Research on Prediction of Liver Disease Based on Machine Learning Models

Jiajun Lu

Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen, 518000, China

Abstract. Liver disease is a disease that has attracted much attention in the world. Liver diseases such as cirrhosis and liver cancer are the common causes of death in the world. Many liver diseases have no obvious symptoms in the early stage of onset, so they are easily overlooked by people. Treatment for liver illnesses is crucial and relies heavily on early diagnosis and management. This study assessed the effectiveness of various machine learning approaches for the identification of liver disease due to the high cost and complexity of the diagnostic process. This research used five machine learning models to predict the presence of liver disease based on a patient's medical records using an Indian liver patient record dataset. The dataset was prepared for model training through data preprocessing and analysis, including handling missing values, encoding categorical variables, and normalizing features. Five machine learning algorithms were evaluated, with Random Forest emerging as the highest performing model with an accuracy of 73.56% on the test set. This study contributes to the field by demonstrating the potential of machine learning to accurately predict liver disease, aiding in early diagnosis and treatment.

Keywords: machine learning; dataset; missing values; encoding categorical variables; normalizing features

1. Introduction

Millions of individuals throughout the world suffer from the common and deadly health condition known as liver disease. As a result of heavy air pollution, poor diet, excessive alcohol use, and irregular drug use, there are more and more people with liver disease every year. Liver disease includes a series of diseases that may have a significant impact on personal health and well-being. Diseases such as cirrhosis, liver cancer, and fatty liver have become worldwide diseases and are among the top causes of death in the world. With an annual incidence of 25% and ongoing growth, nonalcoholic fatty liver disorder has surpassed alcoholic liver disease as the most widespread liver disease worldwide [1]. Effective treatment and improved patient prognosis depend greatly on early diagnosis and precise prediction of liver illnesses.

Timely diagnosis and prediction of liver disease can significantly impact treatment progress, outcomes, and associated healthcare costs [2]. Symptoms are usually not noticeable until the early to severe stages of liver disease development and may go unnoticed. If liver disease is diagnosed at an initial stage, through timely intervention and implementation of an appropriate treatment plan, further complications can be prevented, and the disease state can be managed or even reversed. Furthermore, accurate prediction of liver disease is critical for improved patient stratification, personalized care, and risk management. Prognostic models utilizing clinical data, genetic information, and advanced computational methods can help predict disease progression and thus design optimal treatment strategies. These strategies can then improve patient outcomes. Machine learning techniques are showing great promise in healthcare by harnessing the power of data analytics to enhance disease detection and prediction.

Machine learning has become an important tool in healthcare research and clinical practice. By analyzing large amounts of medical data, machine learning models can identify patterns, detect correlations, and generate forecasts based on observed patterns. Machine Learning techniques are used in multiple healthcare domains such as diagnosis, treatment optimization, disease prediction, and patient monitoring [3]. Machine learning algorithms can help clinicians detect abnormalities, such as tumors, fractures or lesions, with precision that can sometimes exceed human expertise. In

the context of liver disease, machine learning algorithms can leverage patient data to build predictive models that facilitate early diagnosis and prognosis.

The objective of the study was to look into a model based on machine learning that might effectively identify patients suffering from liver illness. By utilizing a dataset of Indian liver patient records from Kaggle, this study aims to explore five machine learning algorithms: Logistic Regression, Random Forest, Support Vector Machines, Gradient Boosting, K-Nearest Neighbors, perform extensive data analysis, and evaluate this Predictive performance of five different models. Furthermore, this study will investigate potential factors that may affect the accuracy of predictive models and compare the findings of the study with the existing literature in this field.

2. Data and method

2.1. Data

2.1.1 Data and variables

Detailed medical records of patients with liver disorders are available in the Indian liver patient records dataset, which can be obtained on Kaggle. The data set provides 583 samples, that is, the medical records of 583 patients, and there are 11 indicators in each sample. The name and description of the indicators are shown in Table 1. When the Dataset column's value is 1, it means the patient suffers liver illness; when it's 2, it indicates the patient doesn't. This dataset serves as the foundation for our research project. Table 2 lists the mean, standard deviation, maximum value, and minimum value of the numeric type indicators in the data set.

