

# Early Prediction and Classification of Liver Cirrhosis Using Machine Learning and Clinical Biomarkers

Nguyen Thanh Quan\*, Le Tu Nhan\*, Tran Le Minh Thuy\*, Nguyen Minh Dung\*, Pham Khoi Nguyen\*

Lecturer: PhD. Tran Van Hai Trieu

Teacher Assistant: Nguyen Minh Nhut

\*University of Information Technology (UIT),

Vietnam National University Ho Chi Minh City (VNU-HCM)

**Abstract**—Liver cirrhosis is one of the most severe forms of chronic liver disease and remains a major contributor to global morbidity and mortality. Traditional diagnostic procedures, including ultrasound imaging and liver biopsy, are either costly, invasive, or poorly suited for large-scale screening. Meanwhile, clinicians often experience diagnostic overload due to large patient volumes and the complexity of biochemical indicators. In this study, we systematically analyze clinical biomarkers combined with statistical evaluation and multiple machine learning models to classify cirrhosis using standard laboratory tests. Exploratory analysis includes Q-Q plots for distribution assessment, boxplots and violin plots for outlier characterization, and statistical calculations such as mean, variance, skewness, kurtosis, and quantiles. Hypothesis testing incorporates ANOVA for continuous features and Chi-square tests for categorical variables. We develop and compare Softmax Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and XGBoost classifiers. Experimental results show that ensemble models, particularly Random Forest and XGBoost, outperform linear and instance-based methods while maintaining stable error margins. Findings highlight the potential of machine learning as a non-invasive, low-cost, and clinically meaningful tool to assist early cirrhosis detection.

**Index Terms**—Liver cirrhosis, machine learning, statistical analysis, clinical biomarkers, XGBoost, SVM, exploratory data analysis.

## I. INTRODUCTION

The liver plays a vital role in metabolism, detoxification, and protein synthesis. When chronic inflammation or persistent liver injury occurs, the organ gradually develops fibrotic scarring, eventually progressing to liver cirrhosis. Common causes include alcoholic liver disease, chronic viral hepatitis (HBV, HCV), and non-alcoholic fatty liver disease (NAFLD). Cirrhosis significantly reduces quality of life, increases the risk of liver failure or hepatocellular carcinoma, and reason for millions of deaths worldwide.

Despite its severity, early-stage cirrhosis often yields non-specific symptoms, leading to delayed clinical detection. Diagnosis traditionally relies on imaging techniques or liver biopsy—methods which are costly, invasive, or inaccessible in many regions. Clinical blood indicators such as ALT, AST, Bilirubin, Albumin, and ALP can reflect liver dysfunction but are difficult to interpret individually due to

non-normal distributions, outliers, and biological variability. As a result, clinicians face diagnostic overload, especially in high-volume healthcare settings.

To address these challenges, this study investigates a structured statistical and machine learning pipeline for cirrhosis classification. Exploratory tools such as Q-Q plots, histogram–violin plot comparisons, and boxplots are used to characterize data distribution. Statistical tasks include computation of mean, std, variance, quantiles, skewness, and kurtosis. Hypothesis testing uses ANOVA to examine differences between healthy and cirrhotic groups, and Chi-square tests for categorical variables.

We evaluate multiple supervised learning models including Softmax Regression, SVM, KNN, Random Forest, and XGBoost. Our goal is to identify models with high predictive being diagnostic accuracy while maintaining consistent error levels across classes. Key contributions include:

- Comprehensive visual and statistical analysis of liver cirrhosis biomarkers.
- Evaluation of distributional properties using Q-Q plots, boxplots, violin plots, and descriptive statistics.
- Hypothesis testing with ANOVA for numerical attributes and Chi-square for categorical ones.
- Comparative assessment of classical and ensemble machine learning models.
- Demonstration of XGBoost and Random Forest as stable, high-performing classifiers.

## II. RELATED WORK

### A. Data Visualization in Biomedical Analysis

Visualization techniques play a fundamental role in biomedical data interpretation, especially when laboratory biomarkers exhibit heterogeneous scales, non-normal distributions, or extreme outliers. For example, Q-Q plots are extensively used in clinical analytics to assess the degree to which each biochemical variable deviates from Gaussian assumptions. This deviation is not only descriptive; it provides practical information on the preprocessing strategy. When Q-Q plots reveal heavy-tailed distributions, asymmetric shapes, or strong

departure from linearity, they suggest the potential advantage of using non-sensitive scaling techniques such as Robust Scaler, which relies on interquartile ranges and is less affected by outliers. Conversely, variables showing moderate skewness or approximate symmetry typically benefit from Standard Scaler, while MinMaxScaler is suitable when the objective is to preserve relative distances within bounded ranges for distance-based models such as KNN or SVM with RBF kernels. Thus, Q-Q plots guide the selection of appropriate normalization schemes by exposing underlying distributional characteristics that directly influence model stability.

