Liver Disease Prediction

Aryan Rohilla	Avinash Barala	Ekansh	Shruti Jha
2021024	2021028	2021044	2021289
IIIT Delhi	IIIT Delhi	IIIT Delhi	IIIT Delhi
aryan21024@iiitd.ac.in	avinash21028@iiitd.ac.in	ekansh21044@iiitd.ac.in	shruti21289@iiitd.ac.in

Abstract

The liver is the largest internal organ of the human body and the only organ with regenerative abilities. Nowadays, the number of liver diseases is causing a million deaths all around the world. There are more than 100 types of liver diseases. Common types of liver diseases are Hepatitis, Fatty Liver Disease, Autoimmune Conditions, Cirrhosis, Liver Cancer, Liver Failure and many more.

Traditional liver diagnostic methods for detecting various liver diseases are very costly, and prediction is necessary. If the prediction is made early, then the damaged part of the liver can be removed, and even if a person is left with only 10% of a healthy liver, he can easily regenerate it. In this project, our objective is to study the dataset for Indian liver patients and apply machine learning models that can accurately predict whether a patient has liver disease or not and can be given an early good treatment.

1. Introduction

Liver diseases are increasing worldwide, year by year, especially in India, and this poses a huge challenge to public health as the detection methods are costly and compromise the treatments. To tackle this, we want to develop an affordable machine-learning model for Liver Disease Prediction so that it will be helpful in the field of healthcare and medicine.

The traditional diagnostic methods for liver diseases are not only expensive but are also limited to the upper class only and are not affordable and require a high amount of resources, so diagnosis is delayed to late-stage, making the condition critical. But even till at least 10% of the liver remains healthy, it can regenerate.

2. Literature Survey

For this, we have studied multiple articles and research papers and also collected data for our ML model. 2 of the Research Papers are

2.1. Liver Disease Detection by Deepika Bhupathi, Christine Nya-Ling Tan and Sayan Kumar Ray of Manukau Institute of Technology in which ML is used for the early detection of liver diseases [1]

There are five major Machine Learning Techniques used in this research, along with some minor ones. These are:

1. Logistic Regression, which is a simple linear classifier.
2. Decision Trees are used to capture non-linear patterns and interactions between features, and Classification and Regression Trees (CART) were also used.
3. A Random Forest is an ensemble of decision trees offering robustness against overfitting.
4. Support Vector Machines (SVM) are Useful for high-dimensional data and complex decision boundaries.
5. Neural Networks are deep learning models that capture intricate patterns in data.
6. Some of the other methods included were K-Nearest Neighbours (K-NN), Linear Discriminant Analysis (LDA), Unsupervised Algorithm (Autoencoders) and Naive Bayes.

After that, evaluation is done using Accuracy, Precision, Sensitivity, F1 Score and ROC curve. Autoencoders and K-NN had the highest accuracy, with 92.1% and 91.7% respectively.

2.2. Liver Disease Prediction using Machine Learning by Vasan Durai, Suyan Ramesh, and Dinesh Kalthireddy of SRM Institute of Science and Technology [2]

1. Introduction

   The paper highlights the importance of digital technologies in medical technologies. It highlights the significant amount of data that is periodically used in medical procedures, including inferential, prognosis, and treatment of disease. All of the three are reliable on the data.

As the Liver is the largest internal and the second largest organ of the human body, it plays a crucial role in the digestion, metabolism, and other important functions of the body. The traditional diagnosis method for the liver is by testing the level of the enzymes in the blood. The research employs the model with the combination of both SVM and the Naive Bayes algorithms for the detection of disease in the liver.

2. Review of Literature

   The study and the methodologies that are discussed are -

   1. The importance of predicting the life expectancy of patients with cirrhosis or other liver diseases.
   2. Detection of liver disease because of excessive alcohol using data mining
   3. Using conditional probability and Bayes theorem to predict liver cancer.

3. Findings

   From the literature review of the research paper, it is found that there are many mechanisms for predicting liver disease that vary in effectiveness and accuracy. The paper aims to improve the current methodology by experimenting with the datasets and the data elements, such as considering different combinations of the data, etc.

