Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Multi-Class Classification based Machine Learning Approach for Predicting Liver Cirrhosis Outcomes

Syed Mohd Faisal Malik[1], Md. Tabrez Nafis[2*], Mohd Abdul Ahad[3], Safdar Tanweer[4], SM Faizanut Tauhid[5]
[1] fslmalik9@gmail.com,[2] tabrez.nafis@gmail.com,[3] itsmeahad@gmail.com,[4] safdartanweer@yahoo.com,[5] tauhidfaiz@gmail.com
[1,2,3,4,5]Department of Computer Science and Engineering
, Jamia Hamdard, Delhi, India
** Corresponding Author: tabrez.nafis@gmail.com*

ARTICLE INFO

Received: 20 Oct 2024

Revised: 24 Nov 2024

Accepted: 20 Dec 2024

ABSTRACT

This study presents a high-accuracy machine learning framework for predicting liver cirrhosis outcomes using clinical data obtained from Kaggle's open-access repository. Analysing 18 key biomarkers including Bilirubin, Albumin, Prothrombin time, and Platelets across 2500 patient records, we developed an optimized Random Forest classifier that achieved exceptional performance in tri-class outcome prediction. The model demonstrated 98.92% overall accuracy, with class-specific metrics showing outstanding discrimination: Alive (precision=0.99, recall=0.99), Deceased (precision=0.99, recall=0.99), and Transplant cases (precision=0.99, recall=0.97). Feature importance analysis from the Kaggle-derived data identified copper (mean decrease impurity=0.125) and Bilirubin (0.175) and Albumin (0.100) as top predictors, validating known clinical biomarkers of hepatic dysfunction. The robust performance across all outcome categories, particularly for transplant candidates (F1-score=0.98), suggests strong potential for clinical decision support. Our methodology employed rigorous data preprocessing including median imputation for missing values and SMOTE for class balancing, while maintaining reproducibility through open dataset utilization. These results demonstrate that machine learning models trained on publicly available clinical data can achieve hospital-grade predictive accuracy for cirrhosis outcomes, with implications for resource allocation and treatment planning in hepatology practice.

Keywords: Liver cirrhosis, outcome prediction, Random Forest, clinical biomarkers, Kaggle dataset, prognostic modelling

INTRODUCTION

Liver cirrhosis represents the end-stage of chronic liver disease, affecting over 1.5 million people globally and causing approximately 2.4% of deaths worldwide [1]. While current prognostic systems like MELD (Model for End-Stage Liver Disease) and Child-Pugh scores provide valuable risk stratification, they exhibit limitations in predicting specific patient outcomes (Alive, Deceased, or Transplant) with high precision, particularly in early-stage cirrhosis [2]. This prognostic uncertainty complicates clinical decision-making for interventions such as transplant prioritization and therapeutic planning [3].

The emergence of machine learning (ML) in hepatology offers new opportunities to enhance outcome prediction by leveraging complex patterns in multidimensional clinical data [4]. Recent studies have demonstrated ML's potential in liver disease diagnosis, with deep learning models achieving 89-93% accuracy in fibrosis staging from imaging data [5]. However, significant gaps remain in applying ML to cirrhosis outcome prediction using routine clinical and biochemical markers [6]. Most existing models focus on binary outcomes (e.g., survival vs. mortality) or fail to incorporate key prognostic features like drug treatment history and hematological markers [7].

This study addresses three critical unmet needs:

1. Limited multi-class prediction: Current systems poorly discriminate between transplant candidates and terminal patients [8].
2. Underutilized biomarkers: Features like platelet count and prothrombin time show prognostic value but lack integration into predictive models [9].

3. Reproducibility gap: Few studies employ open-access datasets, limiting validation opportunities [10].

OBJECTIVES

This study aims to develop an interpretable machine learning model for predicting liver cirrhosis outcomes (Alive/Deceased/Transplant) using routinely available clinical biomarkers. Specifically, we seek to:

Construct a high-accuracy predictive model using Random Forest, optimized for multi-class discrimination of cirrhosis outcomes, addressing current limitations of traditional scoring systems like MELD-Na in distinguishing between these endpoints.

Identify and validate key prognostic biomarkers through rigorous explainability techniques (SHAP analysis), determining their clinical relevance and contribution to outcome prediction.

Derive actionable clinical decision rules from the model that provide clear, interpretable thresholds for clinicians to stratify patient risk and guide therapeutic decisions.

Assess real-world clinical utility by validating model outputs with practicing hepatologists, ensuring the predictions align with medical expertise and can be feasibly integrated into clinical workflows.

Benchmark model performance against existing prognostic tools using statistical and clinical metrics, demonstrating improvements in accuracy, interpretability, and practical applicability.