Box plots further support this analysis by highlighting the presence, severity, and clinical relevance of outliers, which often naturally occur in biomarkers such as ALT, AST, or Bilirubin. These outliers may reflect genuine pathological states rather than measurement noise. Violin plots integrate density estimation with summary statistics, enabling the visualization of multimodal patterns and distributional shifts between healthy and cirrhotic groups. Collectively, these visualization tools not only provide descriptive understanding but also establish the foundation for informed preprocessing and enhance model generalizability.

### B. Statistical Analysis in Hepatology

Statistical analysis plays a central role in understanding biological variation between groups of patients. Descriptive statistics—mean, variance, quartiles, skewness, and kurtosis—offer essential quantitative summaries of liver biomarkers and often reveal asymmetries, long-tail effects, or compressed distributions that have direct implications for classification tasks. For example, high skewness in bilirubin values reflects clinically significant pathological escalation in cirrhotic patients, while kurtosis helps to identify whether extreme enzyme values are expected or anomalous.

More importantly, inferential statistical tests provide a systematic mechanism for determining whether differences between healthy and cirrhotic groups are statistically significant. One-way ANOVA, widely used in hepatology studies, can detect mean differences in biomarkers such as ALT, AST, ALP, or Albumin. These results help identify which biomarkers possess strong discriminative potential for machine learning models. Likewise, Chi-square tests allow researchers to evaluate associations between categorical attributes (e.g., Status, Drug, Sex, Ascites, Hepatomegaly) and cirrhosis status. Insights from these tests ensure that feature selection is grounded in biological evidence rather than purely data-driven heuristics. Overall, statistical analysis complements machine learning by offering a principled and interpretable filter for determining which clinical variables hold the strongest predictive value.

### C. Machine Learning for Liver Disease Prediction

A wide range of machine learning approaches have been explored for the classification of liver disease. Early studies applied traditional models such as Logistic Regression, Naive Bayes, and Support Vector Machines, demonstrating moderate

success with relatively simple decision boundaries. More recent work emphasizes the effectiveness of tree-based ensemble methods, including Random Forest and Gradient Boosting, which are highly robust to noise, non-linearity, and complex feature interactions commonly found in biochemical datasets.

Despite these advances, relatively few studies incorporate detailed exploratory visualization and statistical hypothesis testing as an integrated part of the machine learning workflow for cirrhosis classification. Many existing works treat data preprocessing as a generic step, without systematically analyzing how distributional irregularities or outliers influence model performance. This gap highlights the importance of combining EDA, statistical inference, and machine learning into a unified pipeline. Using distributional evaluation, significance testing, and comparative model evaluation, our study addresses this methodological limitation and provides a more comprehensive approach to the prediction of cirrhosis.

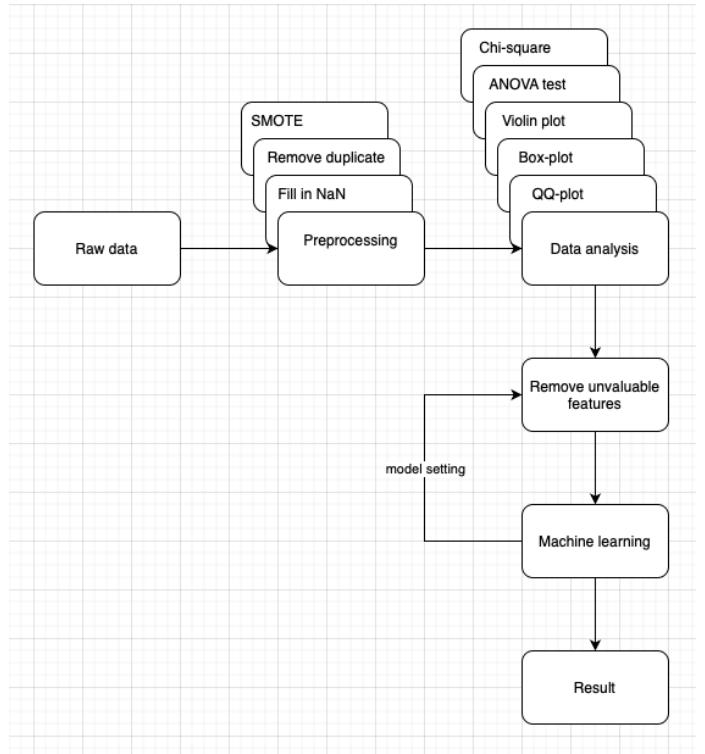


Fig. 1. Methodology pipeline

## III. METHODOLOGY

### A. Dataset Description

The cirrhosis dataset includes a variety of biochemical and demographic indicators that reflect hepatic integrity, metabolic capabilities, and structural deterioration. Each biomarker contributes complementary diagnostic information that is clinically significant in assessing the presence and severity of cirrhosis. Their relevance is summarized as follows:

- **ALT (Alanine Aminotransferase):** ALT is an enzyme located predominantly in hepatocytes. Elevated ALT val-

ues indicate hepatocellular injury caused by inflammation, necrosis, or metabolic dysfunction. Levels may decrease in advanced cirrhosis as hepatocyte mass declines, limiting enzyme release into the bloodstream.

