APPLIED MACHINE LEARNING SIMULATION PROJECT

Multi-Class Classification of Liver Cirrhosis Stages: An Evaluation of Random Forest, SVM, and Logistic Regression for the staging of Primary Biliary Cholangitis

Marco Cuscunà [University of Bologna | *marco.cuscuna@studio.unibo.it*]
GitHub Repository github.com/Markus2409/Applied_Machine_Learning_Project.

Abstract. *P*rimary Biliary Cholangitis (PBC) is a progressive liver disease where accurate staging is critical for determining prognosis and treatment strategies. This study evaluates the performance of three classical machine learning algorithms (Random Forest, Support Vector Machines (SVM), and Logistic Regression) in the multi-class classification of PBC stages (1,2,3,4). The analysis was conducted on a dataset characterized by significant class imbalance and missing values, necessitating the use of data imputation techniques to preserve the limited sample size. Experimental results indicate that while the models demonstrated discrete capability in distinguishing between the extreme stages of the disease (Stage 1 and Stage 4), they faced challenges in correctly classifying intermediate stages. A detailed error analysis revealed that performance was heavily influenced by the scarcity of Stage 1 samples and the presence of an atypical clinical profile; specifically, a severely misclassified Stage 4 patient exhibited biochemical markers (e.g., normal albumin and copper levels) and clinical features (absence of hepatomegaly) typically associated with early stage disease, thereby misleading the classifiers. Among the tested models, Logistic Regression emerged as the most balanced approach, offering robust classification for extreme-stages, while SVM showed more reasonable performance for intermediate ones compared to the other models. The study concludes that while classical machine learning methods provide a baseline for stratification, they are limited by the dataset's complexity. Future work with Deep Learning techniques, particularly representation learning, could be useful to extract more subtle features and overcome the intrinsic limitations of classical approaches in this clinical context.

Dataset: Cirrhosis Prediction Dataset.

Keywords: Primary Biliary Cholangitis; Machine Learning; Multi-Class Classification; Imbalanced Data;

Published: 19th January 2026

1 Introduction

The accurate staging of liver disease is a critical factor in determining patient prognosis and defining appropriate therapeutic strategies. This study focuses on the multi-class classification of disease stages in patients affected by Primary Biliary Cholangitis (PBC), leveraging classical Machine Learning algorithms to predict disease progression based on clinical and biochemical covariates.

1.1 Primary Biliary Cholangitis

Primary Biliary Cholangitis (PBC) is defined as a chronic cholestatic autoimmune liver disease characterized by a destructive, small duct, and lymphocytic cholangitis, and marked by the presence of antimitochondrial antibodies (Trivella *et al*. [2023]).

The incidence and prevalence of PBC vary widely across different regions. Although historically considered disproportionally more common among White non-Hispanic females, contemporary data show a higher prevalence in males and racial minorities than previously described. Clinical outcomes largely depend on early recognition and prompt institution of treatment, which can be influenced by provider bias and age (Trivella *et al*. [2023]).

The progression of the disease is clinically and histologically classified into four stages (Ludwig's classification), which represent the target classes for our predictive models:

- **Stage 1 (Portal Stage):** Characterized by inflammation restricted to the portal triads. The liver structure is largely intact.
- **Stage 2 (Periportal Stage):** The inflammation and fibrosis begin to extend beyond the portal areas into the surrounding liver tissue (periportal fibrosis).
- **Stage 3 (Septal Stage):** Characterized by bridging fibrosis, where scar tissue connects different portal areas, but regenerative nodules are not yet widespread.
- **Stage 4 (Cirrhosis):** The terminal stage, characterized by diffuse fibrosis and the conversion of normal liver architecture into structurally abnormal nodules. (You *et al*. [2023])

1.2 Dataset Description

The dataset employed in this analysis originates from a seminal longitudinal study conducted by the Mayo Clinic between 1974 and 1984 on Primary Biliary Cirrhosis (now Cholangitis). The primary objective of the original trial was to evaluate the efficacy of the drug D-penicillamine in a randomized, placebo-controlled setting.

A total of 424 patients, referred to the Mayo Clinic during the ten-year interval, met the eligibility criteria for the study. The dataset is structured into two distinct cohorts based on their participation in the clinical trial:

- **Randomized Cohort (n=312):** These patients con-

sented to participate in the randomized trial. This subset represents the core of the dataset and contains largely complete data regarding clinical covariates and follow-up.

- **Observational Cohort (n=106):** These patients met the eligibility criteria but did not participate in the randomized trial. However, they consented to have basic measurements recorded and to be monitored for survival analysis.

Consequently, the final dataset available for analysis consists of the 312 randomized participants plus the remaining 106 observational cases, totaling 418 subjects.

Further discussions on the dataset and clinical findings are detailed in Dickson *et al*. [1989] and Markus *et al*. [1989].

The dataset comprises 20 attributes, including the unique identifier, the target variable (Stage), and 18 predictive features covering demographic information, survival data, and clinical measurements.

A detailed description of each attribute, divided by type (Categorical and Numerical), is provided in Table 1.

The variable **Stage** serves as the ground truth for our multi-class classification task, representing the progression of the disease from early inflammation (Stage 1) to cirrhosis (Stage 4).

1.3 Machine Learning Models

To address the multi-class classification problem of PBC staging, we selected three distinct supervised learning algorithms. These models represent different learning paradigms: probabilistic (Logistic Regression), geometric (Support Vector Machines), and ensemble-based (Random Forest). This diversity allows for a comprehensive evaluation of how different mathematical approaches handle the specific challenges of this dataset, such as class imbalance and clinical feature variability.