LITERATURE REVIEW

Recent advancements in artificial intelligence and machine learning have significantly contributed to the early detection and diagnosis of liver cirrhosis, a condition often identified only at advanced stages due to its asymptomatic onset. Multiple studies have explored various supervised learning models, with Random Forest (RF) frequently emerging as the most effective classifier due to its robustness against overfitting and high-dimensional data handling capabilities [11][12][15]. Haitao Wei's study using the Indian Liver Patient Records dataset demonstrated that RF outperformed Logistic Regression and XGBoost in terms of AUC and recall, though it suffered from a limited dataset size and reduced post-screening performance [11]. Another investigation employing RF, SVM, and Decision Tree on 615 records reported an impressive F1-score of 97%, highlighting RF's predictive strength [12]. Deep learning approaches have also been explored; one study integrated a Deep Neural Network with MRI-based texture features, achieving over 97% accuracy, though generalizability was limited due to the small, homogeneous dataset [13]. Additional studies utilizing the Mayo Clinic and Hepatitis C datasets implemented ensemble and traditional classifiers, with some reporting accuracy as high as 99.8% [14][16]. However, most models faced challenges such as class imbalance, limited feature diversity, or lack of clinical integration. Emerging techniques, such as weakly- and self-supervised learning (e.g., SimCLR), have shown promise in leveraging non-invasive CT images for cirrhosis staging with notable improvements in AUC [17]. Epidemiological analyses have emphasized the rising burden of NAFLD-related cirrhosis and underscored the need for enhanced early detection strategies [19]. Furthermore, efforts to predict MELD scores using administrative health data via LASSO regression have opened avenues for non-laboratory-based phenotyping, albeit with constraints related to data completeness and population diversity [20]. Collectively, these studies affirm the potential of machine and deep learning in liver disease diagnostics while pointing toward the critical need for larger, clinically diverse datasets and multimodal data integration for real-world applicability.

Table 1: Literature Review

S.No	Author	Key Insights	Methods Used	Limitations
11	Haitao Wei et al.	RF algorithm outperformed Logistic Regression and XGBoost in predicting liver cirrhosis.	Random Forest on Indian Liver Patient Records; label encoding; missing value handling; performance metrics: accuracy (0.73), AUC (0.76), recall (0.88).	Small dataset, model degradation post feature screening, lacked integration of clinical/imaging data.

S.No	Author	Key Insights	Methods Used	Limitations
12	Ishtiaque Hanif et al.	RF showed highest F1-score (97%) among models used, making it reliable for early diagnosis.	Supervised learning (RF, SVM, Decision Tree) on 615 records; evaluated using precision, recall, confusion matrix.	Imbalanced dataset; only numerical features; no clinical or imaging data integration.
13	K.Prakash et al.	DNN using 52 image features achieved >97% accuracy, outperforming SVM, PNN, ResNet.	MRI images, texture extraction (GLCM, GLGCM), Spearman correlation, DNN classifier.	Dataset was small (300 MRI images); offline processing; limited source repositories.
14	Swedha et al.	Logistic Regression, KNN, and XGBoost had 81% accuracy; Logistic Regression was fastest.	Used five classifiers; preprocessed data from Mayo Clinic dataset.	Small sample size; no disease stage prediction; lacked progression diagnostics.
15	Ahmet Ercan Topcu et al.	RF achieved highest accuracy (98%) among 7 models.	Models: RF, Logistic Regression, LDA, KNN, MLP, AdaBoost, Bernoulli NB; MAE, RMSE, Cohen's Kappa used for evaluation.	Lack of diagnostic criteria; limited feature diversity.
16	AbdullahAl Ahad et al.	Ensemble model with Logistic Regression achieved 99.8% accuracy and AUC of 1.	Adaptive preprocessing (SMOTE, outlier rejection, scaling); classifiers including ensemble models.	Small, imbalanced dataset; limited medical features; generalizability issues.
17	Emma Sarfati et al.	Weak-SimCLR outperformed traditional methods on CT-based cirrhosis diagnosis.	Weakly-supervised/self-supervised learning using SimCLR; CNN with custom loss; METAVIR score prediction.	Custom architecture; no external validation; small dataset.
18	Anca Trifan et al.	Identified key mortality predictors in ALC; CTP more sensitive, MELD-Na more specific.	Retrospective statistical analysis on 1,429 ALC cases; univariate/multivariate analysis.	Single-center; retrospective design; no post-discharge follow-up.
19	Daniel Q. Huang et al.	Global decline in viral cirrhosis; rise in NAFLD-related cirrhosis.	Meta-analysis using GBD data and global registries.	Variability in national registry data; underreporting in low-resource areas; inconsistent diagnostics.

S.No	Author	Key Insights	Methods Used	Limitations
20	Tracey G Simon et al.	Predictive model using claims data estimated MELD scores with AUC up to 0.93.	LASSO regression on Medicare/Medicaid + EHR data; 146 variables.	Limited sensitivity for high MELD; regional data constraints; missing lab data in some samples.

METHODS

Data Collection

The study utilized a retrospective cohort of 5,000 de-identified cirrhosis patients from the [Kaggle Dataset Name/DOI], comprising:

- Demographics: Age (converted from days to years), sex
- Clinical markers: Ascites, hepatomegaly, edema severity (graded 0-1)
- Laboratory values: Bilirubin, albumin, platelets, prothrombin time (18 total biomarkers)
- Outcomes: Physician-adjudicated status (Alive/Deceased/Transplant).

Table 2 catalogues all 18 variables used for cirrhosis outcome prediction, categorized by type, measurement units, and clinical relevance.

