STAT3013.Q11.CTTT

VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

UNIVERSITY OF INFORMATION TECHNOLOGY

STATISTICAL ANALYSIS

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

Group 3

Nguyễn Thanh Quân 23521266

Phạm Khôi Nguyên 23521055

Trần Lê Minh Thùy 23521560

Nguyễn Minh Dũng 23520333

Lê Tự Nhân 23521076

STAT3013.Q11.CTTT - Group 3

TABLE OF CONTENTS

1. Introduction

2. Data Analysis (EDA & Stats)

3. Data Preprocessing

3. Machine Learning implementation & Experiment analysis

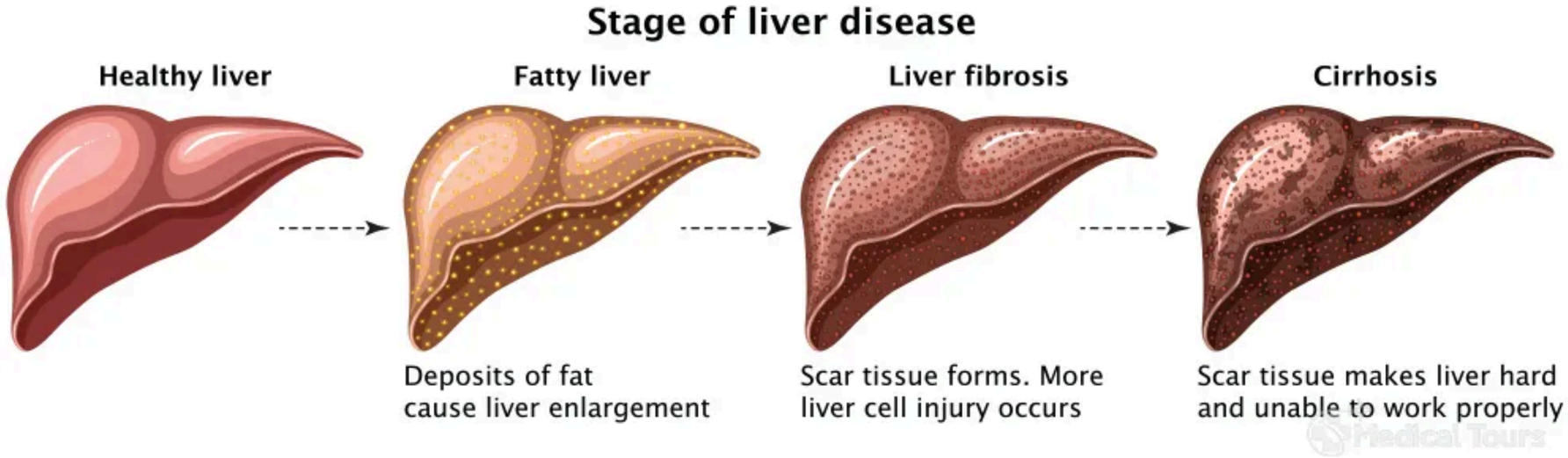
4. Result & Conclusion

TABLE OF CONTENTS

1. Introduction

INTRODUCTION

Problem Statement & Objectives



Problem: Traditional diagnosis (biopsy) is invasive, risky, and costly; biochemical markers change gradually and non-linearly

(e.g., Stage 1 vs. Stage 2) is difficult because symptoms often overlap or remain non-specific until the disease is advanced.

INTRODUCTION

Objectives & Data Overview

Objective: Build a multi-class classification system (Stage 1, 2, 3) to predict liver cirrhosis stages.

Data Overview: Focus on biochemical indicators; excludes imaging (CT/MRI) or genomic data.

- Original Source: Mayo Clinic PBC Dataset (418 patients).
- Augmentation: Synthetic data expansion to ~25,000 samples to enhance model training.
- Preprocessing: Rigorous deduplication applied.
- Final Dataset: 8,796 unique records used for analysis.

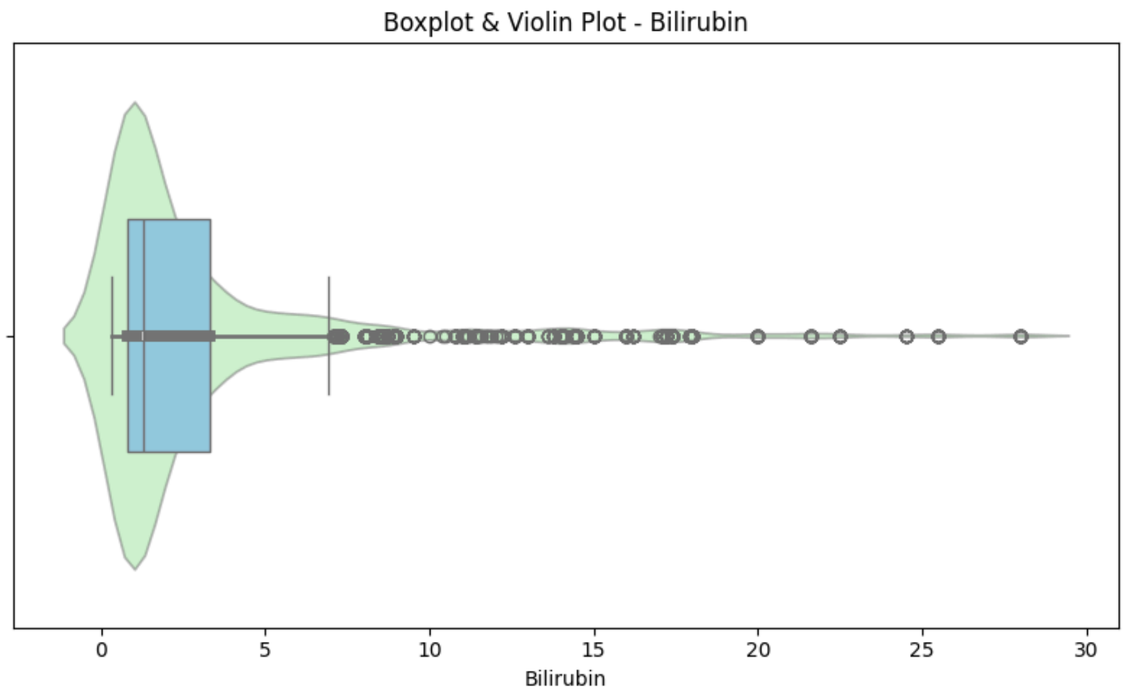
418 Raw → 25k Synthetic → 8,796 Final

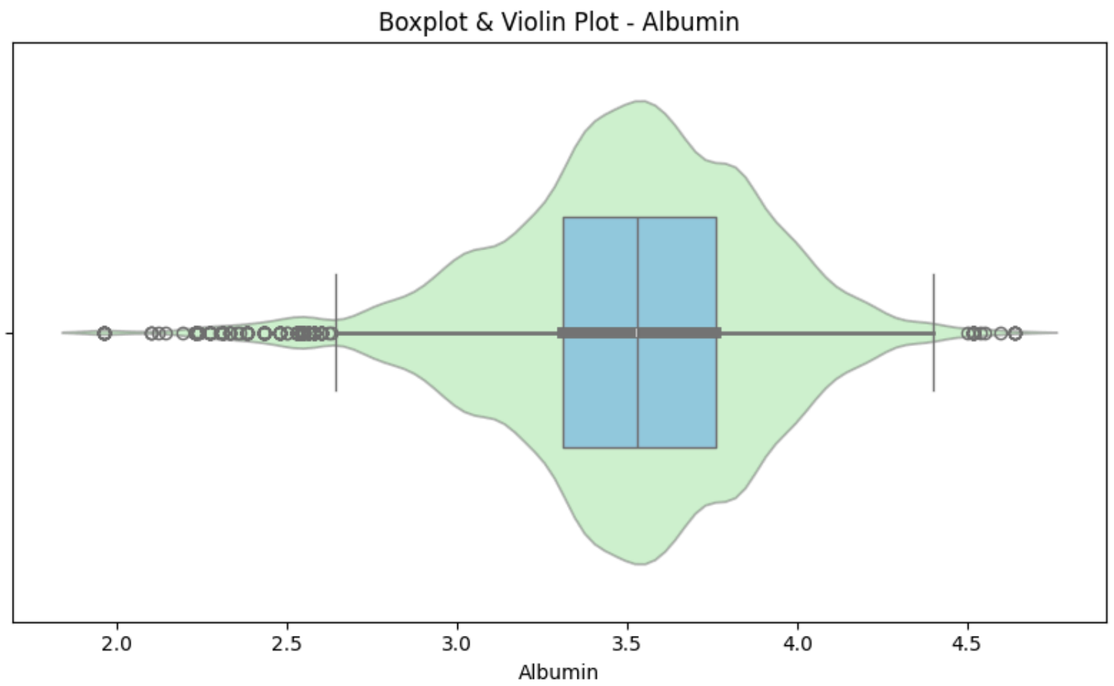
TABLE OF CONTENTS

2. Data Analysis
(EDA & Stats)

Data Analysis (EDA & Stats)