4. Factors affecting accuracy

   There are many factors that affect the accuracy of the algorithms and the model, such as the quantity of the dataset, feature selection, and quality of the data.

5. Experimental Study

   The paper presents an experimental study whose objective is to predict liver disease with high accuracy using different algorithms and finding the most suitable algorithm for it. This system employs the Child-Pugh score to assess the prognosis of chronic liver disease.

After evaluating it on the different algorithms, the J48 algorithm works best while selecting the feature with an accuracy of 95.04%

In conclusion, the research paper uses machine learning techniques to implement models for predicting liver disease and tries to improve the accuracy of the model by solving the challenges that are faced in the present methodologies.

3. Dataset

3.1. Dataset Details

The Liver Disease Patient dataset on Kaggle is used to classify liver disease. It contains records of liver patients from across the world. It is a structured dataset and has a total of 30691 rows and 11 columns. The columns contain features like age, gender of the patient, total bilirubin, direct bilirubin, Alkphos Alkaline Phosphatase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, Total proteins, ALB albumin and A/G Ratio Albumin and Globulin Ratio and the target variable, Result. In the Result column, “1” stands for liver disease and “2” for without liver disease. Median imputation is applied for missing values in the dataset. The numerical values have been normalised using MinMaxScaler, and the outliers have been removed using the IQR method. The gender of the patient is categorical data and has been encoded using BinaryEncoder. Duplicate rows have been dropped. After data preprocessing, the dataset has 19206 rows and 13 columns. The class distribution for ‘Result’ depicts an imbalance in the data. The SMOTE method is used to balance the data.

3.2. Exploratory Data Analysis

The following pie chart [Figure 2] suggests the percentage of male patients having liver disease(50.2%) is nearly equivalent to that of female patients(49.8%).

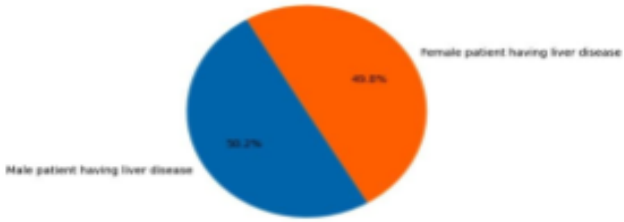


Figure 2: This tells that both male and female have almost same chances for having a liver disease.

From the following frequency distribution [Figure 3], it can be inferred that 13677 patients have liver disease and 5529 do not have liver disease.

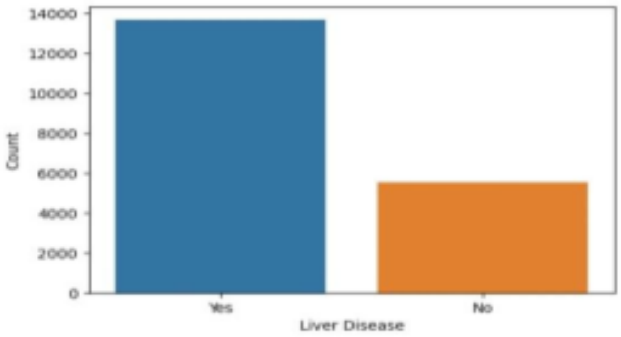


Figure 3: Out of total survey how many people are having a liver disease.

It can be observed from the following histogram [Figure 4] that most patients coming for liver disease diagnosis lie

between the age group of nearly 30 to 60 years.

