Primary Biliary Cholangitis Patient Survival Prediction

James Hu

Brown University DSI

https://github.com/jameshujames/cirrhosis-survival-prediction

Introduction

Primary Biliary Cholangitis, or PBC, is a chronic autoimmune liver disease that can lead to cirrhosis, liver failure, and even death. Disease progression is highly variable, and clinical assessment is challenging because many patients are asymptomatic at the time of diagnosis. Thus, there is great interest in using routinely collected clinical and laboratory data to predict patient survival outcomes and to identify high-risk individuals at an early stage.

In this project, I study a binary classification problem: predict whether a patient diagnosed with PBC survives or dies during the follow-up period (~4 years). The dataset used for analysis is the Cirrhosis Patient Survival Prediction dataset from the UCI Machine Learning Repository, which originates from a Mayo Clinic clinical study on PBC patients (dataset; study). The dataset contains 418 patient records with 17 features, including demographic variables, clinical observations, and standard blood test measurements. A key characteristic of this dataset is missing values across multiple clinically relevant features, reflecting real-world limitations of medical data collection.

Previous work on PBC survival analysis has identified laboratory markers such as bilirubin and albumin as strong predictors of patient outcomes, and traditional clinical risk scores like the Mayo Risk Score have been widely used for prognostic assessment. However, machine learning studies on similar clinical survival datasets have reported moderate predictive performance at best, constrained by the data's nature of small sample sizes, missing data, and measurement noise. As a result, predictive performance in this setting is expected to be limited, so models developed for this type of task typically achieve performance comparable to existing clinical baselines, rather than dramatic improvements.

EDA

Exploratory data analysis was conducted to understand the distributions of key clinical variables, assess relationships between predictors and patient outcomes, and identify patterns relevant to disease severity and survival.

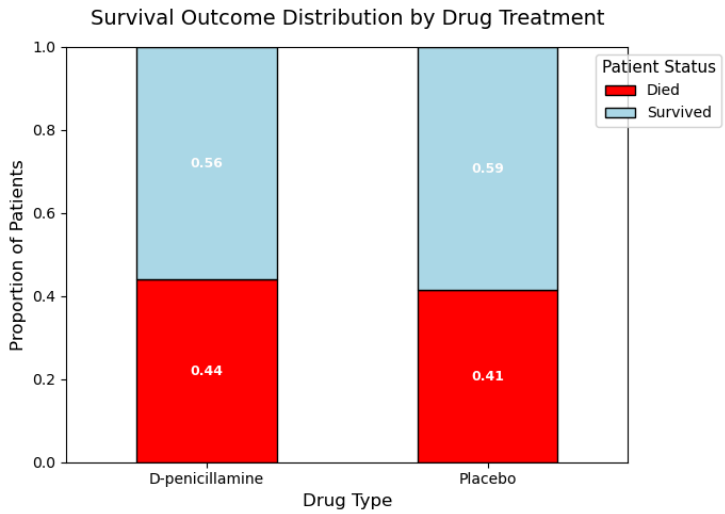


Figure 1 above shows the distribution of survival outcomes stratified by drug treatment group. Patients in the dataset were assigned to either D-penicillamine or placebo. The proportions of patients who survived and died are nearly identical across the two treatment groups, with approximately 56% survival in the D-penicillamine group and 59% survival in the placebo group. A chi-square test for independence yields a non-significant result ($p = 0.75$), indicating no detectable association between treatment assignment and survival outcome in this dataset. This suggests that drug treatment alone is unlikely to be a strong predictor of survival and motivates the use of additional clinical and laboratory features.

