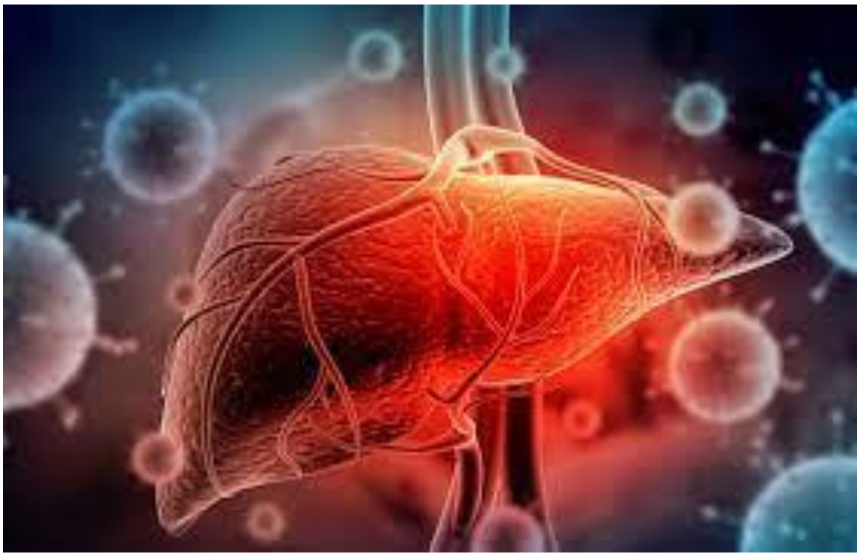
Ben-Gurion University of the Negev

Department of Industrial Engineering and Management

Liver Cirrhosis Stage Prediction



Final Project – Machine Learning and Data Mining

(364-2-5091)

Institute: Ben Gurion University of the Negev

Lecture: Prof. Boaz Lerner

Students: Sivan Raviv, Avital Finanser, Yarden Choen

Date: 26.02.25

Code available at: https://github.com/Yarden231/Liver-Cirrhosis-Stage-Prediction

Abstract

This research project investigates the application of **machine learning techniques to predict cirrhosis disease stages** using a comprehensive dataset from the Mayo Clinic's primary biliary cirrhosis (PBC) study conducted from 1974 to 1984. The dataset comprises clinical parameters and biomarkers from 418 patients, with 20 variables including both categorical and numerical data points.

Our methodology encompasses extensive data preprocessing, including handling missing values, feature transformation, and addressing class imbalance. We implemented and compared multiple machine learning algorithms, including Random Forest, XGBoost, CatBoost, LightGBM, Artificial Neural Networks, and ensemble voting methods.

Results demonstrate that **Random Forest outperformed other models with 72.3% accuracy and an F1-score of 0.722,** exhibiting strength in identifying advanced cirrhosis stages. Feature importance analysis revealed that blood lipid indicators (triglycerides and cholesterol) and copper concentration were the strongest predictors, surpassing conventional liver function markers such as bilirubin and SGOT.

This project confirms **the potential of machine learning in enhancing cirrhosis diagnosis by improving early detection capabilities, identifying key risk factors, and supporting personalized treatment decisions.** Altogether, we acknowledge certain limitation impacting model performance as the relatively small dataset size constrained our ability to achieve higher accuracy rates, particularly for early-stage cirrhosis detection.

Contents

1. Business Understanding

The liver, the body's largest organ, is crucial for detoxification, metabolism, and protein synthesis, with cirrhosis ranking as the 12th leading cause of liver-related mortality in the United States. Cirrhosis, characterized by progressive replacement of healthy liver tissue with fibrotic scarring, compromises organ function and leads to complications including portal hypertension, ascites, jaundice, and hepatic encephalopathy. **Primary causes** include chronic alcohol use, viral hepatitis (B and C), and increasingly, non-alcoholic fatty liver disease (NAFLD). **Disease progression** typically follows an insidious course from compensated (asymptomatic despite scarring) to decompensated stages (marked by clinical complications and hepatic failure). **Diagnostic challenges** arise due to cirrhosis's asymptomatic early stages, though machine learning (ML) has emerged as a promising tool, demonstrating superior diagnostic accuracy and enabling disease trajectory prediction. However, clinical implementation faces challenges due to patient heterogeneity and complex pathophysiology, emphasizing the need for more sophisticated models. **This project's vision** employs machine learning to predict cirrhosis and accurately identify its four distinct stages by analyzing key parameters, addressing a critical healthcare need and aiming to enhance patient outcomes. (Asrani et al., 2019; Sumeet et al., 2013; D'Amico et al., 2006; Singal et al., 2020; Smith et al., 2019; Zhai et al., 2024; Zhang et al., 2021)

2. Data Understanding

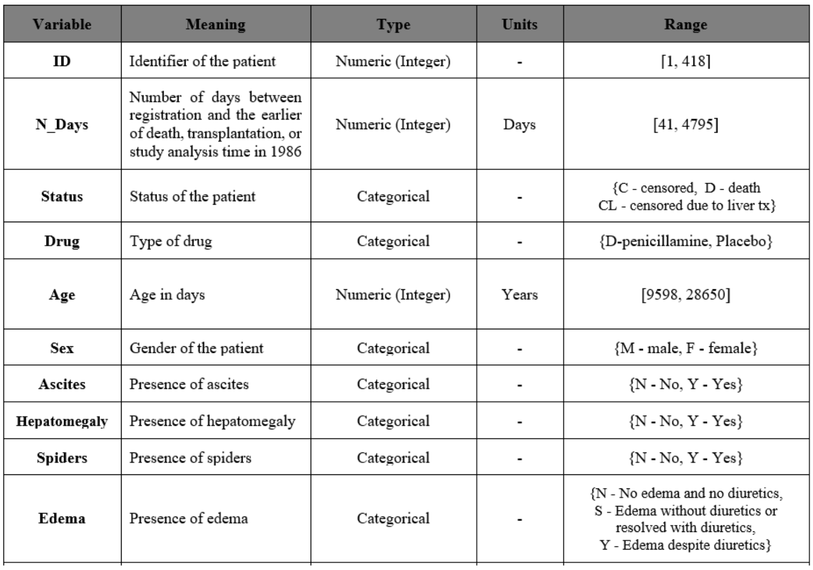
2.1 Data

Data source: Our study is based on the Cirrhosis Prediction Dataset collected between 1974 and 1984 during the Mayo Clinic study on primary biliary Cirrhosis (PBC). The data is available from the UC Irvine Machine Learning Repository (Link).

