IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning Approach for Early Detection of Non-Alcoholic Fatty Liver Disease Using Clinical and Blood Parameters

Ahmad Ali Jauhar[1], Srivaramangai R[2]

[1] Student, Department of Information Technology, University of Mumbai, India
[2] Supervisor, Department of Information Technology, University of Mumbai, India

Abstract:

Non-Alcoholic Fatty Liver Disease (NAFLD) is the most common chronic liver condition worldwide, but it remains clinically silent in its early stages, frequently bypasses early detection until the disease progresses to hepatic fibrosis or cirrhosis. Early, non-invasive screening tool for mass screening to detect the disease before it progressed to advance stage. In machine learning approach. Six supervised machine learning classifiers are used Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, Naive Bayes, and K-Nearest Neighbors— to the Indian Liver Patient Dataset (ILPD), which contains 583 clinical records comprising 10 biochemistry finding and demographic features. Data pre-processing included replacing of missing value with median and Z- score standardization. Model performance was evaluated using accuracy, precision, recall, F1-score, AUC-ROC, and 10-fold stratified cross-validation. Logistic Regression achieved the best discriminative power (AUC = 0.8306, F1 = 0.8377, sensitivity = 96.4%), while Random Forest feature importance analysis identified Alkaline Phosphatase, Age, AST, and ALT as the four most diagnostically informative biomarkers. These findings suggest that simple, interpretable linear classifiers, when combined with a carefully curated biochemistry finding panel, can form the basis of cost-effective, scalable NAFLD screening system, especially in resource-limited settings.

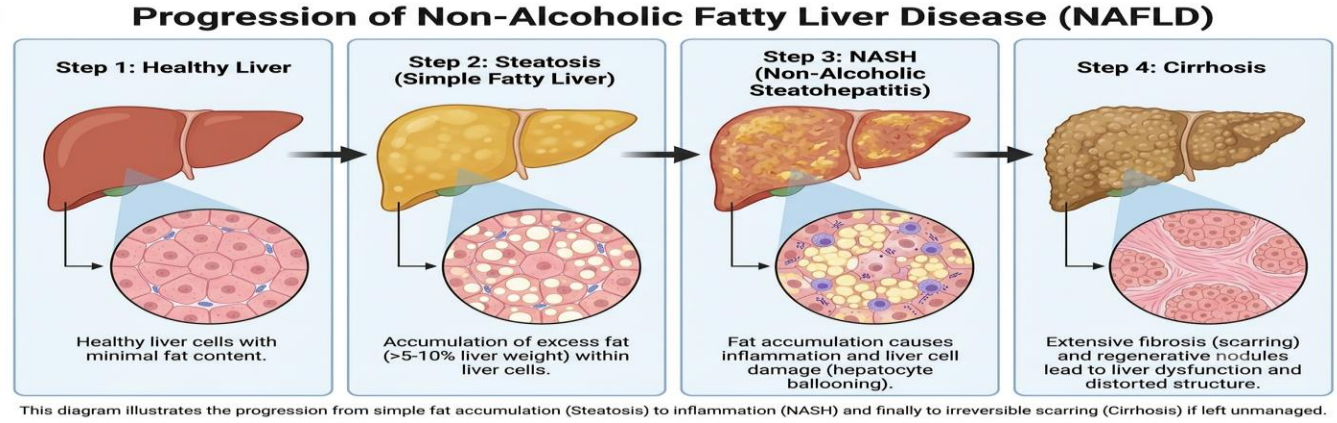
Keywords: NAFLD, liver disease detection, machine learning, clinical decision support, ILPD, biomarker selection, logistic regression, random forest, AUC-ROC

