Predictive Analysis of Liver Cirrhosis using Random Forest Algorithm

K.Santha Sheela
Computer Science and Enginerering
Velammal College of Engineering and Technology
Madurai, India
sheelakodi@gmail.com

Harini S
Computer Science and Enginerering
Velammal College of Engineering and Technology
Madurai, India
21cse008harini.s@gmail.com

Keerthana K S
Computer Science and Enginerering
Velammal College of Engineering and Technology
Madurai, India
21cse015keerthana@gmail.com

Krithiga Ganesh Kumar
Computer Science and Enginerering
Velammal College of Engineering and Technology
Madurai, India
21cse017krithiga@gmail.com

Sowmiya Eswari P
Computer Science and Enginerering
Velammal College of Engineering and Technology
Madurai, India
21cse029sowmiyaeswari@gmail.com

***Abstract*- Liver cirrhosis, characterized by liver tissue scarring, presents diagnostic challenges. This study proposes a Random Forest-based approach for predicting liver cirrhosis risk. Leveraging machine learning, we analyze a dataset encompassing demographic, clinical, and laboratory parameters, including age, gender, liver enzyme levels, and comorbidities. After rigorous preprocessing to address missing values, the model undergoes training on a subset of the data with hyper parameter optimization through cross-validation. Evaluation metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) assess predictive performance. Feature importance analysis identifies key predictors. Results indicate the model's potential as a screening tool for identifying high-risk individuals. Integration into clinical practice could aid early identification and proactive management of liver disease, improving patient outcomes and reducing healthcare burdens.**

Keywords: Liver cirrhosis, Random Forest, Machine learning, Clinical parameters, Preprocessing, Screening tool

I. INTRODUCTION

Liver cirrhosis is a debilitating condition characterized by the progressive deterioration of liver function due to fibrosis and scarring, representing a significant global health challenge. Timely detection of liver cirrhosis is paramount for implementing effective interventions to mitigate disease progression and improve patient outcomes. Bilirubin, a byproduct of heme metabolism, is a widely used biomarker in liver function tests and is often elevated in individuals with liver diseases.

Machine learning techniques, particularly Random Forest algorithms, have demonstrated efficacy in predictive modeling across various medical domains, including liver diseases. Random Forest, an ensemble learning method capable of handling complex datasets and capturing non-linear relationships between features, presents a promising approach for analyzing multifactorial diseases such as liver cirrhosis.

This study aims to investigate the predictive value of bilirubin levels in identifying individuals at risk of developing liver cirrhosis using Random Forest. By leveraging a comprehensive dataset comprising demographic, clinical, and laboratory parameters, including bilirubin levels, we seek to develop a robust predictive model capable of accurately classifying patients based on their liver cirrhosis risk.

II. DATA PREPARATION

A. Data collection

In this experiment, we collected a dataset from the Kaggle Machine Learning Repository.

1) ID: unique identifier
2) N_Days: number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
3) Status: status of the patient C (censored), CL (censored due to liver tx), or D (death)
4) Drug: type of drug D-penicillamine or placebo
5) Age: age in [days]
6) Sex: M (male) or F (female)
7) Ascites: presence of ascites N (No) or Y (Yes)
8) Hepatomegaly: presence of hepatomegaly N (No) or Y (Yes)
9) Spiders: presence of spiders N (No) or Y (Yes)
10) Edema: presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)
11) Bilirubin: serum bilirubin in [mg/dl]
12) Cholesterol: serum cholesterol in [mg/dl]
13) Albumin: albumin in [gm/dl]
14) Copper: urine copper in [ug/day]
15) Alk_Phos: alkaline phosphatase in [U/liter]
16) SGOT: SGOT in [U/ml]
17) Triglycerides: triglicerides in [mg/dl]
18) Platelets: platelets per cubic [ml/1000]
19) Prothrombin: prothrombin time in seconds [s]
20) Stage: histologic stage of disease (1, 2, 3, or 4)

B. *Data preprocessing*

In our research, we utilized the R programming language to preprocess a dataset concerning liver cirrhosis. Initially, we imported the dataset, "cirrhosis.csv," employing the read.table function from the dplyr package. Subsequently, we replaced any "NA" values within the dataset with R's native NA representation to maintain data integrity. Further, to ensure robust analyses, we systematically removed rows containing missing values using the na.omit function from the dplyr package. Finally, we generated a comprehensive summary of the cleaned dataset using the summary function, providing essential insights into the distribution and characteristics of each variable. These preprocessing steps set the stage for subsequent analyses, facilitating the exploration of liver cirrhosis-related factors with accuracy and reliability.