1.3.1 Logistic Regression (LR)

Despite its name, Logistic Regression is a linear classification algorithm widely used in medical research as a baseline due to its interpretability and computational efficiency.

While standard linear regression predicts a continuous value, Logistic Regression maps the output to a probability score between 0 and 1 using the Sigmoid function (or Softmax for multi-class problems, like in the case of the study). The model assigns the patient to the stage with the highest probability. Logistic Regression is particularly robust when the dataset size is small relative to the number of features, limiting the risk of overfitting compared to more complex models.

1.3.2 Support Vector Machine (SVM)

Support Vector Machine is a powerful supervised learning algorithm effective in high-dimensional spaces. The core principle of SVM is to find the optimal hyperplane that separates the data points of different classes with the maximum possible margin.

Since clinical data is rarely linearly separable, we employed the Kernel Trick. This technique projects the data into a higher-dimensional space where separation becomes possible. Specifically, we utilized the Radial Basis Function (RBF) kernel.

The SVM algorithm is particularly useful in this study because it focuses on the data points closest to the decision boundary (the "support vectors"), making it theoretically capable of defining precise boundaries between disease stages.

1.3.3 Random Forest (Ensemble method)

Random Forest is an ensemble learning method that constructs a multitude of Decision Trees at training time. It operates on the principle of *Bagging* (Bootstrap Aggregating).

Instead of relying on a single decision tree, which is prone to overfitting and high variance, Random Forest aggregates the predictions of many individual trees.

- **Randomness:** Each tree is trained on a random subset of the data and uses a random subset of features for splitting at each node.
- **Voting:** For classification tasks, the final prediction is determined by the majority vote of all the trees in the forest.

This approach is inherently robust to noise and outliers, making it a standard choice for tabular medical data. Furthermore, Random Forest provides intrinsic feature importance measures, allowing us to understand which clinical variables (e.g., Bilirubin, Albumin) are most influential in determining the disease stage.

2 Aim of the Study

The primary objective of this research is to develop and evaluate Machine Learning models capable of performing a multi-class classification of the histological stages of Primary Biliary Cholangitis (PBC). Unlike binary classification tasks (e.g., healthy vs. sick), staging requires the model to discriminate between four distinct levels of disease progression (Stage 1 to Stage 4), a task that is clinically crucial for determining appropriate therapeutic interventions.

Specifically, this study aims to evaluate the performance of three different supervised learning approaches: Random Forest, Support Vector Machines (SVM), and Logistic Regression to identify the most suitable model for this specific clinical domain. Investigating the eventual impact of class imbalance. Beyond standard performance metrics (Accuracy, F1-Score, MCC), this study aims to conduct an "error analysis" on severely misclassified patients (e.g., Stage 4 patients predicted as Stage 1 or viceversa). The goal is to understand if these errors are due to model limitations or to "atypical" clinical profiles that deviate from the standard biochemical presentation of the disease. In addition, the hopes is also helps to determine which features drive the models' decisions, providing insights that could support medical professionals in diagnosis. Ultimately, this work seeks to determine whether classical Machine Learning algorithms are sufficient for accurate PBC staging or if the complexity and limitations of the dataset suggest the need for more advanced techniques, such as Deep Learning approaches.

3 Methods

The methodology employed in this study follows a structured Machine Learning pipeline, divided into three main modules:

Table 1. Description of the dataset attributes, separated by feature type.

Variable	Description
Numerical Features	
N_Days	Number of days between registration and the earlier of death, transplantation, or study analysis time.
Age	Age of the patient in days.
Bilirubin	Serum bilirubin level [mg/dL].
Cholesterol	Serum cholesterol level [mg/dL].
Albumin	Albumin level [gm/dL].
Copper	Urine copper excretion [$\mu g/day$].
Alk_Phos	Alkaline phosphatase level [$U/liter$].
SGOT	Serum glutamic-oxaloacetic transaminase [U/ml].
Triglycerides	Triglycerides level [mg/dL].
Platelets	Platelet count per cubic [$ml/1000$].
Prothrombin	Prothrombin time in seconds [s].
Categorical Features	
ID	Unique patient identifier.
Status	Patient status: *C* (censored), *CL* (censored due to liver tx), or *D* (death).
Drug	Type of drug administered: *D-penicillamine* or *Placebo*.
Sex	Gender: *M* (Male) or *F* (Female).
Ascites	Presence of ascites: *N* (No) or *Y* (Yes).
Hepatomegaly	Presence of hepatomegaly: *N* (No) or *Y* (Yes).
Spiders	Presence of spider angiomas: *N* (No) or *Y* (Yes).
Edema	Presence of edema: *N* (No edema/no diuretics), *S* (Edema present without diuretics/resolved), or *Y* (Edema despite diuretics).
Stage	**Target Variable**: Histologic stage of disease (1, 2, 3, or 4).

Data Preprocessing and Exploration, Feature Selection, and Model Optimization. This section details the procedures applied in the first module.

3.1 Data Preprocessing and Exploratory Data Analysis

Full access to the code: 01_EDA_and_Data_Preprocessing.ipynb

The initial phase of the study focused on ensuring data quality and understanding the underlying biological signals within the dataset. The raw data contained 418 samples with 20 clinical features. The preprocessing workflow addressed missing values, data encoding, feature scaling, and a preliminary analysis of class separability. As a preliminary step, the *ID* was set as the dataframe index. This operation ensures data integrity and prevents the unique identifier from being erroneously interpreted by the model as a predictive numerical feature.

