Multi-Class Prediction for Cirrhosis Outcomes

Sneha Jaikumar, Guning (Emily) Shen, Shefali Pai, Annabelle Jiang

Abstract

Cirrhosis of the liver, or hepatic cirrhosis, is a debilitating condition that results in permanent scarring and liver failure. We explore the extent to which four different machine-learning models, Logistic Regression, Random Forest, Neural Network, and XGBoost, can be used as a tool for predicting the survival of patients grappling with liver cirrhosis. Leveraging a dataset derived from a seminal Mayo Clinic study conducted between 1974 and 1984 on primary biliary cirrhosis (PBC), we explore the potential of 17 clinical features as predictive indicators. We use the accuracy yielded by each model in addition to confusion matrices to reach a thorough conclusion on model performance and the impact this can have on cirrhosis prevention and treatment.

1. Introduction

Cirrhosis is a debilitating condition that arises from prolonged liver damage and represents a significant global health concern. Moreover, chronic liver diseases, including cirrhosis, contribute significantly to the global disease burden, with an estimated 2 million deaths annually (Asrani et al. 2019). This further emphasizes the critical need for accurate prognostic tools to guide clinicians in the management of cirrhosis patients. The etiology of cirrhosis is diverse, encompassing chronic infections, such as hepatitis, and lifestyle-related factors, notably chronic alcohol consumption. The consequential impact on liver function can lead to severe complications, affecting the overall survival of individuals diagnosed with cirrhosis. This paper explores using machine learning as a tool for predicting the survival of patients grappling with liver cirrhosis. Through this exploration, we hope to contribute to the refinement of clinical decision-making processes, thereby improving the overall management and care of individuals afflicted by this challenging medical condition.

2. Data Preprocessing and Datasets

We use the Cirrhosis Patient Survival Prediction dataset released from Mayo Clinic (Dickson, E. et al 2023). In this dataset, the researchers collected 18 features from 7905 patients. Among the 18 features, 8 are categorical and 10 are numerical. Patient status is divided into three different categories: C (censored), CL (alive due to liver transplant), or D (death). Our goal is to construct models to predict the status of each patient with one of the three classes based on the 18 features.

Since some data cleaning processes, such as dropping null value dropouts and imputing missing values, have been already conducted before the release of this dataset, our data preprocessing pipeline mainly constitutes three steps. First, we use log transformation to correct skewed numerical features including Cholesterol, Copper, and Prothrombin. Second, we convert the value of each categorical feature into integer representations. The data distribution for 18 features is shown in the table below. Finally, we convert the patient Status into one-hot vector representation: [1, 0, 0] for C, [0, 1, 0] for CL, and [0, 0, 1] for D.

We divide the dataset into train, val, and test set with a ratio of 8:1:1 for model evaluation during and after training.

Feature Table

Feature Name	Type	Description
N_days	Numerical	Number of days between registration and the earlier of death, transplantation, or study analysis time
Drug	Categorical	Type of drug D-penicillamine or placebo
Age	Numerical	Age of patients
Sex	Categorical	Male/female
Ascites	Categorical	Presence or absence of fluid accumulation in the abdominal cavity
Hepatomegaly	Categorical	Enlargement of the liver
Spiders	Categorical	The presence of spider angiomas or spider nevi, visible vascular lesions on the skin associated with liver disease
Edema	Categorical	The abnormal accumulation of fluid in liver. N means no edema; S means present without diuretics; Y means edema despite diuretic therapy
Bilirubin	Numerical	A blood marker indicating liver function
Cholesterol	Numerical	Level of cholestorol in blood
Albumin	Numerical	Albumin level
Cropper	Numerical	Cropper level
Alk_Phos	Numerical	Alkaline phosphate
SGOT	Numerical	Enzyme indicated liver and heart health
Tryglicerides	Numerical	tryglicerides
Platelets	Numerical	platelets per cubic
Prothrombin	Numerical	blood clotting factor
Stage	Categorical	Historical stage of disease (1, 2, 3, 4)

2.1 Related Work. The research for survival prediction in cirrhosis patients has been significantly impactful and has become the focal point for many researchers. The area of survival prediction in cirrhosis patients has shifted under the integration of machine learning techniques and deep learning methods. Utilizing advanced algorithms helps further research in the field. One of the existing scoring systems for liver disease in patients is the Model for End-Stage Liver Disease, or the MELD. This is a mathematical formula-based system that does not support learning new data. Research has been conducted to both improve this system and experiment with machine learning and deep learning models to accomplish data learning to improve the prediction accuracy of a patient's health. Some common models used for the analysis were XGBoost, Random Forest, LightGBM, etc. That often yields a high-accuracy result. The XGBoost model has been used widely on this dataset, as it is known to handle various data types and provide accurate results. The LightGBM model also falls into the same category as XGBoost with similar functionalities. On the other hand, models such as Random Forest, Logistic Regression, and Neural Network work for the analysis but have not been explored much in this context of survival prediction. We aim to close this gap by performing a comprehensive evaluation among models that are used often and rarely used in this context.