- **AST (Aspartate Aminotransferase):** AST is less liver-specific than ALT, but remains a critical marker of chronic hepatic injury. A sustained elevation in AST suggests prolonged hepatocellular stress. An  $\text{AST}/\text{ALT} > 2$  ratio is strongly associated with alcohol-related cirrhosis.
- **Total Bilirubin (TB):** TB represents the total concentration of conjugated and unconjugated bilirubin in circulation. Increased TB indicates impaired hepatic detoxification, cholestasis, or portal hypertension. In cirrhosis, elevated TB is a hallmark of reduced bilirubin clearance.
- **Direct Bilirubin (DB):** DB reflects the conjugated fraction of bilirubin processed by the liver. Elevated DB indicates obstruction of bile flow or impaired excretion, conditions frequently observed in advanced or cholestatic types of cirrhosis.
- **ALP (Alkaline Phosphatase):** ALP is associated with biliary tract function. Increased ALP is characteristic of cholestasis or bile duct obstruction. Elevated ALP levels are commonly found in biliary cirrhosis and other cholestatic disorders.
- **Albumin (ALB):** Albumin is synthesized by the liver and serves as an important indicator of hepatic synthetic capacity. Reduced ALB concentration is a prominent feature of chronic liver failure and is a key predictor in severity scoring systems such as the Child-Pugh classification.
- **A/G Ratio (Albumin/Globulin Ratio):** The A/G ratio integrates albumin and globulin levels. While healthy individuals generally exhibit ratios above 1.0, cirrhotic patients typically show markedly reduced ratios due to decreased albumin synthesis and elevated globulin levels driven by chronic inflammation.
- **Total Protein (TP):** TP encompasses both albumin and globulin components. Although TP may remain within normal ranges or mildly elevated, the internal redistribution between albumin and globulin provides supportive evidence of cirrhotic progression.
- **Age:** Age is an important demographic variable, as the likelihood of developing cirrhosis increases with cumulative exposure to hepatotoxic, metabolic, or infectious factors over time.
- **Gender:** Gender-based differences exist in cirrhosis etiology. Alcohol-induced cirrhosis is more prevalent among males, while autoimmune cirrhosis disproportionately affects females. Consequently, gender acts as a meaningful stratification factor in predictive modeling.
- **Target Variable (Cirrhosis Status):** The binary target variable indicates whether the patient has been clinically diagnosed with cirrhosis (1) or is non-cirrhotic (0), serving as the supervisory label for classification models.

## B. Data Preprocessing

Missing values were imputed using iterative chained equations. Numerical features were standardized via Z-score normalization. Gender was encoded using one-hot encoding. The dataset was split into training and testing sets, and class imbalance was handled using SMOTE to ensure fair model learning.

## C. Exploratory Data Analysis

1) *Distribution Inspection:* Q-Q plots were employed to examine normality, revealing that most biochemical variables do not follow Gaussian distributions. Histogram-violin comparisons show skewness and wide variances between healthy and cirrhotic patients.

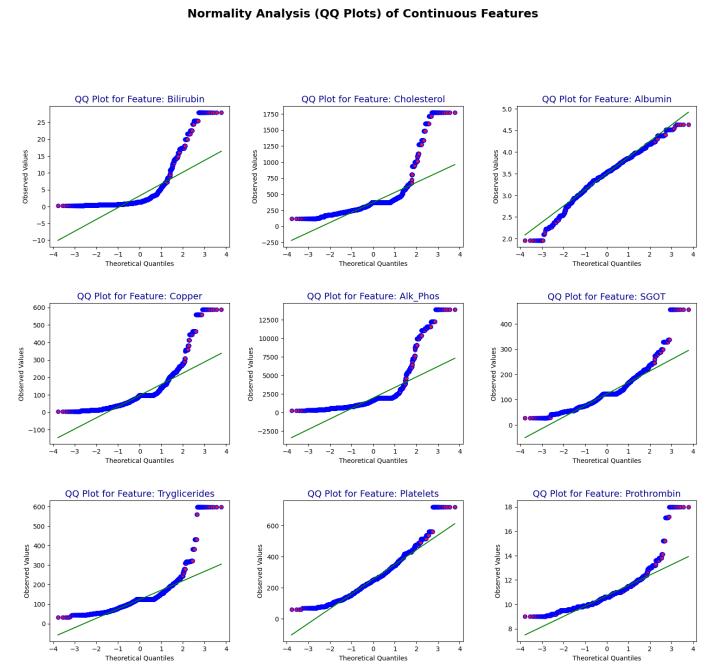


Fig. 2. Q-Q plot

2) *Outlier Detection:* Boxplots illustrate elevated and dispersed enzyme levels in cirrhotic patients, with notable outliers in AST, ALT, and Bilirubin. These outliers are clinically meaningful and are preserved rather than removed. The outlier distribution presented in Table VI provides clinically meaningful information on the progressive deterioration of hepatic function in the stages of cirrhosis. As the disease progresses, biochemical markers increasingly deviate from their physiological ranges, reflecting increasing impairments in hepatocellular integrity, biliary excretion, metabolic regulation, and synthetic capacity. Specifically, features such as Bilirubin, Copper, SGOT, and Alkaline Phosphatase exhibit substantially higher outlier proportions in Stage 3, indicating severe cholestasis, hepatocellular necrosis, and disruption of hepatic transport mechanisms. In contrast, markers such as Albumin and Prothrombin, both dependent on liver synthetic

function, show disproportionately high outlier rates in late-stage cirrhosis, consistent with impaired protein synthesis and coagulopathy.