3.1.1 Data Cleaning and Imputation Strategy

The dataset presented a significant number of missing values across several features (e.g., *Cholesterol*, *Copper*, *Tryglicerides*). A blind removal of these rows would have resulted in an unacceptable loss of information given the limited dataset size. Therefore, a hybrid cleaning and imputation strategy was adopted:

1. **Target Integrity:** Samples with missing values in the target variable (*Stage*) were removed immediately, as they cannot be used for supervised learning (6 samples removed).
2. **Noise Reduction:** A threshold was applied to filter out samples with excessive missing features (more than 2 missing values), identifying and removing highly incomplete records that could introduce noise (106 samples removed).
3. **KNN Imputation:** For the remaining missing numerical values, we utilized the *K-Nearest Neighbors (KNN) Imputer* ($k = 5$). This method estimates the missing value based on the geometric distance of similar patients in the dataset, preserving the multivariate relationships better than mean/median imputation.
4. **Mode Imputation:** Residual missing values in categorical variables were filled using the mode.

3.1.2 Encoding and Feature Scaling

To render the dataset suitable for algebraic operations required by Machine Learning algorithms, categorical variables (e.g., *Sex*, *Edema* [see Table 1]) were transformed into numerical values using *Label Encoding*.

Subsequently, an analysis of the feature magnitudes revealed significant disparities; for instance, *Bilirubin* ranges from 0.3 to 28, while *Age* exceeds 25,000 days. To prevent features with larger ranges from dominating the cost function during model training, we applied *Standard Scaling* (z-score normalization) to all numerical predictors. This transformation ensures that all input features have a mean of 0 and a

standard deviation of 1. The target variable *Stage* was excluded from scaling to preserve its ordinal class meaning (1, 2, 3, 4).

3.1.3 Analysis of Class Imbalance

A critical finding during the Exploratory Data Analysis (EDA) was the severe class imbalance present in the dataset. As visualized in Figure 1, the distribution of patients across stages is highly skewed:

- **Stage 1 (Early):** 5.1% (16 patients)
- **Stage 2:** 21.5% (67 patients)
- **Stage 3:** 38.5% (120 patients)
- **Stage 4 (Late):** 34.9% (109 patients)

This over-representation of advanced stages (3 and 4) is likely due to *Selection Bias* in clinical trials, where patients are typically enrolled only after developing noticeable symptoms. This imbalance poses a risk that the models may bias their predictions towards the majority classes, necessitating careful evaluation metrics (Precision, Recall for each Stage and MCC) rather than simple Accuracy.

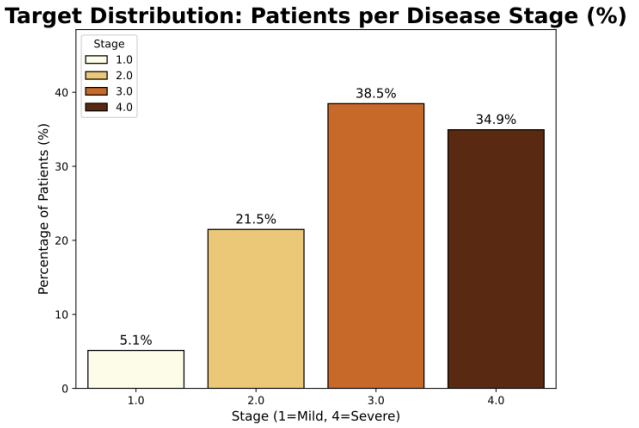

Figure 1. The bar chart illustrates the relative frequency of patients across the four histological stages of PBC. A significant class imbalance is evident, with early-stage disease (Stage 1, 5.1%) being severely underrepresented compared to the other stages.

3.1.4 Feature Correlation and Data Leakage Detection

A Pearson correlation analysis (Figure 2) was conducted to identify key predictors and potential data leakage. The heatmap analysis highlighted two specific variables:

- **N_Days** (correlation $r = -0.37$ with Stage)
- **Status** (correlation $r = 0.34$ with Stage)

Despite their high correlation, these features represent "future outcomes" (survival time and death/transplant event) that are not known at the time of diagnosis. Including them would constitute *Data Leakage*, artificially inflating the model's performance. Consequently, both *N_Days* and *Status* were flagged for removal in the subsequent modeling phase.

3.1.5 Clinical Feature Analysis

Visual inspection of the features via Bar charts and Boxplots (respectively Figure 3 and Figure 4) revealed distinct patterns characterizing the disease progression:

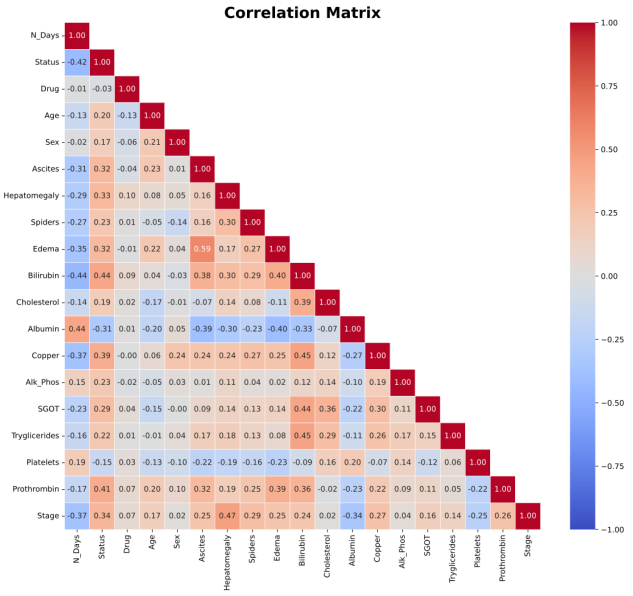

Figure 2. Feature Correlation Heatmap. The plot displays linear dependencies between variables. This analysis is crucial to identify the most likely predictive features for classifying the disease stages and to detect potential multicollinearity.