Characteristics: The dataset contains 418 records, each representing a patient's biological information. It consists of 19 explanatory variables capturing various biological indicators and one target variable indicating disease stage. This structure enables analysis of the relationship between biological markers and disease progression.

2.2 Variables

Table 1 measurement units and interval of each feature in the dataset

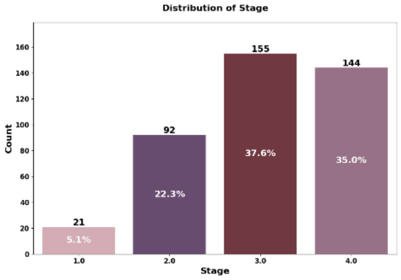


Variable	Meaning	Type	Units	Range
ID	Identifier of the patient	Numeric (Integer)	-	[1, 418]
N_Days	Number of days between registration and the earlier of death, transplantation, or study analysis time in 1986	Numeric (Integer)	Days	[41, 4795]
Status	Status of the patient	Categorical	-	{C - censored, D - death CL - censored due to liver tx}
Drug	Type of drug	Categorical	-	{D-penicillamine, Placebo}
Age	Age in days	Numeric (Integer)	Years	[9598, 28650]
Sex	Gender of the patient	Categorical	-	{M - male, F - female}
Ascites	Presence of ascites	Categorical	-	{N - No, Y - Yes}
Hepatomegaly	Presence of hepatomegaly	Categorical	-	{N - No, Y - Yes}
Spiders	Presence of spiders	Categorical	-	{N - No, Y - Yes}
Edema	Presence of edema	Categorical	-	{N - No edema and no diuretics, S - Edema without diuretics or resolved with diuretics, Y - Edema despite diuretics}

Variable	Meaning	Type	Units	Range
Bilirubin	Serum bilirubin in	Numeric (Continuous)	[mg/dl]	[0.3, 28]
Cholesterol	Serum cholesterol in	Numeric (Integer)	[mg/dl]	[120, 1775]
Albumin	Albumin in	Numeric (Continuous)	[mg/dl]	[1.96, 4.64]
Copper	Urine copper in	Numeric (Integer)	[ug/day]	[4, 588]
Alk_Phos	Alkaline phosphatase in	Numeric (Continuous)	[U/liter]	[289, 13862.4]
SGOT	SGOT in	Numeric (Continuous)	[U/ml]	[26.35, 475.25]
Triglycerides	Triglycerides in	Numeric (Integer)	[mg/dl]	[33, 598]
Platelets	Platelets per cubic	Numeric (Integer)	[ml/1000]	[9, 18]
Prothrombin	Prothrombin time in seconds	Numeric (Continuous)	[sec]	[62, 721]
Stage (Target)	Histological stage of disease	Categorical	-	{1, 2, 3, or 4}

Distributions of Continuous Features (Appendix A)

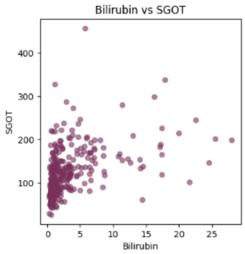
Statistical description of the categorical variables (Appendix B)



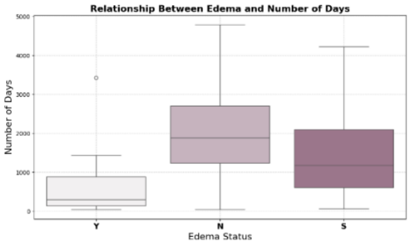
The **target variable, Stage**, is a categorical variable representing the histological stage of liver cirrhosis (values 1-4). Our objective is to train a predictive model to classify this variable accurately in the test set. The dataset is imbalanced, with uneven distribution of observations across the four classes.

Relationships between variables

Pearson correlation - **Bilirubin** showed strong positive correlations with **Copper (0.46)**, **Cholesterol (0.40)**, **Triglycerides (0.44)**, and **SGOT (0.44)**, indicating links between bilirubin, liver dysfunction, and lipid metabolism. Conversely, **Bilirubin** correlates negatively with **Albumin (-0.31)**, reflecting decreased protein synthesis in liver dysfunction. **Copper** correlates negatively with **N_Days (-0.36)**, suggesting elevated copper is associated with shorter survival times. **Albumin** correlates positively with **N_Days (0.43)**, confirming its value as a prognostic marker for longer survival (Appendix C).



Selected scatter plot between variables (highest correlation - above |0.4|) **Bilirubin vs SGOT:** Higher bilirubin levels correspond with increased SGOT values, highlighting both as important markers of liver injury - bilirubin indicating impaired waste clearance while SGOT reflecting hepatocyte damage, with their correlation demonstrating the liver dysfunction characteristic of cirrhosis (Appendix D).



An example of the relationship between variables: **Edema and N_Days -** Patients with Edema (Y) showed fewer days, suggesting more severe disease or shorter survival. Those without Edema (N) or with diuretic-responsive Edema (S) exhibited wider day ranges and higher median values, indicating patients without Edema or responding to treatment survived longer.