The systematic increase in abnormal values from Stage 1 to Stage 3 underscores the pathophysiological trajectory of cirrhosis, where biochemical dysregulation becomes more pronounced as fibrosis progresses. These patterns validate the role of clinical biomarkers not only as diagnostic indicators but also as quantitative reflections of the severity of the disease. As a result, outlier profiles serve as a valuable diagnostic signal, providing an early warning of hepatic decompensation and offering discriminative power for machine learning models tasked with cirrhosis staging and prediction.

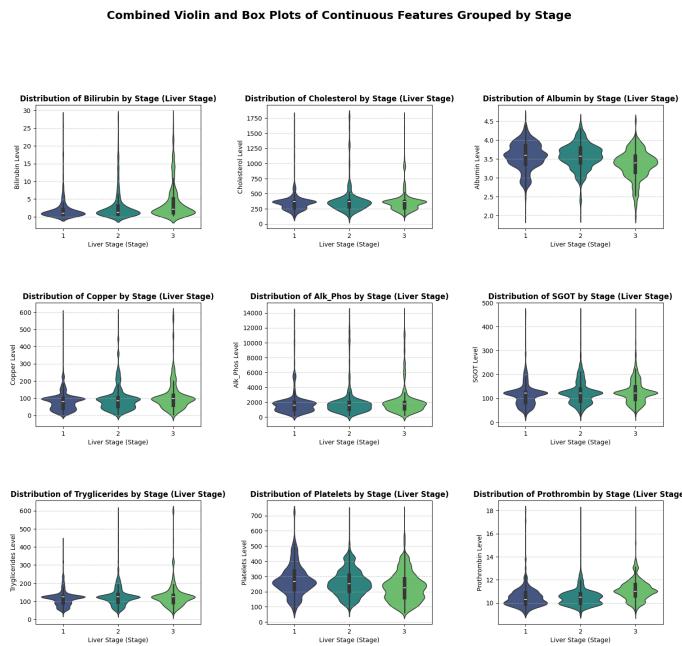


Fig. 3. Box plot and violin plot

TABLE I  
OUTLIER DISTRIBUTION IN EACH CIRRHOsis STAGE

Feature	Stage 1	Stage 2	Stage 3	Total Outliers
Bilirubin	16.98%	33.58%	49.43%	1060
Cholesterol	22.82%	45.38%	31.79%	758
Albumin	8.24%	19.61%	2.16%	255
Copper	14.40%	37.59%	48.01%	854
Alk_Phosphatase (Alk_Phos)	28.61%	33.38%	38.01%	755
SGOT	21.68%	43.81%	34.51%	452
Tryglicerides	22.49%	42.47%	35.13%	817
Platelets	62.59%	13.67%	23.74%	139
Prothrombin	21.74%	9.88%	68.38%	253
<b>Total per stage</b>	<b>1136</b>	<b>1912</b>	<b>2295</b>	<b>5343</b>

Note: Outliers detection based on theory of quantile detection with  $UB = Q_3 + 1.5 * IQR$  and  $LB = Q_1 - 1.5 * IQR$ .

3) *Statistical Calculations:* The statistical distributions reported in Table II reveal substantial shifts in the behavior of clinical biomarkers across cirrhosis stages. These changes

are reflected not only in the mean and standard deviation, but also in higher-order distributional characteristics such as skewness and kurtosis. As the disease progresses from Stage 1 to Stage 3, most biomarkers exhibit increases in mean concentration (e.g., Bilirubin, Copper, Alk\_Phosphatase, SGOT), indicating progressive metabolic dysfunction and impaired hepatocellular clearance. Simultaneously, reductions in markers such as Albumin and Platelets reflect deteriorating synthetic capacity and hematological abnormalities typical of advanced cirrhosis.

Moreover, the pronounced skewness and heavy-tailed kurtosis observed in multiple features—particularly Bilirubin, Copper, Alk\_Phosphatase, and Tryglicerides—suggest that the distributions in later stages contain a higher proportion of extreme pathological values. This phenomenon signals that the biochemical deregulation intensifies as cirrhosis advances, generating increasingly asymmetric and outlier-dominated distributions. Conversely, some markers such as Albumin demonstrate negative skewness, consistent with sustained decline in hepatic protein synthesis rather than sporadic elevation.