- **The "Asymptomatic" Stage 1:** Stage 1 patients showed a 0.0% prevalence for all physical markers (*Ascites, Hepatomegaly, Spiders, Edema*). This confirms that early-stage PBC is physically asymptomatic in this cohort, making it difficult to distinguish from healthy individuals based solely on these markers.
- **Stage 4 Specificity:** *Ascites* and severe *Edema* emerged as high-specificity markers for the terminal stage. While not all Stage 4 patients present these symptoms, their presence is almost exclusively linked to Stage 4.
- **Linear Progression Markers:** *Bilirubin* and *Copper*, demonstrated a "staircase effect," with median values or prevalence rising monotonically from Stage 1 to Stage 4. These are expected to be the most discriminative features for the classifiers.
- **Intermediate Overlap:** Features like *Albumin* showed significant overlap between Stage 2 and Stage 3, suggesting that the classifiers might struggle to separate these intermediate "grey area" classes compared to the distinct Stage 1 and Stage 4 profiles that should be distinguished very well.

3.2 Feature Selection and Model Selection

Full access to the code: 02_Feature_Selection_and_Modeling.ipynb

Following data preprocessing, the focus shifted to identifying the most predictive clinical features and establishing a robust validation framework. This phase addresses the challenge of feature redundancy and ensures the selection of robust biomarkers despite the limited sample size, through a rigorous *Stability Selection* protocol.

3.2.1 Validation Strategy and Benchmark Creation

Given the significant class imbalance observed in the target variable (Stage 1 represents only $\approx 5\%$ of the data), a standard random split would likely result in validation sets lacking minority class samples. To mitigate this, we adopted a

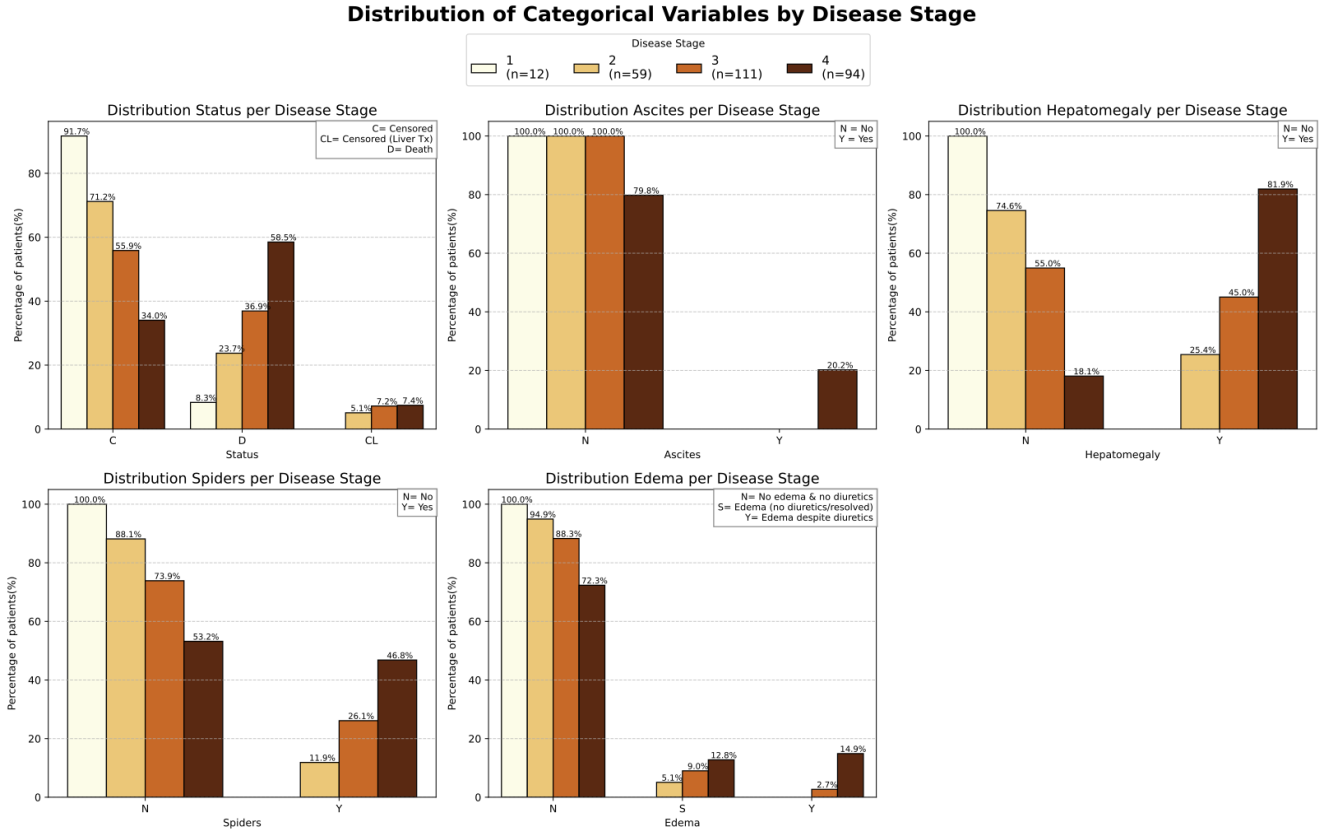

Figure 3. Distribution of categorical features per stage. Bar charts displaying the prevalence of symptoms across disease stages. While Stage 1 is largely asymptomatic, markers like *Ascites* and severe *Edema* appear almost exclusively in Stage 4, acting as strong discriminators for late-stage disease.

