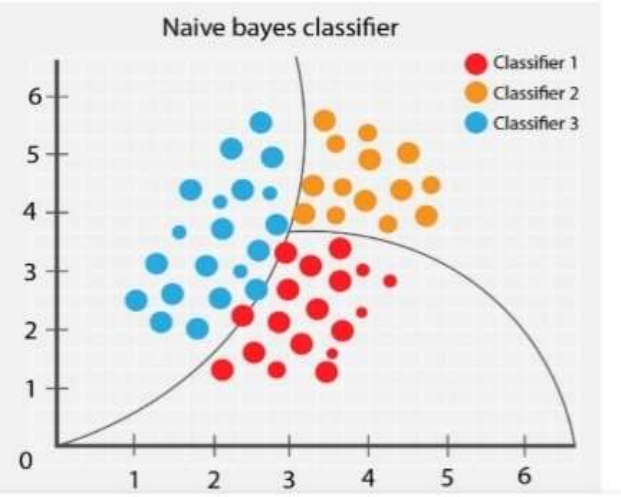

Fig.2 Naïve bayes classifier

Random Forest Algorithm:
Random Forest is an ensemble learning algorithm that operates on the principle of bagging. It comprises multiple independent decision trees functioning as a cohesive ensemble. The selection of variables is carried out randomly, and each tree is trained on bootstrap samples. The final prediction is obtained by aggregating the outputs from these individual trees. To construct multiple trees, the bootstrap sampling technique is employed, wherein random samples of equal size are drawn from the dataset. This method is versatile and can be utilized for both regression

and classification analyses. In regression, it calculates the average

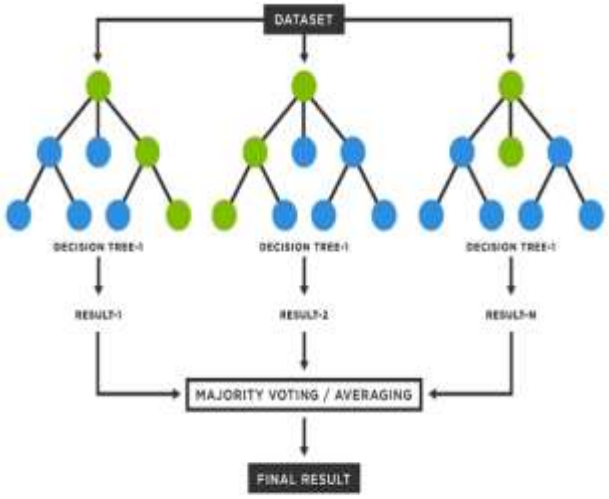

Fig.3 Random Forest

of all predicted values, while in classification, it determines the class with the highest frequency among all predicted classes. By constructing decision trees from random data samples, Random Forest effectively mitigates the issue of overfitting, enhancing the robustness of the algorithm.

IV. EXPERIMENTAL RESULT

In this section, we meticulously assess the efficacy of our model by conducting a comprehensive evaluation using the ILPD dataset, thereby enabling a thorough comparison of the obtained outcomes.

Confusion matrix: It is a tabular representation that allows us to compare the predicted and actual values of a target variable. It provides a clear and concise summary of the performance of a classification model, showing the true positive, true negative, false positive, and false negative values. By examining these values, we can assess the accuracy and effectiveness of the model in making predictions.

Accuracy: Accuracy is a measure that quantifies the correctness of predictions made by a model. It represents the proportion of records that are correctly predicted by the model. A higher accuracy value indicates a superior performance of the model. Accuracy is computed by dividing the number of correctly predicted records by the total number of records. Mathematically, it can be expressed as the ratio of true positives (TP) and true negatives (TN) to the sum of true positives, true negatives, false positives (FP), and false negatives (FN).

Accuracy = (TP+TN)/(TP+TN+FP+FN)

V. CONCLUSION:

The investigation and analysis of liver disease severity detection in patients have been thoroughly examined in this

ALGORTHIMS	ACCURACY
Random Forest	85%
SVM	74%
Naive Bayes	71%

Table.1 Evaluation results

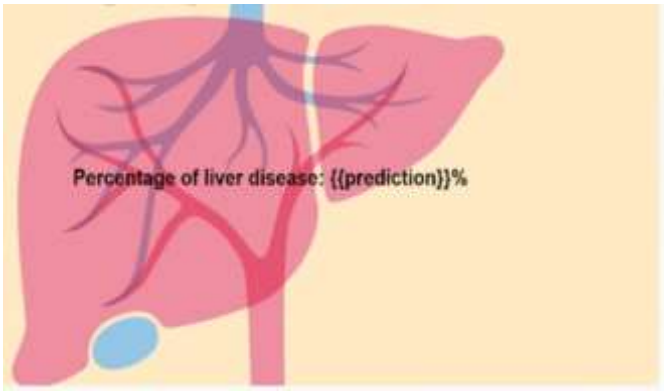

Fig.4 percentage of liver disease

research paper. Multiple sophisticated techniques were employed to ensure the cleanliness of the data, including the imputation of missing values using the median. Various state-of-the-art classification algorithms such as support vector machines, random forest, and Naive Bayes were rigorously applied and evaluated. Remarkably, the results have shown that the Random Forest algorithm surpasses the performance of other classification algorithms in terms of accuracy. Thus, it can be conclusively stated that Random Forest is a highly suitable approach for accurate liver disease prediction. This analysis provided valuable insights into the severity and progression of liver disease, enriching our understanding of the patient's overall health status of liver.

REFERENCES

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

[2]. S. K. B. Sangeetha, N.Afreen, and G. Ahmad, "A Combined Image Segmentation and Classification Approach for COVID-19 Infected Lungs," J. homepage http//iieta. org/journals/rces, vol. 8, no. 3, pp. 71–76, 2021.

[3]. M. Sameer, A. K. Gupta, C. Chakraborty, and B. Gupta, "Epileptical Seizure Detection: Performance

analysis of gamma band in EEG signal Using Short-Time Fourier Transform," in 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC), 2019, pp. 1–6, doi: 10.1109/WPMC48795.2019.9096119.

[4]. A. Mahajan, K.Somaraj, and M. Sameer, "Adopting Artificial Intelligence Powered ConvNetTo Detect Epileptic Seizures," in 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 427–410.1109/IECBES48179.2021.9398832.

[5]. N. Nasir, N. Afreen, R. Patel, S. Kaur, and M. Sameer, "A Transfer Learning Approach for Diabetic Retinopathy and Diabetic Macular Edema Severity Grading," Rev. intelligence Artif., vol. 35, pp. 497–502, Dec. 2021, doi: 10.18280/ria.350608.

[6]. M. Sameer and B. Gupta, "ROC Analysis of EEG Subbands for Epileptic Seizure Detection using Naive Bayes Classifier," J. Mob. Multimed., pp. 299–310, 2021.

[7]. M. Sameer and B. Gupta, "Time–Frequency Statistical Features of Delta Band for Detection of Epileptic Seizures," Wirel. Pers. Commun., 2021, doi: 10.1007/s11277-021-08909-y.

[8]. S. M. Beeraka, A. Kumar, M. Sameer, S. Ghosh, and B. Gupta, "Accuracy Enhancement of Epileptic Seizure Detection: A Deep Learning Approach with Hardware Realization of STFT," Circuits, Syst. Signal Process., 2021, doi: 10.1007/s00034-021-01789-4.

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