Predicting Survival in Cirrhosis Patients Using Data Visualization and Machine Learning in R

ABSTRACT

This research explores predictive modeling of cirrhosis progression and outcome through a thorough analysis approach. Utilizing the R programming language and advanced data visualization techniques, the study investigates a dataset comprising diverse clinical attributes of cirrhosis patients. Initial data preprocessing steps, encompassing missing value imputation and feature engineering, ensure dataset readiness. Through meticulous exploration utilizing visualization techniques such as scatter plots, box plots, heatmaps, and interactive visualizations, relationships between various clinical variables and cirrhosis progression are elucidated. Subsequently, predictive modeling algorithms, including logistic regression, decision trees, and random forests, are employed to forecast cirrhosis outcomes. Rigorous evaluation of model performance provides valuable insights into disease prognosis and management. This research presents a comprehensive analysis framework, contributing to improved clinical decision-making and personalized patient care strategies in hepatology.

CCS CONCEPTS

• Insert CCS text here • Insert CCS text here • Insert CCS text here

KEYWORDS

Cirrhosis, Predictive modeling, Data visualization, R programming, Clinical outcomes

INTRODUCTION

Chronic liver disease cirrhosis, which causes progressive scarring, is a major global health concern that contributes significantly to morbidity and mortality rates. To address this urgent issue, sophisticated data visualization techniques and predictive modeling approaches become essential resources for deciphering the intricacies of cirrhosis progression and prognostication. Using a combination of data preparation, exploration, and modeling methods supported by the powerful capabilities of the R programming language, this study undertakes a thorough investigation of predictive modeling in cirrhosis. The investigation takes place in a dataset that is rich in a variety of clinical characteristics related to individuals with cirrhosis. Our journey begins with careful data pretreatment efforts, which include finding and managing missing values and then carefully calibrating attribute distributions. In this study, we delve into the predictive modeling of cirrhosis progression and outcome, leveraging the R programming language and sophisticated data visualization techniques. By employing a diverse array of statistical analyses and visualization approaches, we aim to uncover intricate relationships between clinical attributes and cirrhosis status. Through meticulous data preprocessing and exploration, followed by the development and evaluation of predictive models, we seek to provide insights that can inform clinical decision-making and enhance patient care in hepatology. This research not only contributes to the advancement of predictive analytics in cirrhosis but also underscores the efficacy of R-based methodologies in unraveling complex clinical phenomena.

LITERATURE REVIEW

This research presents a comprehensive analysis contributing to improved clinical decision-making and personalized patient care strategies. [1]. The Model for End-Stage Liver Disease (MELD) serves as a reliable tool for assessing short-term mortality risk in patients with chronic liver disease. Based on serum creatinine, bilirubin, and INR, it offers advantages over the Child-Turcotte-Pugh classification, including greater discriminatory ability and objectivity. Despite variations in laboratory measurements, the MELD scale's continuous nature and statistical derivation provide a more robust predictor of survival. Complications of portal hypertension, such as ascites and encephalopathy, do not significantly enhance its predictive accuracy. The MELD scale's validity extends across diverse etiologies and disease severities, supporting its use in liver transplant allocation decisions. [2]. Yang et al. (2022) in BMC Gastroenterology highlights the evolving landscape of liver transplantation (LT) and the challenges in predicting short-term survival for acute-on-chronic liver failure (ACLF) patients post-LT. While the Model for End-Stage Liver Disease (MELD) score has traditionally been used for organ allocation, its limitations in predicting ACLF patient outcomes have led to the exploration of alternative models like the Chronic Liver Failure Consortium (CLIF-C) ACLFs. Additionally, machine learning (ML) algorithms, particularly the Random Forest (RF) model, show promise in improving prognostic accuracy. However, further multicenter studies are needed to validate these findings and optimize organ allocation strategies. [3]. The growing interest in applying machine learning (ML) techniques to medical data, particularly in predicting outcomes in liver transplantation (LT). Traditional statistical models like the Cox model are compared with ML methods such as Random Survival Forests (RSF) and Artificial Neural Networks (NNs). While RSF outperformed Cox models in terms of discrimination, NNs showed better predictive performance. Challenges include interpretation and stability of ML models, but they offer promise for improving decision-making in healthcare. Further research is needed to explore the impact of variable selection and dynamic methods on predictive accuracy. [4].The

