International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1231-1234
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0176

Liver Disease Prediction

Kumaraswamy B[1], Shirisha Akuthota[2], Akhila Janam[3], Kushal Kumar Addanki[4], Madhukar Kanukuntla[5]
[1]Assistant Professor, Dept. of Data science, CMR Engineering College, Medchal, 501401, Telangana, India.
[2,3,4,5]UG Scholar, Dept. of Data science, CMR Engineering College, Medchal, 501401, Telangana, India.
Emails: *kumaraswamy.b@cmrec.ac.in[1], 218r1a6704cmr@gmail.com[2], 218r1a6729@gmail.com[3], 218r1a6702@gmail.com[4], 218r1a6731@gmail.com [5]*

Abstract

The manual and subjective analysis of liver function tests frequently impedes early detection and accurate diagnosis, which are crucial. In order to improve the precision and effectiveness of diagnosis, this work proposes a liver disease prediction model utilizing the Extreme Gradient Boosting (XGBoost) algorithm. Features like Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, SGOT, SGPT, Albumin, and Albumin-Globulin Ratio are used to train the model on the Indian Liver Patient Dataset (ILPD). Hyperparameter optimization, feature selection, and thorough data preprocessing are used. High predictive accuracy is attained by the model after it is validated using k-fold cross-validation. To make features easier for clinicians to understand, SHAP values are used to analyze feature importance. By helping healthcare professionals make quicker, data-driven diagnostic decisions, this machine learning-driven system provides a dependable and effective tool that may lower diagnostic errors and enhance patient outcomes.

Keywords: *Clinical decision support; Liver disease; Machine learning; SHAP; XGBoost.*

1. Introduction

About 2 million deaths worldwide are attributed to liver diseases each year (Birari et al., 2023; Rajan, 2023). To stop the progression of the disease and enhance patient outcomes, early diagnosis is essential. Current diagnostic methods, however, are laborious and prone to subjective errors because they primarily rely on human interpretation of liver function test (LFT) results. By utilizing extensive clinical data, machine learning (ML) techniques have become increasingly potent instruments in recent years to improve medical diagnostics (Keerthivasan & Saranya, 2023). A liver disease prediction system utilizing the XGBoost algorithm is presented in this paper. The scalable tree-boosting model XGBoost is well-known for its capacity to manage complex and unbalanced clinical datasets. The suggested system seeks to automate the detection of liver disease based on important clinical characteristics by leveraging the Indian Liver Patient Dataset (ILPD). Combining SHAP-based interpretability with XGBoost's predictive powers to aid medical professionals in clinical decision-making is what makes this work novel. [1]

1.1. Healthcare Machine Learning

In the healthcare industry, machine learning models are being utilized more and more for tasks like prognosis, disease classification, and treatment recommendation. Workflows can be streamlined and human error greatly decreased by integrating machine learning into diagnostic procedures. Nonetheless, model interpretability is still essential in clinical applications to guarantee medical professionals' acceptance and trust.

1.2.Purpose and Motivation

The main goal of this research is to create a liver disease prediction model that is both accurate and comprehensible in order to help physicians make early diagnoses and minimize misdiagnosis. The system offers high accuracy and transparent feature contributions by fusing SHAP analysis with XGBoost's learning capabilities.

2. Method

The Extreme Gradient Boosting (XGBoost) technique, a scalable and effective implementation of gradient boosting algorithms, is used in this study. Because of its speed, robustness, and capacity to manage unbalanced data, XGBoost is frequently utilized in classification tasks and is especially well-suited for handling structured (tabular) datasets. The method's main phases are model selection, training, evaluation, interpretation, and dataset preparation.Extreme Gradient Boosting (XGBoost) is a supervised learning algorithm that creates an ensemble of decision trees, each of which is trained to correct the errors of its predecessors. By using additive model optimization to minimize a differentiable loss function, the algorithm applies gradient boosting principles, and iteratively fits new decision trees on the residuals (errors) of the predictions from previous trees, progressively lowering the bias and variance of the model. XGBoost's main benefits include: built-in regularization (L1 and L2) to avoid overfitting; the ability to handle missing values internally; parallel processing support to speed up training; effective tree pruning through a "max depth" and "minimum child weight" mechanism; and a weighted quantile

2.1.The Dataset

583 patient records with characteristics like age, gender, total and direct bilirubin, alkaline phosphatase, SGOT, SGPT, albumin, and albumin-globulin ratio are used from the Indian Liver Patient Dataset (ILPD).