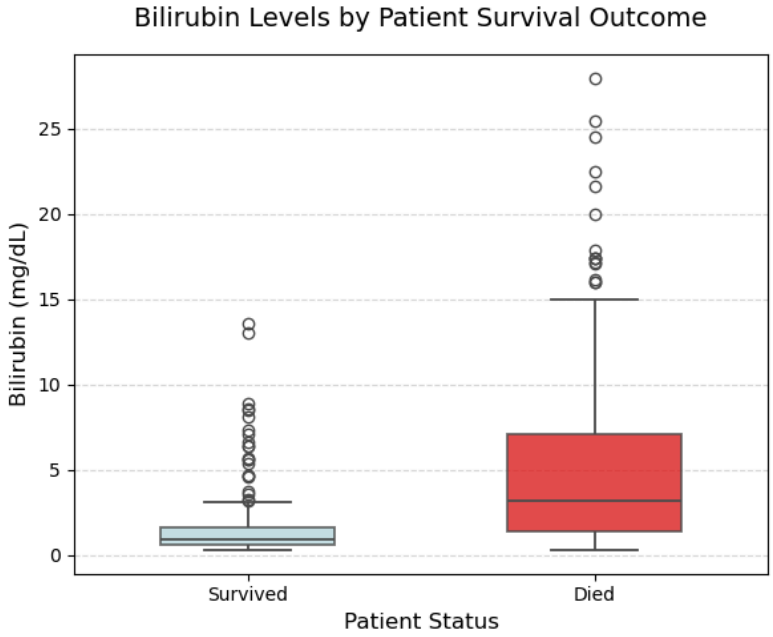


Figure 2 above compares bilirubin levels between patients who survived and those who died. Bilirubin is a byproduct of red blood cell breakdown and is processed and excreted by the liver, so elevated bilirubin levels are indicative of impaired liver function. The boxplot reveals a substantial shift toward higher

bilirubin values among patients who died, with both higher medians and more extreme upper-tail values compared to survivors. This pronounced difference suggests a strong association between elevated bilirubin levels and mortality risk.

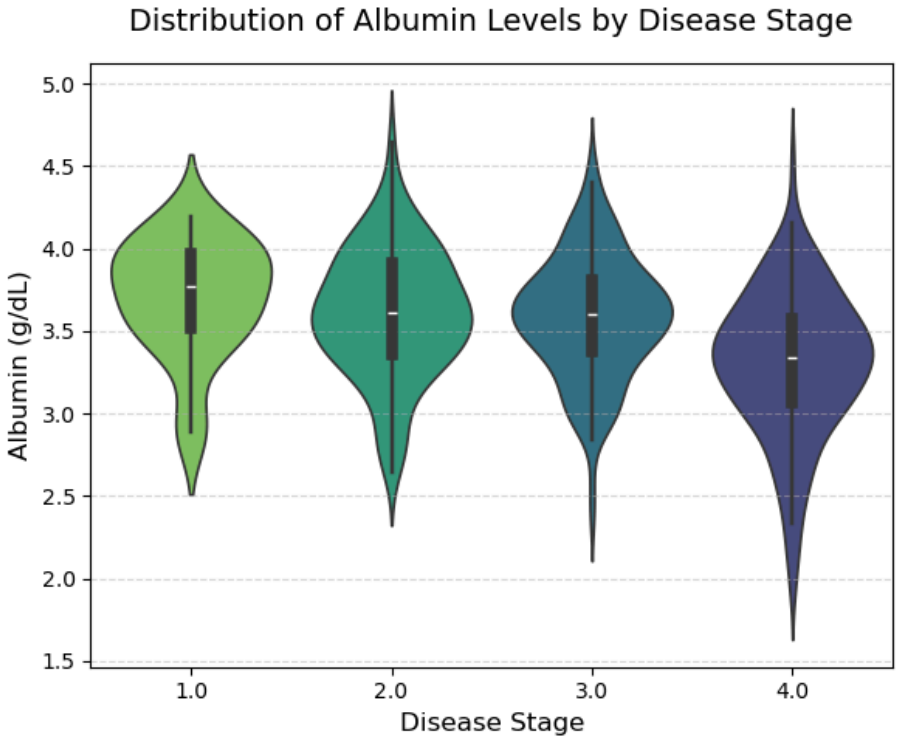


Figure 3 illustrates the distribution of albumin levels across disease stages. Albumin is a protein synthesized by the liver and is an important marker of liver synthetic function. The violin plots show a clear downward trend in albumin levels as disease stage increases, with patients in advanced stages exhibiting lower median albumin values. In addition, the distributions widen at higher disease stages, indicating increased variability in albumin levels among patients with more severe disease. This pattern is consistent with clinical expectations and suggests that albumin may be informative for assessing disease severity and prognosis.