Table. 1. The name and description of the indicators in the dataset.

	count	**mean**	**std**	**min**	**max**
Age	583.00	44.75	16.19	4.00	90.00
Total_ Bilirubin	583.00	3.30	6.21	0.40	75.00
Direct_ Bilirubin	583.00	1.49	2.81	0.10	19.70
Alkaline_ Phosphotase	583.00	290.58	242.94	63.00	2110.00
Alamine_ Aminotran-sferase	583.00	80.71	182.62	10.00	2000.00
Aspartate_ Aminotran-sferase	583.00	109.91	288.92	10.00	4929.00
Total_ Protiens	583.00	6.48	1.09	2.70	9.60
Albumin	583.00	3.14	0.80	0.90	5.50
Albumin_ and_Glo-bulin_ Ratio	579.00	0.95	0.32	0.30	2.80

During the initial data analysis phase, this study investigated the proportion of liver disease cases and non-liver disease cases in the data set to understand the distribution. The proportion of people in the data set with liver disease was 71.5%, and the proportion of people without liver disease was 28.5%.

Furthermore, this study examined the gender distribution and age demographics of the patients to gain insight into the patient population. Through statistics, it is found that the proportion of male patients is significantly larger than that of female patients, and the number of male patients aged 30-

65 accounts for a large proportion. Furthermore, through data visualization, it can be found that the age of patients with liver disease, regardless of gender, basically conforms to a normal distribution. The distribution of gender and age of liver disease patients and general patients is shown in Figure 1.

Table. 2. Evaluation index of each numerical feature in the dataset

Name	Description
Age	The patient's age
Gender	The patient's gender
Total_Bilirubin	The total amount of two classes of bilirubin, direct and indirect bilirubin
Direct_Bilirubin	The amount of Direct Bilirubin
Alkaline_Phosphotase	The amount of Alkaline Phosphotase
Alamine_Aminotransferase	The amount of Alamine Aminotransferase
Aspartate_Aminotransferase	The amount of Aspartate Aminotransferase
Total_Protiens	The total amount of two classes of proteins, albumin and globulin
Albumin	The amount of Albumin
Albumin_and_Globulin_Ratio	Albumin to globulin proportion
Dataset	Field that indicated whether a patient had liver disease or not