literature review highlights the significant mortality rates among cirrhotic patients admitted to the ICU, ranging from 46% to 64%. Severity of illness was assessed using four mortality prediction systems, which indicated a high severity of illness in the study population. Coma and acute renal failure emerged as independent predictors of mortality. The study also questions the utility of invasive procedures such as pulmonary artery catheterization in cirrhotic patients. Furthermore, all patients admitted post-cardiac arrest died, prompting discussion on the appropriateness of CPR in this population. These findings underscore the need for rationalization of critical care delivery and informed decision-making regarding aggressive interventions. [5].The literature on predicting mortality among patients with liver cirrhosis underscores the critical need for accurate prognostic tools due to the condition's high morbidity and mortality rates. While the Model for End Stage Liver Disease (MELD) score has been a standard tool for risk assessment, its limitations in predicting outcomes across various patient populations and timeframes have prompted exploration into alternative approaches. Recent studies have highlighted the potential of machine learning techniques, particularly deep learning algorithms, to leverage electronic health record data for more precise mortality predictions. These advancements aim to improve clinical decision-making and ultimately enhance patient outcomes in cirrhotic populations. [6]. Novel nomogram for predicting in-hospital mortality in patients with liver cirrhosis and sepsis, utilizing LASSO regression analysis on data from the MIMIC database. It addresses the significant mortality risk associated with sepsis in cirrhosis patients, highlighting the scarcity of research in this area. The nomogram, incorporating nine independent variables including age, heart rate, bilirubin levels, glucose, sodium, anion gap, fungal infections, mechanical ventilation, and vasopressin use, demonstrates superior predictive performance compared to established scoring systems like SOFA, MELD, and ABIC. The study underscores the need for prospective validation of the nomogram in clinical settings. 7. The study by Campbell et al. (2015) evaluated prognostic scoring tools for predicting ICU mortality in patients with cirrhotic liver disease. They found that the RFH and CTP + L scoring tools showed similar performance in predicting ICU mortality, with the latter being more practical due to its simplicity. Multivariable analysis identified lactate, bilirubin, and PaO2/FiO2 ratio as significant predictors of mortality. While hepatic encephalopathy scores did not enhance predictive value, further validation is needed. The study highlights the importance of practical, validated scoring tools for assessing ICU mortality in cirrhotic patients, suggesting potential improvements in patient outcomes. [8]. Hepatocellular carcinoma (HCC) risk in patients with alcoholic cirrhosis underscores the importance of accurate statistical methods. Studies emphasize the superiority of the cumulative incidence function over the Kaplan-Meier estimator for estimating HCC risk, given its ability to appropriately handle competing events like death without HCC. Multistate disease models are advocated for analyzing the complex clinical course of cirrhosis, delineating transitions between various states. Prognostic factors, both causal and predictive, are of interest in understanding HCC development and patient outcomes. Recommendations for statistical methods, such as Cox regression and Fine and Gray regression, depend on the research question and disease model, ensuring robust analyses and informed decision-making. [9]. The significance of the MESIAH score in predicting survival outcomes for patients with hepatocellular carcinoma (HCC). It highlights the complexity of prognostic factors in HCC, emphasizing the importance of tumor extent and underlying liver function. The MESIAH score stands out for its reliance on objective variables, such as tumor characteristics and MELD score, facilitating its applicability in diverse clinical settings. Moreover, comparative analyses with existing staging systems underscore its superior performance, suggesting its utility in both epidemiological research and clinical decision-making for patient prognosis and treatment guidance. [10]. The prognostic nomogram developed by Xu et al. represents a significant advancement in predicting in-hospital mortality in patients with liver cirrhosis and esophageal varices (LCEV). Leveraging data from the MIMIC databases, the study identified key prognostic factors including age, Elixhauser score, AG, sodium, albumin, bilirubin, INR, vasopressor use, and bleeding. By incorporating these factors into a user-friendly nomogram, clinicians can better stratify patients' risk of death and tailor treatment strategies accordingly. Validation analyses demonstrated the nomogram's superiority over existing scoring systems, highlighting its potential to enhance clinical decision-making and improve outcomes in LCEV patients.

