VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

UNIVERSITY OF INFORMATION TECHNOLOGY

FACULTY OF INFORMATION SYSTEMS

FINAL PROJECT REPORT
STATISTICAL ANALYSIS

TOPIC:

Advanced Machine Learning Approaches for Early Detection and Staging of Liver Diseases and Cirrhosis Using Clinical and Laboratory Data

Instructor:

Dr. Trần Văn Hải Triều

TA. Nguyễn Minh Nhựt

Group 03:

1	Nguyễn Thanh Quân	23521266
2	Phạm Khôi Nguyên	23521055
3	Trần Lê Minh Thùy	23521560
4	Nguyễn Minh Dũng	23520333
5	Lê Tự Nhân	23521076

HO CHI MINH CITY, MAY 2025

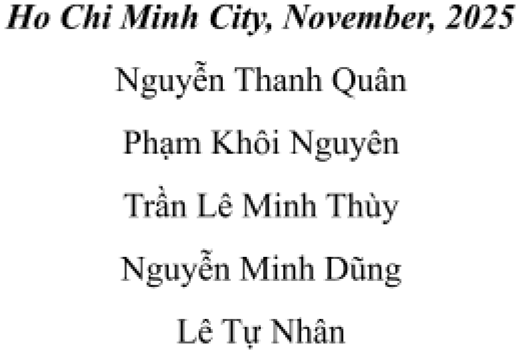
ACKNOWLEDGEMENT

We would like to extend our deepest gratitude to Dr. Tran Van Hai Trieu (PhD) and Engineer Nguyen Minh Nhut for your dedication and commitment in sharing your extensive knowledge and expertise with us throughout our learning and research journey in the Statistical Analysis course. Your guidance has played a crucial role in shaping our academic and professional development.

Participating in the Statistical Analysis course under your instruction has been an intellectually rewarding experience. Your clear explanations, practical insights, and consistent support have motivated us to overcome challenges and continually improve.

Despite the difficulties and limitations we encountered during our project, your steady encouragement enabled us to navigate obstacles with confidence and persistence. With your valuable feedback, we are confident that our project will continue to advance and form a solid foundation for our future work.

We sincerely wish you good health, fulfillment, and continued success in your mission to inspire and educate future generations. Your impact will continue to resonate in the academic journeys of many students.



Ho Chi Minh City, November, 2025

Nguyễn Thanh Quân

Phạm Khôi Nguyên

Trần Lê Minh Thùy

Nguyễn Minh Dũng

Lê Tự Nhân

INSTRUCTOR’S FEEDBACK

Ho Chi Minh City, November, 2025

ASSIGNMENT AND MEMBER EVALUATION TABLE

Name	Student ID	Assignment	Evaluation
Nguyen Thanh Quan	23521266	- Leader (arrange & assign tasks to Team's members) - Writing a paper. - Write Technical Reports on Models & Data Introduction. - Set up & Manage Git Repository. - Implement Softmax Model.	100%
Le Tu Nhan	23521076	- Implement XGBoost Model. - Conduct Statistical Analysis. - Prepare presentation slides	100%
Tran Le Minh Thuy	23521560	- Implement the SVM Model. - Report formatting. - Create an outline of the Report contents. - Create slide outline + prepare content	100%
Pham Khoi Nguyen	23521055	- Implement the Random Forest Model. - Data Preprocessing - Writing paper - Prepare content slide	100%
Nguyen Minh Dung	23520333	- Implement the KNN Model. - Setup Experiment - Prepare content slide - Research Clinical Insight of Data	100%

Table 1: Assignment and member evaluation table

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Key	Meaning
SVM	Support Vector Machine
KNN	K Nearest Neighbors
SMOTE	Synthetic Minority Over-sampling Technique
XGBoost	eXtreme Gradient Boosting
SGOT	Serum Glutamic-Oxaloacetic Transaminase
ALB	Albumin
Alk_Phos	alkaline phosphatase
HBV	Hepatitis B Virus
HBC	Hepatitis C Virus
Q-Q plot	Quantile-Quantile Plot
NASH	Nonalcoholic Steatohepatitis
PBC	Primary Biliary Cholangitis
ANOVA	Analysis Of Variance
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve

CHAPTER 1. GENERAL INTRODUCTION

1.1. Rationale Of The Study

The liver is a vital organ responsible for metabolism, detoxification, and protein synthesis. When chronic inflammation or persistent injury occurs, often due to alcohol, viral hepatitis (HBV, HCV), or fatty liver disease, the organ develops fibrotic scarring, eventually progressing to liver cirrhosis [1].

1.2. Objectives

The primary objective of this study is to build and evaluate a machine learning system for the multi-class classification of liver cirrhosis stages. The objectives are as follows:

1. **Data analysis:** Performing comprehensive statistical analysis (Kruskal, Chi-square) and visualization (Q-Q plots, Boxplots) to characterize the distribution of clinical biomarkers across disease stages.
2. **Data processing:** To implement robust preprocessing pipelines, including the handling of missing values, outliers, and applying SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance.
3. **Model development:** To train and compare the performance of various machine learning algorithms, specifically Softmax Regression, SVM, KNN, Random Forest, and XGBoost, to classify patients into stage 1, stage 2, or stage 3 with high accuracy [7].
4. **Evaluation:** To identify the most effective model based on accuracy, precision, recall, F1-score, and ROC-curve.

1.3. Scope of The Study

Our analysis focuses on clinical biomarkers across varying disease stages. Bilirubin’s ranking as the fifth influential feature reflects a dataset scope predominantly composed of early-to-intermediate stage patients, as this marker typically elevates only in advanced cirrhosis. This confirms the models' alignment with biological progression.

CHAPTER 2: DATA OVERVIEW AND DESCRIPTION

2.1. Dataset Description

2.1.1 Data Origin

The dataset used in this study originated from a well-characterized cohort of patients with primary biliary cirrhosis (PBC) collected at the Mayo Clinic. Between 1974 and 1984, a total of 424 patients diagnosed with PBC were evaluated for satisfying science research demand about liver cirrhosis stage change, as well as D-penicillamine trial efficient research in reducing Wilson's disease. Of these, 312 patients were enrolled and monitored with comprehensive clinical and biochemical assessments. An additional 112 patients who did not participate in the clinical trial were followed for baseline characteristics and survival outcomes; six were lost to follow-up shortly after diagnosis, leaving 106 complete records. Together, these constitute the 418 patient entries now widely used as the “liver cirrhosis” or “PBC” dataset in research and machine learning applications [3].

2.1.2 Data Acquisition And Structure

The data were collected prospectively during routine clinical evaluation and research follow-up visits at the Mayo Clinic [3]. Measurements include demographic characteristics, physical examination findings, laboratory markers, treatment status, and clinical staging. The dataset is structured as a tabular matrix in which each row represents a patient and each column represents a clinical or biochemical variable.

Features	Values	Clinical significance
Status	Categories: -0 = Censored (patient alive at cutoff or lost to follow-up) -1 = Death due to primary biliary cholangitis (PBC) -2 = Liver transplantation	In case evaluation, this column plays a significant role in identifying the current health status which contributes and baseline to cause liver stage changes
Drug	Categories: -0 = Placebo -1 = D-penicillamine	D-penicillamine was investigated as a potential disease-modifying agent in PBC; however, long-term trials demonstrated no significant survival benefit for placebo
Sex	Categories: -0 = Male -1 = Female	Most liver cirrhosis stages changes take the baseline of genetics resource, habitats or food consuming routine for a long time; biological sex also represents the body strength in case but not most

Ascites	Categories: -0 = Absent -1 = Present	Ascites is the accumulation of fluid in the peritoneal cavity, often detectable via physical examination. Cirrhosis is the underlying etiology for approximately 80% of patients presenting with ascites in the United States [4]
Hepatomegaly	Categories: -0 = Absent -1 = Present	Hepatomegaly is a common experiment result by sound scan which determines the enlarged liver size when having damage. In advanced cirrhosis, liver size may decrease because of progressive fibrosis and nodular regeneration
Spider	Categories: -0 = Absent -1 = Present	Spider angiomas reflect hyperestrogenism secondary to impaired hepatic estrogen metabolism and are characteristic of advanced chronic liver disease [5]
Edema	Levels: -0 = No edema and no diuretic therapy -0.5 = Edema present but responsive to diuretics or edema without prior diuretic use -1 = Edema despite diuretic therapy	This happens whenever there is damage to the liver, can not properly produce proteins, like albumin, that maintains fluid balance, and because increased pressure in the portal vein (portal hypertension) causes fluid to leak out of blood vessels into the tissues
Bilirubin	Normal range: 0.1–1.2 mg/dL	evaluates how to properly process and excrete these waste products from red blood cell breakdown, leading to its buildup in the blood and causing symptoms like jaundice (yellowing of the skin and eyes)
Cholesterol	Reference range: -less than 200 mg/dL (desirable) -200–239 mg/dL (borderline high) -240 mg/dL or greater (high)	Represents the status of low procedure of extracting protein in blood and making low nutritional status in patients
Albumin	Normal range: 3.5–5.0 g/dL	Albumin is a crucial protein for controlling blood volume, preventing fluid leakage, and transporting various nutrients or substances; its decline is a significant prognostic marker for the severity of the disease and its complications, such as kidney failure and ascites
Copper	Normal range: less than 55 µg/24 h	Urinary copper excretion is increased in cholestatic disorders, including PBC, secondary to reduced biliary excretion.