Methods

To evaluate model performance while minimizing bias and variance from data splitting, I adopted an 80/20 train-holdout split, reserving 20% of the data as a final test set that was not used at any stage of model selection or tuning. The remaining 80% of the data was used for model training and hyperparameter optimization.

Within the training set, I applied 5-fold StratifiedKFold cross-validation, ensuring each fold preserved the original class distribution. This was particularly important given the moderate class imbalance in the dataset (approximately a 60/40 split between positive/negative outcomes), as stratification reduces variance in performance estimates and prevents folds from being dominated by a single class. Under this

scheme, each cross-validation iteration effectively corresponded to a 64/16/20 train/validation/test split of the full dataset.

All models were implemented using a unified scikit-learn pipeline to prevent data leakage and ensure consistent preprocessing across algorithms. The preprocessing stage included handling missing values (extra category for OneHotEncoder(), IterativeImputer() for numerical features, and being the "first" category for OrdinalEncoder()), standard scaling numerical features, and encoding categorical/ordinal variables as appropriate. Preprocessing steps were always fit on training data and then applied to validation and test data.

Combining preprocessing and model training into a single pipeline ensured that hyperparameter tuning accounted for preprocessing-related variability and prevented information leakage, resulting in a more reliable estimate of performance.

Model performance was primarily evaluated using the F1 score, with accuracy reported as a secondary metric. The F1 score was chosen because it balances precision and recall and is more informative than accuracy in settings with class imbalance, where a naive majority-class predictor can achieve deceptively high accuracy. All hyperparameter optimization was performed using cross-validated F1 score, and final test-set F1 scores were used for model comparison.

Below is a table that summarizes the four machine learning algorithms evaluated and the corresponding hyperparameter ranges explored during tuning.

Logistic Regression	SVM	Random Forest	XGBoost
• Regularization strength C (10^{-6} to 10^{6} , 13 values log-spaced) • Class weight (None, balanced)	• Kernel (Linear, RBF) • Regularization strength C (10^{-4} to 10^{4}, 9 values log-spaced) • Gamma (10^{-5} to 10^{1}, 7 values log-spaced) • Class weight (None, balanced)	• Number of trees (200 - 800, 7 values linearly spaced) • Max tree depth (None, 3, 5, 7, 9) • Max features per split (sqrt, log2, 0.4, 0.6, 0.8)	• Max tree depth (2,4,6,8) • Subsample ratio (0.6 to 1.0, 5 values) • Column subsample ratio (0.6 to 1.0, 5 values) • Minimum child weight (1,5,10) • L2 regularization (10^{-3} to 10^{2}, 6 values log-spaced) • Early-stopping (up to 3000)

To quantify uncertainty in performance, variability from data splitting was assessed via cross-validation. In addition, for all models, including non-deterministic methods like random forest and XGBoost,