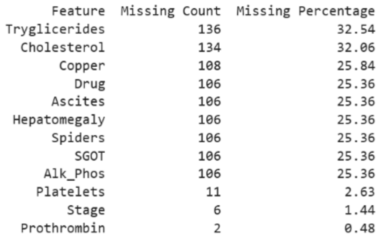
Target with features - Disease progression correlated with increased liver dysfunction markers (Bilirubin, Copper, Alkaline Phosphatase) and deterioration of health indicators (Albumin, Platelets, Prothrombin). Advanced stages (3-4) showed pronounced changes in these markers compared to early stages. While some variables (Triglycerides) showed subtle changes across stages, others (N_Days, Bilirubin) displayed stark contrasts, emphasizing their diagnostic value. The correlation between advanced stages and older age underscored the importance of early diagnosis (Appendix E).

3. Data Preparation

Irrelevant feature removal - Removed the ID column, as it was not analytically relevant.

Duplicate records - We have defined a duplicate record as a record whose fields are identical to another. Accordingly, no duplicates were found.

Missing values - Missing values were noted in several features. We grouped these features based on the proportion of missing data: **Significant Missing Data (≥25%)**: Features like Tryglicerides, Cholesterol, Copper, Drug, Ascites, Hepatomegaly, Spiders, SGOT, and Alk_Phos had over 25% of their values missing.



```
      Feature  Missing Count  Missing Percentage
Tryglicerides            136               32.54
  Cholesterol            134               32.06
       Copper            108               25.84
         Drug            106               25.36
      Ascites            106               25.36
 Hepatomegaly            106               25.36
      Spiders            106               25.36
         SGOT            106               25.36
     Alk_Phos            106               25.36
    Platelets             11                2.63
        Stage              6                1.44
  Prothrombin              2                0.48
```

Drug: Imputed as "non-participant" for 106 patients not in the clinical trial (according to the dataset source) to reflect this distinction.

Other features: Imputed using median values of respective target groups.

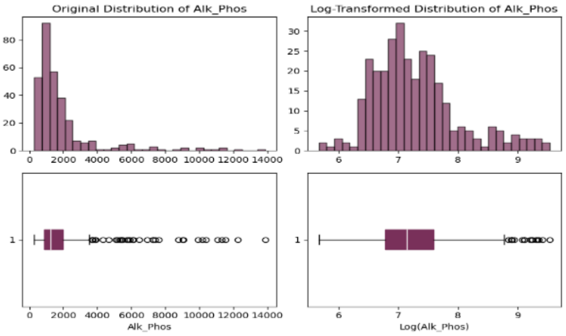
Minimal Missing Data (<3%): Platelets, Stage, and Prothrombin.

Stage (target variable) - 6 samples with missing values removed to avoid introducing noise. Platelets and Prothrombin - Imputed using median values of respective target groups. Note: using the median as imputing is robust to outliers and better represents the central tendency for skewed data.

Data proportion - During EDA, we observed an imbalance in the Cirrhosis stages distribution, with Stages 1 and 2 having significantly fewer samples. This imbalance could lead to biased outcomes and unstable predictions. To ensure more reliable analysis and model training robustness, we merged Stages 1 and 2 into a single category, we merged Stages 1 and 2 into a single category.

Dataset size - Given the dataset's limited size (412), we avoid resorting to down sampling, which could result in data loss, or up sampling, which may introduce bias.

Handling Outliers - Variables such as Bilirubin, Alkaline Phosphatase (Alk_Phos), and SGOT exhibited extreme values that significantly deviated from the primary distribution, distorting central tendency and dispersion measures. We retained these values as they may reflect essential medical conditions for classification. To mitigate their effects, we employed logarithmic transformation to reduce the statistical influence of extreme values while preserving their integrity and relationships within the data.



Data Representation - Categorical Features Encoding (Appendix G)

One-hot encoding - multi-class categorical features (Status, Drug, Edema) were One-Hot encoded to create binary columns for each category (dummy variables). This ensures the model treats each category as distinct without implying ordinal relationships.

Label Encoding - Binary categorical variables labels were encoded into 0 and 1. The binary features we encoded including all dummy variables produced by one-hot encoding and originally binary features (Sex, Ascites, Hepatomegaly, Spiders).

Feature Selection - Multicollinearity Detection via VIF - We employed Variance Inflation Factor analysis to identify multicollinearity among features. Using the standard threshold of VIF≥10, we identified that Prothrombin, Albumin, and Age exhibited high multicollinearity, which corresponded with patterns observed in our correlation matrix. Consequently, these features were removed from the dataset.

Scaling - Both Normalization and Min-Max - The implementation of feature scaling techniques was necessitated primarily by the inclusion of Artificial Neural Networks in our modeling approach. **Normalization (Standardization)**: Transformed features to have a mean of 0 and standard deviation of 1. **Min-Max Scaling**: Rescaled features to the 0-1 range while preserving relative data relationships.

$$X' = \frac{X - \mu}{\sigma}$$

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

4. Modeling

We conducted systematic hyperparameter tuning for each model using **Optuna**, an **efficient optimization framework that dynamically explores parameter space using a Bayesian-inspired approach and prunes suboptimal trials early, conserving computational resources**. We selected Optuna to maximize F1 scores in our multi-class classification task due to its effectiveness with imbalanced data and its ability to deliver optimal results in a practical timeframe.

Random Forest - The Random Forest algorithm is an ensemble learning method that reduces overfitting by incorporating random sampling and variable selection while constructing individual decision trees. Aggregating predictions from multiple trees improves accuracy and generalization, making it effective for our classification task. We tuned parameters including Criterion, N estimators, Max depth, Max features and Minimum samples split.

	Criterion	N estimators	Max depth	Max features	Minimum samples split
Default RF	gini	100	None	sqrt	2
Tuned RF	**gini**	**464**	**31**	**sqrt**	**10**

ANN - A computational model inspired by biological neural networks consisting of interconnected input, hidden, and output layers. Neurons learn by adjusting weights during training to make predictions. While theoretically suitable for capturing complex patterns in our cirrhosis classification task, we anticipated limited performance due to our small dataset, as ANNs typically require substantial training data. We tuned hidden layers size, activation, solver, alpha, and learning rate parameters.