		Levels are markedly higher in Wilson disease than in PBC
Alkaline Phosphatase	Normal range: 44–147 U/L	ALP elevation is a hallmark of cholestasis and biliary injury. In PBC, ALP is typically elevated two- to ten-fold above the upper limit of normal and is used as a diagnostic and therapeutic response marker
Serum Glutamic-Oxaloacetic Transaminase	Normal range: 10–40 U/L	AST elevation reflects hepatocellular injury. An AST/ALT ratio greater than 1 may suggest advanced fibrosis or alcoholic liver disease. In end-stage cirrhosis, transaminases may normalize despite severe disease
Triglycerides	Reference ranges: -less than 150 mg/dL (normal) -150–199 mg/dL (borderline high) -200 mg/dL or greater (high)	Mild-to-moderate hypertriglyceridemia can occur in chronic cholestasis due to reduced lipoprotein lipase activity, although it is less prominent than hypercholesterolemia
Prothrombin	Normal range: 11–13.5 seconds	Prolonged PT reflects deficiency of vitamin K-dependent clotting factors synthesized by the liver. It is a sensitive marker of hepatic synthetic dysfunction and poor prognosis in cirrhosis
Stage (target column)	Categories: 1 = Portal inflammation with bile duct damage -2 = Periportal fibrosis and ductular proliferation -3 = Cirrhosis	Histological stage is the gold-standard measure of disease progression in PBC and serves as a key endpoint in clinical trials

Table 2.1: Description of clinical features, value ranges, and their medical significance.

2.1.3 Data Augmentation & Pre-processing

To support strong machine learning training, we used an expanded version of the original 418-patient Mayo Clinic PBC dataset. We synthetically increased the source data to about 25,000 samples to tackle class scarcity. Then, we applied a strict deduplication process to remove duplicates, leading to a final analytical dataset of 8,796 unique records.

2.2. Data Challenges: Discussion On Missing Values, Outliers, And Class Imbalance Encountered

The liver cirrhosis dataset presents significant features in challenges that may affect the robustness of inefficient statistical analyses and predictive modeling. First, many biochemical variables (e.g., Bilirubin, Alk_Phos, SGOT, Copper, Triglycerides) exhibit strong right-skewed distributions. Additionally, a number of outliers are present across multiple laboratory measurements, reflecting both biological variability and potential measurement inconsistencies; these outliers can disproportionately influence model parameters and necessitate the efficiency of using scalers. Moreover, several clinical variables are moderately to strongly correlated, introducing multicollinearity that may distort feature importance estimation and reduce model stability. Collectively, these challenges highlight the need for careful preprocessing to ensure reliable and clinically meaningful results.

CHAPTER 3: DATA ANALYSIS

Feature	Stage	Mean	Std	Kurtosis	Skewness
Bilirubin	1	2.25	3.64	18.99	4.05
	2	3.07	4.43	9.81	2.98
	3	4.16	4.96	4.22	2.10
Cholesterol	1	355.01	162.19	27.96	4.36
	2	396.79	252.70	13.72	3.52
	3	362.29	168.43	15.70	3.30
Albumin	1	3.58	0.36	0.28	-0.40
	2	3.58	0.34	1.10	-0.47
	3	3.36	0.38	0.99	-0.59
Copper	1	81.84	54.44	12.80	2.29
	2	95.24	71.43	7.05	2.23
	3	110.05	85.71	10.45	2.78
Alk_Phos	1	1843.03	1678.36	18.84	3.85
	2	1955.38	1899.15	14.26	3.61
	3	2086.88	1889.31	10.76	3.11
SGOT	1	114.96	47.80	9.45	1.95
	2	124.02	49.00	4.33	1.36
	3	126.44	46.51	2.31	1.06
Triglycerides	1	116.56	43.30	7.54	1.81
	2	124.92	50.30	12.90	2.24
	3	128.42	68.98	21.97	3.96
Platelets	1	277.24	102.32	2.11	0.96
	2	258.86	85.36	0.32	0.53
	3	230.05	95.02	0.57	0.70
Prothrombin	1	10.49	0.95	22.60	3.61
	2	10.50	0.79	19.15	2.81
	3	11.10	0.86	3.46	1.03

Table 3.1: Description of clinical features, value ranges, and their medical significance.

3.1 Data Distribution Analysis

In our analysis, numerous clinical variables exhibit approximately normal behavior, frequently reflected through a characteristic S-shaped curvature or close adherence to the reference line. This pattern is common in biomedical datasets: individuals without severe disease tend to cluster around physiologically normal values, whereas those with progressive pathology often display a drift of values toward either extreme. Such tendencies create concentrated regions near the upper bounds or near zero, mirroring the biological reality that abnormal measurements in hepatology typically accumulate at the distribution edges rather than at the center.

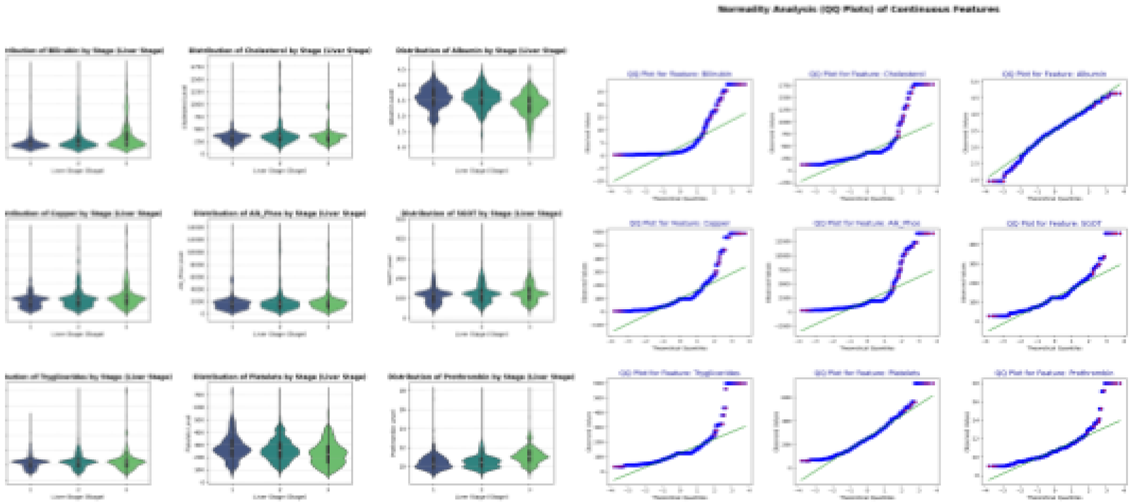


Figure 3.1: Distribution of continuous clinical features across liver cirrhosis stages (Violin and Box plots).

