Machine Learning Techniques in Analysis and Prediction of Liver Disease

Dr. Dattatreya P Mankame[1], Harshitha R[2], Navya N C[3], Nitin Ravichander[4]
[1]*Professor, Dept. of CS & E, Atria Institute of Technology, Bangalore*
[2,3,4]*Student, Dept. of CS & E, Atria Institute of Technology, Bangalore, Karnataka, India*

***Abstract* - Machine Learning features a strong potential in automated diagnosis of varied diseases. The liver plays a very important role in life which supports the removal of toxins from the body. With the recent upscale in various liver diseases, it's necessary to identify liver disease at a preliminary stage. India having a population of 1.33 billion is the second most populated country within the world and every year millions are diagnosed with liver diseases. Some sorts of liver diseases are Jaundice, Hepatitis (A, B, C), Non-Alcoholic liver disease Diseases (NAFLD). These are caused by the consumption of alcohol, contaminated food, and certain factors such as obesity. Thus, we would prefer a system that is reliable and may predict the symptoms of liver diseases. This technique predicts liver diseases using the patient's data and by using Machine Learning algorithms. From the experiments and comparative analysis, it increases classification accuracy and also leads to reduction in classification time and hence it aids for prediction of liver disease more efficiently. The performance is measured in terms of accuracy, auc score, precision, recall and f-measure. Several classification algorithms are used and based on the classification report and performance, the best model is chosen and employed to classify liver patients (Liver patient or not liver patient).**

***Index Terms* - Classification models, Data visualization Feature selection, Liver disease, Machine learning, SVM Algorithm.**

I.INTRODUCTION

The liver is the largest organ of the body and it is essential for digesting food and releasing the toxic elements of the body. The viruses, parasites and alcohol use leads the liver towards liver damage which results in life-threatening conditions. There are many types of liver diseases like hepatitis, cirrhosis, liver tumors, cancer of the liver, and lots of more. Among them are liver diseases and cirrhosis as the main cause of death [1].

Thus, liver disease is one of the major health problems in the world. Every year, around 2 million people die worldwide because of liver disease [2].

According to the worldwide Burden of Disease (GBD) project, published in BMC Medicine, a million people lost their lives in 2010 due to cirrhosis and million are affected by liver cancer [3]. Machine Learning (ML) a part of Artificial intelligence (AI) allows the system to obtain knowledge by using algorithms and statistical models to analyse and draw inferences from patterns in data. Supervised algorithms make use of human inputs and outputs for training process and prediction accuracy, and thus used for different classification applications [1]. Therefore, the application of ML has extended to healthcare as well. One of the major problems in healthcare is the rising number of liver disease patients. Liver is a vital organ with functionalities like production of bile, detoxification of chemicals and production of important proteins for blood clotting and various life sustaining functions [2]. Long term drinking habits have been directly linked to the increased risk of having different liver diseases which may further lead to death which can be prevented if the disease is detected earlier.

II. LITERATURE SURVEY

In paper [1], liver disease prediction has been studied and analysed. The data is cleaned by performing various techniques such as imputation of missing values with median, label encoding to convert categorical into numerical data for easy analysis, duplicate value elimination and outliers are eliminated in order to improve the performance. Different classification algorithms are used to predict the presence or absence of liver disease.

The paper [2] proposes the research work of Naïve Bayes and Support Vector Machine (SVM) classifier algorithms used for liver disease prediction and analysis of the algorithms shows that Liquor is consumed in overabundance by a large number of individuals all over the world. The literature survey from this paper conclude the use of Naive Bayes and Support Vector Machine algorithms for the prediction of liver diseases. There are two major parameters that are involved in understanding the suitability of the respective methodologies and they are - the time taken to execute the prediction process and the accuracy of the predictive result. It is clear through various studies and experimentations that the SVM classifier is the best of all the algorithms owing to the extremely high accuracy rates. But when it comes to the time taken to execute the predictive process, the Naive Bayes classifier reflects higher suitability since it takes the least possible time to execute the process.