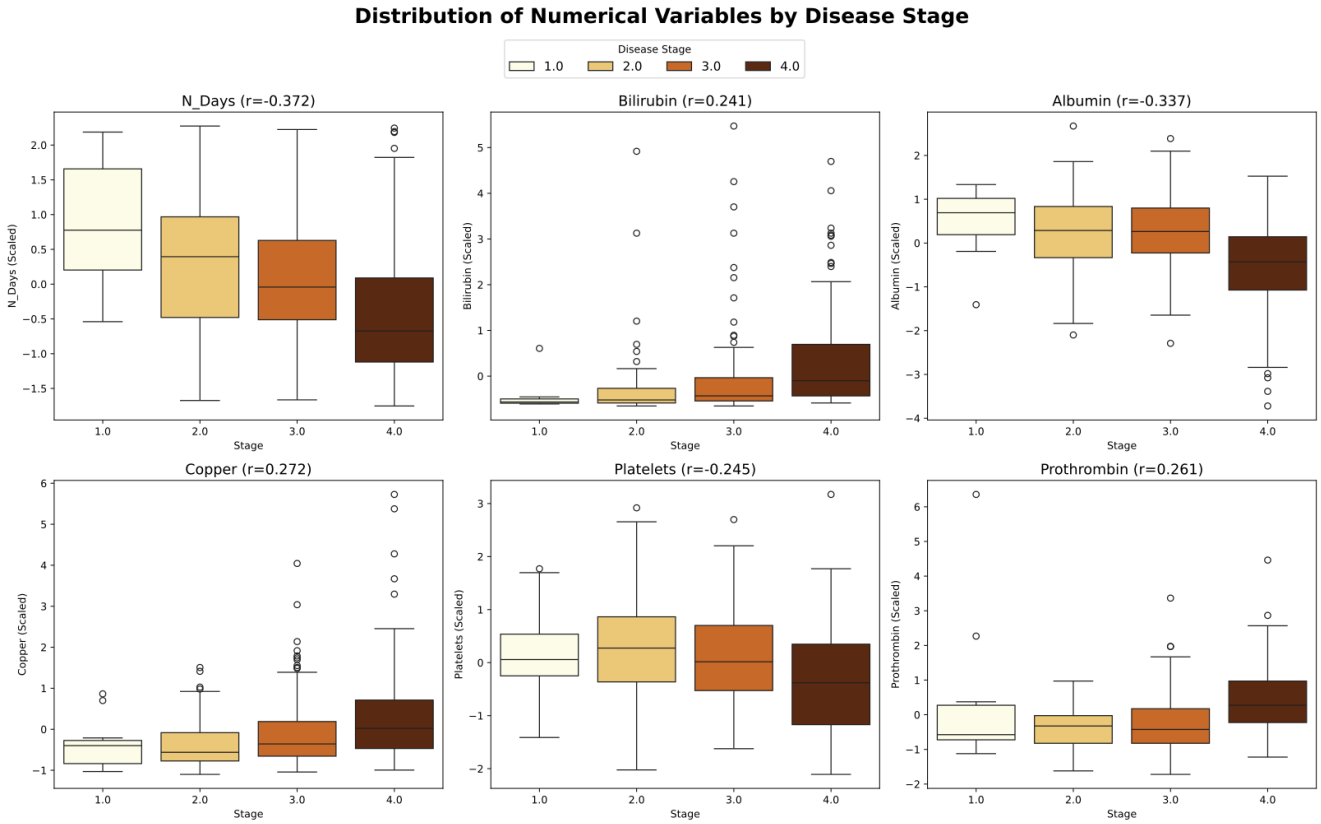

Figure 4. Distribution of numerical features per stage. Boxplots displaying the standardized distribution of the most significant numerical features. A clear "staircase effect" is visible for *Bilirubin* and *Copper*, where median values progressively increase with disease severity (positive correlation). Conversely, *Albumin* and *Platelets* act as protective factors, showing an inverse trend. While the extreme stages (1 and 4) are distinct, significant overlap exists between the intermediate stages (2 and 3), indicating a potential challenge for classification.

Stratified K-Fold Cross-Validation (SCV) strategy ($k = 5$) (Szeghalmy and Fazekas [2023]).

The dataset was partitioned into a *Development Set (80%)* for training and validation, and a held-out *Benchmark Set (20%)* reserved exclusively for final testing. The stratification ensures that the distribution of disease stages remains consistent across all five folds and the independent benchmark set, preventing the model from developing biases due to underrepresented classes during the validation phase.

3.2.2 Preliminary Architecture Selection: Standard RF vs Balanced RF

Before proceeding with feature ranking, it was necessary to identify the most reliable estimator. Standard tree-based algorithms (like Random Forest) typically optimize Gini Impurity based on global accuracy, which often leads to biasing predictions toward the majority class (Stage 3 and 4 in our case) to minimize error.

To address this, we performed a comparative analysis between a Standard Random Forest (RF) and a Balanced Random Forest (BRF). The BRF employs an internal downsampling technique, bootstrapping a balanced subset of data for each tree in the ensemble (Lemaître *et al.* [2017]). We evaluated both architectures across a large range of estimators ($n \in [100, 2000]$) using the Matthews Correlation Coefficient (MCC) as the primary metric (more robust for unbalanced sets than Accuracy).