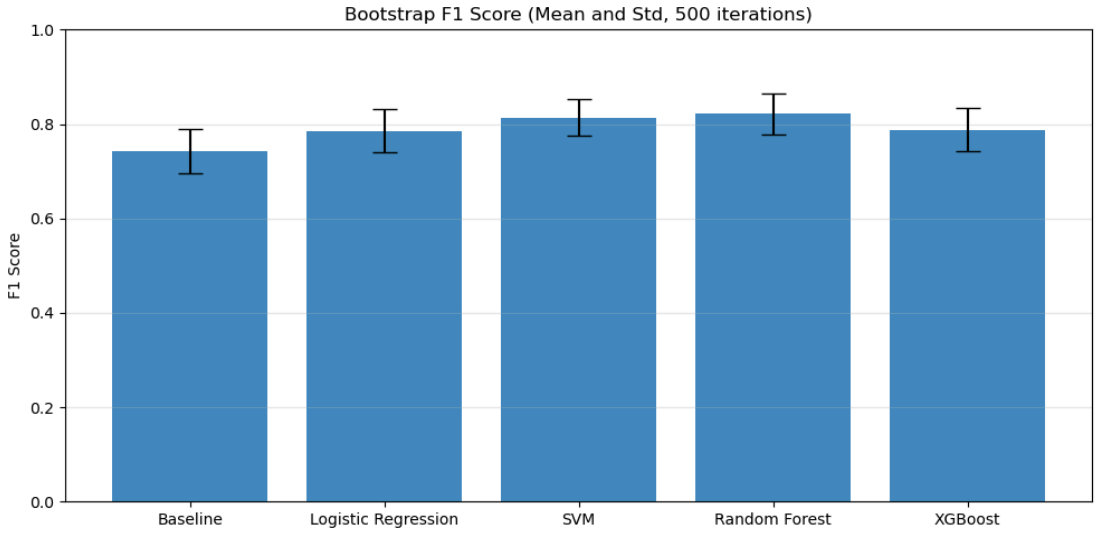
performance was evaluated using repeated bootstrap resampling of the test set. This enabled estimation of the mean and standard deviation of evaluation metrics, providing a more robust comparison of models than single point estimates.

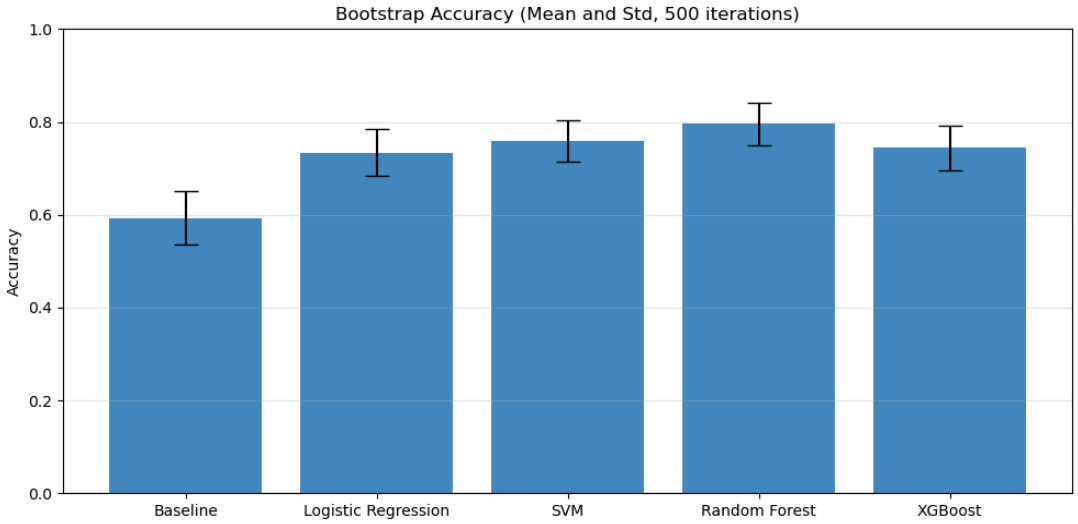
Results

Model performance was evaluated on a held-out test set using bootstrap resampling, with results reported as mean ± standard deviation for both F1 score (primary metric) and accuracy (secondary metric).