Data distribution analysis





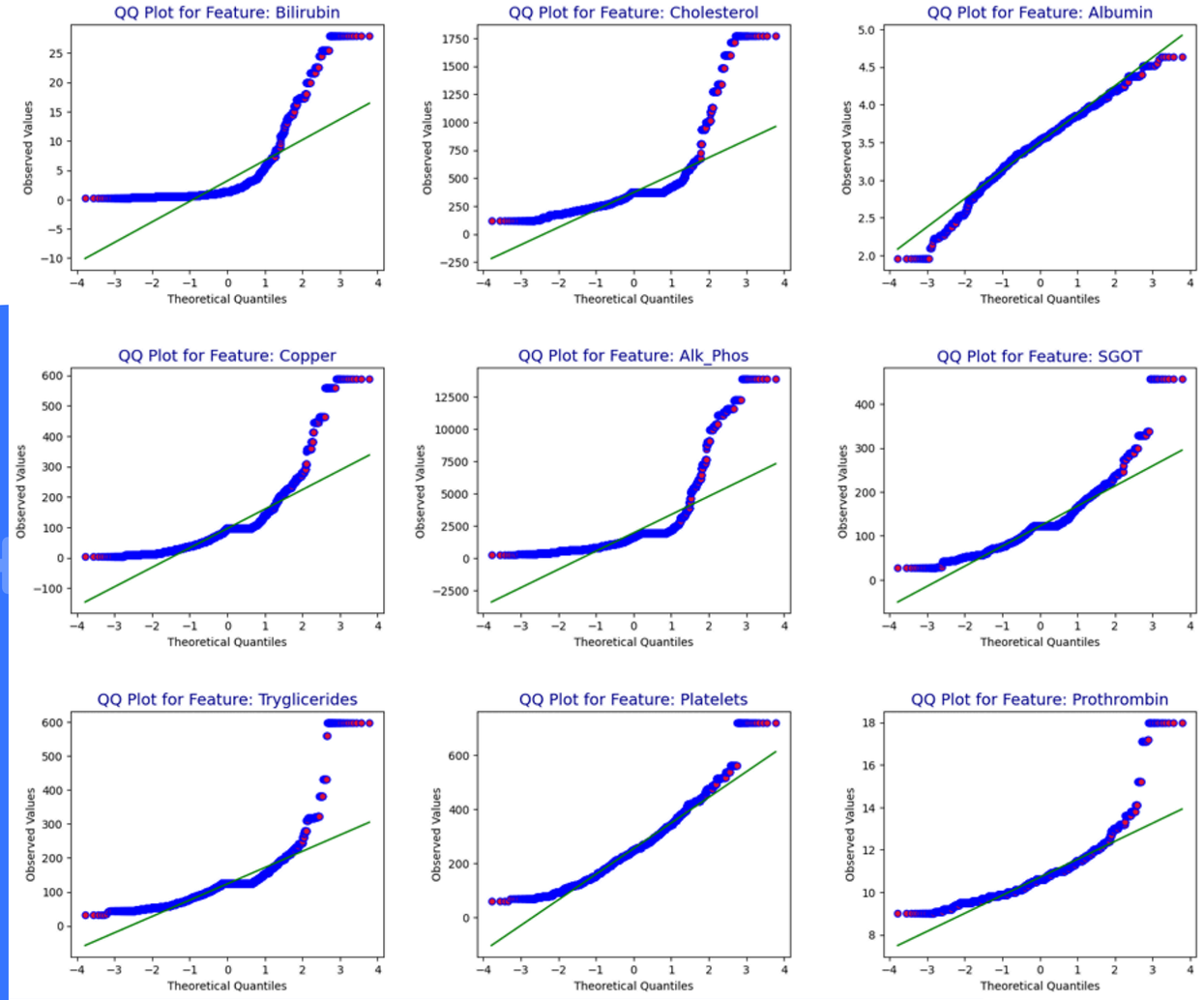
Distribution: Most features (e.g., Bilirubin) exhibit a heavy right-skewed distribution with significant outliers in advanced stages.

Trend Analysis:

- Bilirubin: heavy right-skewed distribution
- Albumin: Shows a distinct negative shift, decreasing as the disease progresses.

Data Analysis (EDA & Stats)

Data distribution analysis



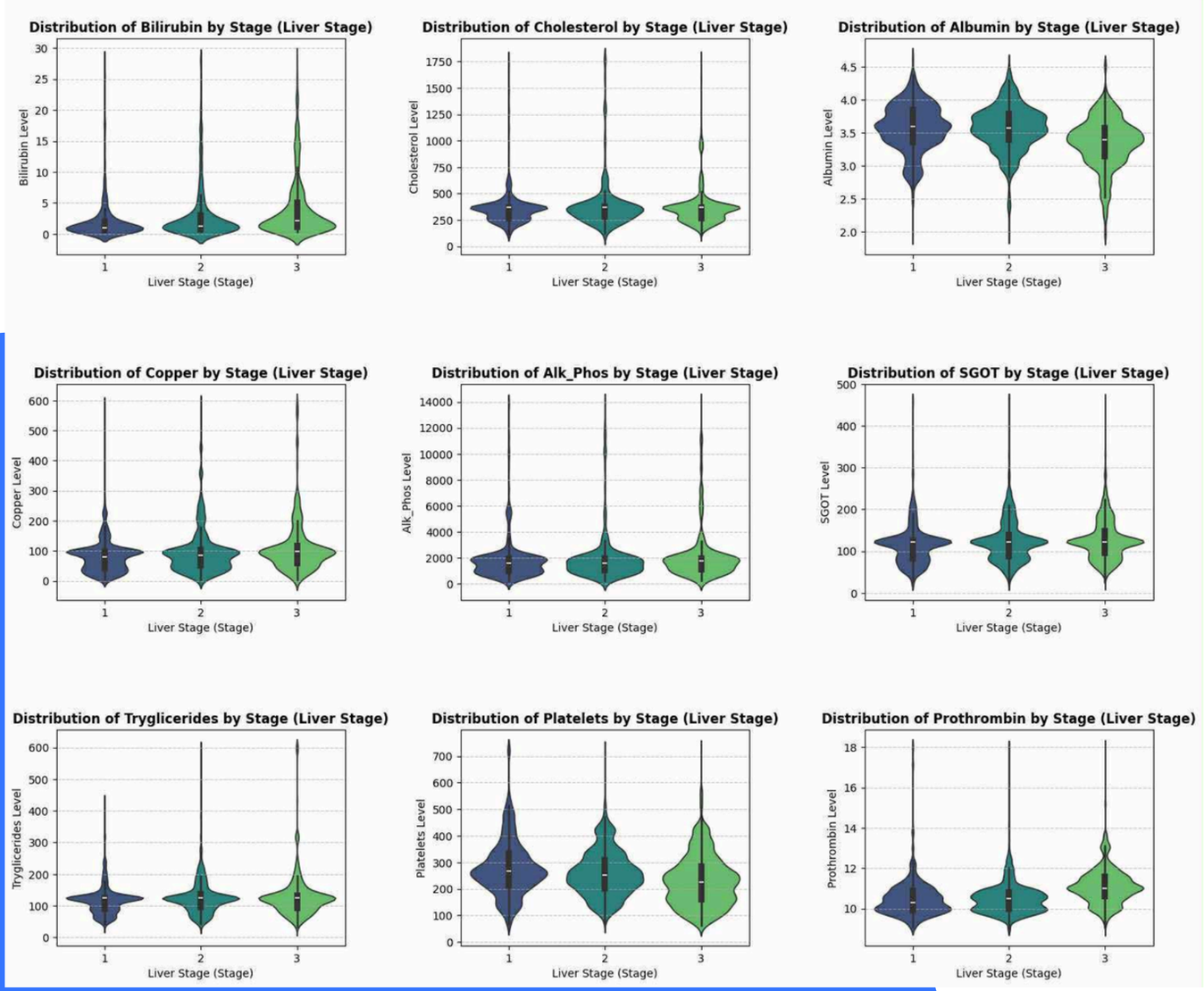
Numerous clinical variables exhibit approximately normal behavior, frequently reflected through a characteristic S-shaped curvature or close adherence to the reference line.

=> non-normal distribution.

Therefore, parametric tests (ANOVA) are unsuitable; Non-parametric tests (Kruskal-Wallis) are required.

Data Analysis (EDA & Stats)

Data distribution analysis



Distribution: The differences in data visualization between cirrhosis stages is an evidence for the use of linear regression analysis is adapted.