	Hidden Layers	Activation	Solver	Alpha	Learning Rate	Max iterations
Default ANN	(64,)	ReLU	adam	0.0001	constant	200
Tuned ANN	**(128, 64)**	**Logistic**	**adam**	**0.0014**	**adaptive**	**249**

XGBoost - Extreme Gradient Boosting is a tree-based algorithm for classification that handles complex datasets effectively. XGBoost builds sequential decision trees, each correcting previous errors. We tuned Number of Estimators, Maximum Depth, Learning Rate, Subsample, and Column Subsampling parameters.

	Number of Estimators	Maximum Depth	Learning Rate	Subsample	Column Subsampling
Default XGBoost	100	6	0.1	1.0	1.0
Tuned XGBoost	**66**	**10**	**0.007**	**0.849**	**0.658**

Light GBM - LightGBM employs gradient boosting with two key techniques: gradient-based one-sided sampling (GOSS) for training each tree using a fraction of data, and exclusive feature Bundle (EFB) for efficient handling of high-dimensional sparse features. We tuned Boosting Type, Learning Rate, Max Depth, Subsample, and Min Gain to Split parameters.

	Boosting Type	Learning Rate	Max depth	Subsample	Min child samples
Default Light GBM	gbdt	0.1	-1	1	20
Tuned Light GBM	**goss**	**0.064**	**2**	**0.915**	**29**

CatBoost - CatBoost is built on the gradient boosting framework, combining multiple weak learners to create a predictive model. It implements this using decision trees with two key innovations: ordered boosting and efficient handling of categorical features. The parameters we tuned are: Bootstrap type, Min data in leaf, Grow policy, Random strength and Learning rate.

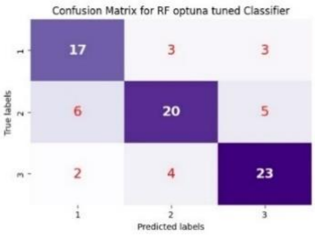
	l2 leaf regularization	**Bootstrap type**	**Min data in leaf**	**Grow policy**	**Random strength**	**Learning rate**
Default CatBoost	3.0	---	1	Symmetric	1	0.1
Tuned CatBoost	**9.349**	**Bayesian**	**10**	**Depthwise**	**7.812**	**0.044**

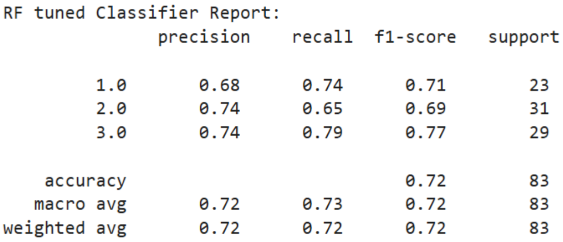
Voting - Voting Classifier is an ensemble method that combines predictions from multiple models to potentially achieve higher performance than any individual classifier. We implemented this approach after tuning all individual models to leverage their optimized configurations. Our voting implementation incorporates: **Voting Type -** Set to "soft" to utilize probability predictions, **Weights -** Assigned proportionally based on each model's performance, **Estimators -** Selected best performing optimized classifiers as base models.

5. Evaluation

We conducted a systematic evaluation of our models using various quality metrics to identify the most suitable one for this problem - producing the best results on the test set. Our approach analyzed accuracy, precision, recall, and F1 scores for a thorough understanding of each model's classification capabilities. We also examined confusion matrices to interpret results across different classes and evaluated confidence intervals (CI) to assess the statistical reliability of our findings.

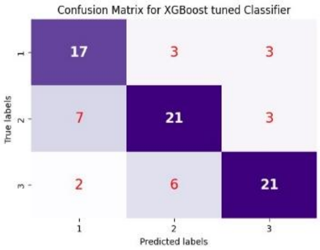
Random Forest - The tuned Random Forest model achieves 72% accuracy, demonstrating strongest performance for advanced cirrhosis (stage 3) with 0.79 recall and 0.77 F1-score. The model maintains effective detection of early-stage disease while showing slightly reduced sensitivity for intermediate stages, creating an improved balance between early detection and advanced case identification.

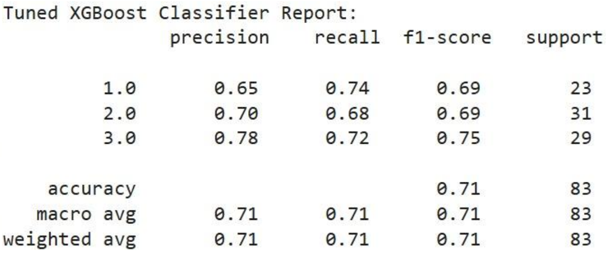




```
RF tuned Classifier Report:
              precision    recall  f1-score   support

         1.0       0.68      0.74      0.71        23
         2.0       0.74      0.65      0.69        31
         3.0       0.74      0.79      0.77        29

    accuracy                           0.72        83
   macro avg       0.72      0.73      0.72        83
weighted avg       0.72      0.72      0.72        83
```

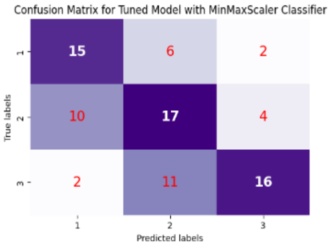
XGBoost - The tuned XGBoost model achieved a 71% accuracy, performing best at identifying advanced cirrhosis cases, while showing improved and balanced effectiveness for early and intermediate stages, demonstrating better overall classification capability than the default model.