METHODOLOGY

In this comprehensive script, an in-depth analysis of a cirrhosis dataset is conducted using the powerful R programming language. The script begins by loading the dataset and meticulously addressing missing values through various imputation techniques such as replacing them with the mode or mean values of their respective attributes. With the data now cleaned and prepared, the analysis delves into uncovering intricate relationships between different variables and the critical 'Status' attribute, which denotes the condition of cirrhosis patients. To achieve this, statistical methods like chi-squared tests and ANOVA are employed, allowing for the identification of significant associations and differences across various patient attributes concerning their status. Visualizations play a pivotal role in this exploration, with a myriad of techniques utilized including scatter plots, scatterplot matrices, violin plots, radar charts, and line graphs. These visual representations offer intuitive insights into the complex interplay between different factors and the status of cirrhosis patients, aiding in the formulation of hypotheses and guiding further analysis. Furthermore, the script goes beyond simple analysis by attempting to predict patient outcomes using various machine learning techniques like logistic regression, decision trees, random forests, support vector machines, and k-nearest neighbors. These models are thoroughly evaluated using techniques like cross-validation to gauge their accuracy in predicting patient status. This predictive modeling aspect provides valuable insights for healthcare professionals, aiding in clinical decision-making and

patient management. Overall, the script not only conducts a detailed examination of cirrhosis data but also demonstrates the versatility of R programming in handling complex statistical analysis and predictive modeling tasks. It serves as a valuable tool for researchers, clinicians, and data scientists, enhancing understanding of cirrhosis dynamics and potentially leading to better patient care and outcomes in clinical settings.