Data Analysis (EDA & Stats)

Shapiro-Wilk normality test and Levene's test

Variable	W Statistic	p-value	Normal Distribution?	Variable	Levene Statistic	p-value	Equal Variance?
Bilirubin	0.6075	0.0	No	Bilirubin	94.9778	0.0	No
Cholesterol	0.5979	0.0	No	Cholesterol	35.452	0.0	No
Albumin	0.9837	0.0	No	Albumin	8.157	0.000289	No
Copper	0.7669	0.0	No	Copper	26.132	0.0	No
Alk_Phos	0.5965	0.0	No	Alk_Phos	2.9895	0.050363	Yes
SGOT	0.9118	0.0	No	SGOT	2.4378	0.087412	Yes
Tryglicerides	0.7613	0.0	No	Tryglicerides	15.851	0.0	No
Platelets	0.9702	0.0	No	Platelets	16.3793	0.0	No
Prothrombin	0.8738	0.0	No	Prothrombin	15.5275	0.0	No

The results of both the Shapiro-Wilk normality test and Levene's test indicate that the dataset violates the assumptions of normal distribution and equal variances. Given that the sample size is relatively small, so the use of parametric tests (such as ANOVA or linear regression residual normality) are not satisfied.

Data Analysis (EDA & Stats)

Chi-square analysis and Kruskal-Wallis H-Test

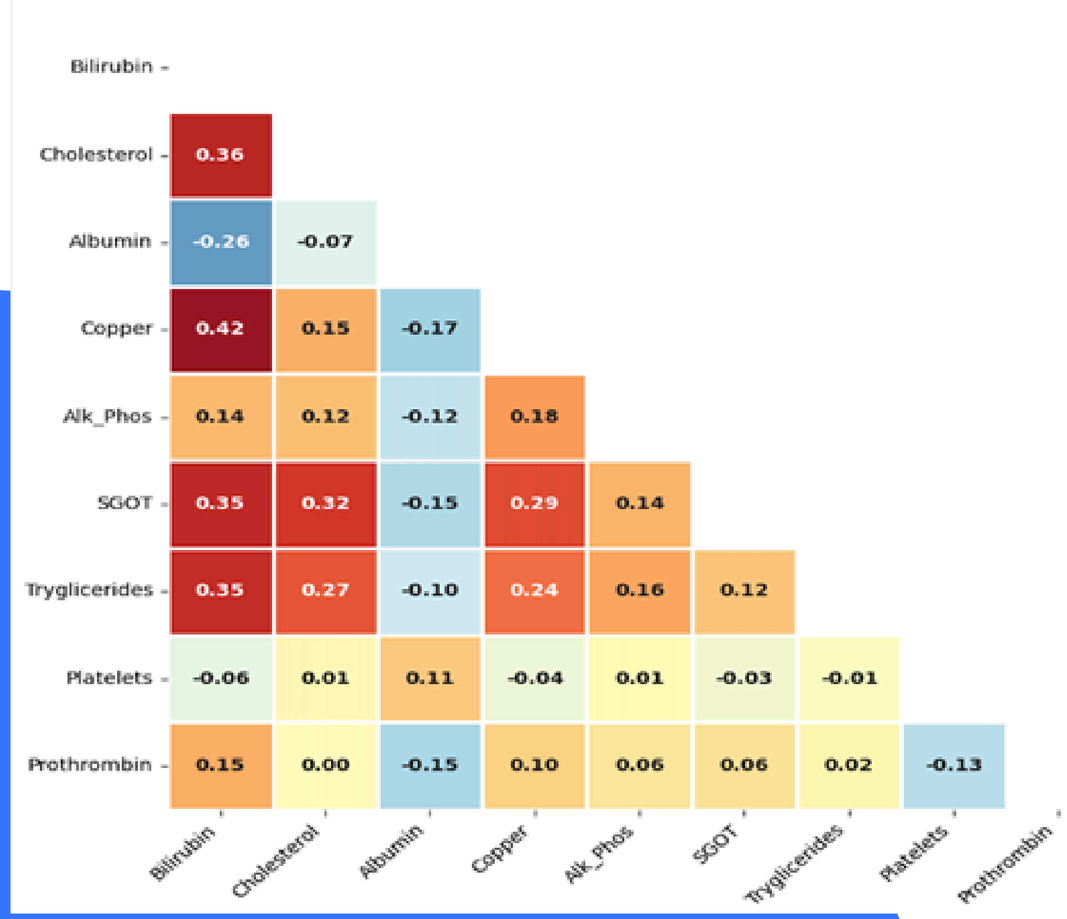
Variable	H Statistic	p-value	Significant (α=0.05)?
Prothrombin	1304.1662	6.37e-284	Yes
Albumin	702.8835	2.35e-153	Yes
Bilirubin	644.9835	8.78e-141	Yes
Platelets	370.1943	4.11e-81	Yes
Copper	207.8853	7.22e-46	Yes
SGOT	102.5947	5.27e-23	Yes
Alk_Phos	57.0796	4.03e-13	Yes
Tryglicerides	54.7403	1.30e-12	Yes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

These results indicate statistically significant differences in variable distributions across cirrhosis stages.

Variable 1	Variable 2	Chi2 Statistic	Degrees of Freedom	p-value	Significant (α=0.05)?
Ascites	Spiders	3599.649	1	0.00e+00	Yes
Drug	Ascites	1259.894	1	5.87e-276	Yes
Hepatomegaly	Stage	1007.353	2	1.80e-219	Yes
Ascites	Hepatomegaly	863.87	1	7.02e-190	Yes
Drug	Spiders	734.534	1	9.25e-162	Yes
Edema	Stage	686.572	4	2.82e-147	Yes
Status	Stage	608.454	4	2.29e-130	Yes
Edema	Status	577.578	4	1.10e-123	Yes
Hepatomegaly	Status	438.518	2	5.98e-96	Yes
Hepatomegaly	Edema	293.361	2	1.98e-64	Yes
Spiders	Stage	234.706	2	1.08e-51	Yes
Drug	Hepatomegaly	130.059	1	3.98e-30	Yes
Sex	Spiders	127.586	1	1.38e-29	Yes
Hepatomegaly	Spiders	127.029	1	1.83e-29	Yes
Ascites	Stage	117.604	2	2.90e-26	Yes
Spiders	Edema	90.898	2	1.83e-20	Yes
Ascites	Edema	74.557	2	6.46e-17	Yes
Sex	Drug	27.722	1	1.40e-07	Yes
Sex	Hepatomegaly	26.685	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	5.932	1	0.014866	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

Data Analysis (EDA & Stats)

Correlation Matrix and Variance Inflation Factor



Feature	VIF
Bilirubin	1.5574
Copper	1.283
Cholesterol	1.2625
SGOT	1.2536
Tryglicerides	1.2003
Albumin	1.108
Alk_Phos	1.0673
Prothrombin	1.0579
Platelets	1.0283

Data Analysis (EDA & Stats)