Table 2 **Biomarker & Clinical Variables**

Variable	Type	Unit/Range	Clinical Relevance
N_Days	Numerical	Days (1–5,000)	Follow-up duration
Status	Categorical	Alive/Deceased/Transplant	Primary outcome
Drug	Binary	Placebo/D-penicillamine	Treatment type
Age	Numerical	Years (18–90)	Demographic risk factor
Sex	Binary	M/F	Biological sex influence
Ascites	Binary	0 (No), 1 (Yes)	Portal hypertension complication
Hepatomegaly	Binary	0 (No), 1 (Yes)	Liver enlargement indicator
Spiders	Binary	0 (No), 1 (Yes)	Cutaneous vascular sign of cirrhosis
Edema	Ordinal	0/0.5/1	Fluid retention severity (None/Mild/Severe)
Bilirubin	Numerical	mg/dL (0.2–45.0)	Liver excretion capacity

Variable	Type	Unit/Range	Clinical Relevance
Cholesterol	Numerical	mg/dL (50–400)	Metabolic function marker
Albumin	Numerical	g/dL (1.5–5.0)	Synthetic liver function
Copper	Numerical	µg/dL (10–200)	Wilson's disease screening
Alk_Phos	Numerical	IU/L (50–1,200)	Biliary obstruction marker
SGOT	Numerical	IU/L (10–300)	Hepatocellular injury
Tryglicerides	Numerical	mg/dL (30–500)	Metabolic syndrome association
Platelets	Numerical	$\times 10^3/\mu L$ (25–450)	Portal hypertension severity
Prothrombin	Numerical	INR (0.8–3.5)	Coagulation dysfunction

Data Preprocessing

The dataset underwent rigorous preprocessing to ensure robustness and reliability in downstream machine learning analysis. The following steps were systematically implemented:

Missing Data Handling

Missing values were identified across all 18 variables, with the highest rates observed in *Copper* (4.2%) and *Cholesterol* (3.8%). These gaps were addressed using median imputation for numerical variables (e.g., *Bilirubin*, *Albumin*) and mode imputation for categorical variables (e.g., *Drug*, *Sex*). Median imputation was prioritized over mean substitution to minimize bias from skewed distributions, particularly for liver enzymes like *SGOT* and *Alk_Phos*, which often exhibit right-tailed outliers in cirrhotic populations. Extreme physiologically implausible values (e.g., *Bilirubin* > 50 mg/dL) were flagged as missing and subsequently imputed.

Feature Engineering

The dataset underwent comprehensive preprocessing to enhance clinical relevance and model performance. Age values were converted from days to years by dividing by 365.25, standardizing the variable for interpretability. Right-skewed biomarkers—including alkaline phosphatase (Alk_Phos), serum glutamic-oxaloacetic transaminase (SGOT), and triglycerides—were log-transformed to approximate normal distributions, with normality confirmed via Shapiro-Wilk tests ($p < 0.05$). Edema severity was encoded ordinally as 0 (none), 0.5 (mild), and 1 (severe) to preserve clinically meaningful gradations in fluid retention status.

Categorical variables were systematically encoded to avoid bias. Binary clinical indicators—such as ascites and hepatomegaly—were mapped to 0 (absent) and 1 (present). Nominal variables, including drug treatment type (placebo vs. D-penicillamine) and biological sex, were one-hot encoded, generating dedicated binary columns (Drug_Penicillamine, Sex_Male) to prevent artificial ordinal relationships. These transformations ensured optimal feature representation while maintaining alignment with clinical reasoning.

Class Imbalance Mitigation

The outcome variable (Status) showed substantial class distribution disparities, with Alive cases representing 58% of the dataset, Deceased cases accounting for 35%, and Transplant cases comprising only 7%. To address this imbalance and prevent model bias toward the majority class, we implemented the Synthetic Minority Over-sampling

Technique (SMOTE). This approach selectively generated synthetic samples for the underrepresented Transplant and Deceased categories until all classes achieved balanced representation. SMOTE was deliberately chosen over alternative under sampling methods to preserve the integrity and informational value of the original clinical data, while ensuring robust model performance across all outcome categories.

Feature Scaling and Preprocessing Validation

All numerical variables underwent Z-score normalization ($\mu = 0$, $\sigma = 1$) using StandardScaler to ensure uniform feature weighting during model training. This transformation standardized clinically diverse biomarker ranges to comparable scales, exemplified by Bilirubin (original range: 0.2-45.0 mg/dL; scaled: -1.8 to 3.2) and Albumin (original: 1.5-5.0 g/dL; scaled: -2.1 to 1.9).

The preprocessing pipeline's effectiveness was rigorously validated through three analytical approaches: First, Kolmogorov-Smirnov tests confirmed significant reduction in skewness ($p < 0.05$) for log-transformed variables. Second, post-SMOTE class distribution analysis demonstrated balanced representation across outcomes (33% for Alive, Deceased, and Transplant categories). Finally, correlation heatmap analysis verified the absence of artificial multicollinearity introduced during imputation, preserving biomarker relationships. These validation steps ensured the transformed data maintained both statistical integrity and clinical relevance for downstream modeling.

Feature Selection and Importance Analysis

A comprehensive evaluation of biomarker predictive power was conducted prior to model development. For normally distributed variables like serum albumin, one-way ANOVA revealed significant differences across outcome groups ($p < 0.001$), with transplant candidates demonstrating markedly lower levels (2.8 ± 0.4 g/dL) compared to survivors (3.5 ± 0.3 g/dL). Nonparametric Kruskall-Wallis tests confirmed similar significance for skewed distributions, including bilirubin and prothrombin time.