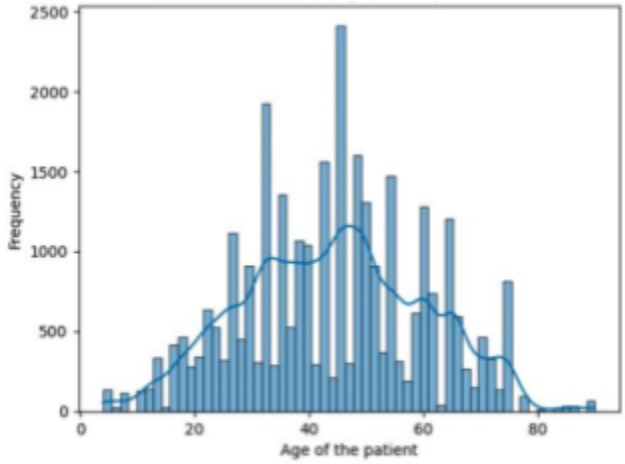


Figure 4: Distribution of Age of the patients

It can be observed from the plot [Figure 5] that the data is positively skewed and prone to outliers. The median of the direct bilirubin is 0.3mg/dl. The value of direct bilirubin should be less than 0.3mg/dl; higher levels of it indicate liver damage.

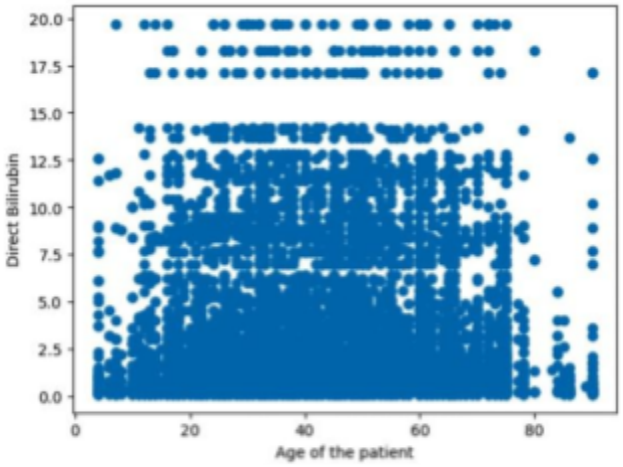


Figure 5: Scatter Plot between Direct Bilirubin & Age of the patient

4. Methodology.

4.1. Feature Selection

1. ANOVA F-test

The numerical feature selection is based on the ANOVA F-test to analyse the relationship between each numerical feature and the target variable, 'Result'. The 'Age of the patient' has been dropped based on the values obtained [Table 1].

2. Chi-Squared Test

The categorical feature selection is based on the Chi-squared test to check the dependence between the 'Gender of the patient' and the target variable, 'Result'. The obtained p-value is 0.47, greater than the significance level(0.05); therefore, it has been dropped.

3. Correlation Heat Map

Based on the values in the graph [Figure 1], ALB Albumin shows a strong correlation with Total Proteins and A/G Ratio Albumin and Globulin Ratio, and Total Bilirubin shows a strong correlation with Direct Bilirubin.ALB Albumin and Total Bilirubin have been dropped to reduce the dependency between the features.

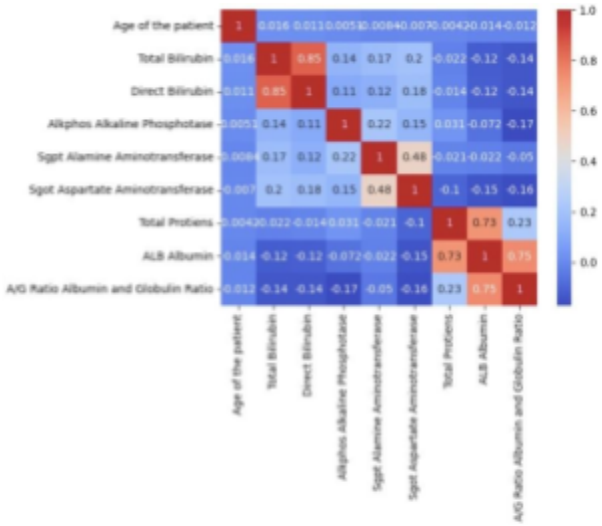


Figure 1: Models are based on dependencies factor given in the above figure

Feature's Name	F-Statistic value	P-value
Age of the patient	0.098	0.753
Total Bilirubin	697.155	6.096e-151
Direct Bilirubin	689.783	2.149e-149
Alkphos Alkaline Phosphotase	621.179	5.763e-135
Sgpt Alamine Aminotransferase	725.170	8.142e-157
Sgot Aspartate Aminotransferase	780.189	2.505e-168
Total proteins	12.036	0.00052
ALB albumin	537.144	3.192e-117
A/G Ratio Albumin and Globulin Ratio	661.806	1.620e-143