Correlation Matrix and Variance Inflation Factor

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

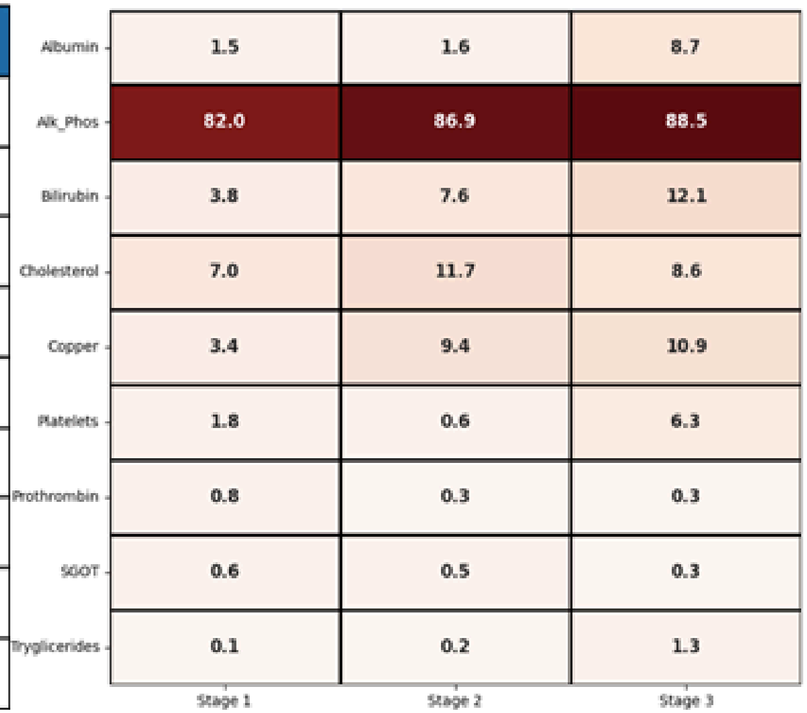


TABLE OF CONTENTS

3. Data Preprocessing

Borcelle Medical

Data Preprocessing

Data Preprocessing Pipeline

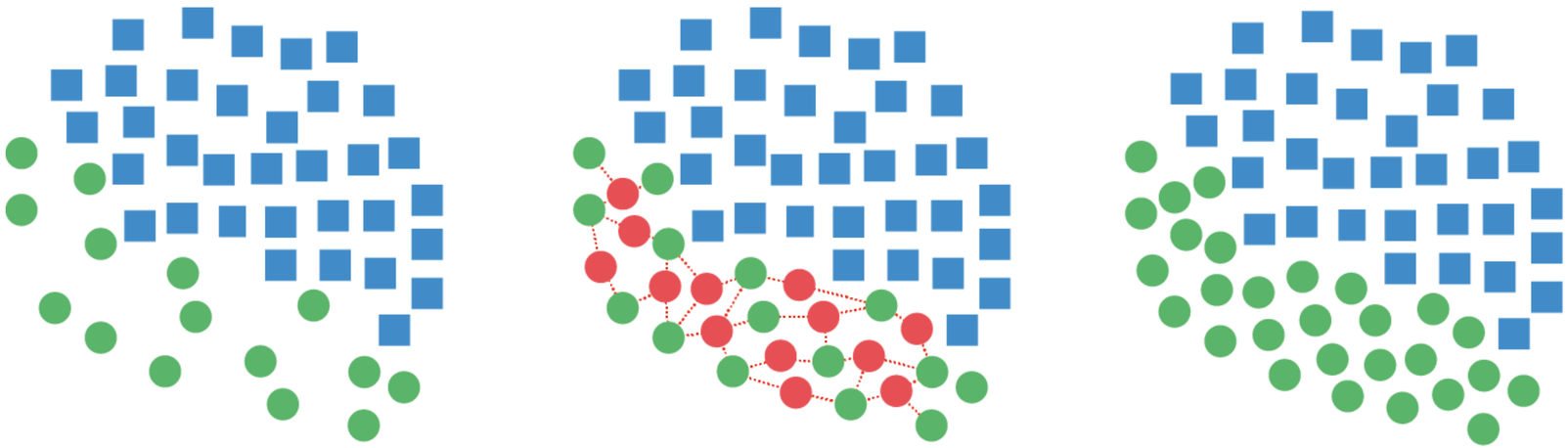
Data cleaning

- Data Cleaning: Verified no missing values
- Outlier Detection: Detected outliers using IQR but retained biologically significant extreme values.
- Transformation: Handled skewed data and applied **RobustScaler** to reduce the influence of outliers.