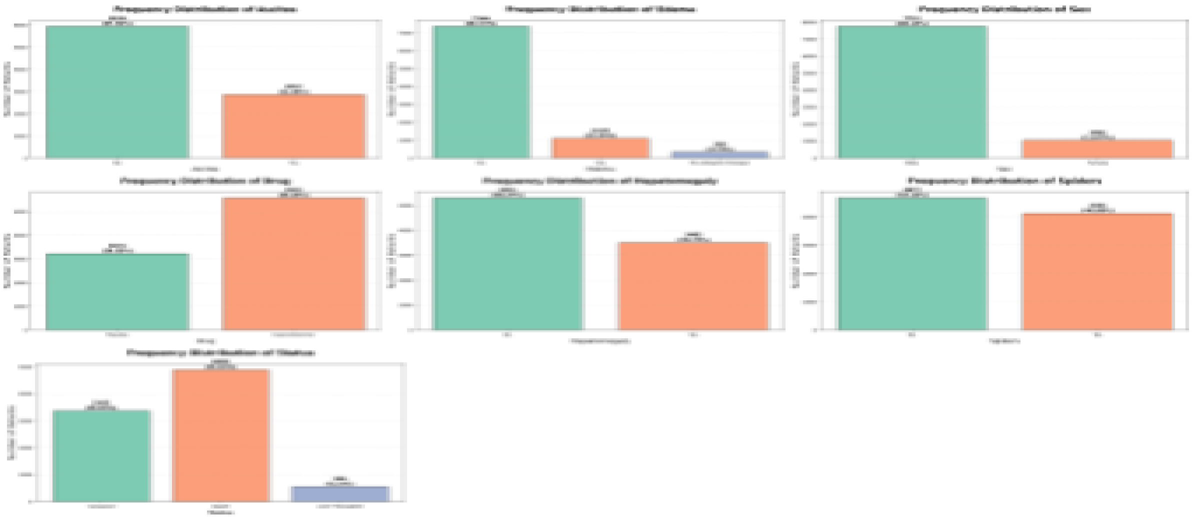


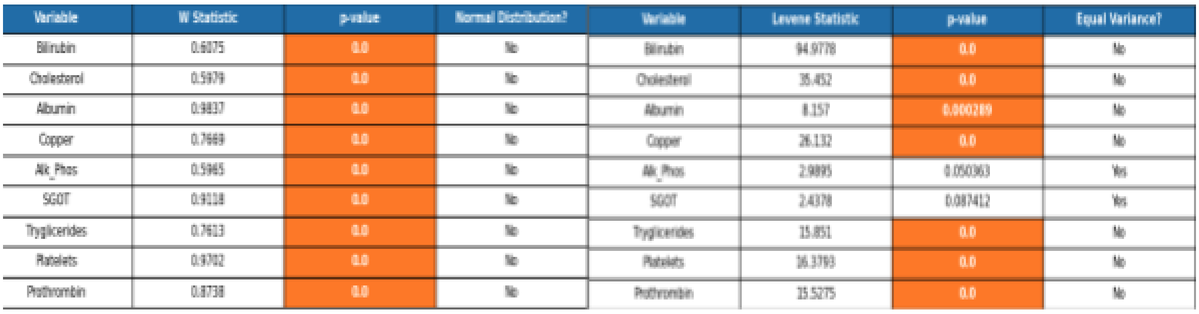
Figure 3.2: Frequency distribution of Ascites in the patient cohort.

3.2 How a data variable potentially affects predictive outcome

The clinical laboratory variables in the liver_cirrhosis dataset provide meaningful insights into disease progression, as their directional trends differ across cirrhosis stages. The descriptive statistics shown above indicate rising trends for variables like Bilirubin, Copper, Alk_Phos, SGOT, Triglycerides, and Prothrombin. In contrast, values such as Albumin and Platelets show a declining trend. This observation leads to the following hypothesis:

H_0: The distribution of each clinical variable is homogeneous across cirrhosis stages (i.e., the variable is independent of disease stage and does not contribute meaningful discriminatory information for predictive modeling).

3.2.1 Normality Assessment



Variable	W Statistic	p-value	Normal Distribution?
Bilirubin	0.6075	0.0	No
Cholesterol	0.5979	0.0	No
Albumin	0.9837	0.0	No
Copper	0.7669	0.0	No
Alk_Phos	0.5965	0.0	No
SGOT	0.9118	0.0	No
Tryglicerides	0.7613	0.0	No
Platelets	0.9702	0.0	No
Prothrombin	0.8738	0.0	No

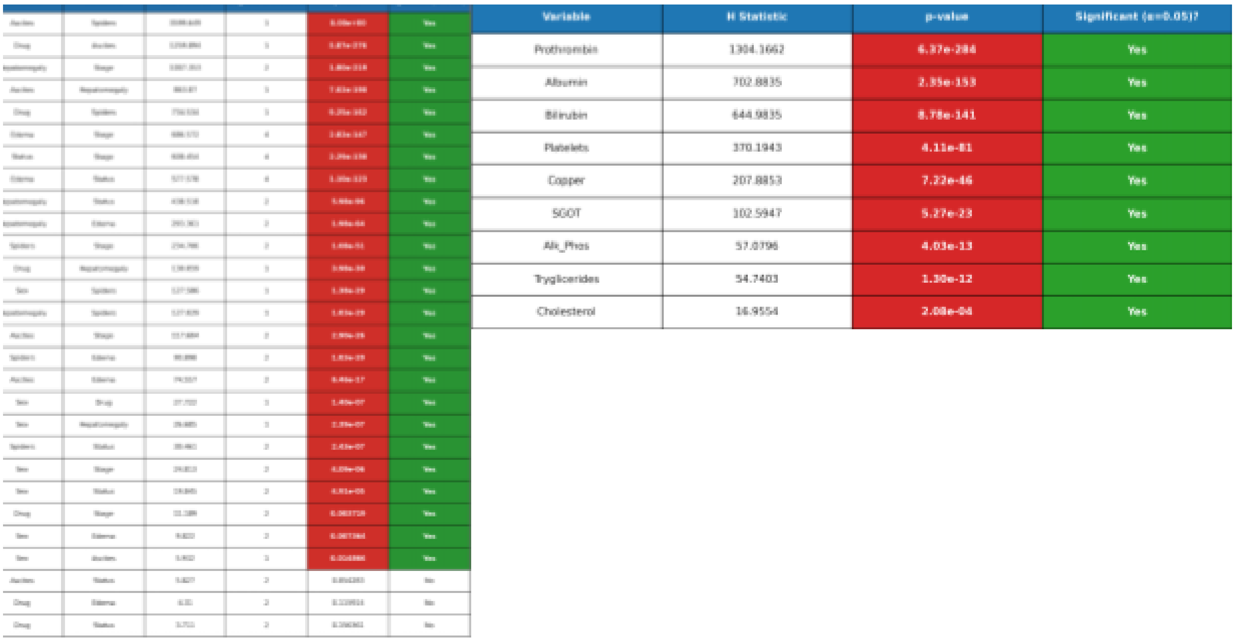
Variable	Levene Statistic	p-value	Equal Variance?
Bilirubin	94.9778	0.0	No
Cholesterol	35.452	0.0	No
Albumin	8.157	0.000289	No
Copper	26.132	0.0	No
Alk_Phos	2.9895	0.050363	Yes
SGOT	2.4378	0.087412	Yes
Tryglicerides	15.851	0.0	No
Platelets	16.3793	0.0	No
Prothrombin	15.5275	0.0	No

Table 3.2: Results of Shapiro-Wilk normality test and Levene's test for homogeneity of variance.

The results of both the Shapiro-Wilk normality test and Levene's test for homogeneity of variance indicate that the dataset violates the assumptions of normal distribution and equal variances. Given that the sample size is relatively small, the asymptotic assumptions supporting the use of parametric tests (such as ANOVA or linear regression residual normality) are not satisfied. Therefore, non-parametric statistical methods are more appropriate for assessing between-group differences.

Consequently, we rely on the Kruskal–Wallis H test, a non-parametric alternative that does not assume normality, and the Chi-square test for categorical variables.

3.2.2 Kruskal-Wallis H-Test And Chi-Square Analysis



Ascites	Spiders	3598.449	1	0.00e+00	Yes
Drug	Ascites	1259.894	1	1.87e-276	Yes
Hepatomegaly	Stage	1007.353	2	1.80e-219	Yes
Ascites	Hepatomegaly	863.87	1	7.61e-190	Yes
Drug	Spiders	756.534	1	9.25e-162	Yes
Edema	Stage	686.172	4	2.83e-147	Yes
Status	Stage	638.454	4	1.29e-136	Yes
Edema	Status	577.578	4	1.20e-123	Yes
Hepatomegaly	Status	438.518	2	5.90e-96	Yes
Hepatomegaly	Edema	293.363	2	1.90e-64	Yes
Spiders	Stage	234.788	2	1.08e-51	Yes
Drug	Hepatomegaly	138.859	1	3.98e-30	Yes
Sex	Spiders	127.586	1	1.39e-29	Yes
Hepatomegaly	Spiders	127.829	1	1.83e-29	Yes
Ascites	Stage	117.884	2	2.90e-26	Yes
Spiders	Edema	90.898	2	1.83e-20	Yes
Ascites	Edema	74.557	2	6.49e-17	Yes
Sex	Drug	27.722	1	1.40e-07	Yes
Sex	Hepatomegaly	26.885	1	2.39e-07	Yes
Spiders	Status	30.461	2	2.43e-07	Yes
Sex	Stage	24.813	2	4.09e-06	Yes
Sex	Status	19.845	2	4.91e-05	Yes
Drug	Stage	11.189	2	0.003719	Yes
Sex	Edema	9.822	2	0.007364	Yes
Sex	Ascites	5.902	1	0.015886	Yes
Ascites	Status	5.827	2	0.054283	No
Drug	Edema	4.31	2	0.115914	No
Drug	Status	3.711	2	0.156361	No

Variable	H Statistic	p-value	Significant (α=0.05)?
Prothrombin	1304.1662	6.37e-284	Yes
Albumin	702.8835	2.35e-153	Yes
Bilirubin	644.9835	8.78e-141	Yes
Platelets	370.1943	4.11e-81	Yes
Copper	207.8853	7.22e-46	Yes
SGOT	102.5947	5.27e-23	Yes
Alk_Phos	57.0796	4.03e-13	Yes
Tryglicerides	54.7403	1.30e-12	Yes
Cholesterol	16.9554	2.08e-04	Yes

Table 3.3: Statistical significance analysis using Kruskal-Wallis and Chi-Square tests across disease stages.

The Kruskal–Wallis and Chi-square analyses yield extremely small p-values for nearly all clinical variables. These results indicate statistically significant differences in variable distributions across cirrhosis stages. Hence, the data do not support the null hypothesis of homogeneous distributions.

3.3 Linear Pairwise And Global Multicollinearity Analysis In The Liver Cirrhosis Dataset

Relationships among variables in biomedical data are often implicit and must be examined directly to avoid informational redundancy or multicollinearity in linear regression analyses.