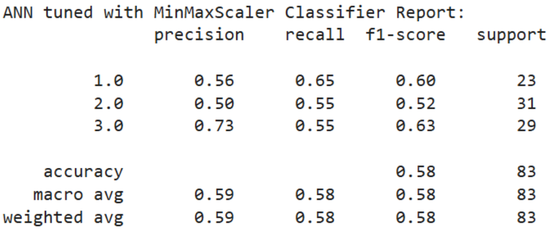




```
Tuned XGBoost Classifier Report:
              precision    recall  f1-score   support

         1.0       0.65      0.74      0.69        23
         2.0       0.70      0.68      0.69        31
         3.0       0.78      0.72      0.75        29

    accuracy                           0.71        83
   macro avg       0.71      0.71      0.71        83
weighted avg       0.71      0.71      0.71        83
```

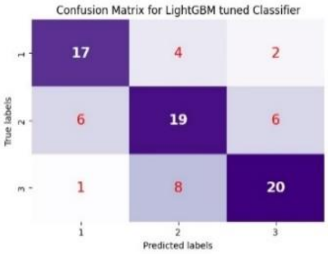
ANN - We selected **Min-Max scaling** due to its uniformity, bias prevention, and improved model performance compared to Standard scaling. We also constructed and evaluated the model by adjusting key hyperparameters to optimize performance. The tuned ANN model with MinMaxScaler achieved 58% accuracy, performing best at identifying advanced cirrhosis cases (precision 0.73) - clinically critical for treatment decisions - while showing weaker performance for early and intermediate stages.

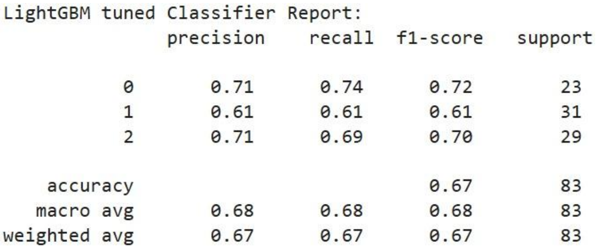




```
ANN tuned with MinMaxScaler Classifier Report:
              precision    recall  f1-score   support

         1.0       0.56      0.65      0.60        23
         2.0       0.50      0.55      0.52        31
         3.0       0.73      0.55      0.63        29

    accuracy                           0.58        83
   macro avg       0.59      0.58      0.58        83
weighted avg       0.59      0.58      0.58        83
```

Light GBM - The LightGBM tuned model achieved 67% accuracy with good performance across cirrhosis stages, especially for Stage 1 and Stage 3 cases. Stage 2 showed relatively lower performance, suggesting intermediate stages are more difficult to classify.

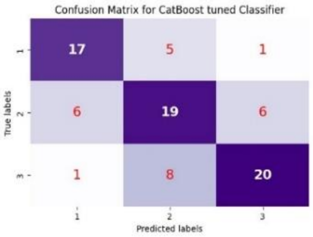


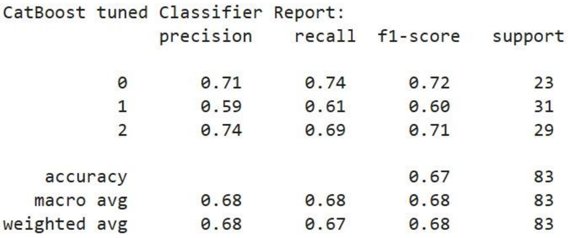


```
LightGBM tuned Classifier Report:
              precision    recall  f1-score   support

           0       0.71      0.74      0.72        23
           1       0.61      0.61      0.61        31
           2       0.71      0.69      0.70        29

    accuracy                           0.67        83
   macro avg       0.68      0.68      0.68        83
weighted avg       0.67      0.67      0.67        83
```

CatBoost

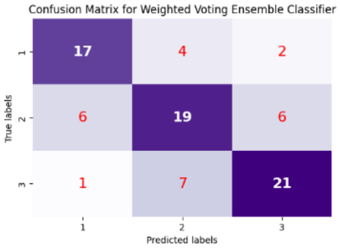
The CatBoost tuned classifier achieved 67% accuracy with good performance across all cirrhosis stages. It performed well for Stage 1 and Stage 3 cases, with Stage 2 showing comparable performance. Most misclassifications occur between adjacent stages, suggesting the model effectively distinguishes between disease progression patterns.

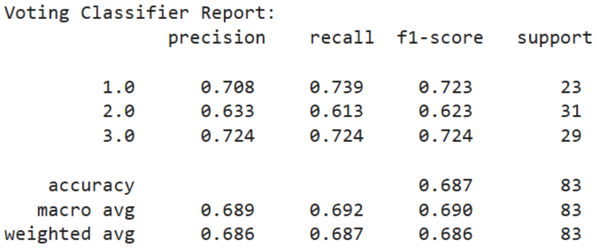




```
CatBoost tuned Classifier Report:
              precision    recall  f1-score   support

           0       0.71      0.74      0.72        23
           1       0.59      0.61      0.60        31
           2       0.74      0.69      0.71        29

    accuracy                           0.67        83
   macro avg       0.68      0.68      0.68        83
weighted avg       0.68      0.67      0.68        83
```

Voting - The Voting Classifier achieved 68.7% accuracy with excellent performance across all cirrhosis stages. It showed particularly strong results for Stage 1 and Stage 3 cases, with Stage 2 performing adequately. Most misclassifications occur between adjacent stages, demonstrating the model's ability to distinguish different phases of disease progression.





```
Voting Classifier Report:
              precision    recall  f1-score   support

         1.0      0.708     0.739     0.723        23
         2.0      0.633     0.613     0.623        31
         3.0      0.724     0.724     0.724        29

    accuracy                          0.687        83
   macro avg      0.689     0.692     0.690        83
weighted avg      0.686     0.687     0.686        83
```