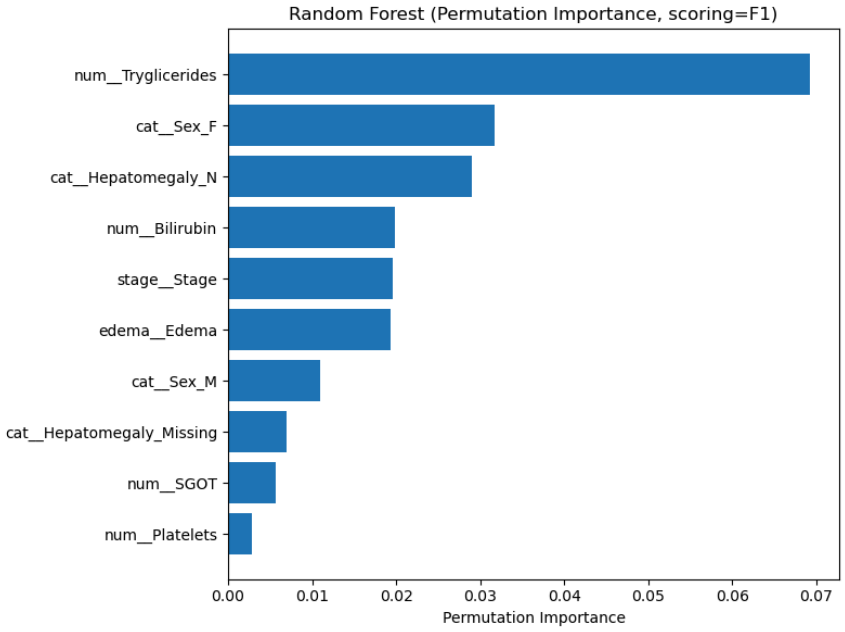
A majority-class baseline achieved an F1 score of 0.743 ± 0.046 and accuracy of 0.593 ± 0.057. All machine learning models outperformed this baseline on both metrics. Random Forest achieved the strongest overall predictive performance, with F1 = 0.822 ± 0.044 and accuracy = 0.796 ± 0.045, followed closely by SVM, which achieved F1 = 0.813 ± 0.039 and accuracy = 0.759 ± 0.045. Logistic Regression showed worse performance with F1 = 0.786 ± 0.046 and accuracy = 0.734 ± 0.051. XGBoost surprisingly underperformed compared to Random Forest and SVM, though beating out logistic regression with F1 = 0.788 ± 0.045 and accuracy = 0.744 ± 0.048.

Relative to the baseline, the Random Forest model achieved an improvement of approximately 1.7 standard deviations in F1 score, while SVM achieved an improvement of approximately 1.5 standard deviations. Logistic Regression and XGBoost achieved improvements of approximately 0.9 and 1 standard deviation above baseline, respectively. These results indicate that non-linear models provided meaningful gains over linear approaches, with Random Forest offering the best balance of precision and recall. The bar charts below summarize the performance of all models in terms of mean ± standard deviation for both F1 score and accuracy.