Implementing a sequential approach, starting with the Correlation Matrix and moving on to the Variance Inflation Factor (VIF), gives a clear view of the data's correlation structure. Together, these methods are essential for making sure that machine learning models, both linear and non-linear, are appropriate and effective.

3.3.1. Correlation Matrix Analysis

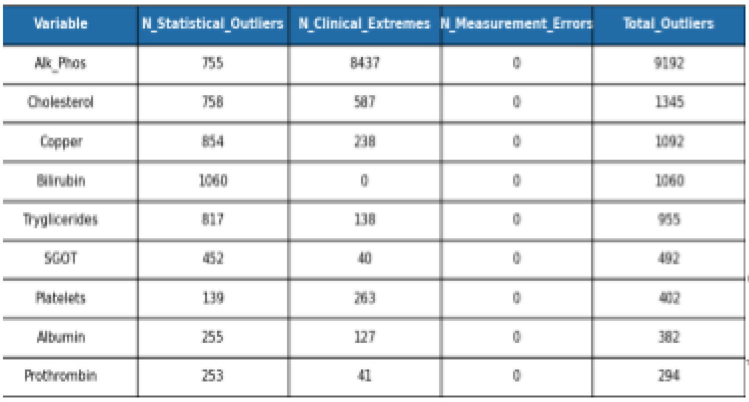
The correlation matrix serves as a foundational diagnostic tool for identifying linear associations between pairs of clinical variables within the liver_cirrhosis dataset from the Mayo Clinic. By examining the pairwise correlation coefficients ranging between, -1 and +1, the matrix allows us to detect hidden duplicated information, redundant predictors, or clinically coherent co-movement patterns among laboratory measurements. Positive coefficients indicate variables that increase together, whereas negative coefficients signify inverse relationships.

3.3.2. Variance Inflation Factor (VIF) For Global Multicollinearity Assessment

In this study, the VIF values for all variables remain below the commonly accepted threshold of 5, indicating an absence of problematic collinearity. This result, combined with the pairwise correlation analysis, suggests that both local (pairwise) and global (multivariate) linear evaluations support the conclusion that predictors are distributed across largely independent directions in the feature space.

3.4 Outlier Detection And Analysis

We looked at outlier analysis from three different angles: data entry errors, statistical anomalies found using the IQR method, and clinically significant extreme values. The results show a clear distinction among the stages of cirrhosis when patient records have biologically extreme measurements. This pattern reflects clinical reality, where abnormal lab results appear as the liver gradually loses its ability to function.



Variable	N_Statistical_Outliers	N_Clinical_Extremes	N_Measurement_Errors	Total_Outliers
Alk_Phos	755	8437	0	9192
Cholesterol	758	587	0	1345
Copper	854	238	0	1092
Bilirubin	1060	0	0	1060
Tryglicerides	817	138	0	955
SGOT	452	40	0	492
Platelets	139	263	0	402
Albumin	255	127	0	382
Prothrombin	253	41	0	294

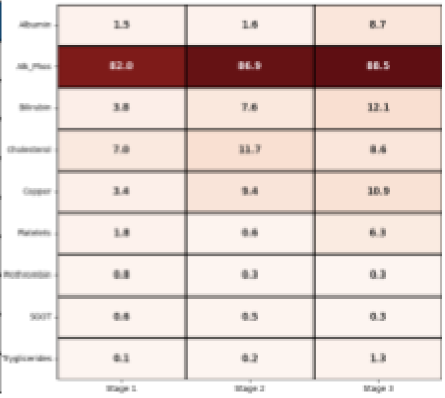


Table 3.4: Summary of outlier detection statistics and their distribution across cirrhosis stages.

CHAPTER 4: DATA PREPROCESSING PIPELINE

4.1 Missing Data Mechanisms

In this study, we evaluated potential imputation strategies, such as median substitution for skewed biochemical markers, regression-based imputation for clinically relevant predictors, or MICE for uncertainty preservation to ensure preparedness for model expansion, integration with external datasets, or longitudinal data merging. This step establishes that the current dataset is structurally sound while also defining a standardized protocol should missingness arise in subsequent phases of the project [6].

4.2 SMOTE Implementation And Evaluation

	Stage 1	Stage 2	Stage 3
Total data points	2744	3038	3014

Table 4.1: Class distribution of liver cirrhosis stages in the original dataset.

PCA shows the main structure of the data by focusing on the primary differences. The PC1-PC2 projection highlights how the variables relate to each other. It acts as a check on the structure's integrity. This allows us to verify that resampling techniques, such as SMOTE, preserved the cirrhosis dataset's inherent covariance patterns without introducing distortion [2].

In contrast to PCA, LDA explicitly maximizes separation between cirrhosis severity groups. We utilize LDA here not as a classifier, but as a structural validation tool. The stability of LDA axes pre- and post-SMOTE confirms that the resampling process preserved clinically meaningful class boundaries without distorting the original data structure.

By using PCA and LDA together, we can verify that the dataset retains its character after balancing. We back this up by checking distributional shifts to ensure no unrealistic patterns were introduced. In cirrhosis research, where specific numbers define disease stages, we cannot afford to distort these values. Therefore, comparing the data before and after SMOTE is a necessary step to guarantee that our medical findings remain grounded in reality.

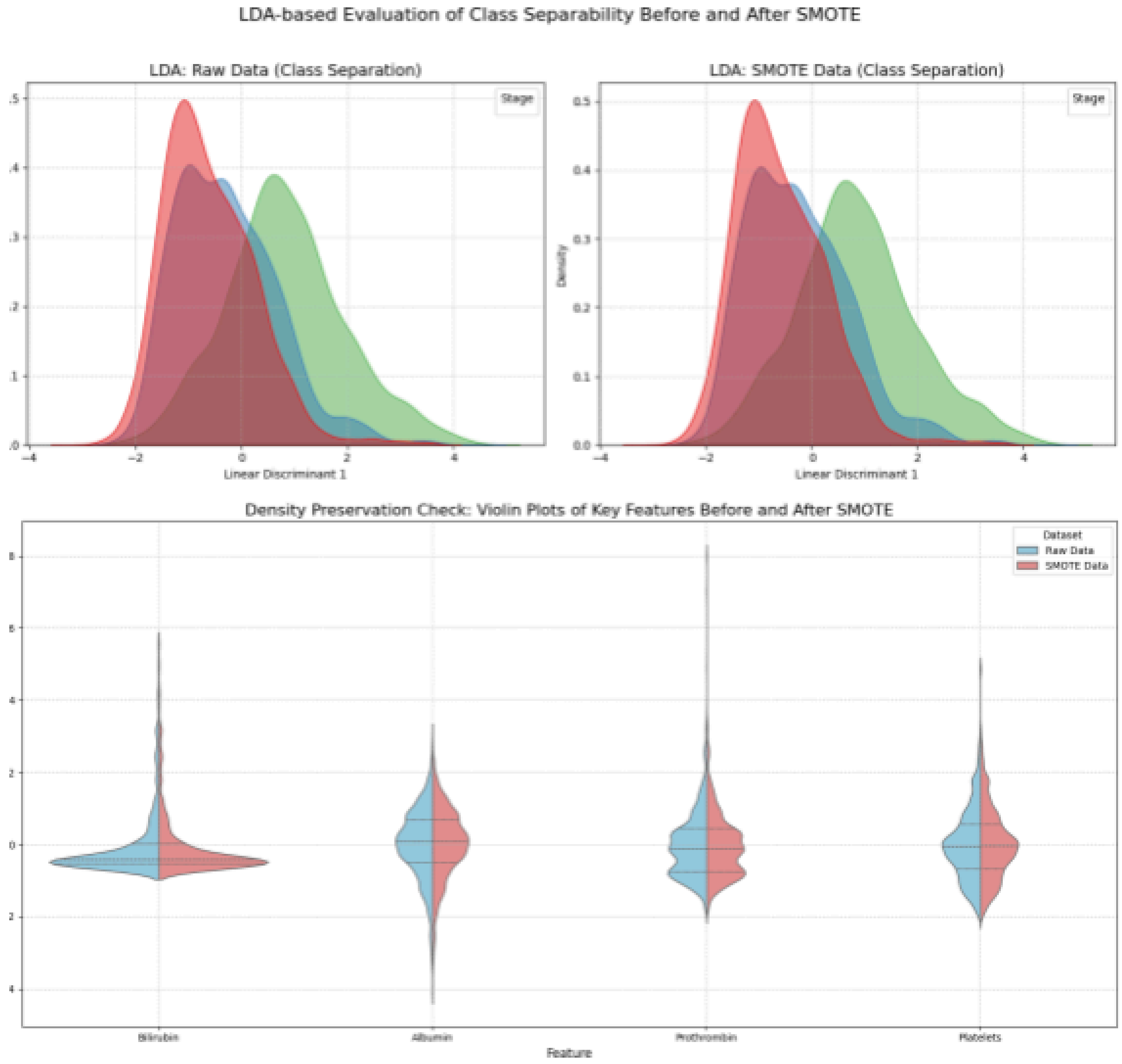


Figure 4.1: Comparison of class separability using Linear Discriminant Analysis (LDA) before and after SMOTE application.