The paper [3] proposes that the liver is the largest organ of the body and it is essential for digesting food and releasing the toxic element of the body. The viruses and alcohol use lead the liver towards liver damage and lead a human to a life-threatening condition. There are many types of liver diseases, such as hepatitis, cirrhosis, liver tumors, liver cancer, and many more. Among them are liver diseases and cirrhosis as the main cause of death. Therefore, liver disease is one of the major health problems in the world. Every year, around 2 million people die worldwide because of liver disease. Machine learning has made a significant impact on the biomedical field for liver disease prediction and diagnosis. Machine learning offers a guarantee for improving the detection and prediction of disease that has made an interest in the biomedical field and they also increase the objectivity of the decision-making process. By using machine learning techniques medical problems can be easily solved and the cost of diagnosis will be reduced. In this study, the main aspect is to predict the results more efficiently and reduce the cost of diagnosis in the medical sector. Therefore, we used different classification techniques for the classification of patients who have liver disease or not.

The paper [4] describes the Classification algorithms are very much suitable and used in different automated medical diagnosis tools. Initial examination of liver disorders will decrease patients' mortality rate. Otoom et al. (2015) proposed a system to detect and monitor coronary artery disease. Two tests with three algorithms- Bayes Net, Support vector machine, and Functional Trees FT were used. The WEKA tool was used for detection. Test was done on 7 best selected features, Bayes Net attained 84.5% of correctness, SVM gave 85.1% accuracy and FT classified 84.5% correctly. et al (2015) used Naive Bayes algorithm for diagnosis of heart disease. Therefore, it has a powerful independence assumption. Weka was used as a tool which executed 70% of percentage split. Naive Bayes had 86.419% of accuracy. Iyer et al. (2015) conducted an experiment to predict diabetes disease with the help of decision tree and Naive Bayes. Different tests were carried out using WEKA data mining tool. Naive Bayes showed 79.5652% correctness by using percentage split test. Maximum accuracy of algorithms was obtained by using percentage split test. Vijayarani and Dhayanand (2015) used Support vector machine and Naive bayes classification algorithms for liver disease prediction. Data analysis was done using MATLAB. Naive bayes gave 61.28% correctness in 1670.00 ms and SVM gave 79.66% accuracy in 3210.00 ms. Gulia et al. (2014) studied on intelligent techniques to classify liver patients using datasets from UCI. The WEKA tool and five algorithms- J48, MLP, Random Forest, SVM and Bayesian Network were for experimentation. After FS, algorithms gave highest accuracy as- J48 70.669%, MLP 70.8405%, SVM 71.3551%, Random forest 71.8696% and Bayes Net gave 69.1252% accuracy.

The study on paper[5] surveyed some data mining techniques to predict liver disease at an earlier stage. The study analysed algorithms such as C4.5, Naive Bayes, Decision Tree, Support Vector Machine, Back Propagation Neural Network and Classification and Regression Tree Algorithms. These algorithms give various results based on speed, accuracy, performance and cost. It is seen that C4.5 gives better results compared to other algorithms. In future an improved C4.5 could be derived with various parameters. This paper gives generalization of various data mining techniques to diagnose liver disease at an earlier stage.

The paper [6] shows the predicting and analysing liver diseases and gives better performance accuracy by comparing various data mining classification algorithms. The performance of the accuracy is measured by confusion matrices with the help of a classification report. In the proposed methodology system of this paper, they decided to use three classification algorithms to predict the liver disorder diseases by comparing the performance accuracy of each classification algorithm. The classifiers are SVM, NB and C4.5 decision tree classifiers. K-fold cross-validation was used in data partitioning based on the Test set, used to test the model and Training set, used to train the data, and concludes that popular classification algorithms such as SVM, NB and C4.5 Decision Tree considered for performance evaluation in liver disorder diseases prediction.

III. PROPOSED METHODOLOGY

The whole ideology of the research paper is to help people detect liver disease at the earliest stage and describe the usage of various classification techniques used in machine learning based on performance metrics such accuracy, f-measure, precision, recall and auc. By this we can get a clear picture of how the algorithm works on liver dataset and how it gives a high efficiency model for predicting liver disease.

A. Data Set Selection

Data selection process involves the selection of appropriate data for analysis to perform data selection techniques. The dataset used is Indian Liver Disease Patients (ILDP) available in UCI repository. The Dataset contains 416 liver patient records and 167 non liver patient records collected from Northeast of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patients (liver disease) or not (no disease). It contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed.

B. Data Visualization and Exploration

Different histograms were plotted to understand the distribution of numerical features of different columns in the dataset as well as to visualize the target variable count and the gender difference. This shows that age is an important factor to be considered during the detection of liver disease. With data visualization by joinplots and scatterplots we could understand that there is a linear relationship between Direct Bilirubin and Total Bilirubin, Aspartate Aminotransferase and Alamine Aminotransferase, Total Proteins and Albumin, Albumin and Globulin Ratio & Albumin, Total Proteins and Albumin and Globulin Ratio.