Table 1: Features available in dataset and their F-Static value and P-value

4.2 Model Details

We have used 6 models :-

- Support Vector Machines (SVM) were used due to their effectiveness in high-dimensional spaces.
- Naive Bayes: Chosen for its simplicity and fast computational speed.
- Decision Trees: To capture complex relationships in the data.
- Random Forest: An ensemble of decision trees to improve generalisation..
- KNN(K-Nearest Neighbors): For classification and regression tasks when data has local patterns or similarity matters
- Neural Networks: For capturing intricate patterns and interactions between features

4.3 Model Evaluation

Model performance will be tabulated and compared using the predefined evaluation metric of :-

- Accuracy: The proportion of true results among the total number of cases examined.
- F1-Score: It is the harmonic mean of precision and recall.
- ROC-AUC Curve: Quantifies a classifier's overall performance, showing its ability to distinguish between classes.
- Confusion matrix: Shows a classifier's performance, summarising true positives, true negatives, false positives, and false negatives in a tabular format.
- Precision: Measures the proportion of true positive predictions among all positive predictions.

4.4 Hyperparameter tuning

The following table[Table 2], depicts the parameters used to optimise the performance of the models:-

Models	Hyperparameters	Optimal Values
Decision Trees	max_depth: [None, 10,20, 30] min_samples_split: [2,5,50,100] min_samples_leaf: [1,2,4]	None 2 1
Random Forest	n_estimators: [50,100] max_depth: [None, 10, 30] min_samples_split: [2,3,5] min_samples_leaf: [1,2,4]	100 30 2 1
KNN	n_neighbors:[3,5,7,9] Weights:['uniform', 'distance'] metric:['euclidean', 'manhattan']	7 distance manhattan

Table 2: Range of hyperparameters used for each model

5. Result and Analysis

5.1. Analysis

The following two plots[Figure 6] represent the learning curves for random forest and KNN classifiers. They provide visual representation of how the models' accuracy evolve with training sizes ranging from 30% to 100% of the total dataset. It can be observed that both the models perform well.

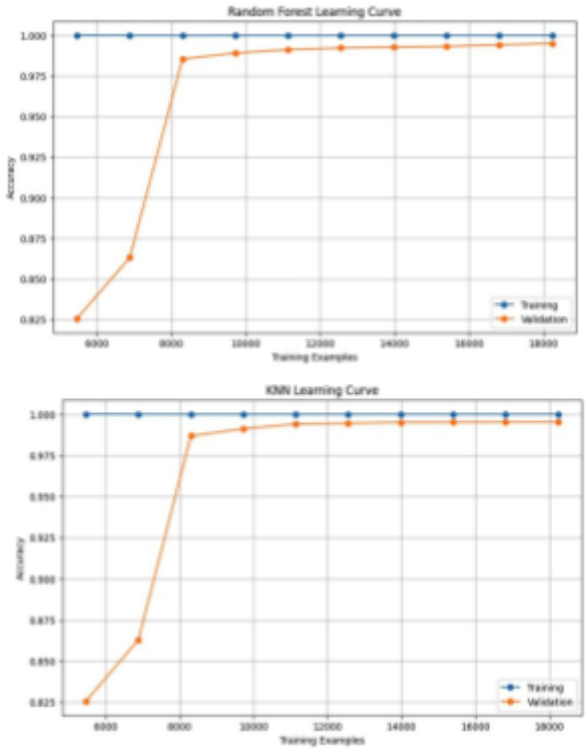
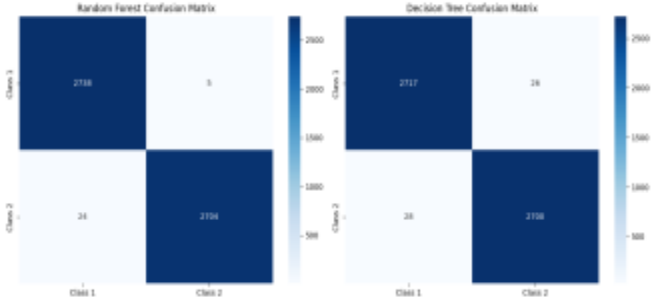