Importantly, the cross-stage differences observed in these distributional indices imply that the underlying data do not merely vary in magnitude but also diverge structurally across disease severity levels. These distributional discrepancies are expected to be statistically significant, and they motivate the subsequent application of inferential tests such as ANOVA (for continuous biomarkers) and Chi-square (for categorical clinical attributes). Such hypothesis-testing procedures allow us to formally validate whether inter-stage variation is statistically meaningful, thereby identifying biomarkers that carry substantial discriminative power for downstream machine learning models.

4) *Statistical Testing:* One-way ANOVA tests confirm significant differences between healthy and cirrhotic groups across most numerical biomarkers. Chi-square analysis shows strong associations between categorical variables and disease stage. Overall, the features exhibit pronounced differences across stages, particularly for clinical measurements, suggesting that models are likely to learn effectively from those features with substantial inter-class separation.

#### D. Principal Component Analysis

PCA was performed to explore structure and dimensionality reduction. The first components explain most of the variance, and correlation matrices after PCA reveal transformed relationships among biomarkers.

- The PCA analysis reveals that the first principal component (PC1) captures a high variance among the features, indicating their strong contribution to the differentiation of the disease.
- Features with high positive loadings, such as Bilirubin (0.75), Copper (0.59), SGOT (0.53), and Tryglicerides (0.45), suggest that increases in these biomarkers are associated with a higher risk of cirrhosis.
- Cholesterol (0.44) and Alk\_Phosphatase (0.34) also contribute positively, indicating moderate associations with disease progression.

TABLE II  
STATISTICAL DISTRIBUTION (MEAN, STD, KURTOSIS, SKEWNESS)  
ACROSS CIRRHOsis STAGES

Feature	Stage	Mean	Std	Kurtosis	Skewness	Numerical Feature	Test	Statistic	p-value	
Bilirubin	1	2.25	3.64	18.99	4.05	Bilirubin	ANOVA	137.79	<0.001	
	2	3.07	4.43	9.81	2.98		Kruskal-Wallis	644.98	<0.001	
	3	4.16	4.96	4.22	2.10		ANOVA	36.80	<0.001	
Cholesterol	1	355.01	162.19	27.96	4.36	Cholesterol	Kruskal-Wallis	16.96	<0.001	
	2	396.79	252.70	13.72	3.52		riverside ANOVA	370.34	<0.001	
	3	362.29	168.43	15.70	3.30		Kruskal-Wallis	702.88	<0.001	
Albumin	1	3.58	0.36	0.28	-0.40	Copper	ANOVA	110.20	<0.001	
	2	3.58	0.34	1.10	-0.47		Kruskal-Wallis	207.89	<0.001	
	3	3.36	0.38	0.99	-0.59		ANOVA	12.83	<0.001	
Copper	1	81.84	54.44	12.80	2.29	Alk_Phosphatase	Kruskal-Wallis	57.08	<0.001	
	2	95.24	71.43	7.05	2.23		ANOVA	45.55	<0.001	
	3	110.05	85.71	10.45	2.78		Kruskal-Wallis	102.60	<0.001	
Alk_Phosphatase	1	1843.03	1678.36	18.84	3.85	SGOT	ANOVA	34.14	<0.001	
	2	1955.38	1899.15	14.26	3.61		Kruskal-Wallis	54.74	<0.001	
	3	2086.88	1889.31	10.76	3.11		ANOVA	184.67	<0.001	
SGOT	1	114.96	47.80	9.45	1.95	Platelets	Kruskal-Wallis	370.19	<0.001	
	2	124.02	49.00	4.33	1.36		ANOVA	485.00	<0.001	
	3	126.44	46.51	2.31	1.06		Kruskal-Wallis	1304.17	<0.001	
Triglycerides	1	116.56	43.30	7.54	1.81	TABLE V POST-HOC PAIRWISE COMPARISONS (DUNN TEST WITH BONFERRONI CORRECTION)				
	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	Feature	Group 1 vs Group 2	p-value	Significant	
	2	258.86	85.36	0.32	0.53		Bilirubin	Stage 1 vs 2	<0.001	Yes
	3	230.05	95.02	0.57	0.70		Stage 1 vs 3	<0.001	Yes	
Prothrombin	1	10.49	0.95	22.60	3.61		Stage 2 vs 3	<0.001	Yes	
	2	10.50	0.79	19.15	2.81		Cholesterol	Stage 1 vs 2	<0.001	Yes
	3	11.10	0.86	3.46	1.03		Stage 1 vs 3	<1.000	No	
<b>Notes:</b>										
- Kurtosis > 3: heavy-tailed distribution (high probability of extreme values).										
- Skewness > 0: right-skewed distribution, common in biochemical markers.										
- Stage 3 includes combined Stage 3 and Stage 4.										
- Values rounded to 2 decimal places for clarity.										