C. *Data Visualization*

We utilized the ggplot2 library to generate boxplots showcasing the distributions of age, bilirubin levels, and albumin levels among patients grouped by their status .Each boxplot provides a concise visualization of the variability in the respective variable across different patient statuses, aiding in the exploration of potential associations with liver cirrhosis outcomes.

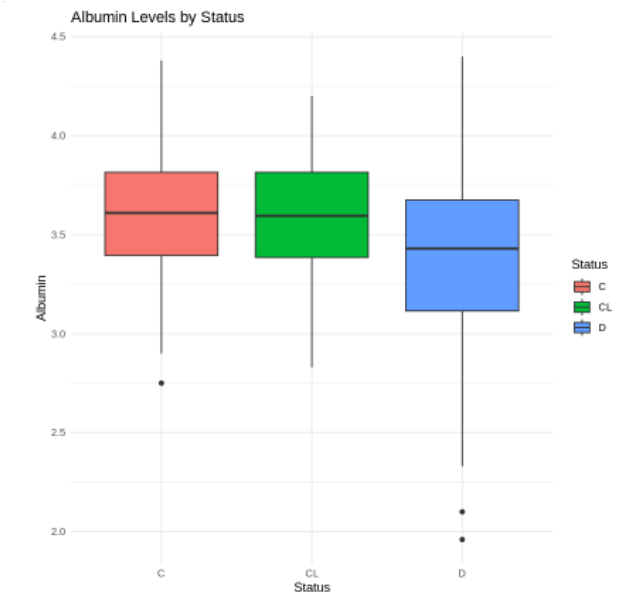

Figure 2.1
Box plot analysis of Albumin

D. *Feature Selection*

The correlation matrix analysis highlights "Albumin" and "Bilirubin" as the most correlated variables among those examined. This finding suggests a potentially strong relationship between these biomarkers in the context of liver cirrhosis. Further exploration of this correlation could provide valuable insights into the disease's progression and prognosis.

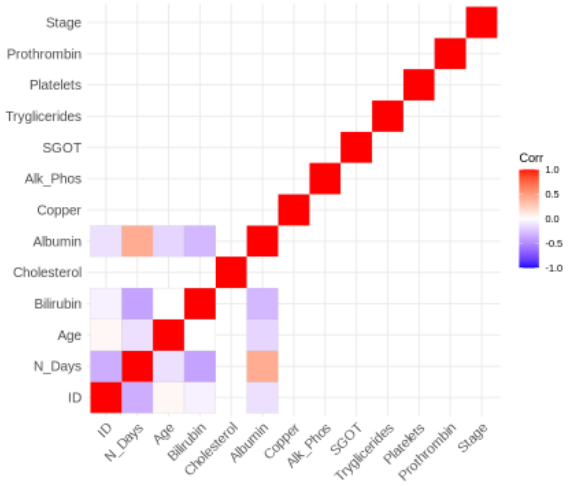

Figure 2.2
Heat map visualization

III. METHODOLGY

1) *Trend of cirrhosis stage over time*

We performed a trend analysis on cirrhosis stage over time. After loading necessary packages and the dataset, a linear regression model is fitted using "N_Days" as the predictor. The model summary reveals coefficient estimates and goodness-of-fit measures. Assumptions are checked via diagnostic plots, and a scatter plot with a regression line visually represents the trend. Finally, the significance of the trend is tested using the coeftest function, providing formal statistical inference.