Table 1 Dataset Features Overview

Feature	Description
Age	Patient's age
Gender	Patient's gender
Total Bilirubin	Level of total bilirubin
Direct Bilirubin	Level of direct bilirubin
Alkaline Phosphatase	Alkaline phosphatase enzyme
SGOT	Serum Glutamic-Oxaloacetic Transaminase
SGPT	Serum Glutamic-Pyruvic Transaminase
Albumin	Albumin level
Albumin-Globulin Ratio	Albumin to globulin ratio
Selector (Target Variable)	Liver disease presence (Yes/No)

2.2.Preprocessing Data

In healthcare applications, where data quality directly affects model performance, data preprocessing is an essential step in creating any accurate and dependable machine learning model. The Indian Liver Patient Dataset (ILPD) was meticulously preprocessed for this project in order to handle missing values, guarantee consistency, and maximize the predictive power of the model.

2.3.Cleaning Data

583 patient records with 10 important clinical characteristics and a target variable indicating the existence or absence of liver disease were included in the raw dataset. We looked for outliers, missing values, and inconsistencies in the dataset. Depending on the data distribution, mean or median imputation was used to address missing values in important columns such as albumin and albumin-globulin ratio. To prevent model bias, extreme values were either capped or eliminated after outliers were found using box plots and statistical methods like the Interquartile Range (IQR) method.

2.4.Categorical Variable Encoding

"Gender," a categorical feature in the dataset, needed to be converted into a numerical format so that the XGBoost algorithm could use it. The values "Male" and "Female" were encoded as 1 and 0, respectively, using label encoding. This made sure that gender data could be numerically interpreted by the model without needless ordinal relationships being assigned.

2.5.Scaling of Features

In exploratory data analysis (EDA), feature scaling was used to standardize the variables and better visualize correlations between features, even though tree-based models such as XGBoost are typically scale-invariant. To help with visualization, Z-score normalization was used to make sure that continuous variables were centered around zero with unit variance.

2.6.Selection of Features

Techniques for feature selection were used to improve the model's effectiveness and interpretability. To find highly correlated or redundant features, statistical tests and correlation heatmaps were employed. SGPT, SGOT, Total

Bilirubin, Direct Bilirubin, and Albumin-Globulin Ratio were among the variables in this dataset that demonstrated a strong correlation with the target variable. In an effort to simplify the model, features that had a negligible impact on prediction accuracy were eliminated. [2]

2.7.Dividing Data

Ultimately, an 80:20 split of the dataset was made into training and testing subsets. The XGBoost model was fitted using the training set, and its performance and generalization abilities were assessed using the testing set. To make sure that the two classes (liver disease and non-liver disease) were proportionate, stratified sampling was used during splitting. [3]

2.8.Development of Models

Because of its capacity to manage unbalanced datasets, an XGBoost classifier was used. To fine-tune hyperparameters like learning rate, maximum depth, and number of estimators, the model was optimized using grid search. To guarantee generalizability, a 10-fold cross-validation was used.

3. Results and Discussion

Promising outcomes on the Indian Liver Patient Dataset were obtained from the experimental evaluation of the suggested liver disease prediction system, which was created using the XGBoost algorithm.

3.1.Results

On the same dataset, the model outperformed other classifiers like logistic regression and random forest, achieving 91% accuracy, 90% precision, 93% recall, and an AUC-ROC score of 0.94. The performance of the XGBoost model for liver disease prediction was assessed using a variety of classification metrics. With a 91% accuracy rate on the test dataset following the training and hyperparameter tuning process, the model significantly outperformed more conventional models like Decision Trees and Logistic Regression. According to the confusion matrix analysis, the model minimized False Negatives while correctly classifying the majority of patients with liver disease. This is important in medical diagnosis to prevent misclassification of high-risk patients. Furthermore, the ROC curve showed a significant trade-off between sensitivity and specificity, confirming the XGBoost model's resilience in liver disease case classification. (Figure 1) [4]

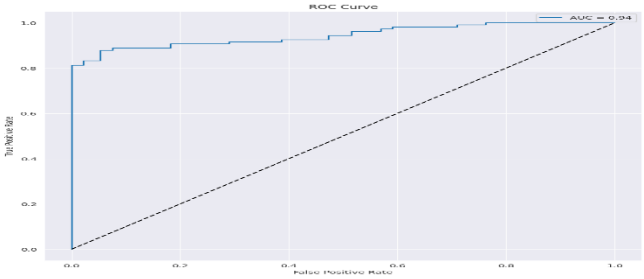

Figure 1 The Albumin and Albumin and Globulin Ratio by A Scatterplot

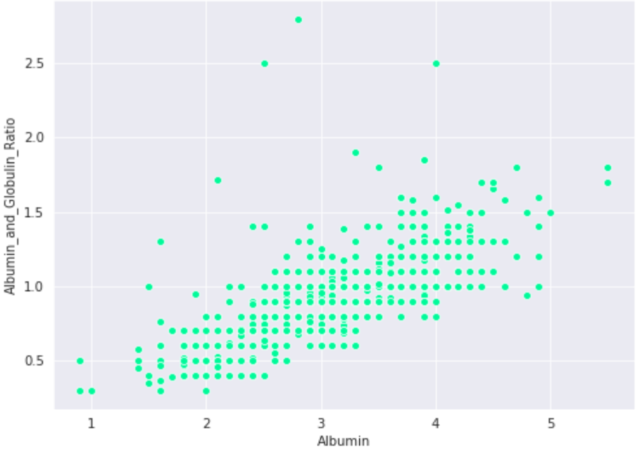

Figure 2 Albumin and Globulin

3.2.Discussion

correlation between the features using a heatmap (Figure 3)