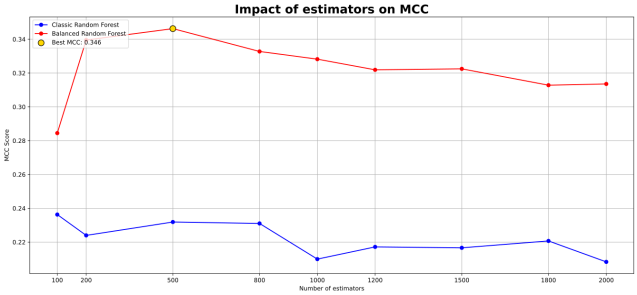

Figure 5. BRF vs RF performance across different number of estimators. BRF (red line) consistently outperforms the RF (blue line) across all estimator configurations. The balanced approach achieves an higher MCC, confirming its superior ability to handle the class imbalance.

As illustrated in Figure 5, BRF demonstrated superior stability and performance. Consequently, this architecture (specifically with $n = 500$ estimators to minimize variance) was selected as the engine for the subsequent feature ranking process.

3.2.3 Stability Feature Selection Protocol

To select a feature subset that is robust to data perturbations, we implemented a *Stability Selection* framework. This protocol avoids relying on a single feature ranking, which can be noisy. The procedure operates as follows for each of the three models (Tree-based, Kernel-based, Linear-based):

1. **Ranking:** For each fold of the SCV, the estimator is trained and tested on the validation test to find the best hyperparameters. After that, the BRF ranks features based on their importance (Gini Importance).
2. **Subset Optimization** To ensure modularity and reproducibility, the optimization logic was encapsulated in a dedicated function (source code available at `stability_feature_sel.py`). We iteratively use the estimator to test subsets of the top-k features. The optimal number of features k that maximize the MCC score on the testing set of the respective fold is chosen.
3. **Voting Mechanism:** Features included in the optimal subset receive a "vote".
4. **Consensus:** Only features selected in at least 4 out of 5 folds (Frequency $\geq$ 80%) are retained for the final model.

3.2.4 Selection Results and Feature Consensus

We applied this stability protocol using three distinct estimators to capture different types of relationships: *Balanced Random Forest* (for orthogonal splits), *SVM with RBF Kernel* (for smooth, non-linear boundaries), and *Logistic Regression* (for linear relationships).

Remarkably, the selection process converged to a unanimous consensus. All three methods identified the exact same set of 11 predictive features, discarding noise and less informative variables (such as Sex and Ascites, which were filtered out due to low contribution).

The final selected feature set consists of: *Bilirubin, Copper, Albumin, Cholesterol, Alk_Phos, SGOT, Tryglicerides, Prothrombin, Platelets, Hepatomegaly, Age*. These 11 features serve as the definitive input for the hyperparameter optimization phase described in the next chapter.

3.3 Hyperparameter Optimization and Benchmark Testing

Full access to code: 03_Optimization_and_Benchmark_Testing.ipynb

This part focused on the fine-tuning of the three candidate classifiers. The objective was to maximize the MCC on the Development Set using SCV, ensuring robustness against class imbalance before the final evaluation on the unseen Benchmark Set.

3.3.1 Bayesian Optimization Strategy

Instead of a traditional Grid Search, which exhaustively evaluates a pre-defined set of hyperparameters, we employed a more robust Bayesian Optimization (via `BayesSearchCV`, 60 iterations). This method models the objective function (MCC) as a Gaussian Process, allowing the algorithm to intelligently explore the hyperparameter space by balancing exploration (apply directly the SCV) and exploitation (refining promising configurations).

This strategy was applied to three distinct pipelines:

Balanced Random Forest Optimization For the tree-based model, we optimized the ensemble structure to prevent overfitting while maintaining high sensitivity for minority classes. The search space included:

- `n_estimators` $\in [100, 1000]$
- `max_depth` $\in [10, 100]$
- `min_samples_leaf` $\in [1, 10]$

The optimal configuration found was a forest of 1000 trees with a maximum depth of 35 and a minimum of 2 samples per leaf.

SVM (RBF Kernel) Optimization For the Support Vector Machine, the optimization focused on the regularization parameter C and the kernel coefficient γ, searching for the ideal non-linear decision boundary in the high-dimensional feature space.

- `C` (Regularization): Log-uniform search $[0.01, 1000]$
- `gamma` (Kernel Coeff.): Log-uniform search $[1e-5, 100]$

The bayesian search converged to $C \approx 3.14$ and $\gamma \approx 0.046$.

Logistic Regression Optimization For the linear baseline, we tuned the inverse regularization strength C to control the penalty on the coefficients.

- `C`: Log-uniform search $[0.01, 1000]$

The optimal value was found at $C \approx 46.79$, indicating a preference for weaker regularization (allowing the coefficients to fit the training data more closely). The final predictions for all three optimized models were serialized and stored. A comprehensive comparative analysis of these results, including confusion matrices and clinical interpretation of the classification performance, is presented in the following Section 4.

4 Analysis of Results and Discussion

Full access to the code: 04_Analysis_Results.ipynb

This section presents the performance evaluation of the three optimized models on the *Benchmark Set*. The analysis focuses not only on global metrics but also on the specific error patterns (Confusion Matrices), sensitivity analysis (FNR), and the investigation of severe misclassification cases to understand the clinical limitations of the proposed approaches.