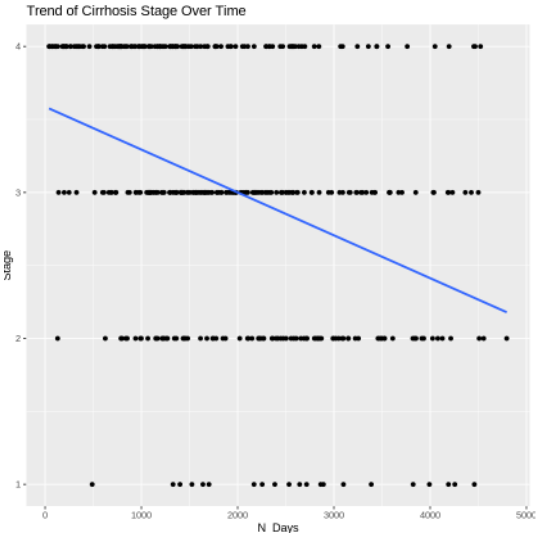

Figure 3.1
Trend of cirrhosis stage over time

Assumptions and diagnostics of the model are checked using diagnostic plots, including residuals vs. fitted values, QQ plot of residuals, scale-location plot, and residuals vs. leverage plot. These plots help assess the model's validity and identify any violations of regression assumptions

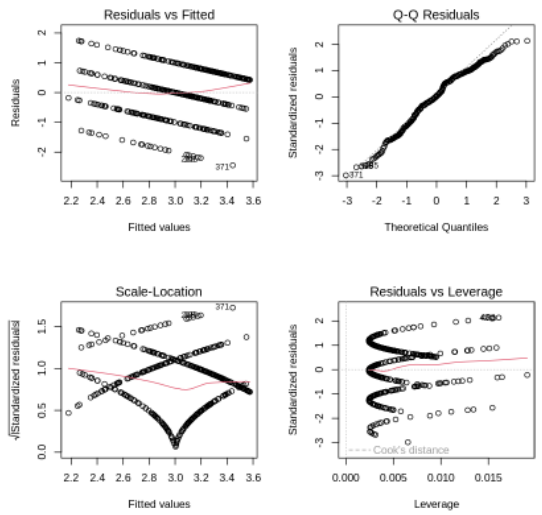

Figure 3.2
Diagnostics of model

2) Random forest Algorithm

We used the construction and assessment of a random forest model for predicting the status of liver cirrhosis patients. The data is first preprocessed, categorical variables are converted to factors, and missing values are handled by removing corresponding rows. The dataset is then split into training and testing sets, and a random forest model is trained using the "train" function from the caret package.

The model is evaluated on the test set, generating a confusion matrix to assess predictive performance. Additionally, the accuracy of the model is calculated as the proportion of correctly predicted instances. This process enables robust modeling and evaluation of liver cirrhosis status prediction using a random forest approach.

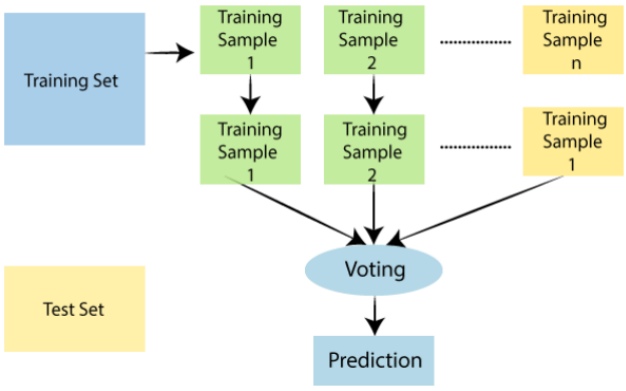

Figure 3.3
Random forest algorithm

3) Predictive analysis:

Confusion Matrix Data Frame Conversion: The code converts the confusion matrix, which summarizes the performance of a classification model by tabulating true positive, false positive, true negative, and false negative predictions, into a data frame format. This conversion facilitates further analysis and visualization of the model's performance metrics.

Confusion Matrix Heatmap: After converting the confusion matrix to a data frame, the code creates a heatmap visualization of the confusion matrix using ggplot2. This heatmap provides a graphical representation of the confusion matrix, with the actual and predicted classes plotted on the x and y axes, respectively. The intensity of color in each cell corresponds to the frequency of predictions, allowing for easy identification of areas of misclassification or confusion.

Accuracy Visualization: Additionally, the code generates a bar plot to visualize the accuracy of the predictive model. Accuracy, a common evaluation metric for classification models, measures the proportion of correctly predicted instances out of the total instances. The bar plot visually represents the accuracy metric, providing a clear indication of the model's overall performance.