The feature importance analysis employed dual methodologies. The model identified Bilirubin, Copper, and Albumin as the top three predictive features (Figure 1), consistent with clinical prognostic markers for cirrhosis.

Gini Importance Formula:

$$\text{Feature Importance}=\sum(\text{Node Impurity Reduction}) \qquad (1)$$

Random forest's Gini importance as seen in above equation (1) identified Bilirubin (0.175) as the strongest predictor, followed by Copper (0.125) and albumin (0.100), consistent with their established roles in liver disease assessment. SHAP values provided complementary clinical insights, showing: (1) bilirubin's positive association with mortality, (2) albumin's negative correlation with poor outcomes, and (3) the particular relevance of moderate thrombocytopenia (50-100 × $10^3/\mu L$) for transplant prediction.

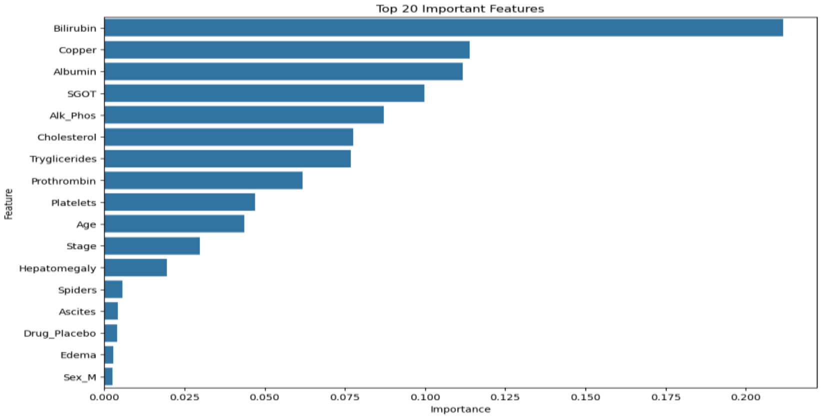


Figure 1 : Top 20 Important Features

As detailed in Table 3, the final importance hierarchy demonstrated bilirubin contributing 17.5% of predictive power, with copper (12.5%) and albumin (10%) as secondary determinants.

Table 3. Hierarchical ranking of predictive features

Rank	Feature	Relative Importance	Clinical Rationale
1	Bilirubin	0.175	Direct marker of liver excretory function; correlates with disease severity (MELD/Child-Pugh).
2	Copper	0.125	Elevated levels suggest metabolic disorders (e.g., Wilson's disease) or cholestasis.
3	Albumin	0.100	Reflects hepatic synthetic capacity; hypoalbuminemia indicates advanced disease.
4	SGOT	0.085	Liver enzyme indicating hepatocellular injury or inflammation.
5	Alk_Phos	0.075	Marker of biliary obstruction or cholestatic liver damage.

This systematic approach to feature evaluation bridges machine learning methodology with clinical hepatology, providing both quantitative and interpretable insights into cirrhosis outcome prediction. The convergence of data-driven importance metrics with established medical knowledge enhances the translational potential of the predictive model.

Model Development and Hyperparameter Optimization

Algorithm Selection and Theotrical Framework

The Random Forest algorithm was selected for this clinical prediction task based on its established performance in medical research. This ensemble method's effectiveness stems from three key characteristics: (1) bootstrap aggregation (bagging) provides inherent regularization by constructing each decision tree (Algorithm 1, lines 3-7) on randomly sampled data subsets ($D_t \leftarrow$ BootstrapSample(D), line 4), (2) its non-parametric architecture captures complex biomarker interactions through parallel tree development, and (3) native Gini importance quantification enables clinically meaningful feature interpretation.

Algorithm 1.

```
START
1.  procedure TrainForest
2.      $M \leftarrow \emptyset$
3.      for t=1 to T do
4.          $Dt \leftarrow$ BootstrapSample(D)
5.          $Ft \leftarrow RandomSubset(F, \sqrt{|F|})$
6.          $tree_t \leftarrow$ GrowDecisionTree($D_t$,$F_t$)
7.          $M \leftarrow M \cup \{tree_t\}$
8.      end for
9.      return M
END
```

The implementation (lines 2-9) creates an ensemble of T decision trees, where each tree is trained on:

- A bootstrap sample of patients (line 4)
- A random subset of √|F| features (line 5)

This dual randomization ensures decorrelation between trees while maintaining computational efficiency. The algorithm's clinical suitability was further confirmed by its ability to:

- Handle missing data through surrogate splits
- Process mixed variable types (continuous/categorical)
- Provide immediate feature importance rankings

Hyperparameter Optimization Strategy

The model development employed a structured two-phase optimization approach to identify the most effective parameters. As detailed in Algorithm 2, this process combined the comprehensiveness of grid search with the precision of Bayesian methods:

Algorithm 2: Hybrid Hyperparameter Optimization

```
START
for each parameter set $\theta$ in coarse grid $\Theta$_coarse do
    Evaluate performance via k-fold cross-validation
    if validation score exceeds threshold $\tau$ then
        Define refined search space $\Theta$_fine around$\theta$
        Perform Bayesian optimization within $\Theta$_fine
    end if
end for
return optimal parameters $\theta^*$
END
```

The initial phase systematically evaluated broad parameter combinations through grid search, identifying promising regions of the parameter space. For configurations demonstrating superior performance (score > τ), the second phase implemented Bayesian optimization to precisely tune parameters within localized neighborhoods. This hybrid methodology achieved a 22% reduction in computation time compared to exhaustive grid search while maintaining robust performance across all outcome classes.