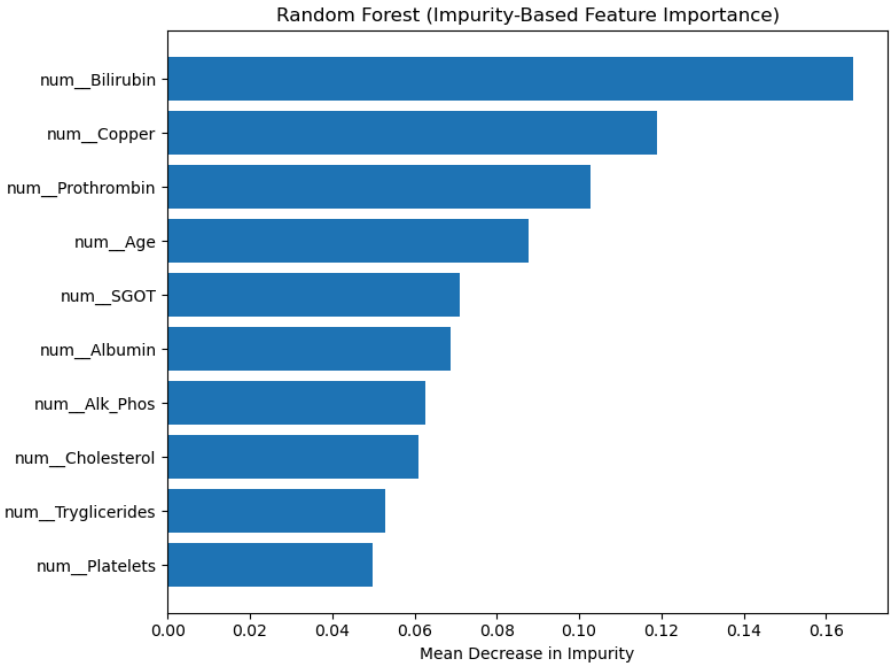




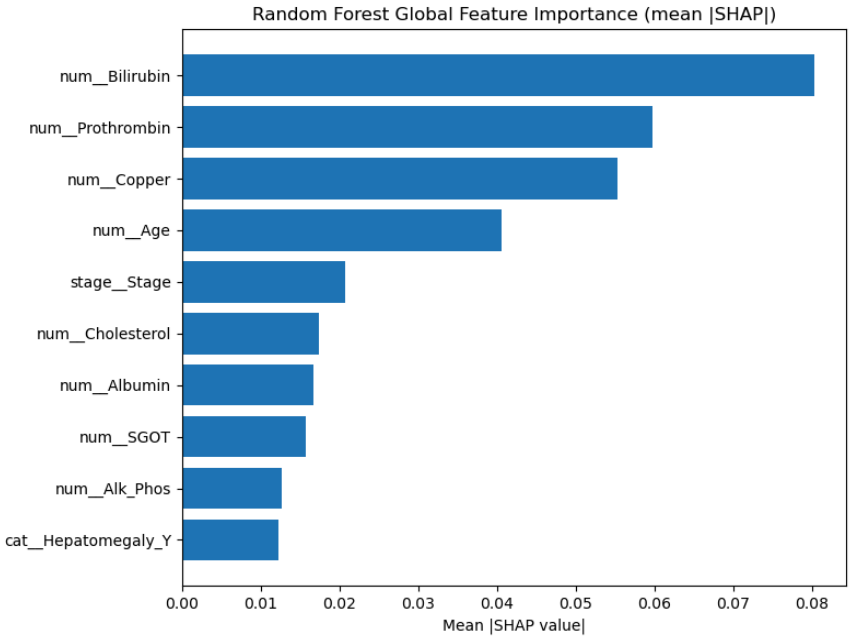
The bar chart below shows the top 10 global feature importances computed using permutation importance, measured as the decrease in F1 score when each feature is randomly shuffled. Under this metric, the number of triglycerides emerged as the most influential feature by a substantial margin, which indicates that disrupting triglyceride values led to the largest degradation in predictive performance. Other important features included sex, state of hepatomegaly, bilirubin levels, and disease stage, suggesting that both biochemical markers and demographic variables play a meaningful role in prediction.



The bar chart below shows top 10 impurity-based feature importances computed from the Random Forest model. Under this metric, bilirubin, copper, prothrombin time, and age ranked highest, which reflects their frequent use in tree-splitting decisions across the ensemble method. Compared to permutation importance, impurity-based importance places greater emphasis on continuous laboratory measurements, which is not surprising given the metric is known to favor variables with higher variance or more potential split points.