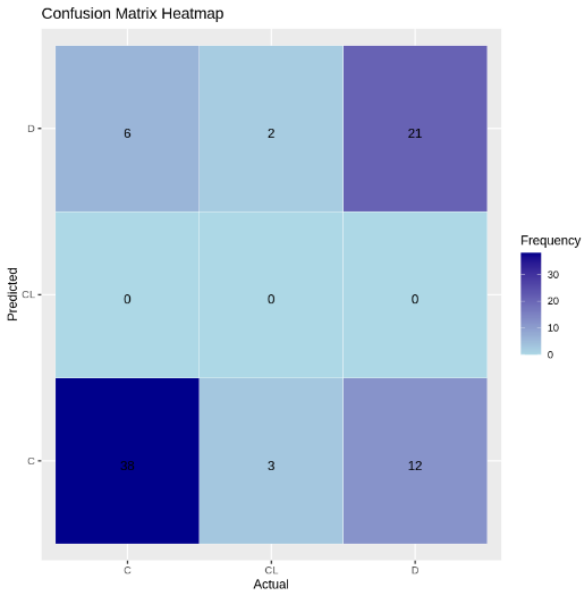

Figure 3.4
Accuracy calculation

IV. CONCLUSION

The visualizations and analysis provided showcase the performance of a model, likely in a classification task. The confusion matrix heatmap offers a comprehensive view of how the model's predictions align with the actual classes. The color gradient aids in identifying patterns, with lighter shades indicating lower frequencies and darker shades representing higher frequencies of correct or incorrect predictions. Additionally, the accuracy plot succinctly summarizes the overall model performance, offering a single metric that stakeholders can easily interpret. The green bar graph, accompanied by percentage labels, provides a clear indication of the model's accuracy, enhancing the understanding of its effectiveness. Overall, these visualizations facilitate both quantitative and

qualitative assessments of the model's performance, aiding in decision-making and further model refinement.

REFERENCES

[1] Zhang, E.-L., Zhang, Z.-Y., Wang, S.-P., Xiao, Z.-Y., Gu, J., Xiong, M., Chen, X.-P., & Huang, Z.-Y. (2016). Predicting the severity of liver cirrhosis through clinical parameters. Journal of Surgical Research, 204(2).

[2] Singh, V., Gourisaria, M. K., & Das, H. (2021). Performance analysis of machine learning algorithms for prediction of liver disease. Publisher: IEEE.

[3] Kumari, S., Singh, M., & Kumar, K. (2021). Prediction of liver disease using grouping of machine learning classifiers. In Conference Proceedings of ICDLAIR2019 (ICDLAIR 2019).

[4] Mostafa, F., Hasan, E., Williamson, M., & Khan, H. (2021). Statistical Machine Learning Approaches to Liver Disease Prediction. Livers, 1.

[5] Dritsas, E., & Trigka, M. (2023). Supervised machine learning models for liver disease risk prediction. Computers, 12(1), 19.

[6] Rawat, P., Bajaj, M., Prerna, P., Vats, S., & Sharma, V. (2023). A study on liver disease using different machine learning algorithms. Publisher: IEEE.

[7] Khan, M. A. R., Afrin, F., Prity, F. S., Ahammad, I., Fatema, S., Prosad, R., Hasan, M. K., Uddin, M., & Salehin, Z.-U.-S. (2023). An effective approach for early liver disease prediction and sensitivity analysis. Iran Journal of Computer Science, 6, 277–295.

[8] Pompili, E., Baldassarre, M., Bedogni, G., Zaccherini, G., Iannone, G., De Venuto, C., Pratelli, D., Palmese, F., Domenicali, M., & Caraceni, P. (2024). Predictors of clinical trajectories of patients with acutely decompensated cirrhosis: An external validation of the PREDICT study. Liver International, 44(1), Jan 2024.

[9] Al Ahad, A., Das, B., Khan, M. R., Saha, N., Zahid, A., & Ahmad, M. (2024). Multiclass liver disease prediction with adaptive data preprocessing and ensemble modeling. Results in Engineering, 22, June 2024.

[10] Tian, J., Cui, R., Song, H., Zhao, Y., & Zhou, T. (2024). Prediction of acute kidney injury in patients with liver cirrhosis using machine learning models: Evidence from the MIMIC-III and MIMIC-IV. International Urology and Nephrology, 2024.