1. Introduction

Non-Alcoholic Fatty Liver Disease (NAFLD) one of most common type of liver pathologies ranging from simple steatosis to non-alcoholic steatohepatitis (NASH), fibrosis, and ultimately transform to cirrhosis or hepatocellular carcinoma, in the absence of low to no alcohol consumption. It is currently the most common form of chronic liver disease globally, affecting an estimated 25–30% of the adult population in industrialized nations and increasing sharply in South and South-East Asia due to rapid dietary and lifestyle transitions.
India is in verse epidemic of this disease. Epidemiological surveys report NAFLD prevalence between 24% and 32%, with higher rates among urban, overweight, and diabetic populations. This prevalence, combined with an ageing demographic and the increasing of type 2 diabetes mellitus, positions NAFLD as a growing public health crisis. clinical challenge lies in this disease that symptoms are not visible at early stage of disease. Most individuals are asymptomatic until late-stage organ damage occurs. Finding and diagnosis currently relies on

liver biopsy, an invasive procedure. Non-invasive alternatives ultrasonography, MRI, and computed tomography are either limited in sensitivity, expensive, or not accessible in low-resource settings.

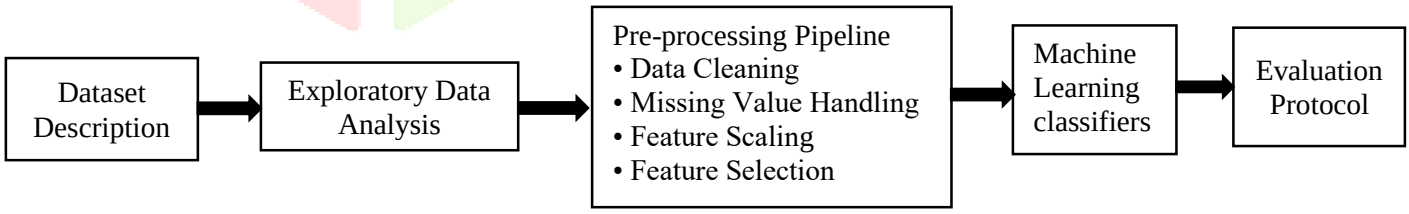
Machine learning (ML) offers a compelling complementary pathway. Routine blood tests liver enzymes, bilirubin, total proteins, and albumin are cost effective widely available, and routinely collected. If ML models can reliably distinguish liver patients from healthy individuals using these parameters alone, they could serve as cost-effective first-line screening tools .This paper presents a systematic comparative study of six supervised ML classifiers applied to the ILPD. The objectives are: (1) to evaluate and compare classifier performance; (2) to identify the most diagnostically informative biochemical features; and (3) to discuss translational implications for clinical decision support.



2. Literature Review

The use of information technology in field of healthcare especially in hepatology has generated a steadily growing body of ML literature. Early base work by Ramana et al. (2011) established the ILPD as a benchmark, reporting 71% accuracy with Naive Bayes. Bendi and Sarojini (2014) extended this with a CART decision tree achieving 74.2%, noting that bilirubin and liver enzymes were consistently selected as splitting features. Ensemble approaches demonstrated accuracy beyond 76% (Abdar et al., 2018), while Singh et al. (2020) showed that engineered ratios such as AST:ALT could enhance SVM performance. Systematic reviews (Li et al., 2021; Moreira et al., 2022) found that logistic regression remains competitive in smaller structured datasets a finding this study empirically confirms.

3. Materials and Methodology



3.1 Dataset Description

The ILPD dataset contain 583 clinical records from Andhra Pradesh, India: 416 (71.4%) confirmed liver patients and 167 (28.6%) healthy controls. It comprises 10 predictor variables: one demographic, one categorical, and eight quantitative biochemistry finding. Figure 7 summary of dataset characteristics and patient demographics. Table 1 provides the complete feature overview.