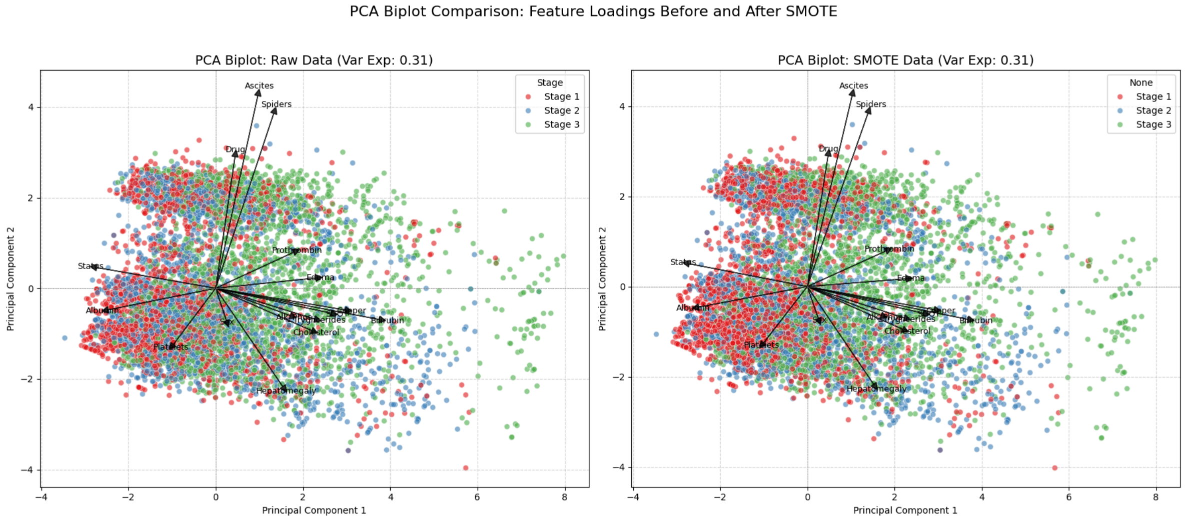


*Figure 4.2***:** *PCA Biplot comparison showing feature loadings and data structure stability before and after SMOTE.*

4.3 Data Transformation

For linear models, it is important to manage right-skewed distributions while keeping the key clinical relationships among biomarkers. Methods such as StandardScaler, MinMaxScaler, and RobustScaler have proven effective in this area.

1. **StandardScaler** standardizes variables to zero mean and unit variance, facilitating stable optimization for algorithms such as Logistic Regression or linear SVM.
2. **MinMaxScaler** preserves the relative distance structure of the original distribution while compressing values into a bounded interval, which is beneficial when variables differ significantly in scale.
3. **RobustScaler**, which relies on the median and interquartile range, effectively handles biomarkers with sporadic high values (e.g., Bilirubin spikes or elevated liver enzymes).

These methods help linear models operate under more idealized statistical conditions without distorting the inherent biological meaning of the features.

For non-linear models, proper scaling is important, especially when working with large clinical datasets that have hundreds of thousands of records. Scaling puts measurements on similar ranges, which prevents distance-based kernels or tree-based methods from fitting too closely to extreme values. It is also important to keep the integrity of the underlying clinical signal. Otherwise, even small distortions can result in unreliable or misleading diagnostic predictions.

CHAPTER 5: MACHINE LEARNING IMPLEMENTATION AND EXPERIMENT ANALYSIS

5.1 Predictive Model Implementation

We conduct experiments using linear classifiers (multinomial logistic regression, linear SVM) to evaluate the hypothesis: ‘Do individual clinical measurements contribute to predicting stage transition?’ Simultaneously, we apply instance-based and ensemble methods (k-nearest neighbors, random forest, XGBoost) to test the complementary hypothesis: ‘Does the joint relationship among multiple clinical indices increase the accuracy of cirrhosis stage prediction?’

It is important to note that these analyses capture predictive associations rather than causal relationships. Establishing causal mechanisms underlying stage transitions would require longitudinal data and appropriate causal inference methods. We therefore evaluate both single-feature predictive power and multivariate model performance using stratified cross-validation, feature importance and explainability tools (e.g., permutation importance, SHAP), and rigorous post-resampling integrity checks when synthetic balancing methods (e.g., SMOTE) are applied.

Model	Core Mechanism	Main Advantages	Limitations and Trade-offs	Medical Significance
Softmax Regression	A multiclass extension of logistic regression that models class probabilities through a set of linear weight vectors. The softmax function normalizes these probabilities to sum to one.	High interpretability: clinical variables contribute via explicit coefficients, allowing physicians to understand how each laboratory indicator affects stage probability. Stable baseline model for tabular medical data.	Limited ability to learn nonlinear relationships or interactions among clinical variables. Performance decreases when cirrhosis stages exhibit complex physiological boundaries not captured by linear functions. Sensitive to multicollinearity and scaling.	Useful for establishing clinically interpretable baseline behavior and identifying key laboratory markers associated with each cirrhosis stage.

Support Vector Machine (SVM)	Identifies the optimal separating hyperplane that maximizes the margin between classes. Incorporates kernel functions to model nonlinear boundaries in the clinical feature space.	Effective in high-dimensional settings. The kernel trick allows SVM to capture complex nonlinear clinical patterns often present in hepatology. Good performance when samples are limited but features are many.	Hard to interpret: difficult to explain why a patient lies on one side of the hyperplane. Strong dependence on kernel selection and hyperparameters. Training can be computationally expensive.	Suitable when cirrhosis staging involves subtle, nonlinear clinical boundaries. However, limited explainability may reduce clinical acceptance.
K-Nearest Neighbors (KNN)	Classifies a new patient by identifying the K most similar patients in the training set based on a distance metric (commonly Euclidean). The predicted stage is determined by majority vote.	Conceptually simple and intuitive. Mimics clinical reasoning based on similarity between patients. No training phase; adapts to new data immediately.	Strongly affected by noise and outliers. Suffers from the curse of dimensionality, making distance metrics unreliable when many laboratory variables are involved. Requires careful feature scaling.	Valuable for exploratory analysis of patient similarity landscapes, but less reliable as a primary classifier for complex liver disease data.
Random Forest	An ensemble of decision trees built on bootstrap samples with randomized feature selection. Final prediction is obtained by majority vote across all trees.	High robustness to noise and variable interactions. Naturally captures nonlinear relationships and complex dependencies among clinical variables. Handles diverse data types well.	Lower interpretability: although feature importance can be extracted, tracing individual predictions through hundreds of trees is difficult. Requires more memory and computation compared with linear models.	Strong candidate for clinical prediction tasks where interactions between biochemical markers are expected. Provides stable performance in imbalanced or noisy medical datasets.

XGBoost	A gradient boosting framework that sequentially builds weak decision trees, with each tree correcting the errors of the previous one. Incorporates regularization and efficient handling of missing values.	State-of-the-art performance for tabular clinical data. Excellent control of overfitting. Handles missing values natively. Frequently achieves highest predictive accuracy among ensemble models.	High complexity with difficult interpretability. Requires extensive hyperparameter tuning and substantial computational resources. Interpreting individual patient predictions typically requires SHAP or other XAI tools.	Highly effective for cirrhosis stage prediction when accuracy is prioritized, especially in datasets with nonlinear interactions among clinical indicators. Must be paired with explainability tools for clinical deployment.

Table 5.1: Comparative overview of machine learning algorithms: mechanisms, advantages, and limitations.

5.2 Experiment Analysis

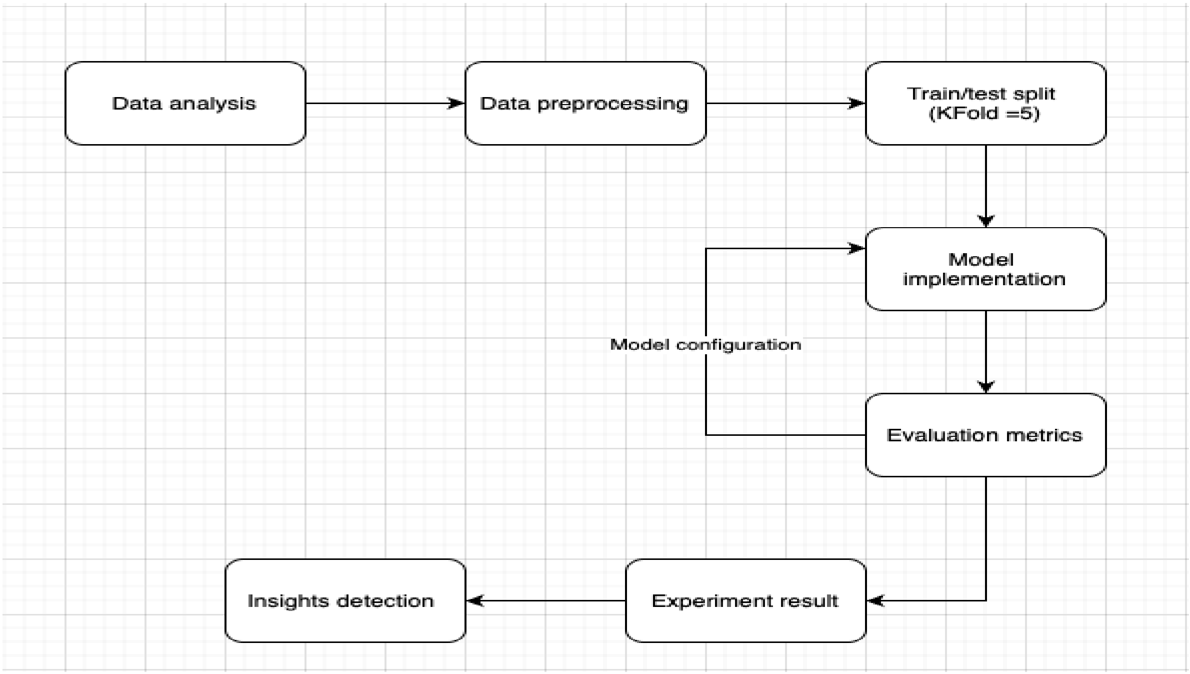


Figure 5.1: Schematic overview of the experimental machine learning workflow.