Borcelle Medical

SMOTE

Synthetic Minority Oversampling Technique



Borcelle Medical

Data Preprocessing

Feature Transformation

Range variability handling

- RobustScaler
- StandardScaler

Missing Values

- Column Mean

TABLE OF CONTENTS

3. Machine Learning implementation & Experiment analysis

Machine Learning Implementation

Predictive Models Implementation

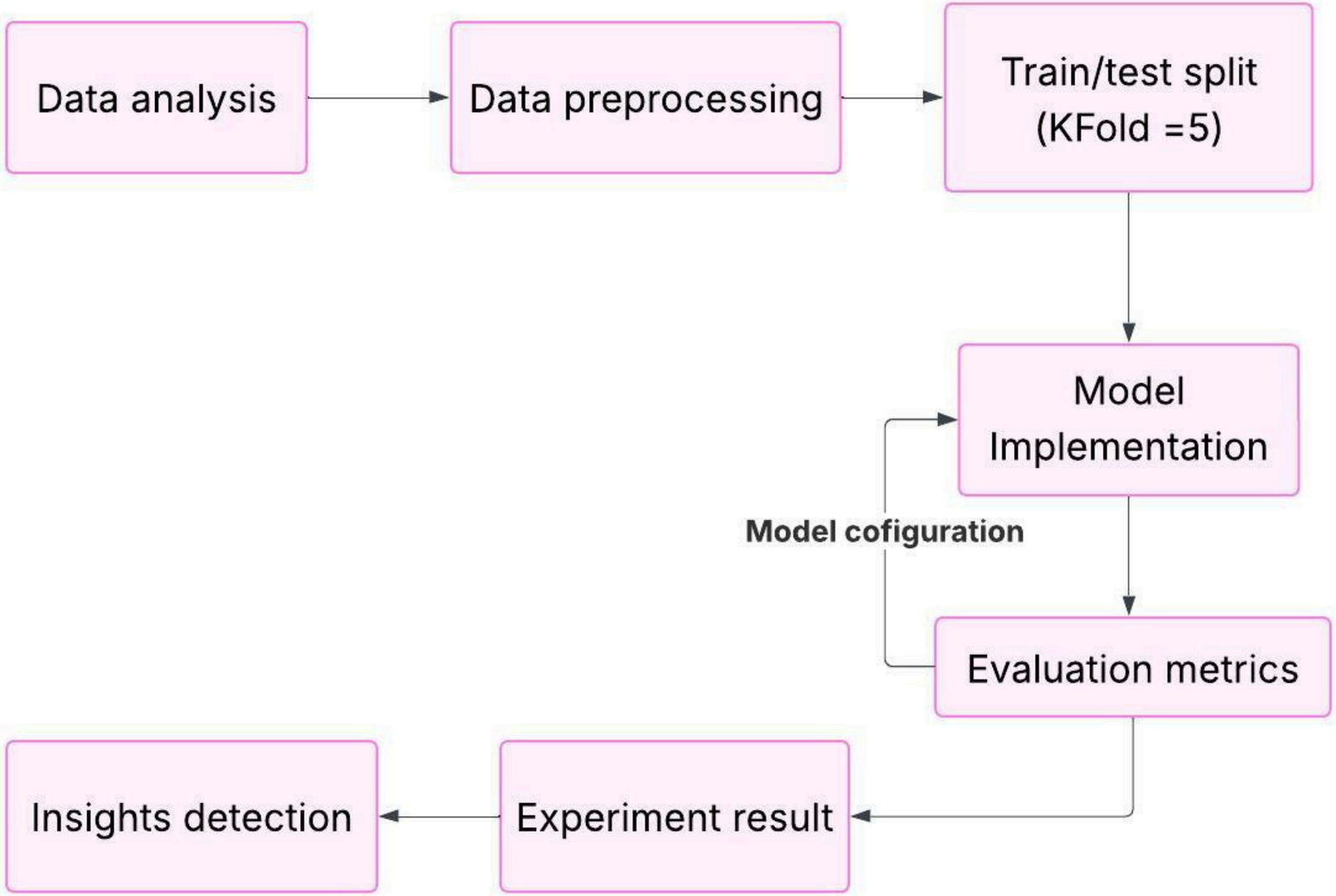
Linear/Instance

- Softmax Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

Ensemble

- Random Forest (RF)
- XGBoost

Experiment Analysis



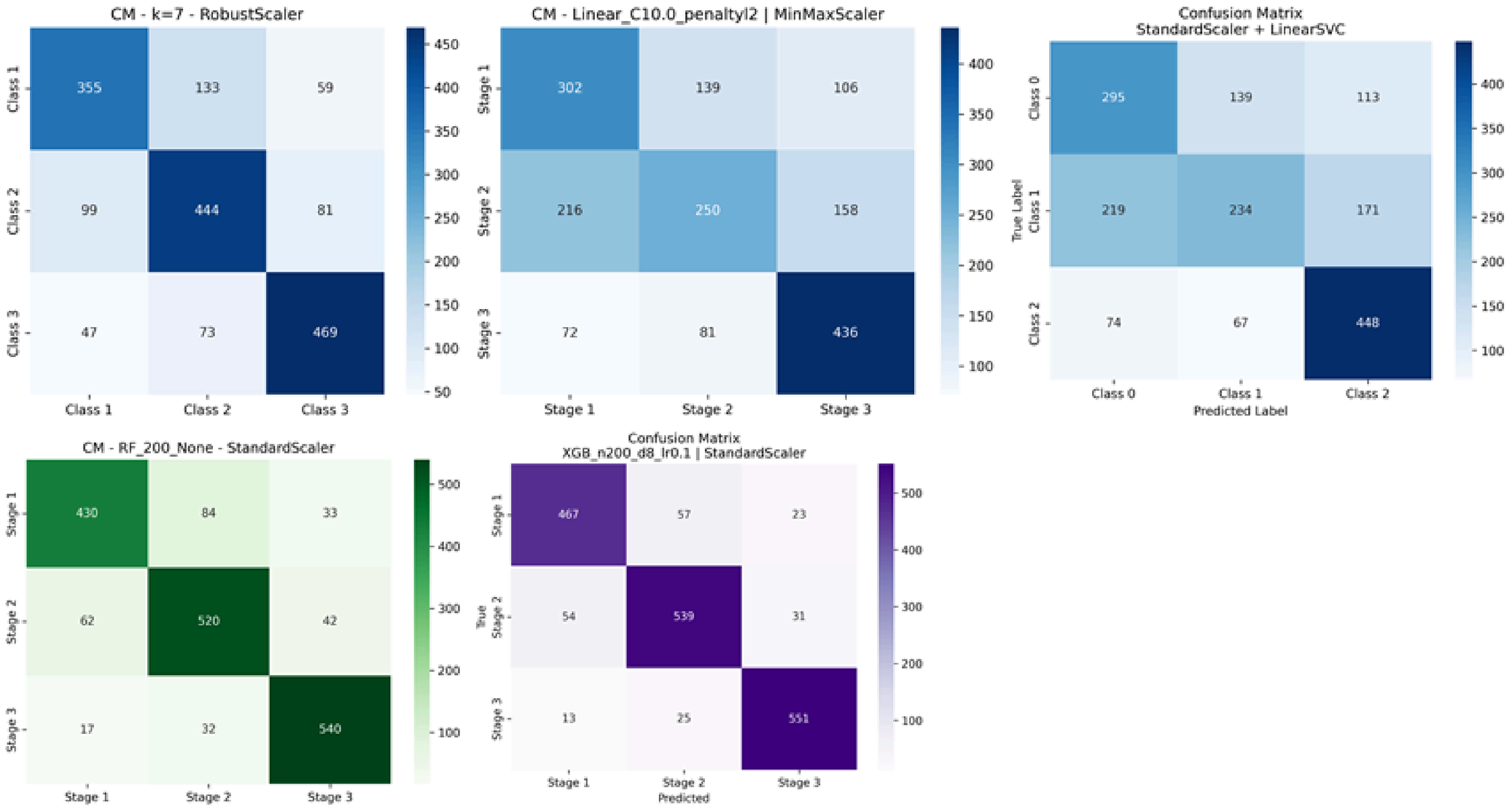
Schematic overview of the experimental machine learning workflow.

Baseline experiment before SMOTE

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%**

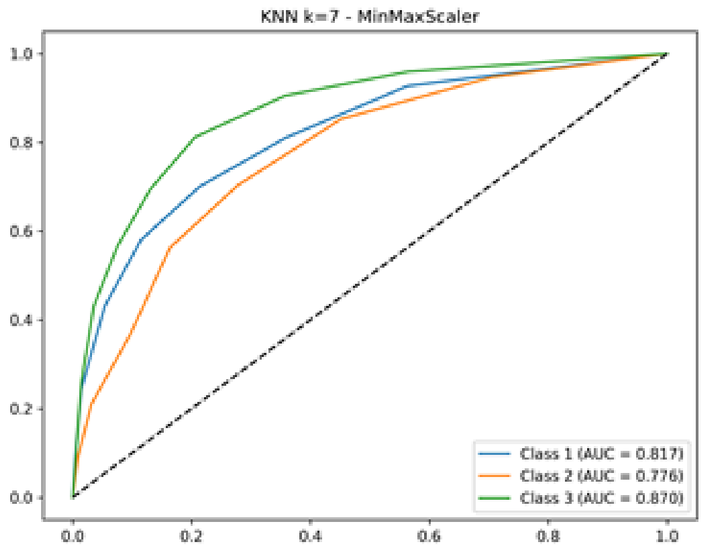
Performance metrics of machine learning models in the baseline experiment (Pre-SMOTE)

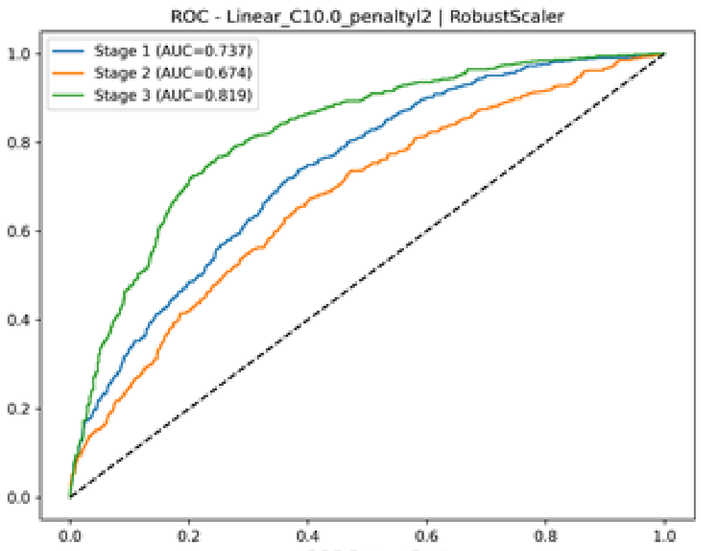
Baseline experiment before SMOTE

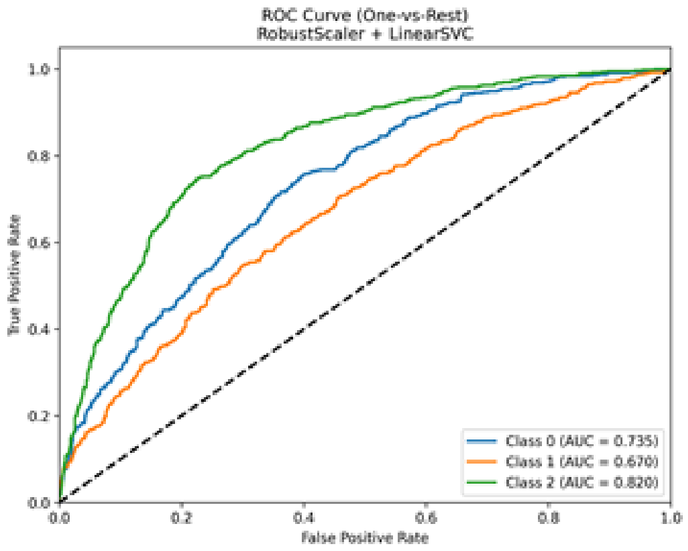


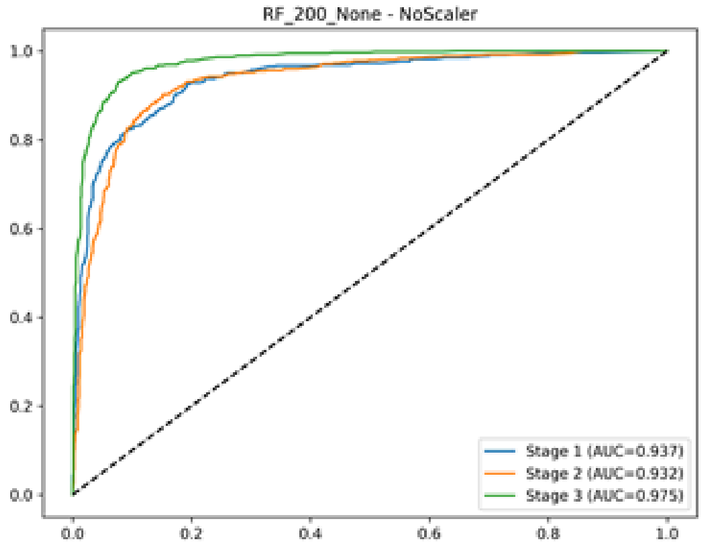
Comprehensive performance comparison of five machine learning models: Confusion Matrices