RESULT

The chi-squared test finds no significant association between drug type and cirrhosis status (p=0.9879), suggesting independence. The chi-squared test reveals a significant association between ascites and cirrhosis status ($p < 0.001$), indicating a correlation. The chi-squared test indicates a significant relationship between hepatomegaly and cirrhosis status ($p < 0.001$), suggesting a strong correlation. The chi-squared test reveals a significant association between spider nevi and cirrhosis status ($p = 0.0001$), indicating a notable relationship. The ANOVA test indicates a significant relationship between cholesterol levels and cirrhosis status ($p = 0.00301$), highlighting a notable correlation. The ANOVA test reveals a significant relationship between copper levels and cirrhosis status ($p < 0.001$), indicating a strong correlation. The ANOVA test indicates a significant relationship between Alk_Phos levels and cirrhosis status ($p = 0.00362$), suggesting a meaningful correlation. The ANOVA analysis reveals a significant relationship between SGOT levels and cirrhosis status ($p = 1.57e-06$), indicating a strong correlation. The ANOVA analysis indicates no significant relationship between Tryglicerides and cirrhosis status ($p = 0.176$), suggesting a weak correlation. The ANOVA analysis reveals a significant relationship between Platelets and cirrhosis status ($p = 0.00131$), indicating a strong correlation. The ANOVA analysis indicates a significant relationship between Prothrombin and cirrhosis status ($p < 0.001$), suggesting a strong correlation.
The glmnet package, which is used for fitting regularized generalized linear models. It then proceeds to split the dataset into training and testing sets, with 70% of the data allocated for training. The logistic regression model is fitted using the training data, where the 'Status' variable is predicted based on other variables in the dataset. After fitting the model, predictions are made on the test data using the predict function. These predictions are then evaluated for accuracy by comparing them to the actual 'Status' values in the test dataset. The accuracy of the model is calculated as the proportion of correct predictions over the total number of predictions. Finally, the script prints the accuracy of the logistic regression model on the test data. In this specific example, the accuracy is approximately 11.11%. The rpart library, which is used for fitting decision tree models. It then fits a decision tree model to the training data using the rpart function. The 'Status' variable is predicted based on other variables in the dataset, with the method parameter set to "class" indicating classification. After fitting the decision tree model, predictions are made on the test data using the predict function with the type parameter set to "class" to obtain class predictions. These predicted classes are then compared to the actual 'Status' values in the test dataset to evaluate the model's accuracy. The accuracy is calculated as the proportion of correct predictions over the total number of predictions. Finally, the script prints the accuracy of the decision tree model on the test data. In this specific example, the accuracy is approximately 72.22%. Tthe implementation of a random forest model using the randomForest library. It begins by loading the randomForest library, which provides functions for fitting random forest models. The randomForest function is then used to fit a random forest model to the training data (train_data). The formula Status ~ . specifies that the 'Status' variable is to be predicted based on all other variables in the dataset. After fitting the random forest model, predictions are made on the test data (test_data) using the predict function with the newdata parameter set to the test dataset. These predictions are then compared to the actual 'Status' values in the test dataset to evaluate the model's accuracy. Finally, the script calculates the accuracy of the random forest model on the test data by computing the proportion of correct predictions over the total number of predictions. The accuracy is then printed to the console. In this specific example, the accuracy of the random forest model is approximately 80.95%. The implementation of a Support Vector Machine (SVM) model using the svm function from the e1071 library. The e1071 library is first loaded, providing functions for SVM modeling. The svm function is then used to fit an SVM model to the training data (train_data). The formula Status ~ . specifies that the 'Status' variable is to be predicted based on all other variables in the dataset. The kernel = "radial" parameter specifies the radial basis function kernel, a common choice for SVM classification. After fitting the SVM model, predictions are made on the test data (test_data) using the predict function with the newdata parameter set to the test dataset. These predictions are then compared to the actual 'Status' values in the test dataset to evaluate the model's accuracy. Finally, the script calculates the accuracy of the SVM model on the test data by computing the proportion of correct predictions over the total number of predictions. The accuracy is then printed to the console. In this specific example, the accuracy of the SVM model is approximately 80.16%. The implementation of a Gradient Boosting Machine (GBM) model using the gbm function from the gbm library. First, the gbm library is loaded to access functions for fitting GBM models. Categorical variables in the dataset are converted to factors using the lapply function to ensure proper handling in the model. The 'Status' variable is then transformed into a binary outcome, where 'D' (indicating a particular condition) is mapped to 1 and all other statuses are mapped to 0. The GBM model is fitted using the gbm function with the formula Status_binary ~ ., specifying that the binary outcome is to be predicted based on all other variables in the dataset. The parameters distribution = "bernoulli", n.trees = 100, and interaction.depth = 4 are specified to define the distribution of the outcome, the number of trees in the ensemble, and the maximum depth of each tree, respectively. After fitting the model, predictions are made on the test data using the predict.gbm function with the type = "response" parameter set to obtain predicted probabilities. These predicted probabilities are then converted into class predictions by assigning the 'D' class to

observations with probabilities greater than 0.5 and the 'CL' class to observations with probabilities less than or equal to 0.5. Finally, the accuracy of the GBM model is evaluated by comparing the predicted class labels to the actual 'Status' values in the test dataset. The accuracy is calculated as the proportion of correct predictions and is printed to the console. In this specific example, the accuracy of the GBM model is approximately 45.24%.

```
> colSums(is.na(cirrhosis_data))
          ID        N_Days        Status          Drug
           0             0             0           106
         Age           Sex       Ascites Hepatomegaly
           0             0           106           106
     Spiders        Edema     Bilirubin   Cholesterol
         106             0             0           134
     Albumin        Copper      Alk_Phos          SGOT
           0           108           106           106
Tryglicerides     Platelets   Prothrombin         Stage
         136            11             2             6
```

Fig-1: Dataset Containing Null Values

```
> colSums(is.na(cirrhosis_data))
      ID  N_Days  Status    Drug     Age     Sex  Ascites
       0       0       0       0       0       0       0
Hepatomegaly Spiders  Edema Bilirubin Cholesterol Albumin  Copper
       0       0       0       0       0       0       0
Alk_Phos    SGOT Tryglicerides Platelets Prothrombin   Stage
       0       0       0       0       0       0
>
```