Using a clinically validated and medically grounded dataset, we first evaluated the discriminative capacity of the input variables through statistical tests such as the Chi-square test and the Kruskal–Wallis test. The extremely low p-values indicate that the clinical indicators differ significantly across cirrhosis stages, supporting their suitability as predictive features. Missing values were minimal and class-balance diagnostics were fully examined.

We then conducted a 5-fold cross-validation procedure across all regression-based models under two experimental scenarios: before SMOTE and after SMOTE. The underlying assumption was as follows: if the performance difference between the two scenarios is negligible, the predictive behavior of the models can be considered stable, allowing us to identify the most effective model for this dataset.

5.2.1 Baseline Experiment Before SMOTE

Linear-origin models showed low predictive performance. The confusion matrix indicates that staging cirrhosis needs to capture complex, nonlinear interactions among clinical indicators. This pattern matches the detailed decision-making seen in clinical practice. This outcome suggests that linear decision boundaries are not enough to model the subtle changes between disease stages.

In contrast, the nonlinear and non-parametric models, KNN, Random Forest, and XGBoost, showed high and stable predictive accuracy across folds. Their strong class-wise F1 scores indicate solid discriminative ability, helping to reduce misclassification between adjacent stages, particularly Stage 1 and Stage 2, which often present overlapping clinical features.

Additionally, the choice of scaling strategy had a big impact on model performance. StandardScaler helped make optimization more stable for models that use gradient-based learning. In contrast, RobustScaler worked especially well for distance-based methods like KNN and SVM by lessening the effect of outliers often found in clinical measurements.

Model	Accuracy	F1 score	Precision	Recall
Softmax Regression	56.14%	55.35%	55.61%	56.14%
Support Vector Machine	55.51%	54.45%	54.81%	55.51%
K Nearest Neighbors	72.05%	71.97%	71.85%	72.05%
Random Forest	84.66%	84.59%	84.63%	84.66%
XGBoost	88.47%	88.44%	88.43%	88.47%

Table 5.2: Performance metrics of machine learning models in the baseline experiment (Pre-SMOTE).

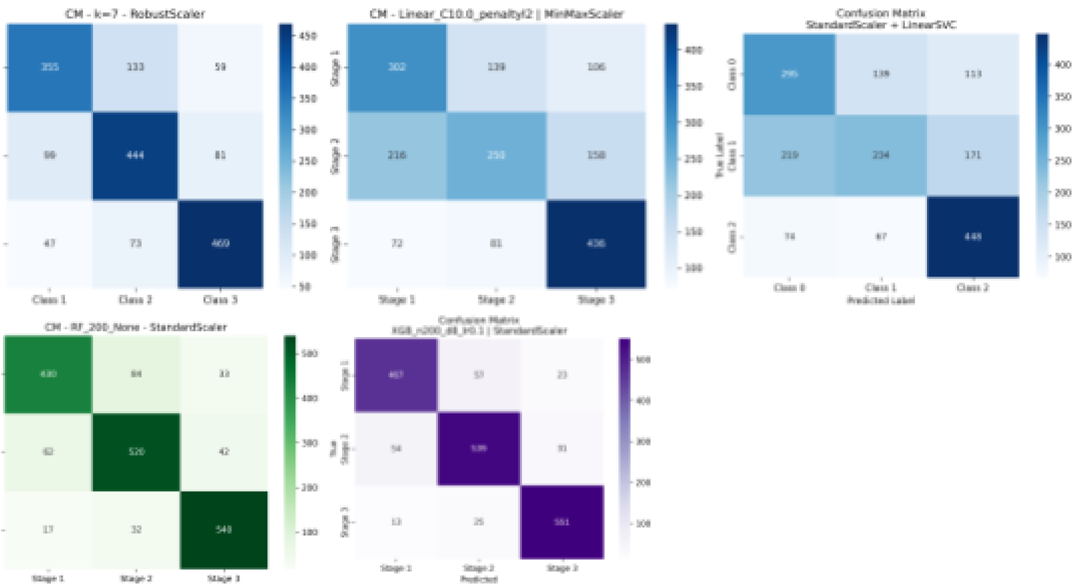


Figure 5.2*: Comprehensive performance comparison of five machine learning models: Confusion Matrices*

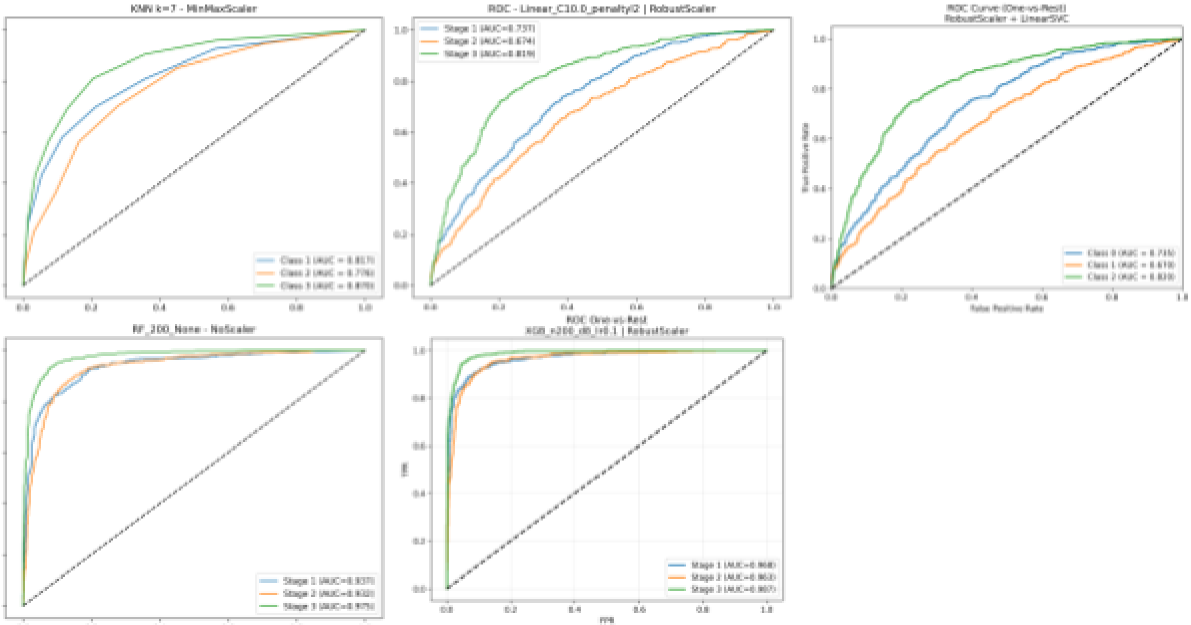


Figure 5.3: Comprehensive performance comparison of five machine learning models: ROC Curves

5.2.2 Experiment After SMOTE

The application of SMOTE, even with a very small oversampling ratio of only 1%, means that for every 1,000 original samples, just 100 synthetic minority samples are added. Which still produced a measurable effect on the best-performing model, XGBoost, improving its performance by approximately 0.24% on average across ACC, F1, Precision, and Recall. This suggests that, in real-world scenarios with larger datasets, SMOTE can be considered a reliable technique for handling class imbalance. At the same time, the remaining models showed results that were largely consistent with their pre-SMOTE performance, indicating that SMOTE did not introduce instability or degrade model robustness.

Model	Accuracy	F1 score	Precision	Recall
Softmax Regression	55.40%	54.36%	55.07%	55.4%
Support Vector Machine	55.45%	54.11%	55.25%	55.45%
K Nearest Neighbors	71.48%	71.42%	71.41%	71.48%
Random Forest	84.20%	84.17%	84.15%	84.20%
XGBoost	88.69%	88.68%	88.67%	88.69%

Table 5.3: Performance evaluation of machine learning models after applying SMOTE.