3. Approach

We experiment with four models - Logistic Regression, Random Forest, XGBoost, and Neural Network, to perform a multi-class classification task on the Cirrhosis dataset. Each model had their functionalities and specifications that further analyzed the dataset.

3.1 Logistic Regression. Logistic regression is a statistical method usually used for binary classification, which means it is employed to predict the probability of an instance belonging to one of two classes. It transforms the output of a linear equation into a value between 0 and 1. In our problem, we employ the one-vs-rest strategy to expand the functionality of logistic regression model to multi-class classification problem. Three binary classification models are trained for each class.

To start, we first split the original data set into X and y and further also create training sets (X_train, y_train), test sets (X_test, y_test), and validation sets (X_val, y_val).

From this, we can create an instance of the logistic regression model. Then using the logreg.fit(X_train, y_train) statement, we can fit the logistic regression model to the training data. The model that we trained is used to make predictions on the testing set, which is X_test. The accuracy of the model on the test set is calculated using the score method, which computes the accuracy by comparing the predicted labels (y_pred) with the actual labels (y_test). The accuracy value that we are given is 0.815. The value that is given portrays the number of accurately predicted instances out of the total number of instances in the data set. Since the value is closer to one, we can assume that the model performs well in terms of correctness.

3.2 Random Forest. The Random Forest Classification model (RFC) creates a collection of decision trees, where each tree classifies independently. The final prediction of the model is determined when combining the individual tree predictions. The advantage of this model is that it supports both regression and classification tasks.

We approach the Random Forest model in two ways. One way was by creating ten decision trees by the default Random Forest model. The other was by having the hyperparameter setting of 500 decision trees with the seed of 42. By having two different predictions, we could evaluate the performance of the Random Forest model functioned on both tests. We first train the data using the fit function. Then we use the model to predict new unseen data and evaluate its performance. We ran this with both Random Forest models. With the prediction of both configurations, we were able to forge two distinct confusion matrices. An increasing number of decision trees enhanced the data prediction.

After we predicted the data with the Random Forest model. A feature score chart and a classification report were generated to help provide further detail about the analysis. The Random Forest model allows users to remove or add features, to be more selective, and to enhance better outcomes.

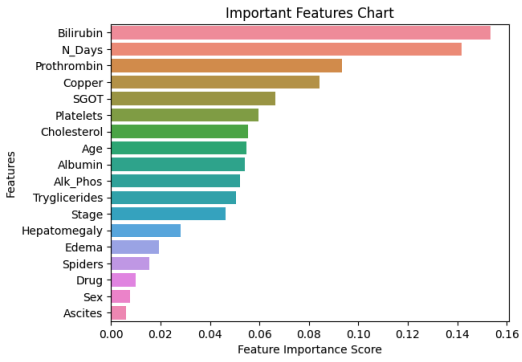

Fig. 1. Important Features

The pivotal features that impacted the outcome of the dataset identified by the model were Bilirubin, the number of days between registration and the earlier death, transplantation, or study analysis time in July 1986, and the Prothrombin time.

3.3 Neural Network. Neural network, inspired by the human brain, is an effective machine learning method for multi-class classification. In our neural network approach, we employ a straightforward yet effective feedforward neural network comprising three layers. The architecture of the model, as illustrated in the figure below.

1. The first layer is a fully connected layer with 32 units with ReLU activation. L2 regularization is applied to

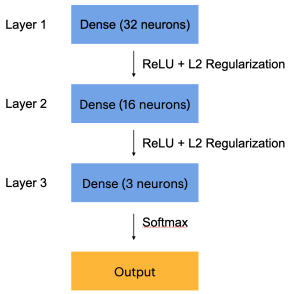

Fig. 2. Model Architecture

avoid overfitting. Given the input size of (batch size, 18), the output size for this layer is (batch size, 32).

2. The second layer is a fully connected layer with 16 units, using ReLU activation and L2 regularization. After passing this layer, the size of the output vector becomes (batch size, 16).

3. The final layer of our model serves as the decision-making component. It takes the 16-neuron output from the previous layer and feeds it into 3 neurons, each corresponding to a class in our classification task. It applies a softmax function to assign probability to each class.