Figure 6: Learning curves for Random forest and KNN (Accuracy vs Training size)

The following matrices[Figure 7], depict the confusion matrix for Random forest, Decision Tree, SVM and KNN respectively. The rows correspond to the actual class while the columns correspond to the predicted class. It can be inferred that random forest, decision tree and KNN have high true positives and true negatives.



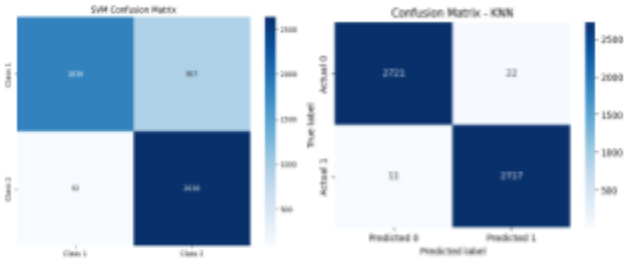


Figure 7: Confusion Matrix of the models

The following plot[Figure 8] depicts the ROC-AUC curve which measures the performance of the models. Random forest has an AUC score of 0.9997 implying a nearly perfect true positive rate and nearly zero false positive rate.

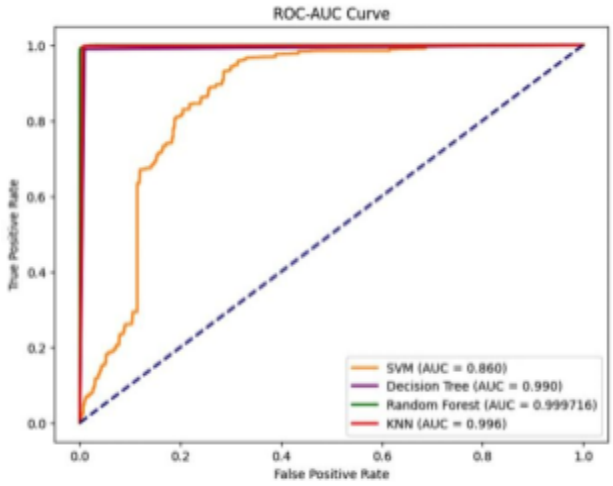


Figure 8: ROC-AUC Curve

5.2 Result

The model yielding the highest overall performance across evaluation metrics among the models implemented is Random Forest, with an accuracy of 99.43%.

Models	Accuracy (in %)	F1-score (in %)	Precision (in %)
Naive Bayes	65.47	58.18	74.06
SVM	81.74	78.6	95.22
Decision Trees	99.01	99.01	98.97
Random Forest	99.43	99.43	99.09
KNN	99.39	99.39	99.59
Neural Network	88.52	88.92	85.71

Table 3 : Scores obtained during model evaluation

Three fold cross-validation was performed to assess the model performance. The performance metric used was accuracy. The mean accuracy obtained across all the folds was equivalent to 99%.

6. Conclusion

The machine learning project for liver disease prediction showcased notable success, with the Random Forest model achieving the highest accuracy of 99.43%. This approach not only offers substantial savings in diagnostic costs but also marks a significant advancement in early detection. The project's findings affirm the potential of machine learning in enhancing healthcare outcomes, especially in settings with limited resources, and set a promising direction for future research in predictive medicine. For future work, further validation on different datasets is required to evaluate the model's performance on unseen data.

7. References

[1](PDF) Liver disease detection using machine learning techniques (researchgate.net)

[2]https://www.academia.edu/39051667/Liver_disease_prediction_using_machine_learning

[3]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9953600/