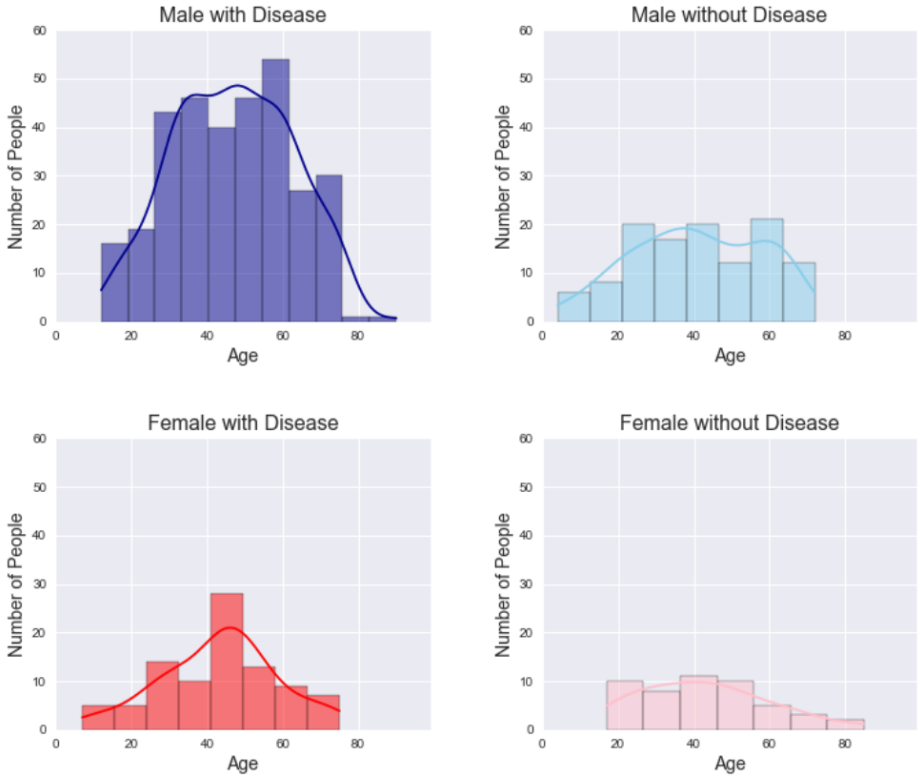

Fig. 1. Distribution of Liver Disease Patients by Age and Gender.

2.1.2 Correlation analysis

This study explored the associations between various characteristics and the presence of liver disease, identifying key variables with the highest positive and negative associations. The study found that the positive correlation coefficient of Albumin and Albumin_and_Globulin_Ratio is the largest (0.16), while the negative correlation coefficient of Direct_Bilirubin is the largest (-0.25), indicating that these indicators have a greater impact on the diagnosis of liver disease in patients, and should be paid attention to in the process of disease diagnosis and treatment. The size of these indicators can

help professionals make better treatment plans. To visualize these relationships, the authors created a correlation coefficient heat map shown in Figure 2.

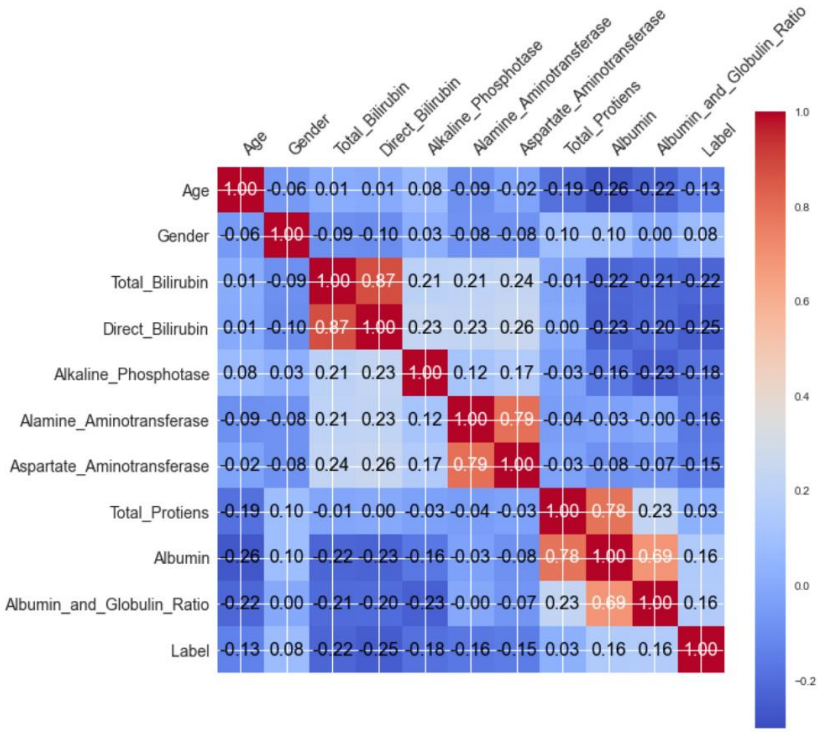

Fig. 2. The correlation coefficient heat map of each feature in the dataset

2.2. Method

By using historical patient data, Machine Learning models can be trained to identify patterns and make predictions. Five popular machine learning algorithms were used to develop predictive models for liver disease detection: Logistic Regression, Support Vector Machines, Random Forests, K-Nearest Neighbors, and Gradient Boosting. Each algorithm has its unique characteristics and is suitable for different types of data sets and classification tasks. The research will explore which model can achieve better prediction performance on the dataset.