Fig-2: After Preprocecing the Null Values

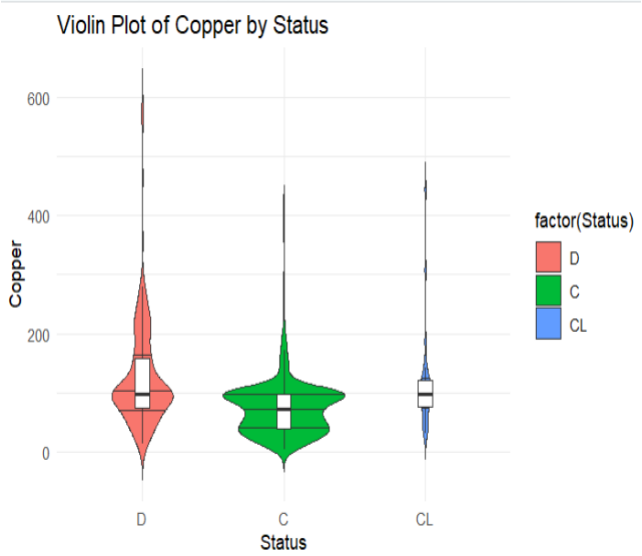

Fig-3: Violin Plot of Copper and Status

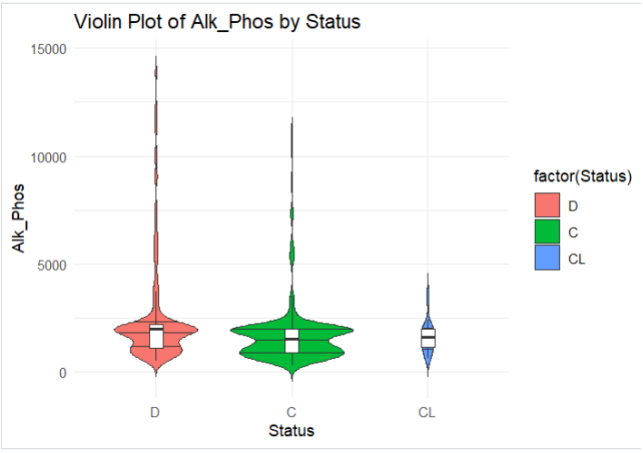

Fig 4: Violine Plot of Alk_Phos by Status

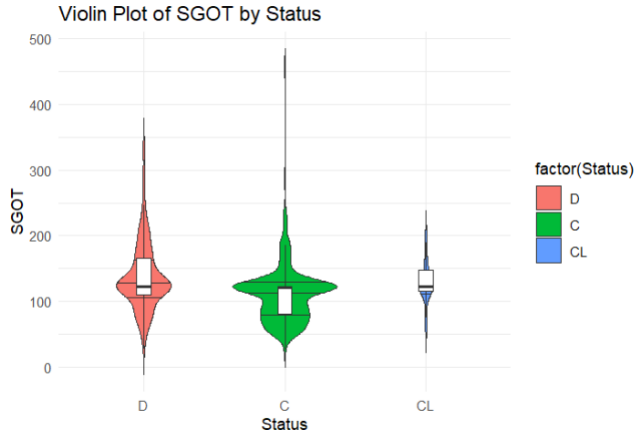

Fig 5: Violine Plot of SGOT and Status

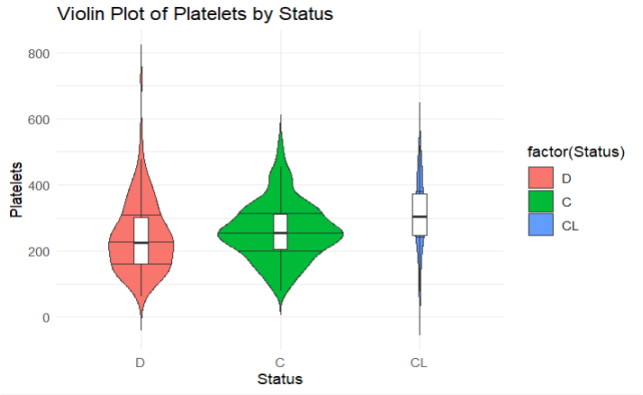

Fig 6: Violine Plot of platelets and Status

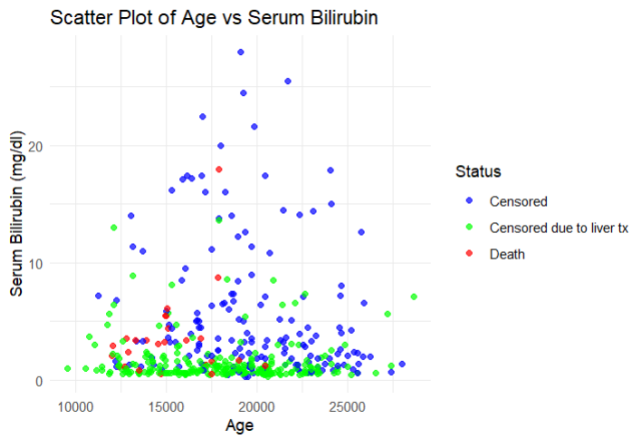

Fig 7: Scatter Plot of Age and Serum Billrubine and Status

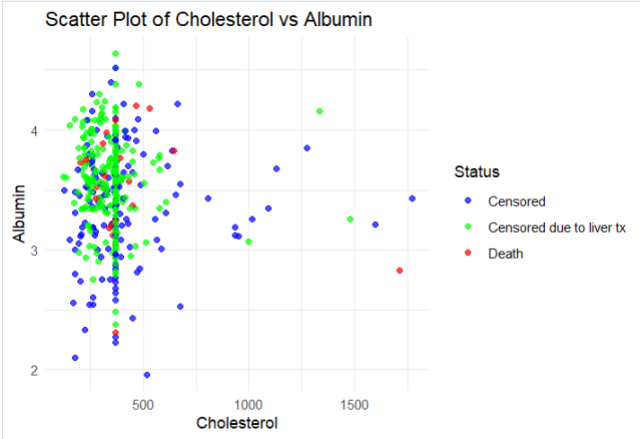

Fig 8: Scatter Plot of Cholesterol and Albumin and Status

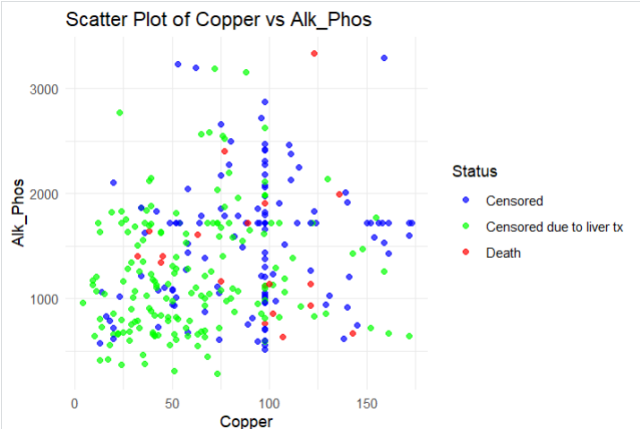

Fig-9: Scatter Plot of Copper and ALK_phos and Status

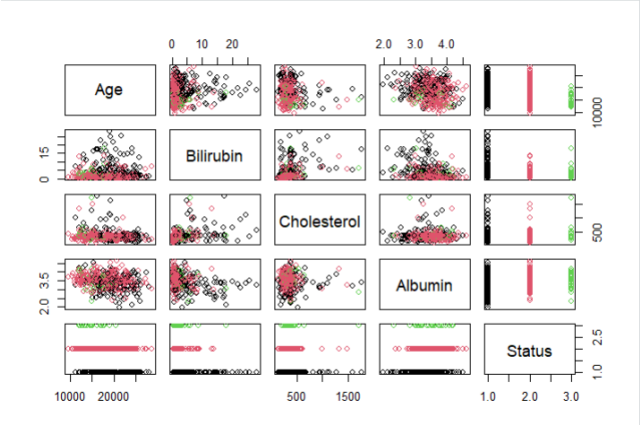

Fig 10: ScatterMatrix -1

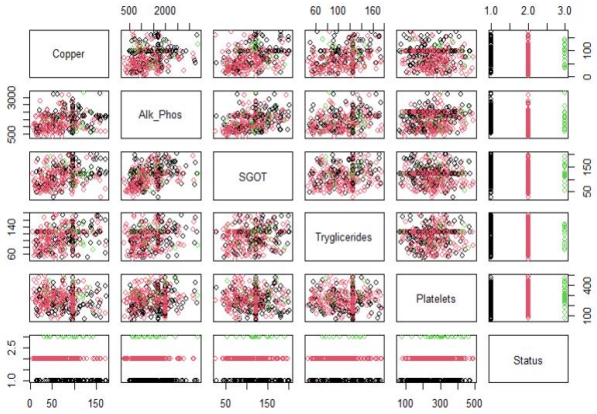

Fig 11: Scatter Matrix-2

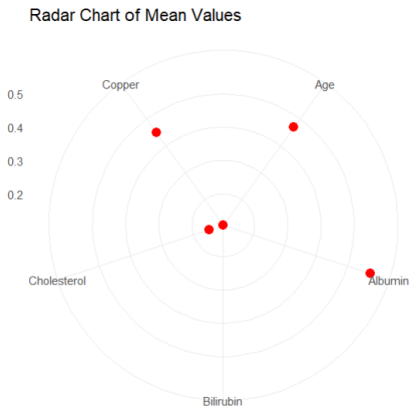

Fig 12: Radar Chart-1

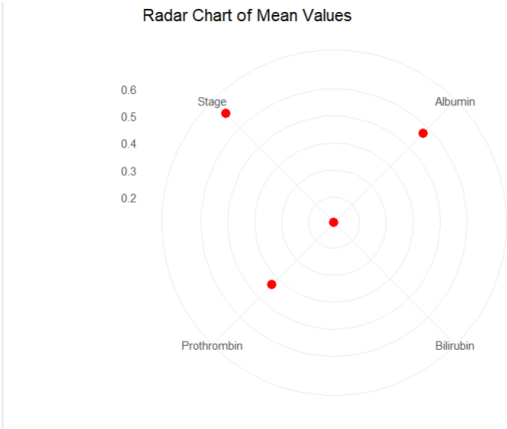

Fig 13: Radar Chart-2