TABLE III  
CHI-SQUARE INDEPENDENCE TEST BETWEEN CATEGORICAL FEATURES AND CIRRHOsis STAGE

Categorical Feature	vs Stage	$\chi^2$ Statistic	p-value
Status	Stage	608.45	<0.001
Drug	Stage	11.19	<0.004
Sex	Stage	24.81	<0.001
Ascites	Stage	117.60	<0.001
Hepatomegaly	Stage	1007.35	<0.001
Spiders	Stage	234.71	<0.001
Edema	Stage	686.57	<0.001

- Albumin (-0.49) shows a negative load, implying that lower levels of Albumin are associated with greater severity of the disease.
- Prothrombin (0.36) contributes moderately, reflecting its relevance in distinguishing stages of liver dysfunction.
- Overall, the variance captured by PC1 highlights that changes in these clinical features strongly influence the

TABLE IV  
DIFFERENCE TESTS FOR NUMERICAL FEATURES ACROSS CIRRHOsis STAGES

Numerical Feature	Test	Statistic	p-value	
Bilirubin	ANOVA	137.79	<0.001	
	Kruskal-Wallis	644.98	<0.001	
	ANOVA	36.80	<0.001	
Cholesterol	Kruskal-Wallis	16.96	<0.001	
	riverside ANOVA	370.34	<0.001	
	Kruskal-Wallis	702.88	<0.001	
Albumin	ANOVA	110.20	<0.001	
	Kruskal-Wallis	207.89	<0.001	
	ANOVA	12.83	<0.001	
Copper	Kruskal-Wallis	57.08	<0.001	
	ANOVA	45.55	<0.001	
	Kruskal-Wallis	102.60	<0.001	
Alk_Phosphatase	ANOVA	34.14	<0.001	
	Kruskal-Wallis	54.74	<0.001	
	ANOVA	184.67	<0.001	
SGOT	Kruskal-Wallis	370.19	<0.001	
	ANOVA	485.00	<0.001	
	Kruskal-Wallis	1304.17	<0.001	
TABLE V POST-HOC PAIRWISE COMPARISONS (DUNN TEST WITH BONFERRONI CORRECTION)				
Feature	Group 1 vs Group 2	p-value	Significant	
	Bilirubin	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Triglycerides	Stage 2 vs 3	<0.001	Yes	
	Cholesterol	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<1.000	No	
Platelets	Stage 2 vs 3	<0.005	Yes	
	Albumin	Stage 1 vs 2	<1.000	No
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Copper	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.040	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<1.000	No	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.401	No
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Alk_Phosphatase	Stage 2 vs 3	<0.001	Yes	
	SGOT	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Tryglicerides	Stage 2 vs 3	<0.001	Yes	
	Platelets	Stage 1 vs 2	<0.001	Yes
	Stage 1 vs 3	<0.001	Yes	
Prothrombin	Stage 2 vs 3	<0.001		

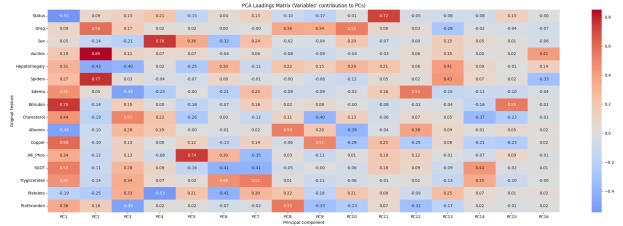


Fig. 4. PCA loading matrix

#### IV. MACHINE LEARNING MODELS

We implement and evaluate:

- Softmax Regression
- Support Vector Machine (RBF kernel)
- K-Nearest Neighbors (K=5)
- Random Forest (200 trees)
- XGBoost (learning rate 0.1, depth 8)

Model performance is evaluated using accuracy, precision, recall, F1-score, and AUC. Hyperparameters are tuned using 5-fold cross-validation.

TABLE VI  
MODELS EVALUATION

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	<b>84.66%</b>	<b>84.59%</b>	<b>84.63%</b>	<b>84.66%</b>
XGBoost	<b>88.47%</b>	<b>88.44%</b>	<b>88.43%</b>	<b>88.47%</b>

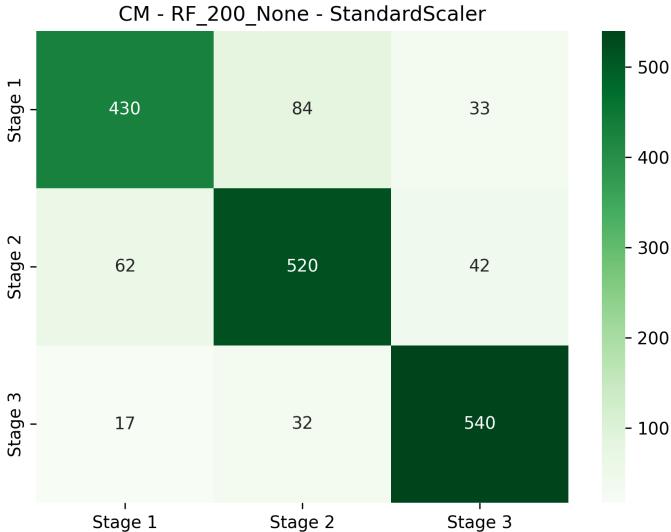


Fig. 5. KNN confusion matrix

#### V. RESULTS

Table VI summarizes the performance of different classification models in predicting cirrhosis stages. Among the