2.2.1 Logistic Regression

An efficient approach for binary classification tasks is logistic regression. Using the logistic function, it simulates the relationship between the dependent variable (the diagnosis of liver disease) and the independent factors (the indicaters) [4]. It can provide insights into the importance and significance of individual features in the prediction process.The goal of logistic regression is to find the best parameters, for which the likelihood of the observed output is maximum. This is done through a process called Maximum Likelihood Estimation (MLE).

Linear regression formula:

$$f(x) = b + w_1 x_1 + w_2 x_2 + ... + w_n x_n = w^T x \tag{1}$$

The basic idea of logistic regression is linear regression, and its formula is as follows:

$$\mathrm{h}(x) = \frac{1}{1+e^{-f(x)}} = \frac{1}{1+e^{-w^T x}} \tag{2}$$

Among them, h(x) is an activation function, which is called the sigmoid function, and its image is shown in Figure 3. Its shape is similar to S-type, and it can be known from the image that its function is to map the result between 0-1 [5].

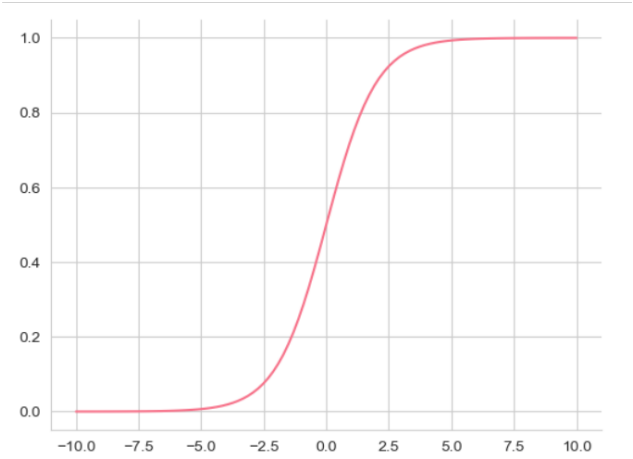

Fig. 3. Sigmoid Function.

2.2.2 Support Vector Machines (SVM)

Support Vector Machines is a potent method that can handle issues involving both linear and nonlinear classification. The best hyperplane that best categorizes the dataset is what the SVM seeks to achieve. A hyperplane is a decision boundary that separates data points (vectors) of one class from data points (vectors) of another class. For an SVM, the best hyperplane is the one that separates the samples with the largest margin, which is the sum of the distances from the samples on either side of the hyperplane. To find the best hyperplane, calculate the distances d_1 and d_2 from the samples on either side of the hyperplane. The hyperplane with the largest value of margin=d_1+d_2 is the best hyperplane. SVM can handle high-dimensional data and has the ability to capture complex decision boundaries [6]. It is worth noting that SVMs are not only used for binary classification tasks but can also be applied to multi-class classification and regression tasks.

2.2.3 Random Forest

Classification and regression issues are both amenable to Random Forest. Through the use of ensemble learning, a number of weak models are joined to create a powerful model. In random forests, the weak models are decision trees. Therefore, for more precise and reliable forecasts, Random Forest constructs many decision trees and blends them. It creates forests by selecting a random subset of characteristics for each split and training each tree on a separate collection of data [7]. Intuitively, a decision tree can only produce one classification result, and random forest solves the disadvantage of weak generalization ability of decision tree. There will be m classification results produced if there are m trees; ultimately, the classification result that receives the most votes will be used. Random forests are known for their robustness, scalability, and ability to handle high-dimensional data.