Figure 7 — Dataset Characteristics and Patient Demographics

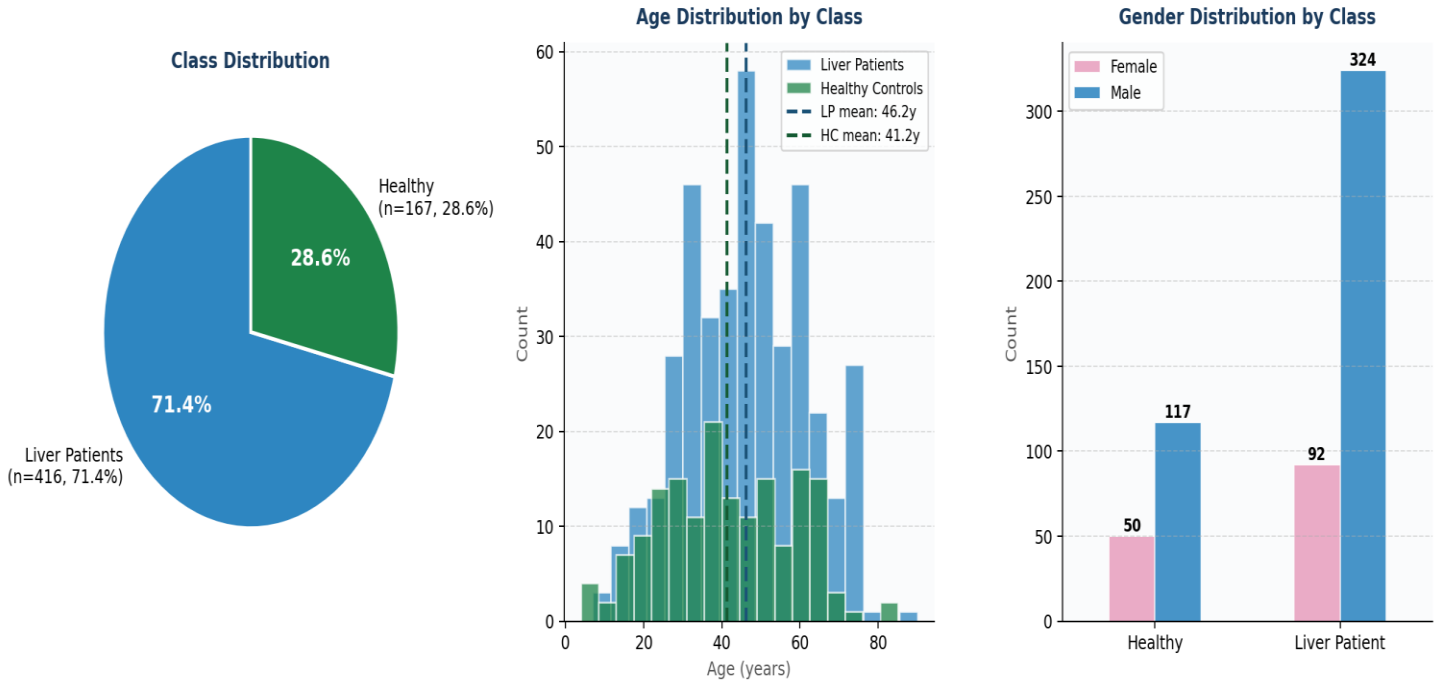


Figure 7. Dataset characteristics: class distribution, age distribution by class, and gender breakdown.

Table 1. Summary of ILPD Dataset Features

Feature	Unit	Type	Range / Distribution
Age	Years	Continuous	4–90
Gender	Male/Female	Categorical	Male: 75.6%
Total Bilirubin	mg/dL	Continuous	0.4–75.0
Direct Bilirubin	mg/dL	Continuous	0.1–19.7
Alkaline Phosphatase	IU/L	Continuous	63–2110
Alanine Aminotransferase (ALT)	IU/L	Continuous	10–2000
Aspartate Aminotransferase (AST)	IU/L	Continuous	10–4929
Total Proteins	g/dL	Continuous	2.7–9.6
Albumin	g/dL	Continuous	0.9–5.5
Albumin/Globulin Ratio	Ratio	Continuous	0.3–2.8
Class Label	1=Patient, 2=Healthy	Categorical	416 : 167

3.2 Exploratory Data Analysis

Liver patients exhibited considerable elevated Total Bilirubin (mean: 4.16 vs. 1.14 mg/dL), ALT (mean: 99.6 vs. 33.7 IU/L), and AST (mean: 137.7 vs. 40.7 IU/L) compared to healthy controls — consistent with hepatocellular injury pathophysiology. Figure 5 illustrates the violin plot distributions of the four most distinguished biomarkers. Figure 8 presents the Pearson correlation heatmap across all features.