C. Data Pre-processing

- Filling of Missing Values - It is the process of identifying the missing variables and adding the mean values. For Indian Liver Disease Patients dataset the Albumin and Globulin ratio has four missing values which is replaced by considering the mean of that column and, once checked after filling the mean values to those rows, Albumin and Globulin ratio had no null values.
- Identifying Duplicate Values - Duplicate values were identified and by the observations we can see around 13 duplicate values but for a medical dataset duplicate values can exist and thus we are not dropping any of the duplicate values.
- Outlier Detection - Outliers are very high values that have an extreme range compared to other values of that particular column which can cause improper measurements in data. In Indian Liver Disease Dataset we can observe an outlier in Aspartate Aminotransferase even though the value is high as we can consider this outlier as earlier predictions have seen such a range. Thus, no outlier has been removed.
- Label Encoding - Encoding of data being very vital for Data Pre-processing is a process of converting textual data into machine readable format which are numbers. Label Encoding was done for Male and Female converting into 1 and 0 as well as the target variable which is the dataset where the values were encoded from 1 and 2 to 1 and 0 for better analysis.
- Resampling - Because of the imbalance in the dataset where we can observe a majority in liver disease patients and a minority in non-liver disease patients, smote is used to synthesize new samples for minority.

D. Feature Selection

Feature Selection is a process of figuring out which inputs are the best for the model and checking if there is a possibility of eliminating certain inputs. Considering Indian Liver Dataset, we can see a very high linear relationship between Total and Direct Bilirubin and by considering this linear relationship, Direct Bilirubin can be opted to be dropped, But by as per medical analysis Direct Bilirubin constitutes to almost 10% of the Total Bilirubin and this 10% may prove crucial in obtaining higher accuracy for the model, thus none of the features are removed.

E. Applying machine learning models

Several classification algorithms were used on the liver disease dataset which includes

- Logistic Regression – Used for solving classification problems. Here, we predict the values of categorical variables using a set of independent variables and the dependent features that are the target variables are in the form for categorical data. For liver disease prediction the target variables are nothing but presence or absence of liver disease. The output here must be a categorical value such as 0 or 1.
- KNN – K-Nearest Neighbours – It is a supervised learning technique which basically stores the data and classifies a new data point based on the similarity. In other words it checks for close proximity of similar data points.
- Decision Tree – It is a learning algorithm which involves making a tree-shaped diagram to chart out a statistical probability analysis. It is a good predictive model which produces better accuracy and ease of interpretation.
- Random Forest Tree – Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- SVM Classifier – They are used generally in classification problems; they have the ability to handle multiple continuous and categorical variables. They have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

F. Hyper Parameter Tuning

Hyper parameter tuning or optimization in machine learning is a technique performed by considering a set of optimal hyperparameters specified for a particular classifier which is used to control the learning process in the model.

- Manual Hyper parameter tuning - All the classification models considered were hyper-tuned by best defined parameters for a classifier and these tuned parameters resulted in giving a high accuracy classification model. The parameters considered for hyper-tuning the model is -

TABLE 1 Hyper-tuned Parameters Used

Machine Learning Algorithm	Hyper-tuned parameters
Logistic Regression	Fit intercept, C
KNN	N-Neighbours, Weights
Decision Tree	N-Estimators, Random State, Max Depth
Random Forest	C, Gamma, Random State, Probability
SVM Classifier	C, Gamma, Random State, Probability

G. Performance measure and analysis

Performance measure of different machine learning algorithms is analysed by considering measures such as -

- Confusion Matrix - Confusion Matrix is a table used in performance measure which helps in easy visualization as well as in distinguishing true positives, false negatives, true negatives, false positives.
- Accuracy - Accuracy measure is calculated by considering the ratio of correctly predicted observations to the total number of observations.
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
- Precision - It is the percentage of true positives out of all the predictions.
$$Precision = \frac{TP}{TP + FP}$$
- Recall - Out of the total positive, what percentage are predicted positive.
$$Recall = \frac{TP}{TP + FN}$$
- F1 Score - F1 Score is the harmonic mean of precision and recall.
$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