4.1 Classification Performance

Table 2 summarizes the global performance metrics. The *Logistic Regression* emerged as the most robust model, achieving the highest MCC (0.312) and Weighted F1-Score (0.47). While the overall accuracy hovers around 48%, this is pretty decent with the difficulty of the 4-class multiclass problem on a small, imbalanced dataset. However SVM remains better to classify mid-stage disease.

4.1.1 Confusion Matrix Analysis

The confusion matrices (Figure 6) reveal distinct behavioral patterns:

- **Balanced Random Forest:** Effectively identifies Stage 4 but struggles significantly with early stages, often confusing Stage 1 and 2.
- **SVM:** Shows a "hybrid" behavior, performing decently on intermediate stages but failing to construct valid support vectors for the minority class (Stage 1).
- **Logistic Regression:** It is the only model capable of detecting Stage 1 patients (2 out of 3). However, it exhibits a strong linear bias: it frequently misclassifies Stage 3 patients as Stage 2, likely due to the proximity of their feature distributions in the linear space.

Model	Stage	Support	Precision	Recall	F1-Score	MCC
Balanced Random Forest	1	3	0.12	0.33	0.18	0.269
	2	14	0.31	0.36	0.33	
	3	24	**0.53**	**0.38**	**0.44**	
	4	22	0.68	**0.68**	0.68	
	W. Avg	*63*	*0.52*	*0.48*	*0.49*	
SVM (RBF)	1	3	0.11	0.33	0.17	0.302
	2	14	**0.41**	0.50	0.45	
	3	24	**0.53**	0.33	0.41	
	4	22	0.68	**0.68**	0.68	
	W. Avg	*63*	***0.54***	***0.49***	***0.50***	
Logistic Regression	1	3	**0.17**	**0.67**	**0.27**	**0.312**
	2	14	0.40	**0.57**	**0.47**	
	3	24	0.50	0.21	0.29	
	4	22	**0.71**	**0.68**	**0.70**	
	W. Avg	*63*	***0.54***	*0.48*	*0.47*	

Table 2. Model-wise Performance Summary. Red bold values indicate the best result for each metric. The table highlights a critical trade-off for the Logistic Regression: in Stage 1, it prioritizes Recall (0.67) over Precision (0.17). Clinically, this is favorable as it minimizes missed diagnoses (False Negatives) for early-stage patients. Conversely, the SVM achieves slightly higher weighted averages due to mid-stage consistency but lacks sensitivity for Stage 1. Finally, the BRF struggles with both early and intermediate stages, resulting in the lowest overall performance.

4.2 Error Rate Analysis

To further dissect the models' reliability, we analyzed the False Negative Rate (FNR) across disease stages (Figure 7).

A crucial clinical finding is that Stage 4 (Cirrhosis) consistently showed the lowest FNR ($\approx 31.8\%$) across all three models. This confirms that the biological signal for decompensated cirrhosis is distinct and recognizable regardless of the mathematical approach used. Conversely, the high FNR for Stage 3 in the Logistic Regression model (79.2%) highlights the trade-off of the linear approach: while it captures the extremes (Stage 1 and 4), it loses resolution on the intermediate decision boundaries, where SVM shows a more consistent performance.

4.3 Investigation of Severe Misclassification

We identified a critical outlier case: *Patient ID 275*. This subject was clinically diagnosed with Stage 4 (Severe) but was predicted as Stage 1 (Mild) by the classifiers. This represents the most dangerous type of error in a clinical setting ("Severe Misclassification").

To understand the root cause, we compared the patient's biomarker profile against the population distributions (Figure 8). The analysis revealed that Patient 275 presented an atypical biochemical profile for a cirrhotic patient:

- **Albumin:** Unexpectedly high (normal range), whereas typical Stage 4 patients show hypoalbuminemia (Laschtowitz *et al.* [2020]).
- **Copper & SGOT:** Notably lower than the Stage 4 median.
- **Hepatomegaly:** The patient belongs to the minority subgroup ($\approx 18\%$) of Stage 4 patients who do not exhibit Hepatomegaly (Figure 9).

This suggests that the models did not failed randomly in this severe missclassification; rather, they correctly interpreted the input features, which in this specific patient mimicked a healthy profile. This underscores the heterogeneity

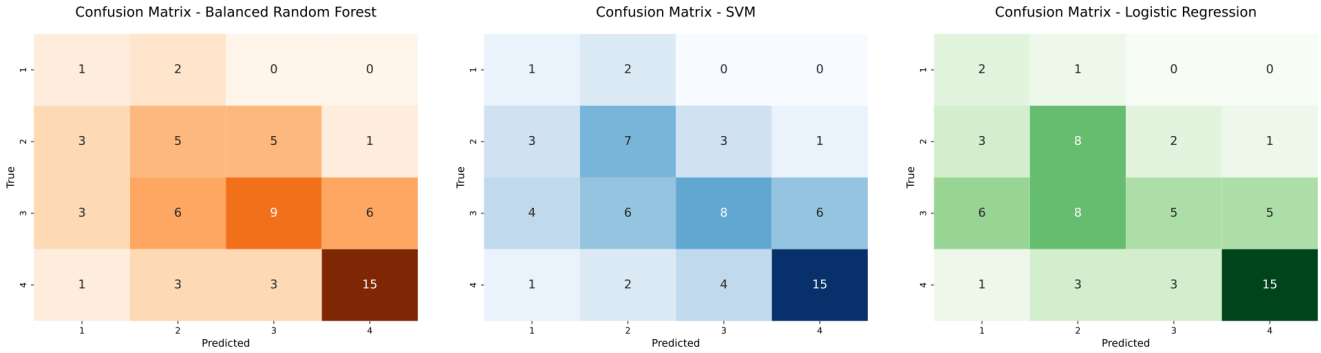

Figure 6. Confusion Matrices. From left to right: Balanced Random Forest, SVM, and Logistic Regression. The linear model (right) shows the best diagonal consistency for the extreme stages (1 and 4), while SVM is the most consistent for mid-stages.