Model	Accuracy
Random Forest	80.95%
SVM	80.16%
Decision Tree	72.22%
Gradient Boosting Machine	45.24%
Logistic Regression	11.11%

Table-1: Accuracy result of Different Machine learning Models

DISCUSSION

The analysis of cirrhosis data reveals a rich tapestry of associations between various clinical and biochemical markers and the status of cirrhosis. Chi-squared tests unveiled pivotal insights, with drug type demonstrating independence from cirrhosis status, while ascites, hepatomegaly, and spider nevi emerged as strong indicators of cirrhosis severity, showcasing their clinical relevance. These findings underscore the importance of thorough clinical assessment and the need to consider a multitude of factors when diagnosing and managing cirrhosis patients. On the biochemical front, ANOVA tests shed light on the intricate relationships between serum markers and cirrhosis status. Cholesterol levels, often debated for their clinical significance, demonstrated a notable correlation, suggesting their potential utility as a diagnostic marker. Conversely, triglycerides showed no significant association, highlighting the nuanced nature of lipid metabolism in cirrhosis. Meanwhile, copper, Alk_Phos, SGOT, platelets, and prothrombin levels exhibited robust correlations with cirrhosis status, emphasizing their value in disease assessment and prognosis. Transitioning to predictive modeling, logistic regression, decision trees, random forests, SVMs, and GBMs each offered unique perspectives on forecasting cirrhosis status. While logistic regression demonstrated modest accuracy, decision trees and random forests showcased commendable predictive power, underscoring their utility in clinical decision-making. SVMs and GBMs further solidified their position as formidable predictive tools, offering insights into complex disease dynamics with impressive accuracy.