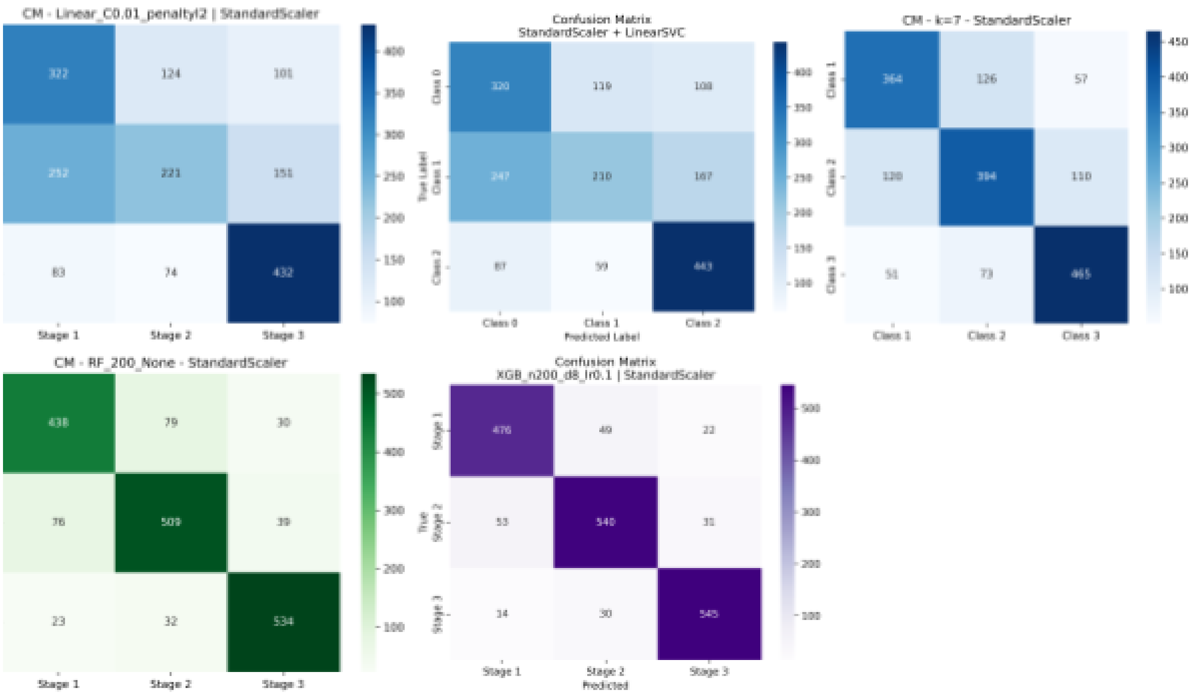


Figure 5.4: Comprehensive performance comparison of five machine learning models: Confusion Matrices

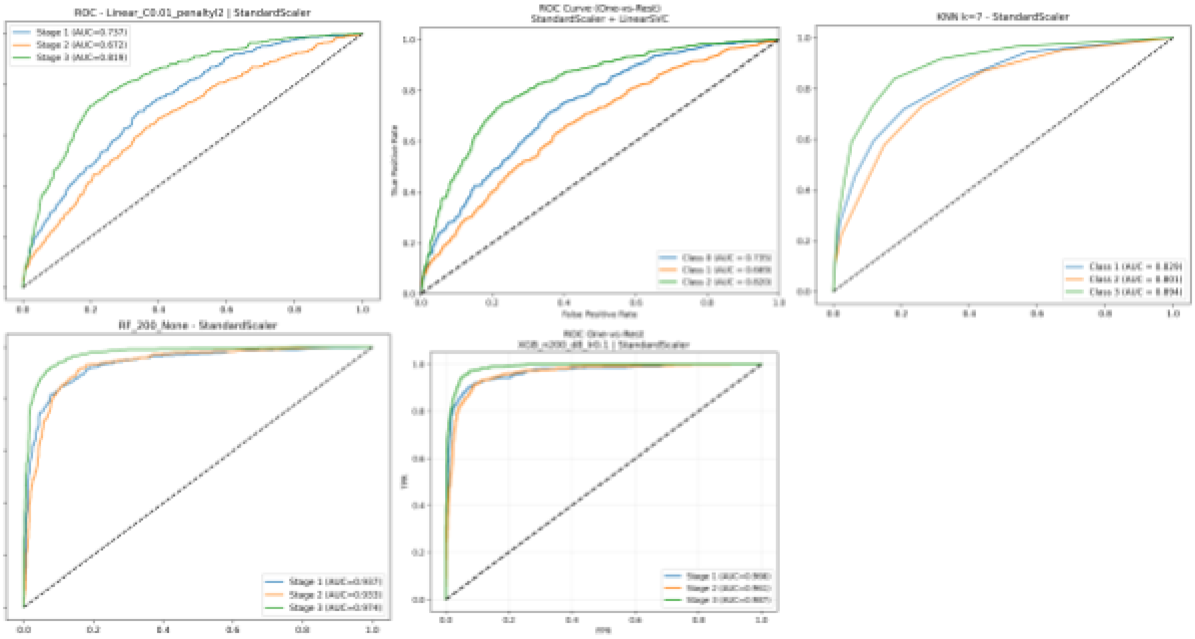


Figure 5.5: Comprehensive performance comparison of five machine learning models: ROC Curves

5.3 How Scaler Affects On Experiment

Model	Scaler	Accuracy	Precision	Recall	F1-score	AUC-ROC
Softmax Regression	None	54.83%	54.39%	54.83%	54.32%	73.14%
	StandardScaler	55.8%	55.24%	55.8%	55.01%	74.22%
Support Vector Machine	None	47.56%	41.24%	47.56%	40.86%	67.26%
	StandardScaler	55.51%	54.93%	55.51%	54.45%	74.06%
K Nearest Neighbors	None	64.66%	64.53%	64.66%	64.55%	78.7%
	RobustScaler	72.05%	72.01%	72.05%	71.97%	85.4%
Random Forest	None	79.72%	79.93%	79.72%	79.72%	93.16%
	RobustScaler	84.66%	84.63%	84.66%	84.59%	94.81%
XGBoost	None	88.47%	88.43%	88.47%	88.44%	97.25%
	RobustScaler	88.47%	88.43%	88.47%	88.44%	97.25%

Table 5.4: Impact of different feature scaling techniques (StandardScaler vs. RobustScaler) on model performance.

As expected, the choice of scaler significantly affected the performance of linear models. Proper scaling helps reduce the penalty effects caused by outliers and different units of measurement. This improvement is clear in the data: the AUC for Softmax Regression increased from 0.7314 to 0.7422, and the AUC for SVM rose from 0.6726 to 0.7406 after applying the right feature scaling. These results indicate that linear decision boundaries are especially sensitive to unscaled or distorted clinical variables. They achieve notable performance improvements with standardized or normalized inputs.

For non-linear, interaction-based models, the role of scaling is more nuanced. In this study, RobustScaler proved especially effective in stabilizing the amplitude of the feature space and preventing the models from overemphasizing outliers. This is particularly important in datasets containing duplicated entries or irregular measurement spikes, conditions that are common in clinical data collection. The improvements further support this observation:

1. KNN increased from 0.787 to 0.854
2. Random Forest improved from 0.9316 to 0.9481
3. XGBoost maintained its strong performance at 0.9725.

These results suggest that models relying on local similarity (e.g., KNN) or ensemble learning (e.g., Random Forest) greatly benefit from the robustness provided by outlier-resistant scaling methods. In large-scale clinical datasets typically found in hospitals or research institutes, where measurement inconsistencies and duplicated records are unavoidable, RobustScaler should be considered a key preprocessing technique. Its ability to regulate extreme values without distorting the core data distribution provides a reliable pathway for further model performance improvements, especially when scaling up to broader, more heterogeneous patient populations.

5.4 SHAP Analysis

The investigation of each clinical indicator’s contribution to the predictive performance and stability of the machine-learning model is essential for validating both the statistical relevance and the clinical interpretability of the results. After empirical verification and comparison with our initial assumptions, the instance-based models produced highly impressive performance in the three-class cirrhosis classification task. This encouraged a more in-depth exploration of feature contribution using SHAP values.

The correlation matrix and VIF analysis showed that most variables offer useful information and have low multicollinearity. This means that each feature adds a unique contribution to the model. However, these methods only reveal pairwise relationships and overall structural redundancy. Additionally, PCA indicated a clear linear structure in the data. The orientations of feature vectors matched known clinical information about biomarkers that usually increase or decrease together. Nonetheless, PCA does not show how changes in specific clinical measures affect the risk of cirrhosis progression. Given these limitations, using SHAP with the Random Forest model (chosen instead of XGBoost to lower computational costs while maintaining good predictive performance) offered a clearer and more clinically relevant view of feature importance. SHAP visualizations helped us identify which biomarkers significantly influenced the classification outcome, which features had little effect, and which variables might add noise. This level of understanding is crucial for turning machine-learning results into useful medical insights that help explain cirrhosis progression.