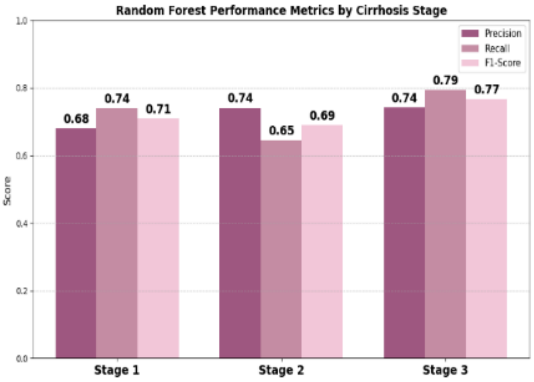
Results and Model Comparison

We evaluated multiple tuned cirrhosis classification algorithms using weighted metrics. F1-score served as our primary criterion due to our dataset's moderate imbalance, with confidence intervals (CI) ensuring that our selection was based on both performance and statistical reliability.

Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score	CI
Random Forest	0.723	0.724	0.723	0.722	0.7128 ± 0.0134 = [0.7027, 0.7229]
XGBoost	0.711	0.714	0.711	0.711	0.6991 ± 0.0190 = [0.6848, 0.7135]
CatBoost	0.67	0.68	0.67	0.68	0.6597 ± 0.0055 = [0.6556, 0.6638]
ANN	0.578	0.595	0.578	0.581	0.5787 ± 0.0176 = [0.5655, 0.5920]
Light GBM	0.67	0.67	0.67	0.67	0.6858 ± 0.0183 = [0.6720, 0.6995]
Voting	0.687	0.686	0.687	0.686	0.6749 ± 0.0105= [0.6669, 0.6828]

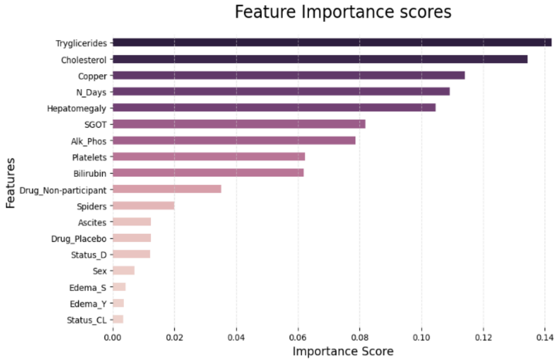
The Chosen Model - Random Forest

Our evaluation identified Random Forest as the optimal model for cirrhosis stage classification, with 0.723 accuracy and 0.722 F1-score. The model's accuracy increases with cirrhosis severity, achieving the highest F1-score (0.77) for Stage 3. This performance ensures patients with advanced cirrhosis requiring immediate intervention are correctly identified, while maintaining balanced overall performance despite relatively lower recall (0.65%) for Stage 2 disease.



Feature Importance Analysis

Triglycerides and Cholesterol emerged as strongest liver disease classification predictors, followed by Copper concentration, treatment duration (N_Days), and Hepatomegaly. Conventional markers like SGOT and Bilirubin showed moderate predictive value, while demographics had minimal impact. This suggests blood lipid panels and copper measurements could enhance liver disease staging efficiency.



Discussion and Conclusions

Model Selection Insights - Ensemble learning algorithms, notably Random Forest, outperformed alternatives for cirrhosis classification (0.72 F1-score), with all models showing higher accuracy for advanced disease stages.

Dataset Limitations - Preserving natural distribution in our limited dataset (412 samples) constrained early-stage classification performance, reflecting a trade-off between data integrity and balanced class performance.

Clinical Applications - Blood lipid indicators as primary predictors suggest potential for metabolic biomarker-focused diagnostic panels over conventional liver function tests, enabling earlier detection of cirrhosis progression.

Bibliography

Asrani, S. K., Devarbhavi, H., Eaton, J., & Kamath, P. S. (2019). Burden of liver diseases in the world. *Journal of Hepatology*, 70(1), 151–171.

D'Amico, G., Garcia-Tsao, G., & Pagliaro, L. (2006). Natural history and prognostic indicators of survival in cirrhosis: A systematic review of 118 studies. *Journal of Hepatology*, 44(1), 217–231.

Singal, A. K., Bataller, R., Ahn, J., Kamath, P. S., & Shah, V. H. (2020). ACG Clinical Guideline: Alcoholic Liver Disease. *The American Journal of Gastroenterology*, 115(3), 277–294.

Smith, A., Baumgartner, K., & Bositis, C. (2019). Cirrhosis: diagnosis and management. *American Family Physician*, 100(12), 759-770.

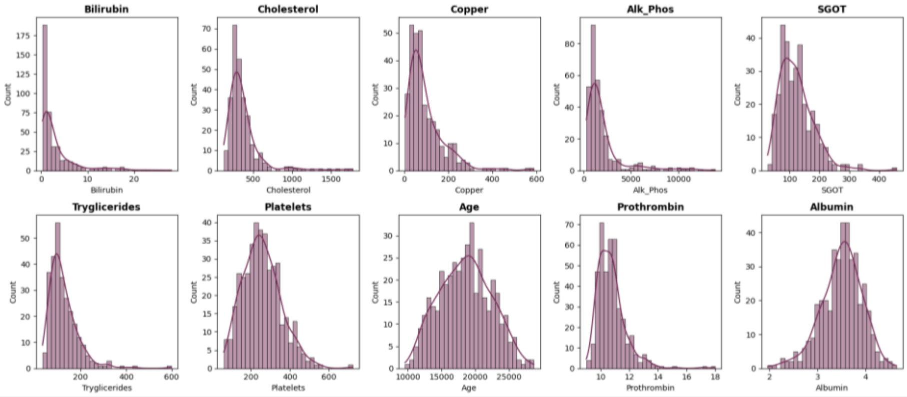
Zhai, Y., Hai, D., Zeng, L., Lin, C., Tan, X., Mo, Z., Tao, Q., Li, W., Xu, X., Zhao, Q., Shuai, J., & Pan, J. (2024). Artificial intelligence-based evaluation of prognosis in cirrhosis. *Journal of Translational Medicine*, 22(1), 933.