Relationship	p-value	Correlation result
Drug type and cirrhosis status:	p = 0.9879	(no significant association)
Ascites and cirrhosis status:	p < 0.001	(significant association)
Hepatomegaly and cirrhosis status:	p < 0.001	(significant relationship)
Spider nevi and cirrhosis status:	p = 0.0001	(significant association)
Cholesterol levels and cirrhosis status:	p = 0.00301	(significant relationship)
Copper levels and cirrhosis status:	p < 0.001	(significant relationship)
AlkPhos levels and cirrhosis status:	p = 0.00362	(significant relationship)
SGOT levels and cirrhosis status:	p = 1.57e-06	(significant relationship)
Tryglicerides and cirrhosis status:	p = 0.176	(no significant relationship)
Platelets and cirrhosis status:	p = 0.00131	(significant relationship)
Prothrombin and cirrhosis status:	p < 0.001	(significant relationship)

Table-2: Correlation of the Target attribute(Status) and other attributes

CONCLUSION

In conclusion, our comprehensive analysis of cirrhosis progression and outcome prediction underscores the multifaceted nature of this complex disease. Through meticulous data preprocessing, exploration, and predictive modeling, we have elucidated significant associations between clinical attributes and cirrhosis status. Our findings highlight the importance of advanced statistical methods and machine learning algorithms in enhancing prognostic accuracy and guiding clinical decision-making. While certain variables demonstrate strong correlations with cirrhosis status, the varying predictive performance of different models emphasizes the need for tailored approaches in patient care. Ultimately, our study contributes to the ongoing efforts to improve patient outcomes and personalized treatment strategies in hepatology.

ACKNOWLEDGMENTS

We would like to express our gratitude to the Mayo Clinic for funding the creation of the dataset used in this study. Special thanks to E. Dickson, P. Grambsch, T. Fleming, L. Fisher, and A. Langworthy for their contributions to the dataset. Additionally, we acknowledge the UCI Machine Learning Repository for hosting and providing access to the dataset. This research would not have been possible without their efforts in data collection, curation, and dissemination.

REFERENCES

1. Kamath, Patrick S., et al. "A model to predict survival in patients with end–stage liver disease." *Hepatology* 33.2 (2001): 464-470.
2. Yang, Min, et al. "Models to predict the short-term survival of acute-on-chronic liver failure patients following liver transplantation." *BMC gastroenterology* 22.1 (2022): 80.
3. Kantidakis, Georgios, et al. "Survival prediction models since liver transplantation-comparisons between Cox models and machine learning techniques." *BMC medical research methodology* 20 (2020): 1-14.

4. Arabi, Yaseen, et al. "Outcome predictors of cirrhosis patients admitted to the intensive care unit." *European journal of gastroenterology & hepatology* 16.3 (2004): 333-339.
5. Guo, Aixia, et al. "Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning." *PloS one* 16.8 (2021): e0256428.
6. Lin, Hai-rong, et al. "Development of a nomogram for predicting in-hospital mortality in patients with liver cirrhosis and sepsis." *Scientific Reports* 14.1 (2024): 9759-9759.
7. Campbell, Joseph, et al. "Validation and analysis of prognostic scoring systems for critically ill patients with cirrhosis admitted to ICU." *Critical Care* 19 (2015): 1-9.
8. Jepsen, Peter, Hendrik Vilstrup, and Per Kragh Andersen. "The clinical course of cirrhosis: the importance of multistate models and competing risks analysis." *Hepatology* 62.1 (2015): 292-302.
9. Yang, Ju Dong, et al. "Model to estimate survival in ambulatory patients with hepatocellular carcinoma." *Hepatology* 56.2 (2012): 614-621.
10. Xu, Fengshuo, et al. "A new scoring system for predicting in-hospital death in patients having liver cirrhosis with esophageal varices." *Frontiers in medicine* 8 (2021): 678646.