We use cross-entropy as our loss function.

$$Loss = -\sum_{i}^{C} t_i \log(f(s_i)), \tag{1}$$

where $f(s)_i$ refers to the softmax score for each class, t_i is the target vector in one-hot representation, and C means the number of classes. Therefore, categorical cross-entropy sums out the cross-entropy loss for each actual class. During training, we apply the Nadam optimizer with an initial learning rate of 0.005 to adaptively adjust the learning rate. We also use early stopping to ensure the model stops to fit after 10 unimproved validation loss.

3.4 XGBoost. Extreme Gradient Boosting, or XGBoost, is a powerful gradient boosting algorithm effective for structured data widely used for tasks such as classification, regression, and ranking. XGBoost introduces a regularized objective function that consists of two parts: a loss term measuring the model's performance on the training data and a regularization term to control the complexity of the model.

The objective function is optimized during the training process to find the best combination of weak learners. XGBoost builds trees sequentially, and each new tree corrects the errors of the combined model so far. The gradient of the loss function is calculated for each instance in the training data. XGBoost also includes regularization terms in the objective function to control the complexity of the individual trees and the overall model. Regularization helps prevent overfitting and contributes to the algorithm's ability to generalize well.

Our XGBoost model, trained for predicting survival rates of patients with liver cirrhosis using 17 clinical features, yielded an overall accuracy of approximately 93.9%, given a maximum tree depth of 4, step size shrinkage of 0.3, and an objective of logistic regression for binary classification. We choose binary classification as our objective because each of the three labels—death (D), censored (C), and censored due to liver transplantation (CL)—had two classes (0 and 1). The model's performance here highlights the efficacy of the model in capturing intricate patterns within the dataset.

A distinctive strength of this model lies in its capacity to handle the multiclass nature of the survival states. The inclusion of three distinct labels highlighted the model's versatility. Its ability to differentiate between these nuanced states attests to its capacity to address the complexity inherent in cirrhosis patient outcomes.

4. Results

We evaluate our models based on two criteria: accuracy and confusion matrix analysis.

4.1 Accuracy. Out of the four models, the two Random Forest Models had the highest accuracy score, and Logistic Regression with the lowest accuracy score. Our results showed that the Random Forest Model would be the best choice for this dataset. Many others chose to apply XGBoost for this dataset, which is also a good option with an accuracy of over 90%. That can be applied to various data types compared to the Random Forest Model.

After fine-tuning, the best accuracy of the feedforward neural network on the test dataset is 0.81701. The two Random Forest Models both yield an accuracy score of 0.95302. The feature score chart shows Bilirubin was the most significant factor in predicting the data. The XGBoost generates an accuracy score of 0.93067, and Logistic Regression with a score of 0.81587.

Model Name	Accuracy
Logistic Regression	0.81587
Random Forest	0.95302
Random Forest-500	0.95302
XGBoost	0.93067
Neural Network	0.81701

4.2 Confusion Matrix. We are able to generate confusion matrices for each of our models. The confusion matrix for logistic regression shows the model correctly identifies 3439 patients with status CL, 868 patients with status C, and 0 with status D. In Random Forest Classification, the model recognized 3657 patients with status CL, 1358 patients with status C, and 159 with status D. In the confusion matrix for Neural Network, we could see the model correctly identified 3400 patients with status CL, 973 patients with status C, and 0 with

status D. For XGBoost, 3567 patients were correctly identified by the model with status CL, 1238 were status C, and 108 with status D. Overall, the performance of the Random Forest Classification model accurately identified most patients in all status out of the four models tested.