AUC - The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

IV. RESULTS

With the help of performance measure and analysis the performance of various machine learning algorithms are evaluated. The dataset was obtained from UCI repository on which data pre-processing techniques such as label encoding was performed. Oversampling was performed using SMOTE and with the help of data visualization the model was trained to understand the duplicate values and outliers present. Feature selection showed a linear relationship on certain attributes of the dataset. Hyper parameter tuning was performed to obtain higher accuracy of models. The highest accuracy was obtained by using SVM Classifier and thus the performance was measured based on a classification report and performance measures such as accuracy, precision, recall, f1- score, roc-auc curve and auc. The roc-auc curve for the SVM Classifier is given below –

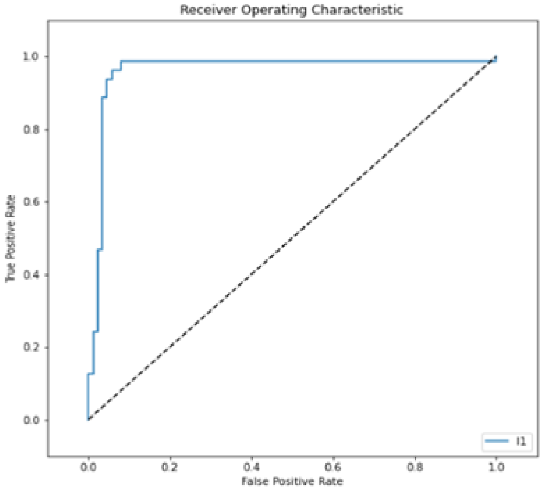

Fig .1 A graph which shows the roc curve based on true positive and false positive rate.

The results based on performance measure of all the classifiers used is given below in the table-

TABLE 2 Results based on classification report

Algorithms	Precision	Recall	f1-Score	Accuracy	roc-auc score
Logistic Regression	71%	71%	70%	71%	77%
KNN	85%	81%	81%	81%	96%
Decision Tree	88%	86%	86%	86%	86%
Random Forest	91%	90%	90%	90%	98%
SVM Classifier	93%	93%	93%	93%	96%

Clearly we can see that SVM Classifier has the highest accuracy compared to the rest.

V. CONCLUSION

This solution gives a comprehensive analysis of “Indian Liver Patient Records' ' dataset with Liver patient and Not Liver patient as classification is performed and this relies upon various machine learning algorithms which provides high accuracy and consumes very less time for entire processing. The process includes data analysis, data pre-processing which includes filling of missing values with mean, label encoding, identifying duplicate value, outlier detection and resampling to improve the performance. Accuracy is effectively utilized to analyze the performance of various classification algorithms. Thus, we can conclude that SVM classifier proved its worthiness in prediction of liver patients by achieving high accuracy amongst the other classifiers.

REFERENCES

[1] 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), 2020.

[2] J. S. H. Adil, M. Ebrahim, K. Raza, S. S. Azhar Ali and M. Ahmed Hashmani, "Liver Patient Classification using Logistic Regression," 2018 4th International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2018.

[3] M. F. Rabbi, S. M. Mahedy Hasan, A. I. Champa, M. AsifZaman and M. K. Hasan, "Prediction of Liver Disorders using Machine Learning Algorithms: A Comparative Study," 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), 2020.

[4] G. Shaheamlung, H. Kaur and M. Kaur, "A Survey on machine learning techniques for the diagnosis of liver disease," 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020.

[5] V. J. Gogi and V. M.N., "Prognosis of Liver Disease: Using Machine Learning Algorithms,” 2018 International Conference on Recent Innovations in Electrical, Electronics &

Communication Engineering (GREECE), 2018, pp. 875-879.

[6] Ma, Han, Cheng-fu Xu, Zhe Shen, Chao-hui Yu, and You-ming Li. "Application of machine learning techniques for clinical predictive modeling: a cross sectional study on nonalcoholic fatty liver disease in China." BioMed research international 2018 (2018).

[7] M.Sivakumar D, Manjunath Varchagall , and Ambika L Gusha S “Chronic Liver Disease Prediction Analysis Based on the Impact of Life Quality Attributes.” (2019). International Journal of Recent Technology and Engineering (IJRTE).

[8] 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).

[9] Mehtaj Banu H’’ Liver Disease Prediction using Machine-Learning Algorithms’ International Journal of Engineering and Advanced Technology (IJEAT)