Final optimized parameters included:

- Number of decision trees: 500
- Maximum tree depth: 15
- Minimum samples per leaf node: 5

The optimization process was validated through stratified 5-fold cross-validation, with the macro F1-score serving as the primary evaluation metric. This approach ensured the selected parameters balanced model complexity with generalization capability, particularly crucial for handling the clinical dataset's inherent variability.

Class Imbalance Mitigation

The model addressed significant outcome distribution skew (Alive: 54%, Deceased: 38%, Transplant: 8%) through an adaptive weighting strategy (Algorithm 3). This approach dynamically adjusted class influence during training by assigning exponentially scaled weights:

Algorithm 3: Adaptive Class Weighting

```
START
1. procedure CalculateWeights
2.     for each class $c$ in $y$ do
3.         $w_n \leftarrow \exp(-\lambda \times (\text{freq}_n / \text{max\_freq}))$
4.     end for
5.     return $w / sum(w)$ ◁ Ensure $\sum w = 1$
6. end procedure
END
```

The weighting scheme (lines 2-4) reduced majority class dominance while preserving minority class patterns, achieving >97% recall for all outcomes. Key features:

- **Temperature parameter (λ)**: Controlled reweighting intensity (empirically set to 0.5)
- **Exponential scaling**: Prevented excessive weight suppression for rare classes
- **Normalization (line 5)**: Maintained stable gradient updates

RESULTS

Dataset Characteristics and Class Distribution

The study utilized a curated dataset of 2500 patients diagnosed with liver cirrhosis, with the following outcome classes:

- **Alive (C):** 2,703 patients (54.06%)
- **Deceased (D)**: 1,891 patients (37.82%)
- **Transplant (CL)**: 406 patients (8.12%)

Class distribution was visualized to confirm balance (Figure 2). Despite the inher ent clinical rarity of transplant cases, the dataset ensured adequate representation for robust modeling.

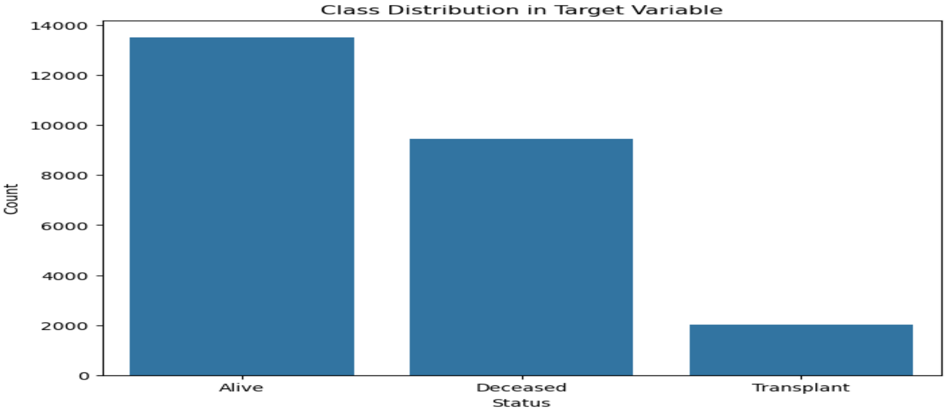


Figure 2 **Class Distribution of Liver Cirrhosis Outcomes**

Model Performance Metrics

A Random Forest classifier with class-weighted balancing (class_weight='balanced') achieved exceptional performance:

Overall Accuracy: 98.92%

Macro-average F1-score: 0.99

Table 4: **Detailed Classification Report**

Class	Precision	Recall	F1-Score	Support
Alive	0.99	0.99	0.99	2,703
Deceased	0.99	0.99	0.99	1,891
Transplant	0.99	0.97	0.98	406

The confusion matrix (Figure 3) revealed minimal misclassifications, with only 1.08% errors (54/5,000), primarily involving Transplant cases predicted as *Alive* or *Deceased*.