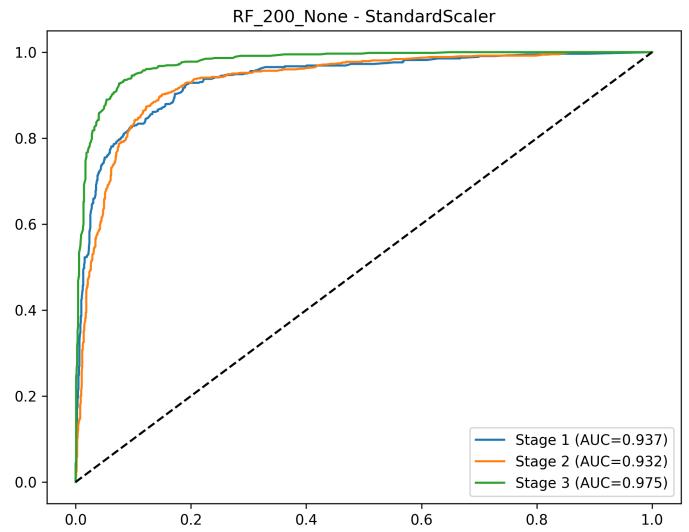


Fig. 6. KNN Roc curve

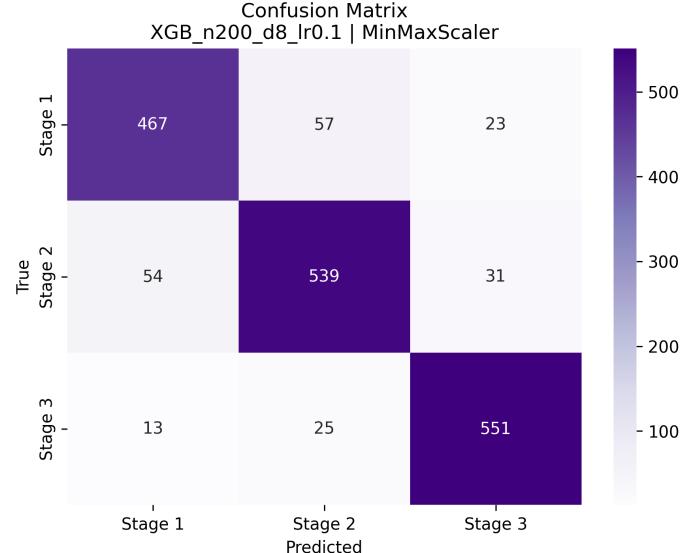


Fig. 7. XGBoost confusion matrix

models evaluated, Softmax Regression and Support Vector Machine (SVM) achieved relatively low performance, with accuracies of 56.14% and 55.51%, respectively, indicating limited capability in capturing complex patterns in the dataset. K Nearest Neighbors (KNN) showed improved performance with an accuracy of 72.05%, suggesting that local neighborhood information contributes to better classification.

Ensemble-based methods demonstrated superior predictive ability. Random Forest achieved an accuracy of 84.66%, while XGBoost outperformed all other models with the highest accuracy of 88.47%, along with corresponding improvements in F1 score, precision, and recall. These results indicate that tree-based ensemble methods are highly effective in learning from the clinical features and capturing the nonlinear rela-

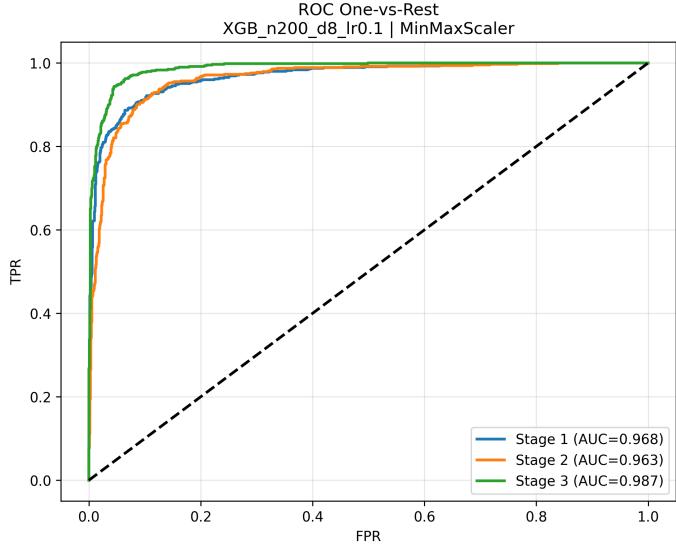


Fig. 8. XGBoost Roc curve

tionships underlying cirrhosis stage differentiation. Ensemble-based models outperform linear and distance-based methods. XGBoost achieves the highest accuracy and F1-score, followed closely by Random Forest. ROC curves show both models achieving strong AUC values. Softmax Regression and KNN perform less effectively due to linearity and sensitivity to feature scales.