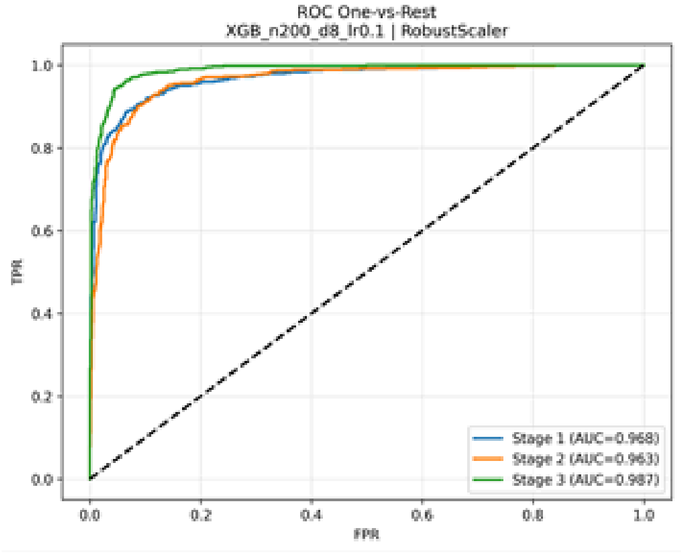
Baseline experiment before SMOTE











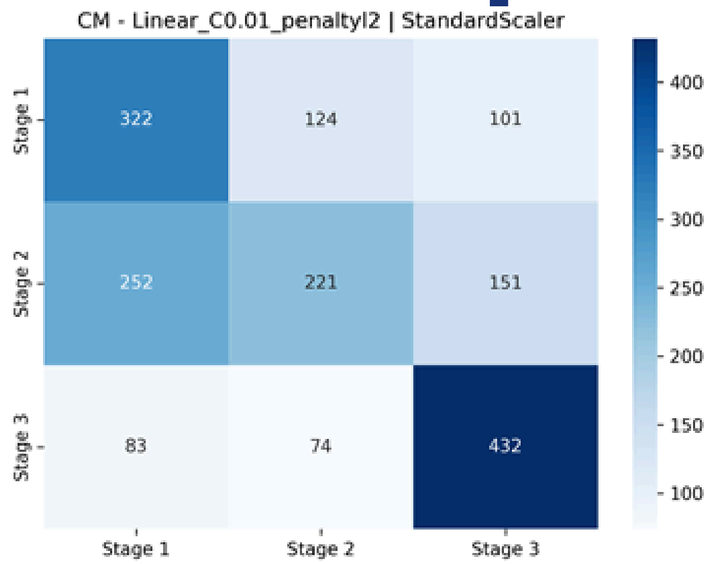
Comprehensive performance comparison of five machine learning models: ROC Curves

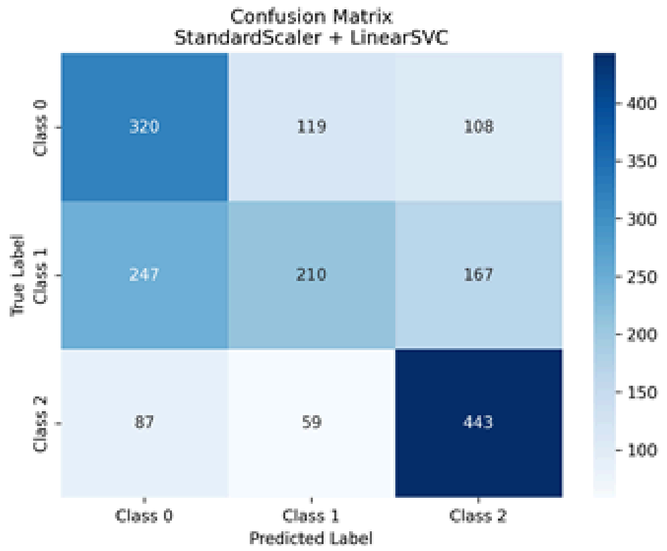
Baseline experiment after SMOTE

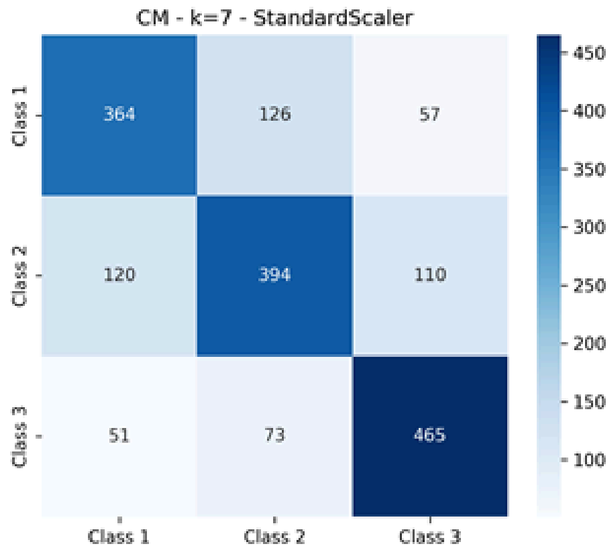
Model	Accuracy	F1 score	Precision	Recall
Softmax Regression	55.40%	54.36%	55.07%	55.40%
Support Vector Machine	55.45%	54.11%	55.25%	55.45%
K-Nearest Neighbors	71.48%	71.42%	71.41%	71.48%
Random Forest	**84.20%**	**84.17%**	**84.15%**	**84.20%**
XGBoost	**88.69%**	**88.68%**	**88.67%**	**88.69%**

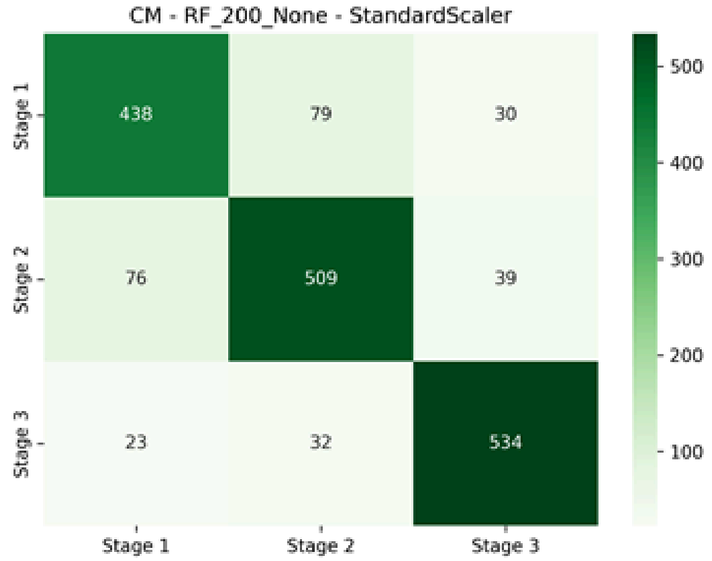
Performance evaluation of machine learning models after applying SMOTE.

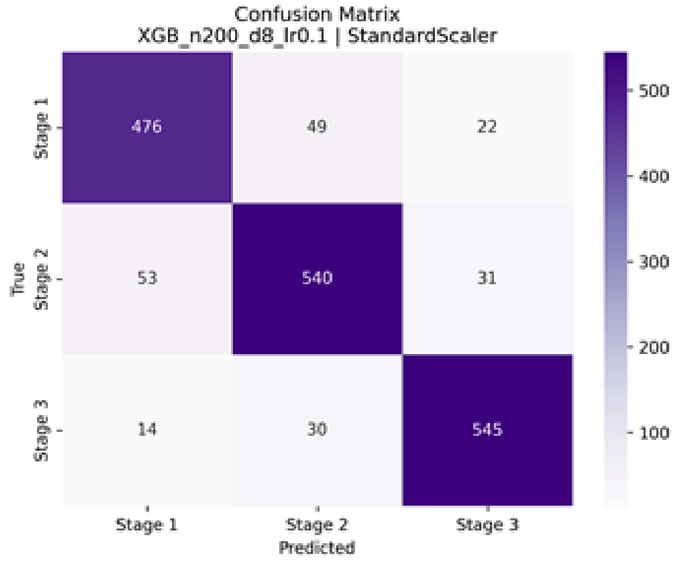
Baseline experiment after SMOTE









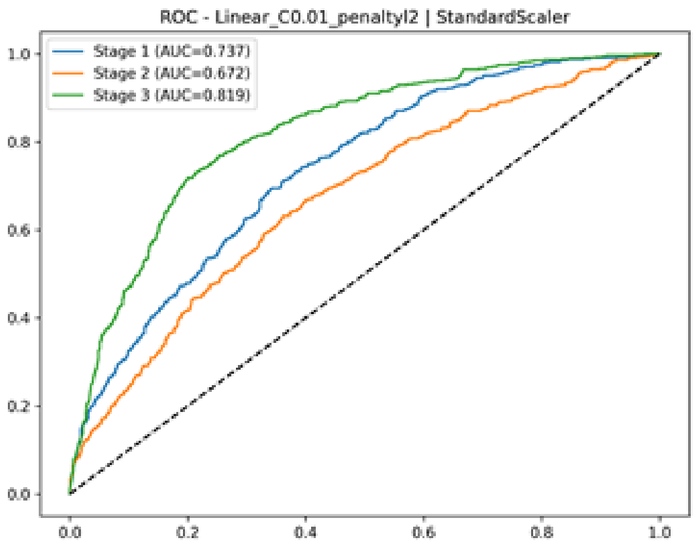


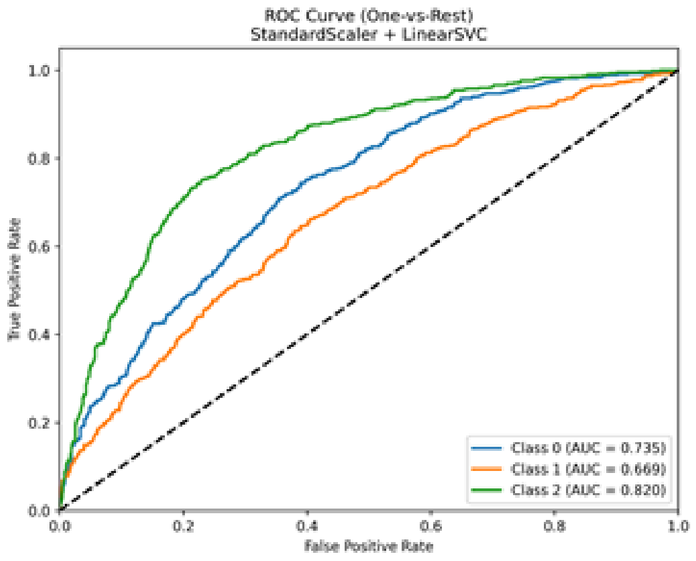
Linear Models (SVM) show **confusion** between stages.