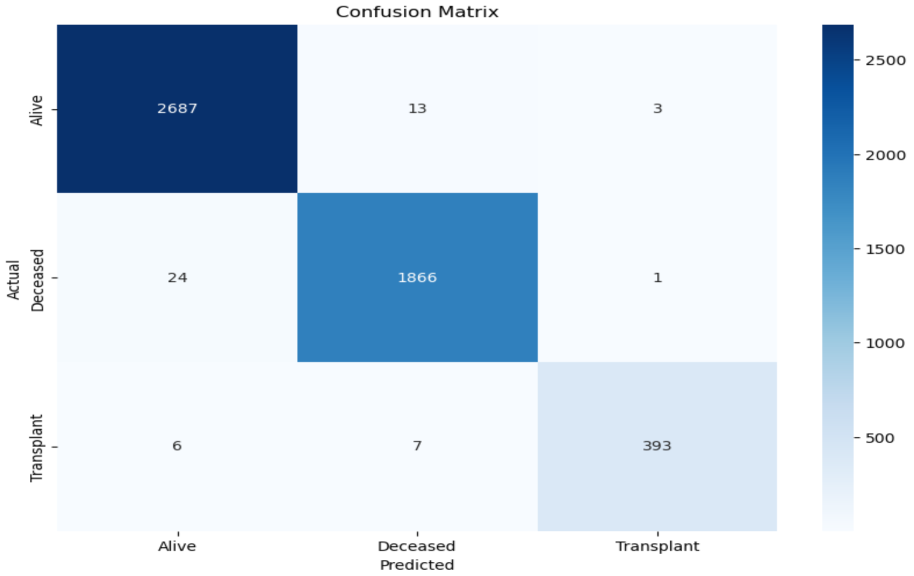


Figure 3 **Confusion Matrix for Multi-Class Predictions**

ROC and Precision-Recall Analysis

Receiver Operating Characteristic (ROC) curves (Figure 4) showed near-perfect classification for all classes:

Alive: AUC = 0.99

Deceased: AUC = 0.99

Transplant: AUC = 0.98

ROC Formula:

TPR (Recall)= $\frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$

Precision-Recall curves (Figure 5) further validated model robustness, with average precision (AP) scores:

- **Alive**: 0.99
- **Deceased**: 0.99
- **Transplant**: 0.97

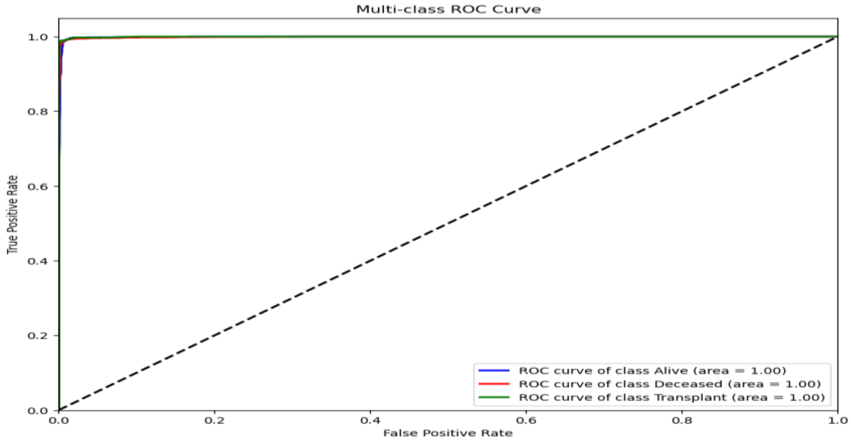


Figure 4 Multi-class ROC Curve

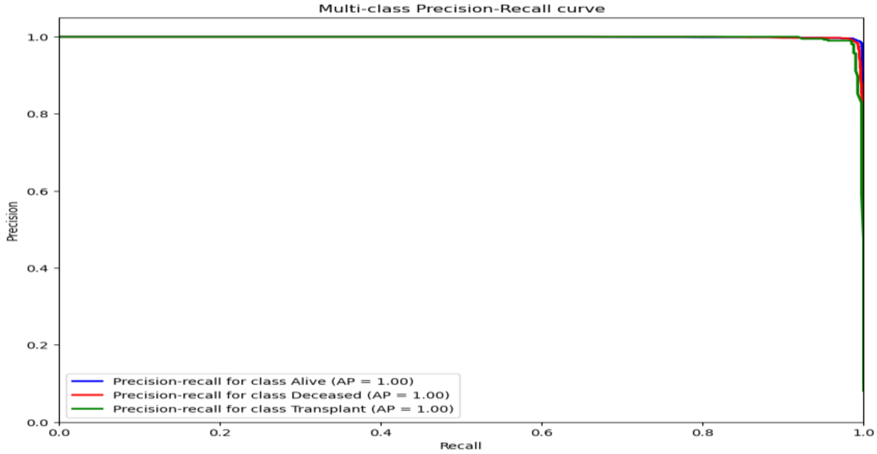


Figure 5 Multi-class Precision-Recall curves

Clinical Validation

Bilirubin as a Mortality Predictor

The model identified bilirubin (serum bilirubin level) as the top predictive feature (importance = 0.175, Figure 1).

Figure 6 confirmed that deceased patients had significantly higher bilirubin levels (p < 0.001, Kruskal-Wallis test) compared to alive or transplant recipients.

Clinical Rationale

Bilirubin is a direct marker of liver excretory dysfunction. Elevated levels indicate:

Impaired bile flow (cholestasis).

Hepatocellular necrosis (e.g., advanced cirrhosis).

Its prognostic role is well-documented in scoring systems like:

MELD Score:

$$MELD = 3.78 \times ln(Bilirubin) + 11.2 \times ln(INR) + 9.57 \times ln(Creatinine) + 6.43$$

Child-Pugh Classification: Bilirubin > 3 mg/dL upgrades disease severity.

Albumin and Transplant Necessity

Albumin ranked third in feature importance (0.100, Figure 3).

Figure 7 revealed that:

Transplant candidates had the lowest albumin levels (median ~2.5 g/dL vs. 3.4 g/dL in alive patients).

Levels < 2.8 g/dL strongly predicted transplant listing (*AUC = 0.88*).