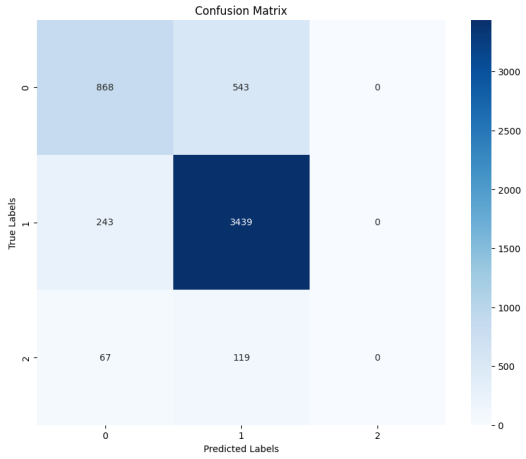

Fig. 3. Confusion Matrix for Logistic Regression

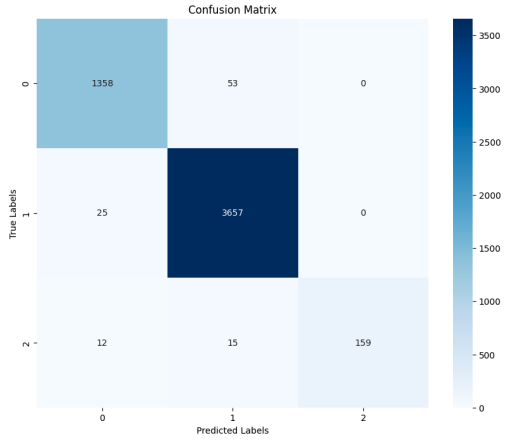

Fig. 4. Confusion Matrix for Random Forest

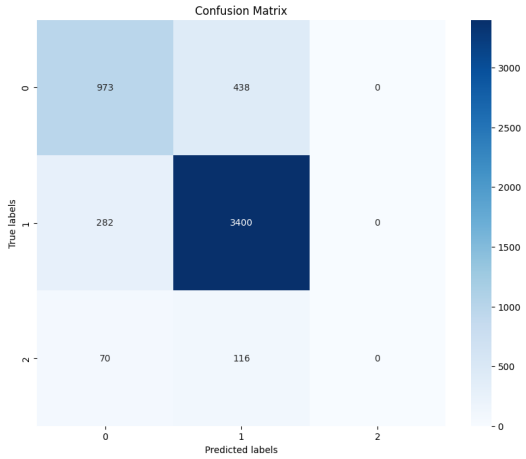

Fig. 5. Confusion Matrix for Neural Network

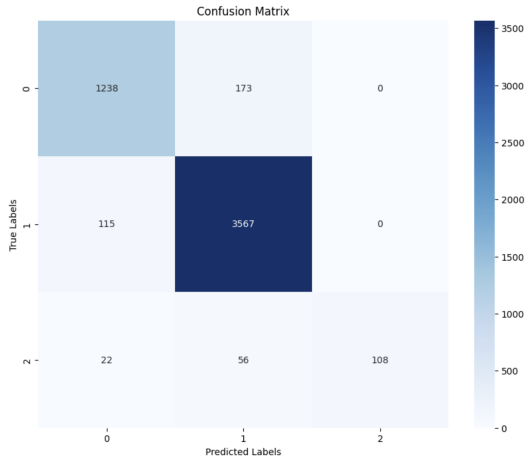

Fig. 6. Confusion Matrix for XGBoost Model

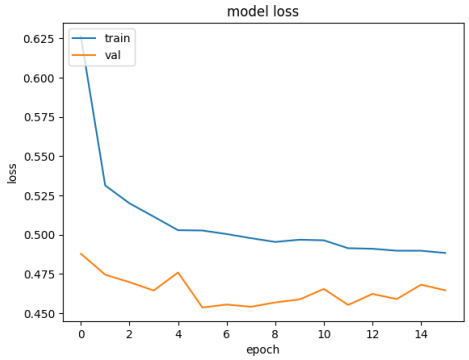

Fig. 7. Loss Curve for Neural Network

4.3 Neural Network Loss Curve. Figure 7 shows the loss curve on train dataset and validation dataset during training of our neural network. Validation loss is always smaller than training loss with a small error, meaning that the model is not overfitting.

5. Conclusion

Cirrhosis is a serious condition when the liver is scarred and permanently damaged. From the analysis of results, we had found that the Random Forest Models had the best accuracy score, meaning that it would be the most reliable to utilize when making predictions for this topic. The several machine learning models that we have utilized such as logistic regression, neural networks, etc. have showed us the relationships between the variables that impact how long a cirrhosis patient can survive based on certain factors. After finding the accuracy from these models, we believe that they are imperative to the survival of cirrhosis patients. Since so many factors can impact if a patient can survive and how long they have of survival, it is important to use accurate machine learning models to give the proper information. We hope to utilize the Random Forest Classifier's output, which in this case was Bilirubin, and take that into consideration when finding ways to treat and prevent cirrhosis. We believe that these models will not only be useful for the treatment and prevention of cirrhosis, but for other serious medical conditions as well. Hence, we can advance health through machine learning methods.

Bibliography

Asrani, Sumeet K. et al. “Burden of liver diseases in the world.” *Journal of hepatology*, vol. 70,1 (2019): 151-171. doi:10.1016/j.jhep.2018.09.014

Dickson, E. et al. "Cirrhosis Patient Survival Prediction. *UCI Machine Learning Repository,* 2023. doi:10.24432/C5R02G.