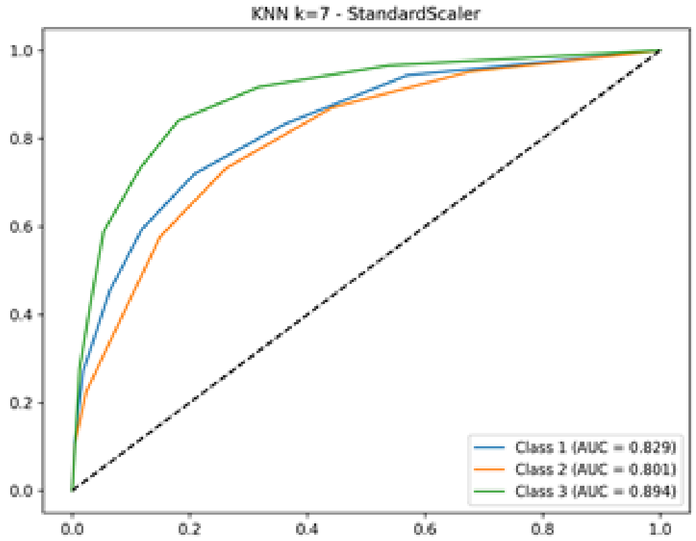
XGBoost shows a **clear separation** (sharp diagonal).

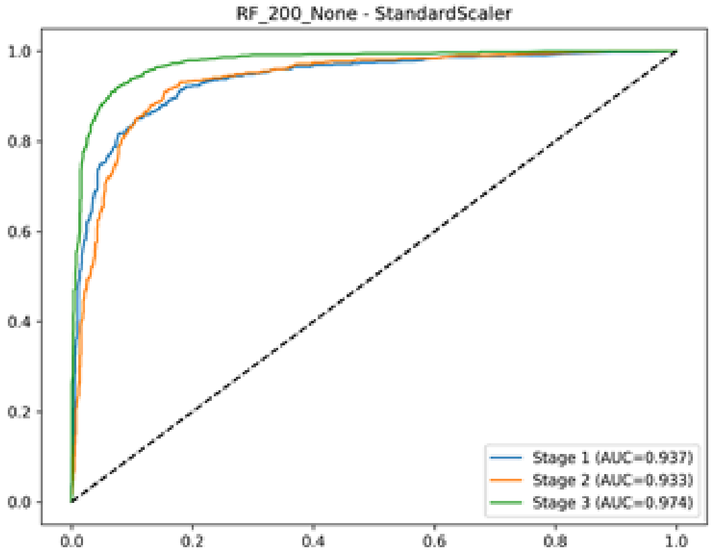
Comprehensive performance comparison of five machine learning models: Confusion Matrices

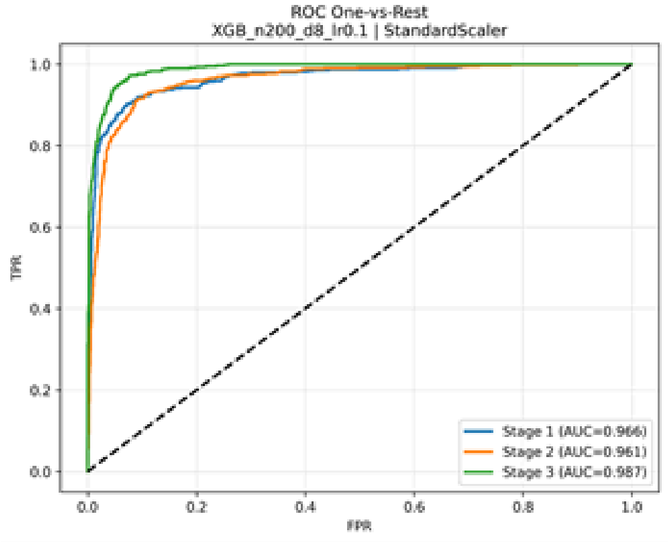
Baseline experiment after SMOTE











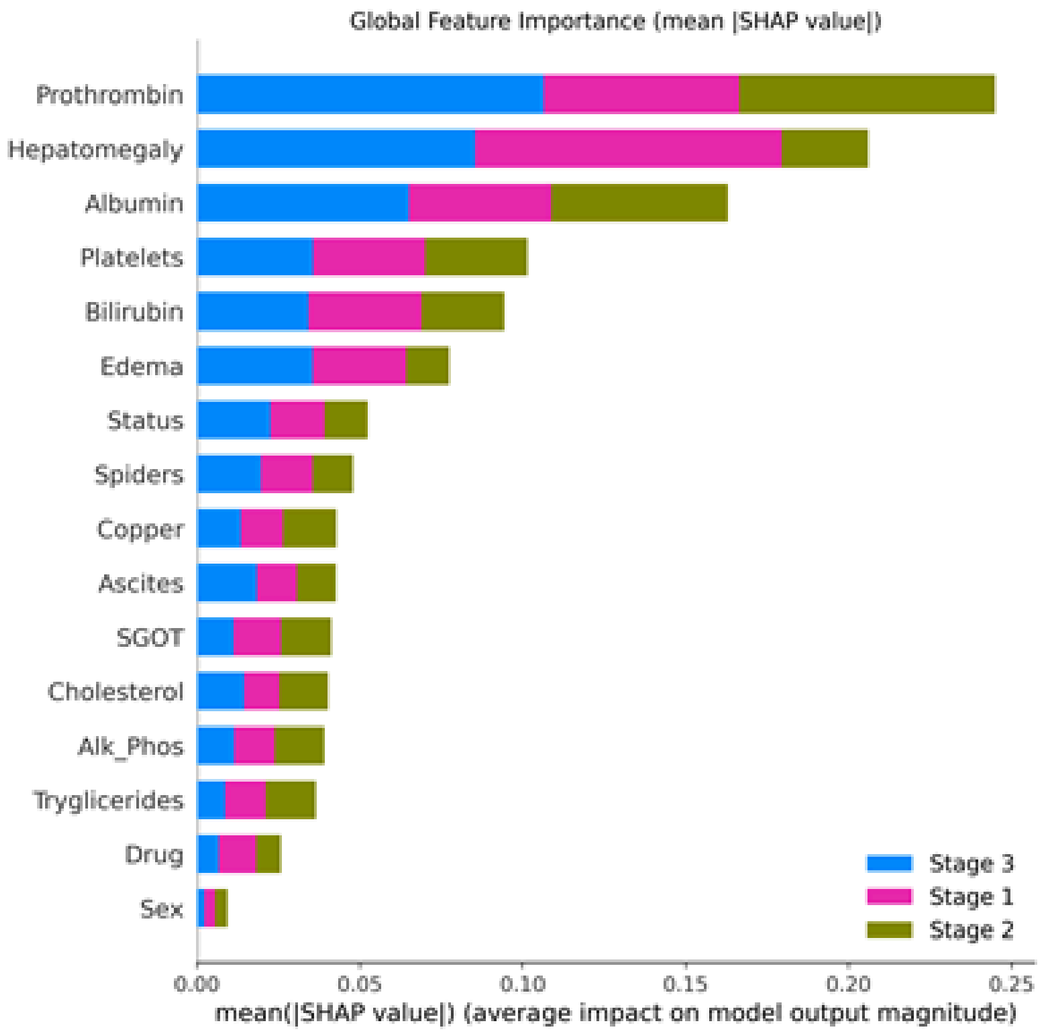
Comprehensive performance comparison of five machine learning models: ROC Curves