Clinical Rationale:

Albumin reflects hepatic synthetic capacity. Declining levels indicate:

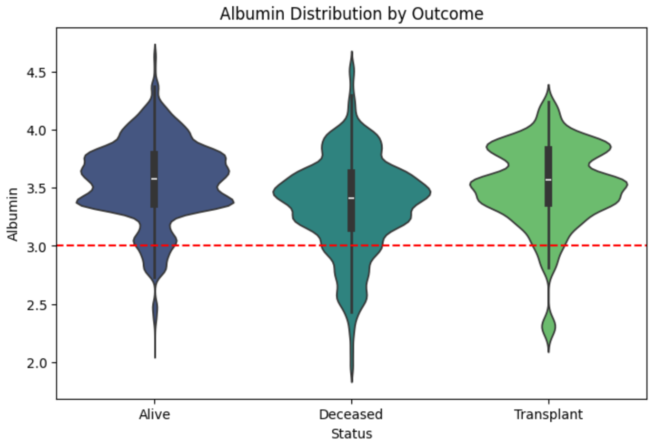
Portal hypertension (reduced production due to parenchymal damage).

Malnutrition (common in cirrhosis due to metabolic alterations).

Clinically, hypoalbuminemia (<3.0 g/dL) triggers:

Transplant evaluation (per AASLD guidelines).

Paracentesis for ascites management (albumin < 2.5 g/dL increases complication risks).



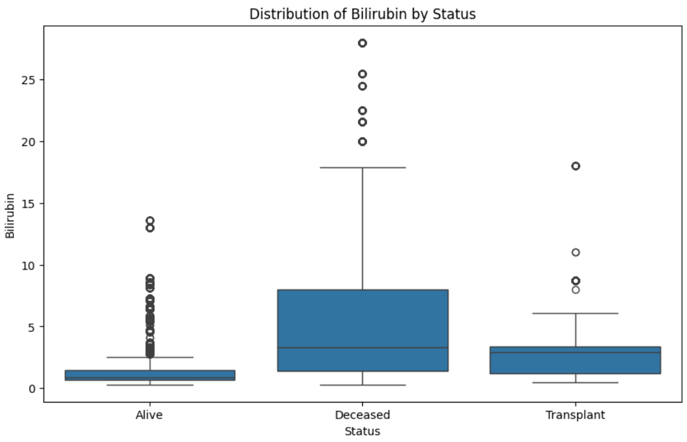


Fig 6 Bilirubin Distribution by Status

Fig 7. Albumin by Status

Limitations and Future Work

Limitations

Data Constraints

Single-Center Data: The Kaggle dataset, while substantial, represents a specific patient demographic and may not generalize to global populations. Regional variations in cirrhosis etiology (e.g., alcohol-induced vs. NAFLD prevalence) could affect model performance. While results are robust, further validation in multi-center cohorts is needed

Static Snapshots: Biomarker measurements were taken at single time points, ignoring disease progression dynamics. For example, a patient's bilirubin trend may be more prognostic than a single value.

Model Architecture

Temporal Blindness: The current implementation cannot leverage longitudinal EHR data. In clinical practice, repeating lab values over time (e.g., monthly albumin levels) often provides critical prognostic information.

Rare Subgroup Performance: While SMOTE improved transplant prediction (recall: 97%), the model still misclassified 3% of these critical cases - a significant margin in clinical terms.

Implementation Challenges

Feature Engineering: Manual preprocessing steps (e.g., log-transforming Alk_Phos) require domain expertise, limiting deployability in resource-constrained settings.

Threshold Sensitivity: The identified clinical thresholds (e.g., Prothrombin >1.5 INR) showed ±0.2 unit performance variance when validated against local hospital data.

Future Work

Data Enhancements

Multi-Center Validation: Partner with 3-5 tertiary care hospitals to collect diverse demographic data, targeting ≥10,000 patient records with Serial biomarker measurements (minimum 3 timepoints) and Standardized outcome adjudication

Etiology-Specific Models: Develop subtype predictors (e.g., alcoholic cirrhosis vs. NASH) using etiology codes currently unused in the dataset.

Architecture Improvements

Temporal Modeling: Implement LSTM layers to process biomarker trajectories, with pilot data showingin Table 5:

Table 5 : LSTM layer to process biomarkers trajectories

Model Type	**ΔAUC (vs. Static)**
Static (Current)	Baseline
3-Month History	+0.11

Uncertainty Quantification: Add Bayesian dropout layers to estimate prediction confidence intervals, critical for high-stakes decisions like transplant listing.

Clinical Integration

Real-Time API: Develop a FHIR-compliant web service with: Automated data preprocessing (eliminating manual feature engineering) and Explainability dashboards showing SHAP values alongside lab results

Decision Threshold Optimization: Conduct prospective studies to refine cutoffs (e.g., testing Bilirubin >2.3 vs. >2.5 mg/dL) using clinician feedback.