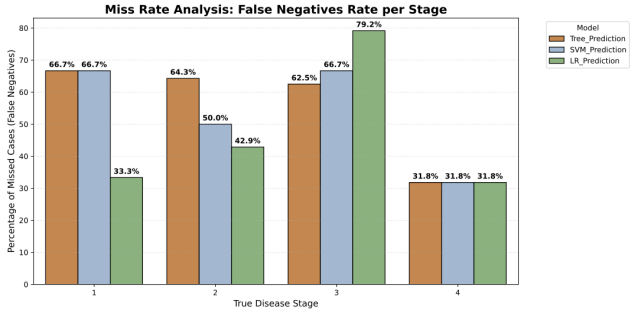

Figure 7. Miss Rate Analysis. False Negative Rate per stage. Note the spike in error for Stage 3 in the LR Model (Green).

of PBC synthoms and the limitations of using only these 11 markers for atypical cases.

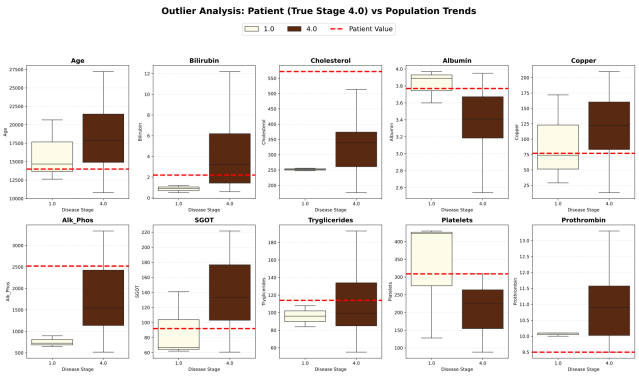

Figure 8. Outlier Profiling (Patient 275). The red dashed line indicates the patient's values compared to the Stage 4-1 population distribution.

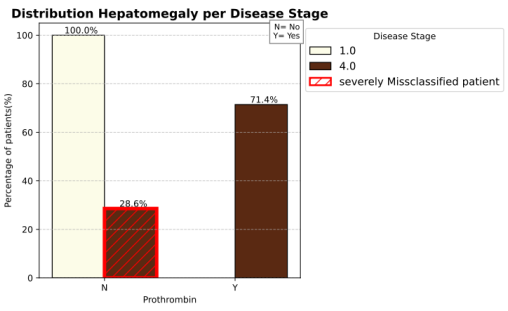

Figure 9. Categorical Mismatch (Patient 275). The misclassified Stage 4 patient (red hatch) falls into the "No Hepatomegaly" category, which is more typical of early-stage disease.

5 Conclusion

In conclusion, while the overall accuracy is limited by the small sample size, the significant feature overlap between intermediate stages (2 and 3), and the limited discriminatory power of the available biomarkers, the Logistic Regression and SVM models proved to be the most capable of generalizing across the disease spectrum.

Specifically, Logistic Regression excelled in recognizing the extreme stages of the disease (1 and 4). However, its high sensitivity for Stage 1 must be interpreted with caution due to the extremely limited sample size of this subgroup. Conversely, the SVM demonstrated greater consistency in classifying mid-stage and Stage 4 patients, whereas it struggled to detect early-stage cases. Notably, all models successfully minimized severe misclassification errors (confusing Stage 1 with Stage 4), with the exception of the single atypical outlier identified in the analysis.

However, the general inability to reliably distinguish intermediate stages (2 vs 3) and the presence of atypical outliers (like ID 275) indicate that classical ML approaches may have reached a performance ceiling on this dataset. Future iterations could benefit from Deep Learning techniques (such as Representation Learning) or the inclusion of additional histological features to capture non-linear nuances that current biomarkers fail to represent.

References

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10. DOI: 10.1002/hep.1840100102.

Laschtowitz, A., de Veer, R. C., der Meer, J. A. V., and Schramm, C. (2020). Diagnosis and treatment of primary biliary cholangitis.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18.

Markus, B. H., Dickson, E. R., Grambsch, P. M., Fleming, T. R., Mazzaferro, V., Klintmalm, G. B. G., Wiesner, R. H., Thiel, D. H. V., and Starzl, T. E. (1989). Efficacy of liver transplantation in patients with primary biliary cirrhosis. *New England Journal of Medicine*, 320. DOI: 10.1056/nejm198906293202602.

Szeghalmy, S. and Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. *Sensors*, 23. DOI: 10.3390/s23042333.

Trivella, J., John, B. V., and Levy, C. (2023). Primary biliary cholangitis: Epidemiology, prognosis, and treatment.

You, H., Duan, W., Li, S., Lv, T., Chen, S., Lu, L., Ma, X., Han, Y., Nan, Y., Xu, X., Duan, Z., Wei, L., Jia, J., and Zhuang, H. (2023). Guidelines on the diagnosis and management of primary biliary cholangitis (2021). *Journal of Clinical and Translational Hepatology*, 11. DOI: 10.14218/JCTH.2022.00347.