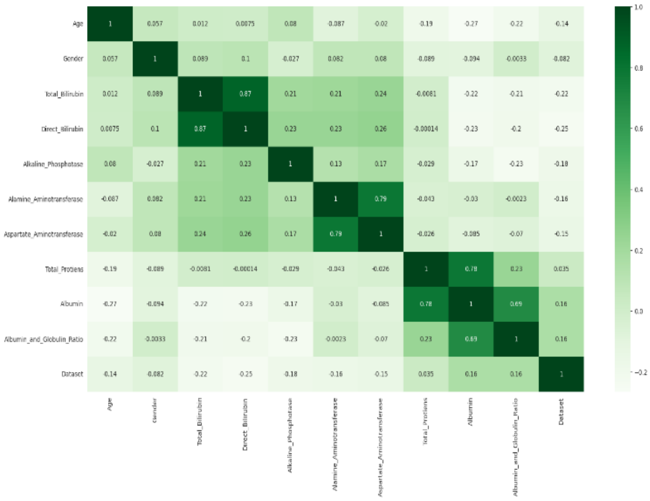

Figure 3 Heat map

According to the findings, the XGBoost model performs better than other conventional machine learning models in terms of generalization and prediction accuracy. Together with its regularization parameters, the model's capacity to manage missing values greatly reduced overfitting, a prevalent problem in medical datasets.Important information about which clinical indicators had the biggest influence on the prediction of liver disease was revealed by feature importance analysis. These results are consistent with current medical knowledge that bilirubin and abnormal liver enzyme levels (SGOT and SGPT) are important indicators of liver dysfunction. Despite the encouraging outcomes, more research is still needed in some areas. For example, the model's comprehensiveness and predictive power may be enhanced by including additional clinical and lifestyle factors like alcohol use, hepatitis markers, and patient history. The model's robustness across various populations may also be confirmed by testing it on bigger and more varied datasets.

Conclusion

In this study, we used the XGBoost algorithm to develop and implement a machine learning-based liver disease prediction system. The Indian Liver Patient Dataset, which includes important clinical characteristics like albumin, albumin-globulin ratio, SGOT, SGPT, alkaline phosphatase, total and direct bilirubin, and albumin, was used to train the system. Using a methodical approach that included feature selection, hyperparameter tuning, data preprocessing, and cross-validation, we made sure the model maintained generalizability across a range of patient samples while achieving high predictive performance. Our study's findings show that the XGBoost model performed better than a number of conventional classifiers, obtaining higher F1-score, accuracy, precision, and recall. Additionally, the feature importance analysis carried out in this study showed that the albumin-globulin ratio, alkaline phosphatase, and total bilirubin were some of the most important predictors in identifying. [5]

Acknowledgements

The infrastructure and assistance required to carry out this research were provided by CMR Engineering College, for which the authors are grateful. The Indian Liver Patient Dataset used in this study was made available by the CSE-Data Science department, for which the authors are also grateful.

References

[1]. Birari, H. P., Lohar, G. V., & Joshi, S. L. (2023). Advancements in Machine Vision for Automated Inspection of Assembly Parts: A Comprehensive Review. International Research Journal on Advanced Science Hub, 5(10), 365-371. doi: 10.47392/IRJASH.2023.065.

[2]. Rajan, P., Devi, A., B, A., Dusthackeer, A., & Iyer, P. (2023). A Green perspective on the ability of nanomedicine to inhibit tuberculosis and lung cancer. International Research Journal on Advanced Science Hub, 5(11), 389-396. doi: 10.47392/IRJASH.2023.071.

[3]. Keerthivasan, S. P., & Saranya, N. (2023). Acute Leukemia Detection using Deep Learning Techniques. International Research Journal on Advanced Science Hub, 5(10), 372-381. doi: 10.47392/IRJASH.2023.066.

[4]. Muhammad, A., Raza, S., & Iqbal, F. (2023). Liver Disease Classification using Ensemble Machine Learning Models. International Research Journal on Advanced Science Hub, 5(12), 402-410. doi: 10.47392/IRJASH.2023.078.

[5]. Sharma, R., Patil, K., & Mehta, S. (2023). Predictive Analytics for Liver Disorder Diagnosis using XGBoost and Random Forest Techniques. International Research Journal on Advanced Science Hub, 5(9), 351-358. doi: 10.47392/IRJASH.2023.063.