2.2.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is an effective machine learning technique for applications including regression, classification, and other tasks. However, it is most commonly used for classification tasks. The algorithm operates on the simple principle of similarity, i.e. similar things are close to each other. KNN operates by determining the K closest training samples (denoted K) and predicting a label based on those samples [8]. The approach will determine the separation between each sample and the unknown sample. The K known samples that are most similar to the unknown sample will then be selected. The category with the most votes among the K known samples is chosen as the category of the unknown sample under the majority voting rule. The simplicity and intuitiveness of KNN make it a popular choice for classification tasks.

2.2.5 Gradient Boosting

A potent ensemble machine learning approach called gradient boosting is well renowned for its prowess in addressing classification and regression issues [9]. It is part of a larger class of augmentation methods that focus on minimizing the error of previous models by assigning higher

weights to misclassified instances, iteratively learning and improving them to build strong predictive models. Gradient Boosting can effectively handle compound interactions between special signs and has achieved great success in various fields. The gradient boosting algorithm creates models in stages. To forecast the residuals or errors of earlier models, it creates new models, which are then combined to produce the final forecasts. It is often used as the base model along with decision trees (especially shallow trees).

3. Results and discussion

To train the machine learning model, this study used a training set derived from a dataset of Indian liver patient records. These models are trained on this data, learning patterns and relationships between input features and the presence of liver disease.

To prepare the data for model training, the author standardized the features to ensure fair comparisons and alleviate any potential bias introduced by varying scales. Additionally, the dataset was separated into a set used for training (70%) and a set for testing (30%), which allowed the author to assess how well various machine learning models performed.

3.1. Performance evaluation indicators

The following performance metrics were used for evaluation:
Accuracy: A measurement of how accurately the model's predictions were made overall, accuracy is measured as the proportion of cases that were correctly categorized in relation to all of the instances.

Precision: The percentage of accurately predicted positive incidences (liver disease) among all anticipated positive instances. It evaluates how well the model is able to keep from producing falsely positive outcomes.

Recall (Sensitivity): This statistic shows what percentage of positive examples were accurately predicted out of all the positive examples that actually occurred. It assesses the model's capacity to identify real positives and evade false negatives.

F1-score: The harmonious average of recall and precision, which balances the model's performance by taking into account both recall and accuracy.

3.2. Results of performance evaluation on test set

The following outcomes are shown in Table 3 as a result of the author testing the trained models on the test set.

Table 3. Five different machine learning models' test set evaluation findings

	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.73	0.76	0.91	0.83
Support Vector Machines	0.71	0.71	1.00	0.83
Random Forest	0.74	0.79	0.85	0.82
K-Nearest Neighbors	0.70	0.73	0.93	0.82
Gradient Boosting	0.68	0.75	0.84	0.79

Based on these results, the random forest model achieves the highest accuracy of 0.82 on the test set. Additionally, it exhibits strong performance in terms of F1-score, recall, and precision. This shows that random forest models can predict the existence of liver illness more accurately.

3.3. Results of performance evaluation on training set

The five machine learning models' prediction scores were compared between the training set and the test set by the author. Table 4 displays the results, and Figure 4 presents a graphic representation of the comparison results.

Table 4. Model prediction scores on training set and test set

Model	Train Score	Test Score
Logistic Regression	73.09	72.99
Support Vector Machines	100.00	73.56
Random Forest	80.25	70.11
K-Nearest Neighbors	71.60	71.26
Gradient Boosting	95.56	68.39

Fig. 4. Comparison of Model Scores on training set and test set.

3.4. Discussion

In a study conducted by Ashwani Kumar and Neelam Sahu, they achieved an accuracy of 79.22% using the Random Forest algorithm with an 80-20% data split and six features [10]. In comparison, our Random Forest model in this research project obtained an accuracy of 73.56% on the test set.

The findings from Kumar and Sahu's study indicate a higher accuracy for the Random Forest model compared to our results. This discrepancy in performance could be attributed to several factors, including variations in the dataset, differences in data preprocessing techniques, and the selection of features.