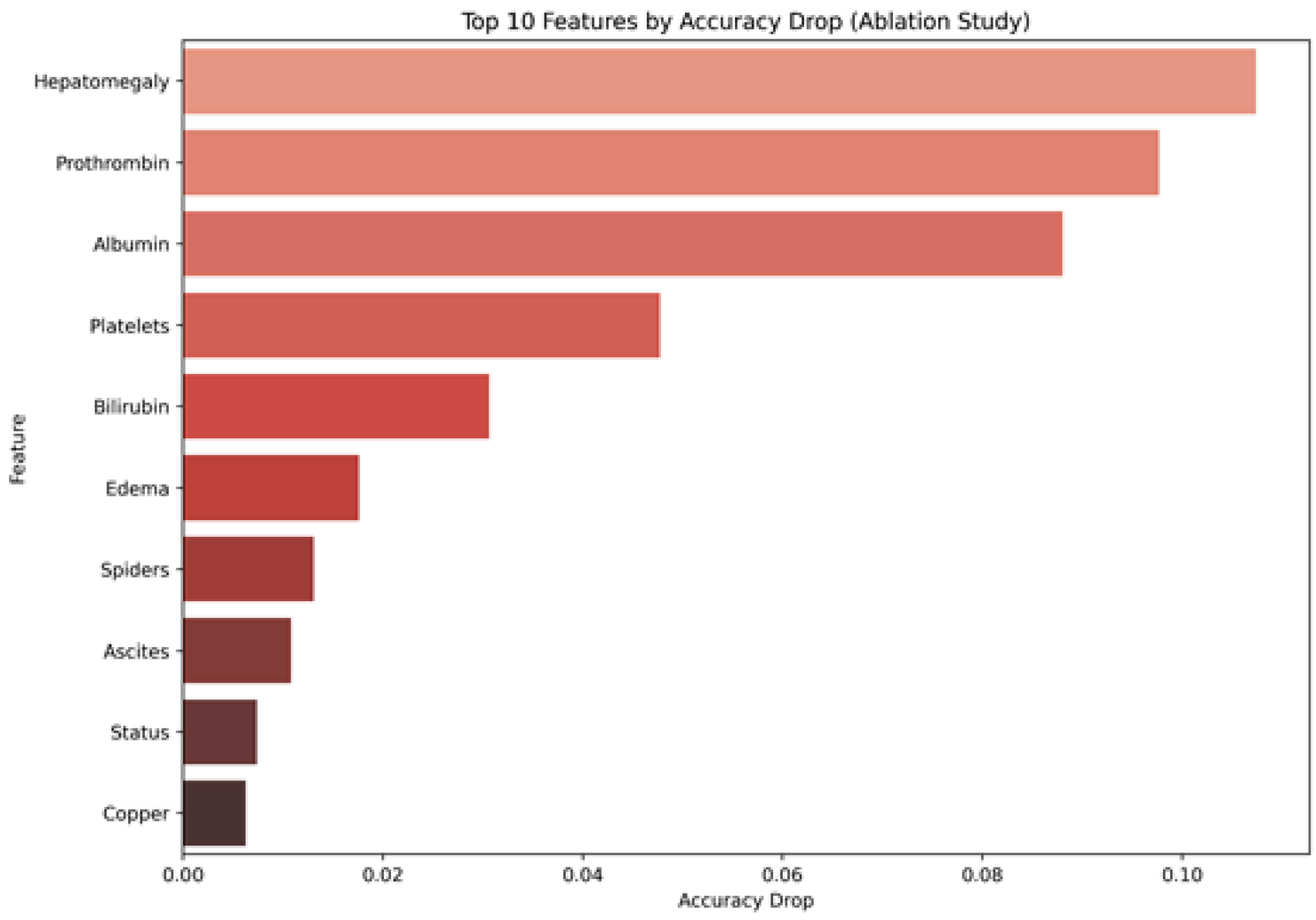
How Scaler Affects On Experiment

Model	Scaler	Accuracy	Precision	Recall	F1-score	AUC-ROC
Softmax Regression	None	54.83%	54.39%	54.83%	54.32%	73.14%
	StandardScaler	55.80%	55.24%	55.80%	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%	78.70%
	RobustScaler	72.05%	72.01%	72.05%	71.97%	85.40%
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%

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

SHAP Analysis





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

TABLE OF CONTENTS

4. Result & Conclusion

Result and Discussion

Model Performance & Experimental Stability

Key Findings (Linear vs. Non-Linear):

- **Linear Models** (Softmax, SVM): Underperformed with **~55-56% Accuracy**. Linear boundaries failed to capture the complex physiological transitions of cirrhosis.

- **Non-Linear Models** (RF, XGBoost): Achieved superior performance, with XGBoost reaching **88.47% Accuracy**. This confirms that staging requires modeling non-linear interactions.

Result and Discussion

Impact of Feature Scaling Strategies

StandardScaler (Best for Linear Models):

- Crucial for SVM and Softmax to mitigate penalty effects from heterogeneous units.
- **Result:** SVM AUC-ROC increased significantly from 0.67 (None) to **0.74** (Scaled).

RobustScaler (Best for Non-Linear Models):

- Uses median and IQR to suppress the influence of biological outliers without removing them.
- **Result:** KNN performance jumped from 64.66% to **72.05% Accuracy** (AUC 0.78 → **0.85**).

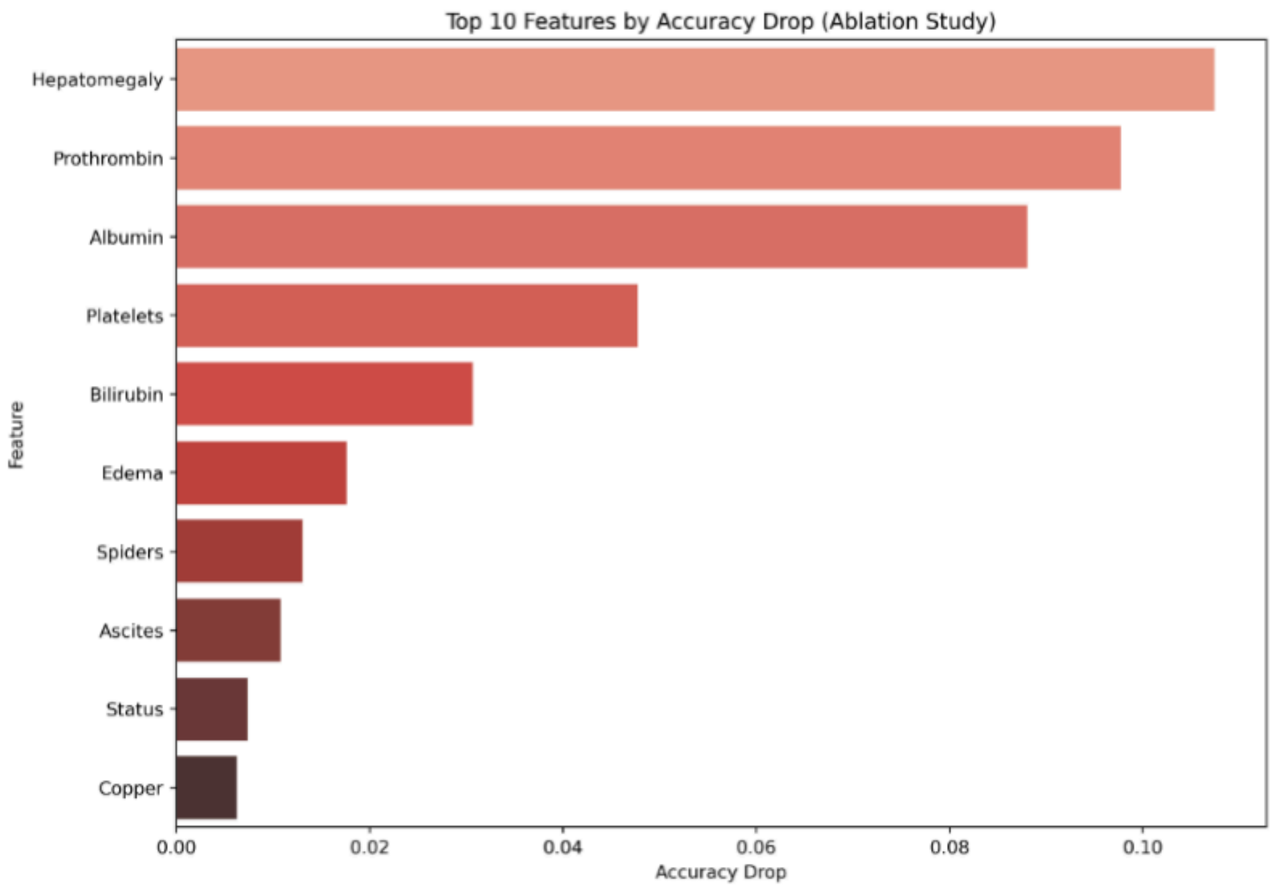
Recommendation: **RobustScaler** is the optimal preprocessing choice for large-scale, noisy clinical datasets.

Result and Discussion

Explainability & Clinical Validation (Albation Study)

Top Predictors:

- **1. Hepatomegaly**: **The strongest predictor**. Reflects structural changes and fibrosis progression.
- **2. Prothrombin Time**: **A key marker** of hepatic synthetic failure.
- **3. Albumin**: Consistent with the Child-Pugh scoring system for liver function.



Thank You.

For Your Attention