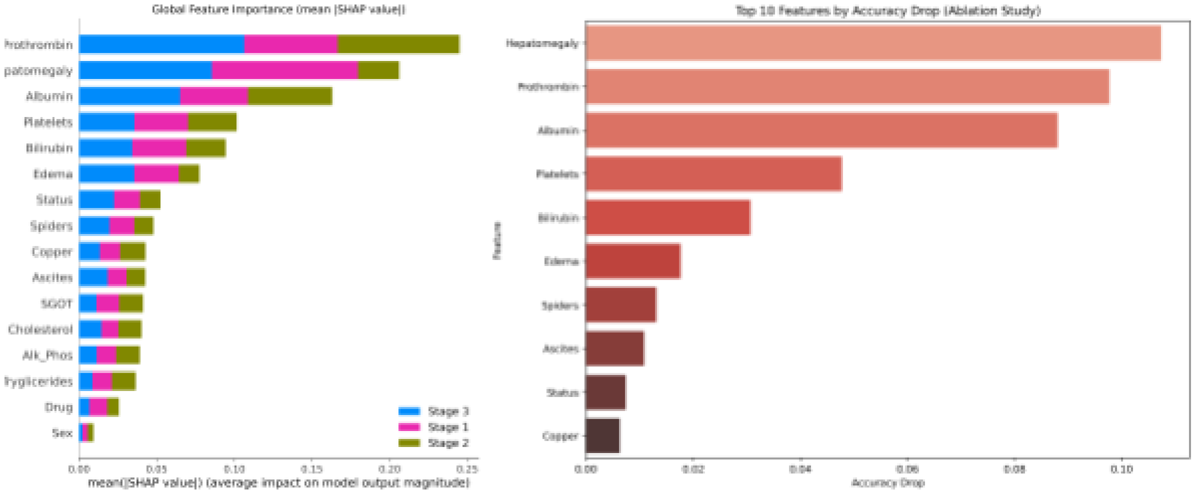


Figure 5.6: Comparative assessment of feature importance using Global SHAP analysis (left) and Ablation Study (right).

5.5 Key Insight And Clinical Evidence Validation

The strong performance of the non-linear machine-learning models highlights the need to account for complex, non-linear interactions among clinical biomarkers when predicting cirrhosis stages. This result is consistent with the biological behavior of hepatic dysfunction, where multiple physiological processes interact in non-linear and sometimes compensatory patterns. Consequently, the predictive capacity of the models suggests that cirrhosis progression cannot be adequately understood through linear correlation alone but instead requires interpretability methods that account for nonlinear patterns.

To explore the role of individual biomarkers, we combined SHAP analysis with Ablation Studies. These methods helped us measure how each variable affects the model output and assess its clinical importance. The findings below offer essential insights related to hepatology, backed by recognized medical literature:

1. **Hepatomegaly as the most influential predictor**: SHAP identified hepatomegaly as the strongest contributor to model predictions. This finding aligns with ***Kaplan et al. [8]***, who established liver enlargement as a cardinal

physical sign of inflammatory activity in Primary Biliary Cholangitis (PBC), preceding the atrophic changes seen in end-stage disease.

2. **Prothrombin time as a highly predictive functional marker**: Prothrombin time emerged as the second most influential feature in our analysis. Clinically, this serves as a direct measure of hepatic synthetic capacity, where a prolonged prothrombin time is strongly associated with impaired liver function and worsening hepatocellular failure. This finding aligns with the work of ***Kamath et al. [9]***, who established the MELD score, identifying the International Normalized Ratio (derived from prothrombin time) as a primary predictor of survival in end-stage liver disease. The model's high ranking of this feature reinforces its documented role in assessing cirrhosis severity.
3. **Albumin as a central indicator of hepatic synthetic decline**: Albumin ranked third in feature importance, a result that is consistent with its longstanding application in clinical scoring systems. Reduced serum albumin levels indicate impaired protein synthesis, which is a hallmark of chronic liver dysfunction. The prominence of albumin in the SHAP analysis validates the model's alignment with the ***Child-Pugh classification established by Pugh et al. [12]***, which utilizes albumin as a key variable for assessing hepatic reserve and prognosis.
4. **Platelet count as an early marker of portal hypertension**: The model identified Platelet count as the fourth most important feature. Thrombocytopenia (low platelet count) is widely recognized as an early sign of portal hypertension that often precedes the clinical presentation of overt cirrhosis. This contribution is clinically supported by the ***Baveno VI Consensus [10]***, which define platelet count as a sensitive non-invasive marker for portal hypertension and spleen stiffness. The model's reliance on this variable shows its sensitivity to early pathological changes.
5. **Bilirubin as a late-stage indicator with lower SHAP impact:** Bilirubin ranked fifth, exhibiting a lower SHAP contribution compared to the structural and functional markers mentioned earlier. While bilirubin is an important indicator of liver excretory function, ***Lammers et al. [11]*** demonstrated that its levels typically elevate significantly only in the advanced or decompensated stages of liver disease, remaining relatively stable during early progression. Given that the dataset encompasses patients across early to mid-stages, the lower weight assigned to Bilirubin reflects an accurate capture of the underlying biology rather than a modeling deficiency.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1. Main Contributions of The Study

Our work confirms that machine learning is effective for early, non-invasive liver screening, as we developed a pipeline using MICE and SMOTE to address real-world data issues like missing values and imbalance. After testing various models including KNN and XGBoost, we found Random Forest delivered the best performance. Crucially, our models didn't just predict; they aligned with medical reality by correctly identifying bilirubin and albumin as key markers, adding clinical weight to our findings.

6.2. Limitations

We acknowledge three main limitations. First, using only the Mayo Clinic PBC dataset restricts how well our results apply to other populations. Second, while SMOTE helps balance the data, synthetic samples are not a perfect substitute for real rare cases. Lastly, these results come from computer simulations; we have not yet validated the models in an actual hospital setting.

6.3. Future Recommendations

Moving from an academic study to a practical tool requires two major steps. First, we need to look beyond the current dataset by collecting records from diverse hospitals and following patients over time, rather than relying on single snapshots. Second, we must make the tool usable; creating a simple app or integrating directly with hospital EHR systems will allow doctors to get automated risk assessments without adding to their workload.

LIST OF REFERENCES

[1] P. Zhen, “A statistical analysis of chronic liver disease diagnosis with noninvasive biomarkers,” in *Proceedings of the 1st International Conference on Health Big Data and Intelligent Healthcare (ICHIH 2022)*, 2022, pp. 76–82.

[2] M. N. Raihen, M. I. Hossain, and V. Chellamuthu, “Predicting clinical outcomes in liver cirrhosis using machine learning and data balancing technique,” in *2022 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, 2022, pp. 318–324.

[3] E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy, “Prognosis in primary biliary cirrhosis: Model for decision making,” *Hepatology*, vol. 10, no. 1, pp. 1–7, Jul. 1989.

[4] Y. A. Patel and A. J. Muir, “Evaluation of new-onset ascites,” *JAMA*, vol. 316, no. 3, pp. 340–341, Jul. 2016.

[5] C. P. Li, P. Y. Lee, H. Chang, *et al.*, “Spider angiomas in patients with liver cirrhosis: Role of vascular endothelial growth factor and basic fibroblast growth factor,” *World J. Gastroenterol.*, vol. 9, no. 12, pp. 2832–2835, Dec. 2003.

[6] A. K. Kale and D. R. Pandey, “Data pre-processing technique for enhancing healthcare data quality using artificial intelligence,” *Int. J. Sci. Res. Sci. Technol.*, vol. 11, no. 2, pp. 299–305, 2024.

[7] M. A. Konerman *et al.*, “Application of supervised and semi-supervised learning prediction models to predict progression to cirrhosis in chronic hepatitis C,” *Sci. Rep.*, vol. 9, no. 1, p. 10824, Jul. 2019.

[8] M. M. Kaplan and M. E. Gershwin, "Primary biliary cirrhosis," *N. Engl. J. Med.*, vol. 353, no. 12, pp. 1261–1273, Sep. 2005.

[9] P. S. Kamath *et al.*, "A model to predict survival in patients with end-stage liver disease," *Hepatology*, vol. 33, no. 2, pp. 464–470, Feb. 2001.

[10] R. de Franchis, "Expanding consensus in portal hypertension: Report of the Baveno VI Consensus Workshop: Stratifying risk and individualizing care for portal hypertension," *J. Hepatol.*, vol. 63, no. 3, pp. 743–752, Sep. 2015.

[11] W. J. Lammers *et al.*, "Levels of alkaline phosphatase and bilirubin are surrogate endpoints of outcomes of patients with primary biliary cirrhosis: An international follow-up study," *Gastroenterology*, vol. 147, no. 6, pp. 1338–1349, Dec. 2014.

[12] R. N. Pugh, I. M. Murray-Lyon, J. L. Dawson, M. C. Pietroni, and R. Williams, "Transection of the oesophagus for bleeding oesophageal varices," *Brit. J. Surg.*, vol. 60, no. 8, pp. 646–649, Aug. 1973.