Zhang, Z., Wang, J., Han, W., & Zhao, L. (2021). Using machine learning methods to predict 28-day mortality in patients with hepatic encephalopathy. *BMC Gastroenterology*, 23, 111.

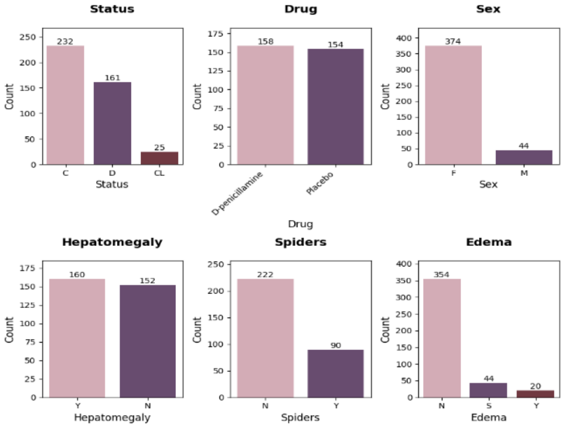
Appendices

Appendix A – Distributions of Continuous Features

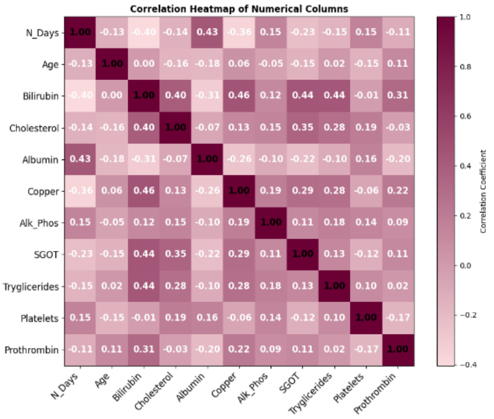
Liver function markers (bilirubin, alkaline phosphatase, SGOT, copper) significantly exceed normal ranges, with elevated cholesterol/triglycerides indicating metabolic abnormalities and borderline albumin suggesting early protein synthesis decline. Mildly prolonged prothrombin times show coagulation impairment, though platelet production remains normal. The middle-aged cohort shows variable disease progression and survival. Right-skewed distributions across biomarkers indicate most patients have moderate disease, with a subset showing severe liver dysfunction affecting multiple systems requiring intensive management.



Appendix B - Statistical description of the categorical variables

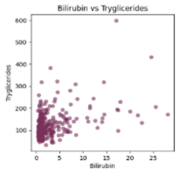


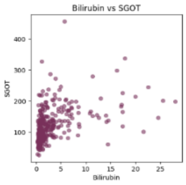
Appendix C - Relationships between variables

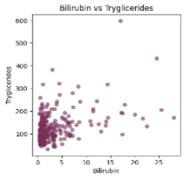


Appendix D – Selected scatter plots between variables (|Correlation| > |0.4|)

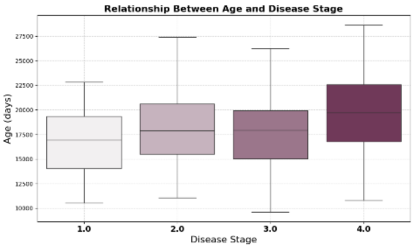
Bilirubin vs Triglycerides: Elevated bilirubin correlated with increased triglycerides despite variability, indicating liver dysfunction disrupts lipid metabolism, linking this liver injury marker to metabolic changes in advanced disease. **Bilirubin vs Copper:** Higher bilirubin levels correlated with increased copper despite variability, reflecting liver's diminished copper excretion capacity in advanced diseases like Wilson's or cirrhosis as dysfunction progressed. **Albumin vs N_Days:** Higher albumin levels correlated with longer survival times, reflecting albumin's role as a liver-produced protein and marker of hepatic function. Low levels indicate advanced liver disease and compromised protein synthesis, confirming albumin's value as a prognostic indicator.



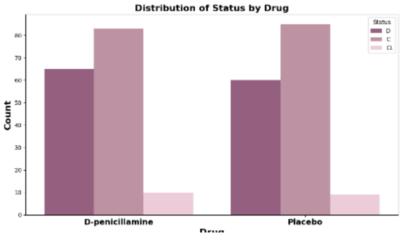




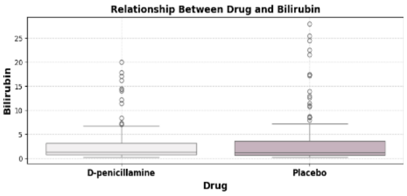
Appendix E - Selected relationships between variables

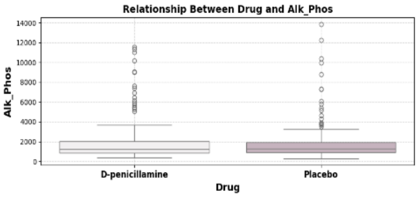

Relationship Between Age and Disease Stage

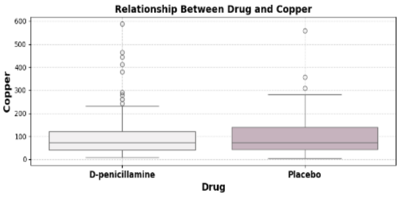
Age and Stage - Advanced disease stages (3-4) featured older patients than earlier stages (1-2), suggesting association with longer disease duration or gradual organ function decline typical with aging.