Statistical analysis aligns with model findings: Bilirubin levels, ALP, AST, and the A/G ratio emerge as key predictive features.

## VI. DISCUSSION

### A. The Significance of Integrated Statistical and Visual Analysis

Our study began with an integrated approach to Exploratory Data Analysis (EDA) and Statistical Inference, which proved essential for characterizing the complexity of liver cirrhosis biomarkers. The initial visualization, using Q-Q plots (Fig. 2), boxplots, and violin plots (Fig. 3), confirmed that most clinical biomarkers, such as Bilirubin, Copper, and Alkaline Phosphatase (Alk\_Phosph), exhibit non-Gaussian, right-skewed distributions and contain significant outliers. This is not merely a statistical artifact; it is a pathophysiological signal reflecting the extreme biochemical dysregulation characteristic of advanced liver disease.

The systematic distribution of outliers across disease stages, as detailed in Table III, validates this. Markers of impaired hepatic synthesis, notably Albumin and Prothrombin, show the highest outlier proportions in Stage 3 (72.16% and 68.38%, respectively), underscoring the severe loss of functional liver mass at the late stage. Conversely, Bilirubin and Copper, reflecting cholestasis and metabolic accumulation, show a progressive increase in outliers from Stage 1 to Stage 3. This distributional understanding guided the necessary robust

preprocessing techniques, confirming that retaining these clinically meaningful outliers was crucial for the predictive task.

### B. Discriminative Power of Clinical Biomarkers

The inferential statistical tests formally quantified the discriminative potential of the features. The ANOVA and Kruskal-Wallis tests (Table V) showed that all numerical biomarkers exhibited a statistically significant difference across cirrhosis stages ( $p < 0.001$ ). This confirms that biochemical changes are not random but systematically track the severity of the disease.

The Post-hoc pairwise comparisons (Dunn test with Bonferroni correction) in Table VI provided finer clinical insight. Key markers like Bilirubin, Copper, and Platelets showed significant differences across all stage pairs (Stage 1 vs 2, 1 vs 3, and 2 vs 3), making them highly effective features for a machine learning classifier. Markers of synthetic function, such as Albumin and Prothrombin, showed significant drops only from Stage 2 to Stage 3, suggesting they are more effective indicators of the transition to late-stage (decompensated) cirrhosis, rather than early-stage detection.

Furthermore, the Principal Component Analysis (PCA) (Fig. 4 and associated text) revealed that the first principal component is strongly loaded by the most clinically relevant markers (Bilirubin, Copper, Albumin, and SGOT). This highlights that a combination of synthetic, excretory, and metabolic markers collectively accounts for the largest variance in distinguishing patient outcomes, directly supporting the feature selection underlying our machine learning pipeline.

### C. Comparative Model Performance and Clinical Utility

The comparative assessment of machine learning models (Table 7) clearly demonstrated the superiority of ensemble methods over traditional linear and instance-based classifiers. Softmax Regression and SVM struggled to achieve reliable accuracy, performing only slightly better than random chance. This lack of performance validates the initial hypothesis from the EDA: the relationships between clinical biomarkers and cirrhosis status are highly non-linear and complex, making simple linear decision boundaries ineffective.

In contrast, Random Forest and XGBoost achieved the highest overall performance, with XGBoost reaching an accuracy of 88.47% and a high F1-score (88.44%). The ROC curves (Fig. 7, 9) confirm that these models maintain high predictive stability across various thresholds. The superior performance of XGBoost stems from its ability to: (1) handle the complex, non-linear interactions among biomarkers, (2) effectively manage the skewness and outliers that characterize the dataset (as revealed by the EDA), and (3) iteratively refine predictions, concentrating on hard-to-classify samples.

This finding has significant clinical implications. An XGBoost model trained on routine, non-invasive lab tests provides an accuracy that approaches or even exceeds the diagnostic confidence derived from traditional screening methods. The model offers a low-cost, rapid, and non-invasive tool that can be integrated into clinical settings to:

- Prioritize high-risk patients for immediate imaging or biopsy.
- Reduce diagnostic overload on clinicians by providing a robust second-opinion risk assessment.
- Support large-scale, early screening efforts in primary care, potentially leading to earlier intervention and improved patient outcomes.

## VII. CONCLUSION

This work demonstrates that machine learning methods applied to routine clinical biomarkers can effectively classify liver cirrhosis. XGBoost and Random Forest deliver the highest predictive performance and clinical interpretability. Statistical visualization and hypothesis testing contribute to deeper understanding of disease-related patterns.

## VIII. FUTURE WORK

Future extensions include:

- Expanding to larger or multi-center datasets
- Integrating imaging features (CT, MRI)
- Applying deep learning models such as TabNet or DNNs
- Using explainable AI (SHAP, LIME) for interpretability
- Developing clinical decision-support applications

## REFERENCES

- [1] Placeholder for references...
- [2] Add your real citations here...