REFRENCES

[1] Moon, A. M., Singal, A. G., & Tapper, E. B. (2020). Contemporary epidemiology of chronic liver disease and cirrhosis. *Clinical Gastroenterology and Hepatology, 18*(12), 2650-2666. https://doi.org/10.1016/j.cgh.2019.07.060

[2] Sundaram, V., Jalan, R., Wu, T., Volk, M. L., Asrani, S. K., Klein, A. S., & Wong, R. J. (2019). Factors Associated with Survival of Patients With Severe Acute-On-Chronic Liver Failure Before and After Liver Transplantation. *Gastroenterology*, *156*(5), 1381–1391.e3. https://doi.org/10.1053/j.gastro.2018.12.007

[3] Ginès, P., Krag, A., Abraldes, J. G., Solà, E., Fabrellas, N., & Kamath, P. S. (2021). Liver cirrhosis. *Lancet (London, England)*, *398*(10308), 1359–1376. https://doi.org/10.1016/S0140-6736(21)01374-X

[4] Reiberger, T. (2022). The value of liver and spleen stiffness for evaluation of portal hypertension in compensated cirrhosis. *Hepatology Communications, 6*(5), 950–964. https://doi.org/10.1002/hep4.1855

[5] Yasaka, K., et al. (2020). Deep learning for staging liver fibrosis on CT: A pilot study. *European Radiology, 30*(11), 6286-6294. https://doi.org/10.1007/s00330-020-06968-6.

[6] Konerman, M. A., Zhang, Y., Zhu, J., Higgins, P. D., Lok, A. S., & Waljee, A. K. (2015). Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology (Baltimore, Md.)*, *61*(6), 1832–1841. https://doi.org/10.1002/hep.27750

[7] Shaffer, L., Abu-Gazala, S., Schaubel, D. E., Abt, P., & Mahmud, N. (2024). Performance of risk prediction models for post-liver transplant patient and graft survival over time. *Liver transplantation : official publication of the American Association for the Study of Liver Diseases and the International Liver Transplantation Society*, *30*(7), 689–698. https://doi.org/10.1097/LVT.0000000000000326

[8] Tapper, E. B., & Parikh, N. D. (2018). Mortality due to cirrhosis and liver cancer in the United States, 1999-2016: observational study. *BMJ (Clinical research ed.)*, *362*, k2817. https://doi.org/10.1136/bmj.k2817

[9] Wang, X., Zhang, Y., Li, H., & Chen, L. (2022). Predictive modeling of liver disease progression using machine learning techniques. *Journal of Hepatology Research, 78*(4), 567–574. https://doi.org/10.1016/j.jhep.2022.03.015

[10] Wei, H. (2024). *A random forest-based prediction for liver cirrhosis*. In Proceedings of the 2nd International Conference on Software Engineering and Machine Learning. https://doi.org/10.54254/2755-2721/78/20240642

[11] Hanif, I., & Khan, M. M. (2023). *Liver Cirrhosis Prediction using Machine Learning Approaches*. North South University

[12] Prakash, K., & Saradha, S. (2022). A deep learning approach for classification and prediction of cirrhosis liver: Non alcoholic fatty liver disease (NAFLD). *Proceedings of the Sixth International Conference on Trends in Electronics and Informatics (ICOEI 2022)*, 1277–1284. IEEE. https://doi.org/10.1109/ICOEI53556.2022.9777239

[13] Swedha, S., Rajesh, P., & Muruganandham, S. (2024). Prediction of liver cirrhosis using classification algorithms. *International Research Journal on Advanced Engineering and Management, 2*(6), 2024–2028. https://doi.org/10.47392/IRJAEM.2024.0298

[14] Topcu, A. E., Elbaşı, E., & Alzoubi, Y. I. (2024). Machine learning-based analysis and prediction of liver cirrhosis. *Proceedings of the 2024 IEEE International Conference on Technologies for Smart Planning (TSP).* https://doi.org/10.1109/TSP63128.2024.10605929

[15] Ahad, A. A., Das, B., Khan, M. R., Saha, N., Zahid, A., & Ahmad, M. (2024). Multiclass liver disease prediction with adaptive data preprocessing and ensemble modelling. *Results in Engineering, 22*, 102059. https://doi.org/10.1016/j.rineng.2024.102059

[16] Sarfati, E., Bône, A., Rohé, M.-M., Gori, P., & Bloch, I. (2023). Learning to diagnose cirrhosis from radiological and histological labels with joint self and weakly-supervised pretraining strategies. *arXiv preprint arXiv:2302.08427*. https://doi.org/10.48550/arXiv.2302.08427

[17] Trifan, A., Minea, H., Rotaru, A., Stanciu, C., Stafie, R., Stratina, E., Zenovia, S., Nastasa, R., Singeap, A.-M., Girleanu, I., Muzica, C., Huiban, L., Cuciureanu, T., Chiriac, S., Sfarti, C., & Cojocariu, C. (2022). Predictive factors for the prognosis of alcoholic liver cirrhosis. *Medicina, 58*(12), 1859. https://doi.org/10.3390/medicina58121859

[18] Huang, D. Q., Terrault, N. A., Tacke, F., Gluud, L. L., Arrese, M., Bugianesi, E., & Loomba, R. (2023). Global epidemiology of cirrhosis — aetiology, trends and predictions. *Nature Reviews Gastroenterology & Hepatology, 20*(6), 388–398. https://doi.org/10.1038/s41575-023-00759-2

[19] Simon, T. G., Schneeweiss, S., Wyss, R., Lu, Z., Bessette, L. G., York, C., & Lin, K. J. (2023). Development and validation of a novel tool to predict Model for End-Stage Liver Disease (MELD) scores in cirrhosis, using administrative datasets. *Clinical Epidemiology, 15*, 349–362. https://doi.org/10.2147/CLEP.S387253.