It is important to note that the composition of the datasets used in the two studies might differ, potentially leading to variations in the distribution and characteristics of the data. The effectiveness of the models could also be affected by variations in the data preprocessing stages, such as how missing values are handled, how features are scaled, and how categorical variables are encoded.

Additionally, the choice of features has a significant impact on how well machine learning models predict outcomes. Kumar and Sahu utilized six specific features [10]. The characteristics used can have a big impact on how well the model can identify important correlations and patterns in the data.

4. Conclusion

With the help of a dataset of Indian liver patient records, this study attempts to develop machine learning models for the prognosis of liver disease. The project involved multiple stages, including dataset description and preprocessing, data analysis, machine learning model selection, training and evaluation.

Through the investigation, the study explored the dataset, studied the distribution of liver disease cases, and examined the correlation of various features with the presence of liver disease. This paper also performed data preprocessing steps to handle missing values, rename columns, encode categorical variables, and standardize features. For the purpose of training and assessing the model, this study also divided the dataset into training and testing sets. For the choice of machine learning

model, we considered five popular algorithms: Logistic Regression, Random Forest, Support Vector Machines, Gradient Boosting, and K-Nearest Neighbors. Among these models, Random Forest emerged as the highest performing model, achieving the highest accuracy on the test set. This paper compared the performance of the model's using accuracy, precision, recall, and F1-score metrics. Although our random forest model demonstrated a test set accuracy of 73.56%, this result differs from a related work by Kumar and Sahu [10], which achieved 79.22% accuracy using a similar approach. Observed variations can be attributed to differences in datasets, data preprocessing techniques, feature selection, and other experimental factors. Further analysis is required to fully understand these changes and identify potential areas for improvement.

In conclusion, this study demonstrates how machine learning methods can be used to infer liver illness from patient medical records. The findings show the potential of machine learning models, especially random forests, in accurately classifying liver disease cases. The findings also emphasize careful data preprocessing, feature engineering and model selection to improve prediction accuracy. Future research directions may include exploring alternative feature selection methods, investigating advanced machine learning algorithms, and incorporating other relevant clinical data to enhance the predictive performance of models. By continuously refining and improving liver disease prediction models, it contributes to early detection and intervention, thereby improving healthcare outcomes for patients.

References

[1] T.G.Cotter, M.Rinella, *Nonalcoholic fatty liver disease 2020: the state of the disease*, Gastroenterology, **158**, 1851-1864 (2020).

[2] S.Lee, H.Huang, M.Zelen, *Early detection of disease and scheduling of screening examinations*, Statistical Methods in Medical Research, **13**, 443-456 (2004).

[3] S. Samarpita and R. N. Satpathy, *Applications of Machine Learning in Healthcare: An Overview*, 2022 1st ICIDeA, Bhubaneswar, India, 51-56 (2022).

[4] K. Sellamuthu, S. P, P. K and R. S, *Liver Disease Prediction using Logistic Regression*, 2022 8th ICSSS, Chennai, India, 01-06 (2022).

[5] C.C. Wu, W.C. Yeh, W.D. Hsu. et al. *Prediction of fatty liver disease using machine learning algorithms*, Computer Methods and Programs in Biomedicine, **170**, 23-29 (2019).

[6] W. Noble, *what is a support vector machine.* Nat Biotechnol, **24**, 1565–1567 (2006).

[7] L. Breiman, *Random Forests*, Machine Learning, **45**, 5–32 (2001).

[8] W. Xing and Y. Bei, *Medical Health Big Data Classification Based on KNN Classification Algorithm*, in IEEE Access, **8**, 28808-28819 (2020).

[9] G. Shobana and K. Umamaheswari, *Prediction of Liver Disease using Gradient Boost Machine Learning Techniques with Feature Scaling*, 2021 5th ICCMC, Erode, India, 1223-1229 (2021).

[10] A. Kumar, N. Sahu, *Categorization of Liver Disease Using Classification Techniques*, IJRASET, **5** (2017)