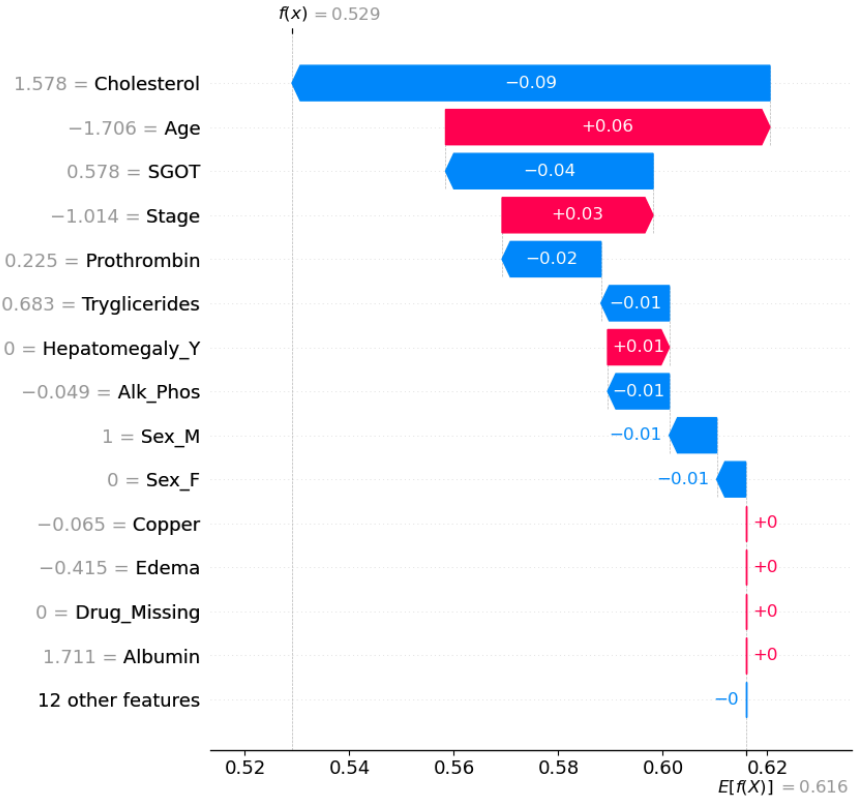


The bar chart below shows global SHAP importance as the mean absolute SHAP value across all test datapoints. Under this metric, bilirubin was the most influential feature overall, followed by prothrombin time, copper, age and disease stage.



To interpret individual predictions, SHAP values were also computed at the local level. The SHAP waterfall plot below is a local SHAP explanation for one test instance at idx = 13. In this example, the individual has a predicted survival of 0.529, which is significantly lower than the average predicted

survival of 0.616. The biggest risk factor is that this person has very high cholesterol compared to the average (660 mg/dL compared to average 358 mg/dL). What is helping boost this individual's survival is that they are much younger than the average (33 years old compared to average 51 years old) and at an earlier disease stage than the average. Overall, most features had near zero SHAP values, indicating that most of the features had minimal influence on this specific prediction. This local explanation demonstrates that, while certain biomarkers are globally important, individual predictions depend on a subset of features whose effects vary across patients.



In summary, across all global importance methods, bilirubin, prothrombin time, copper, and age were the most important features, consistently ranking highly under impurity-based importance and global SHAP values. These variables show that they are clinically meaningful markers of liver function and disease severity. In contrast, platelet count, alkaline phosphatase, and several categorical indicators exhibited consistently low importance, indicating limited marginal contribution to predictive performance.

An unexpected finding was the behavior of triglycerides, which ranked highest under permutation importance but lower under impurity-based and SHAP-based measures. This suggests that triglycerides contain predictive information that is difficult for the model to replace once removed, even if it is not frequently used in tree-splitting decisions, highlighting the value of using multiple complementary importance methods.

Outlook

While the Random Forest model achieved the strongest performance, several avenues remain to improve both predictive power and interpretability. Predictive performance could be improved by expanding the

hyperparameter search space for top-performing models, particularly SVM and Random Forest, to explore a wider range of model complexity and regularization settings. Additionally, the weaker performance of XGBoost was somewhat surprising given its strong performance in many learning tasks. This suggests further investigation, including better tuning and exploration of alternative gradient boosting frameworks like LightGBM and CatBoost, which may handle this clinical feature space more effectively.

Beyond model selection, interpretability could be strengthened through targeted error analysis. Examining false positives and false negatives may reveal systematic patterns or subpopulations where the model underperforms, helping identify blind spots in the current feature set or motivating additional data collection.

Finally, while the F1 score was an appropriate primary metric given class imbalance, future work could explore alternative evaluation metrics tailored to specific priorities. For example, an Fβ score could place greater emphasis on recall or precision depending on whether minimizing false negatives or false positives is more clinically relevant, enabling a more nuanced assessment of model performance.

References

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., & Langworthy, A. (1989). *Prognosis in primary biliary cirrhosis: Model for decision making*. *Hepatology*. PubMed

University of California, Irvine Machine Learning Repository. (2023). *Cirrhosis Patient Survival Prediction Dataset*. Retrieved from https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1