Distribution of Status by Drug

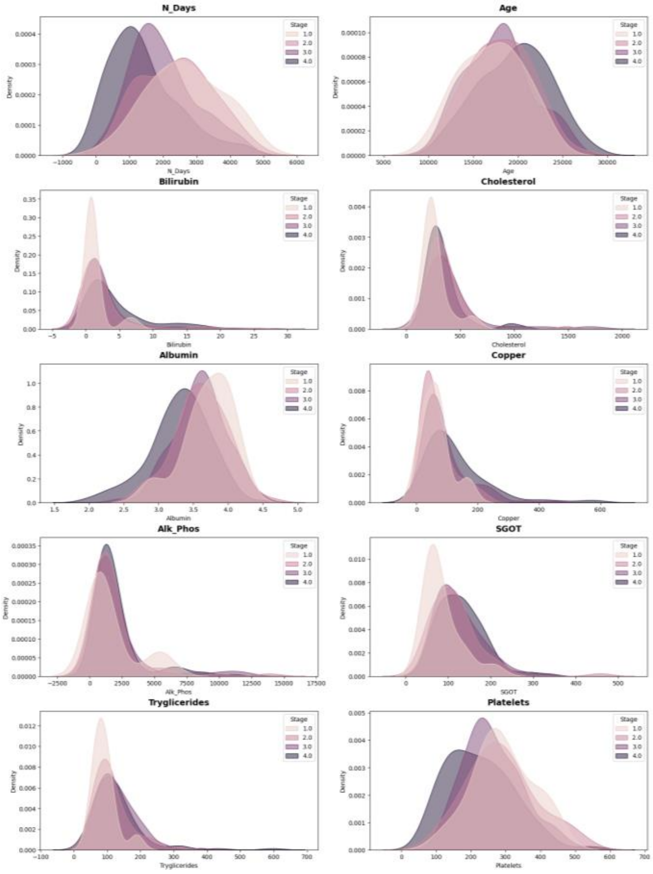
Drug and Status - Patient outcomes (alive, transplant, death) were similar between D-penicillamine and placebo groups, indicating no significant impact on survival or transplant rates. While mortality showed no difference, those taking the drug exhibited lower copper, phosphatase, and bilirubin levels.


Relationship Between Drug and Bilirubin


Relationship Between Drug and Alk_Phos


Relationship Between Drug and Copper

Appendix F – Distribution of Continuous Features by Stage


N_Days

Age

Bilirubin

Cholesterol

Albumin

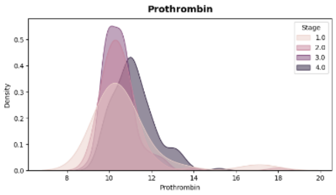
Copper

Alk_Phos

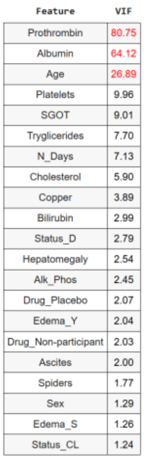
SGOT

Tryglicerides

Platelets

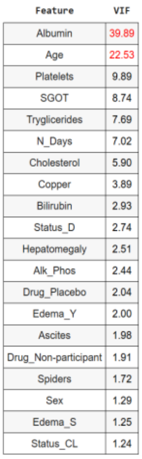

Prothrombin

Appendix G - Categorical Features Encoding



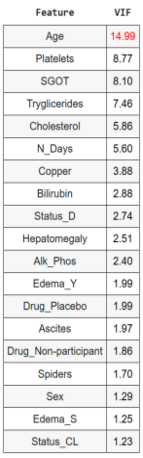
Feature	VIF
Prothrombin	80.75
Albumin	64.12
Age	26.89
Platelets	9.96
SGOT	9.01
Tryglicerides	7.70
N_Days	7.13
Cholesterol	5.90
Copper	3.89
Bilirubin	2.99
Status_D	2.79
Hepatomegaly	2.54
Alk_Phos	2.45
Drug_Placebo	2.07
Edema_Y	2.04
Drug_Non-participant	2.03
Ascites	2.00
Spiders	1.77
Sex	1.29
Edema_S	1.26
Status_CL	1.24

Figure 1 - all features



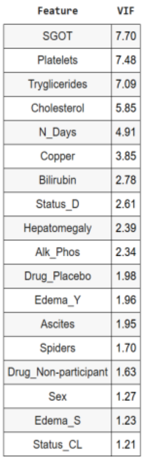
Feature	VIF
Albumin	39.89
Age	22.53
Platelets	9.89
SGOT	8.74
Tryglicerides	7.69
N_Days	7.02
Cholesterol	5.90
Copper	3.89
Bilirubin	2.93
Status_D	2.74
Hepatomegaly	2.51
Alk_Phos	2.44
Drug_Placebo	2.04
Edema_Y	2.00
Ascites	1.98
Drug_Non-participant	1.91
Spiders	1.72
Sex	1.29
Edema_S	1.25
Status_CL	1.24

Figure 2 - after removing



Feature	VIF
Age	14.99
Platelets	8.77
SGOT	8.10
Tryglicerides	7.46
Cholesterol	5.86
N_Days	5.60
Copper	3.88
Bilirubin	2.88
Status_D	2.74
Hepatomegaly	2.51
Alk_Phos	2.40
Edema_Y	1.99
Drug_Placebo	1.99
Ascites	1.97
Drug_Non-participant	1.86
Spiders	1.70
Sex	1.29
Edema_S	1.25
Status_CL	1.23

Figure 3 - after removing Albumin



Feature	VIF
SGOT	7.70
Platelets	7.48
Tryglicerides	7.09
Cholesterol	5.85
N_Days	4.91
Copper	3.85
Bilirubin	2.78
Status_D	2.61
Hepatomegaly	2.39
Alk_Phos	2.34
Drug_Placebo	1.98
Edema_Y	1.96
Ascites	1.95
Spiders	1.70
Drug_Non-participant	1.63
Sex	1.27
Edema_S	1.23
Status_CL	1.21

Figure 4 - after removing Age