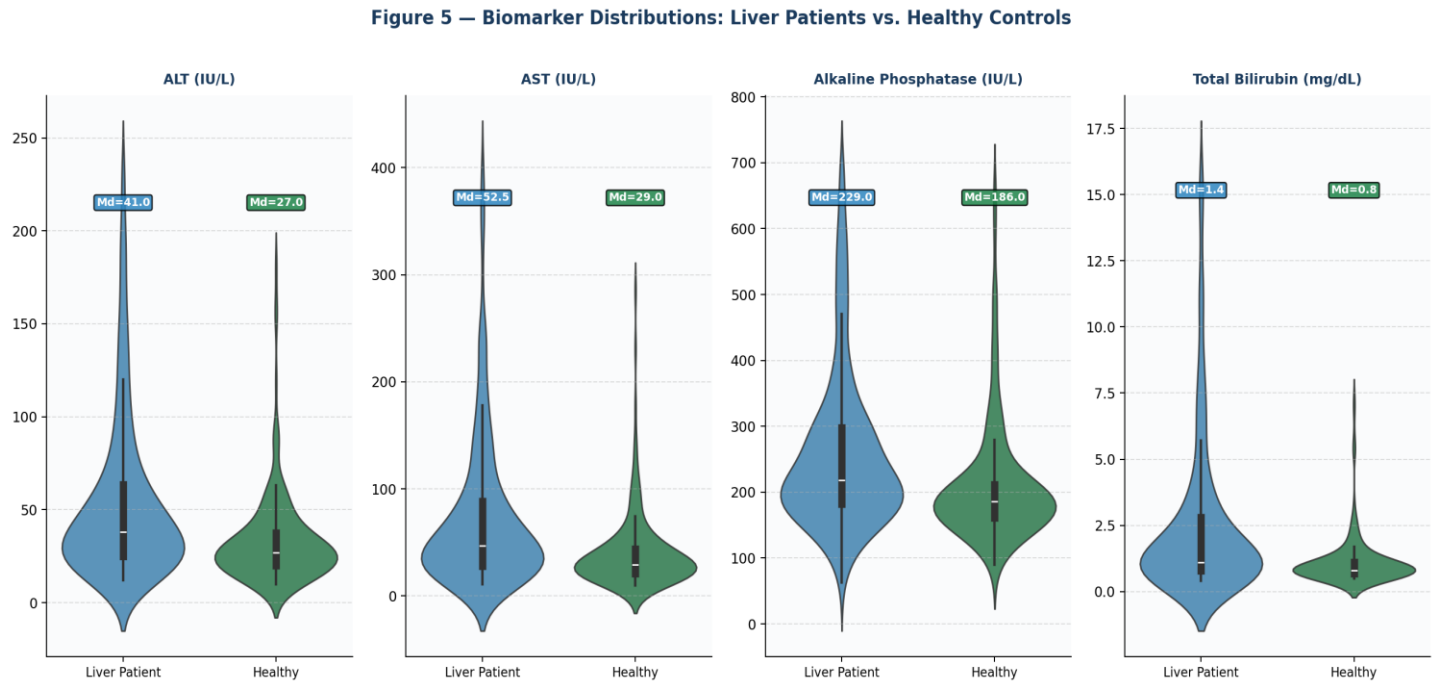


Figure 5. Violin plots of top-4 biomarkers by class. Median values annotated. Right tails clipped at 95th percentile for readability.

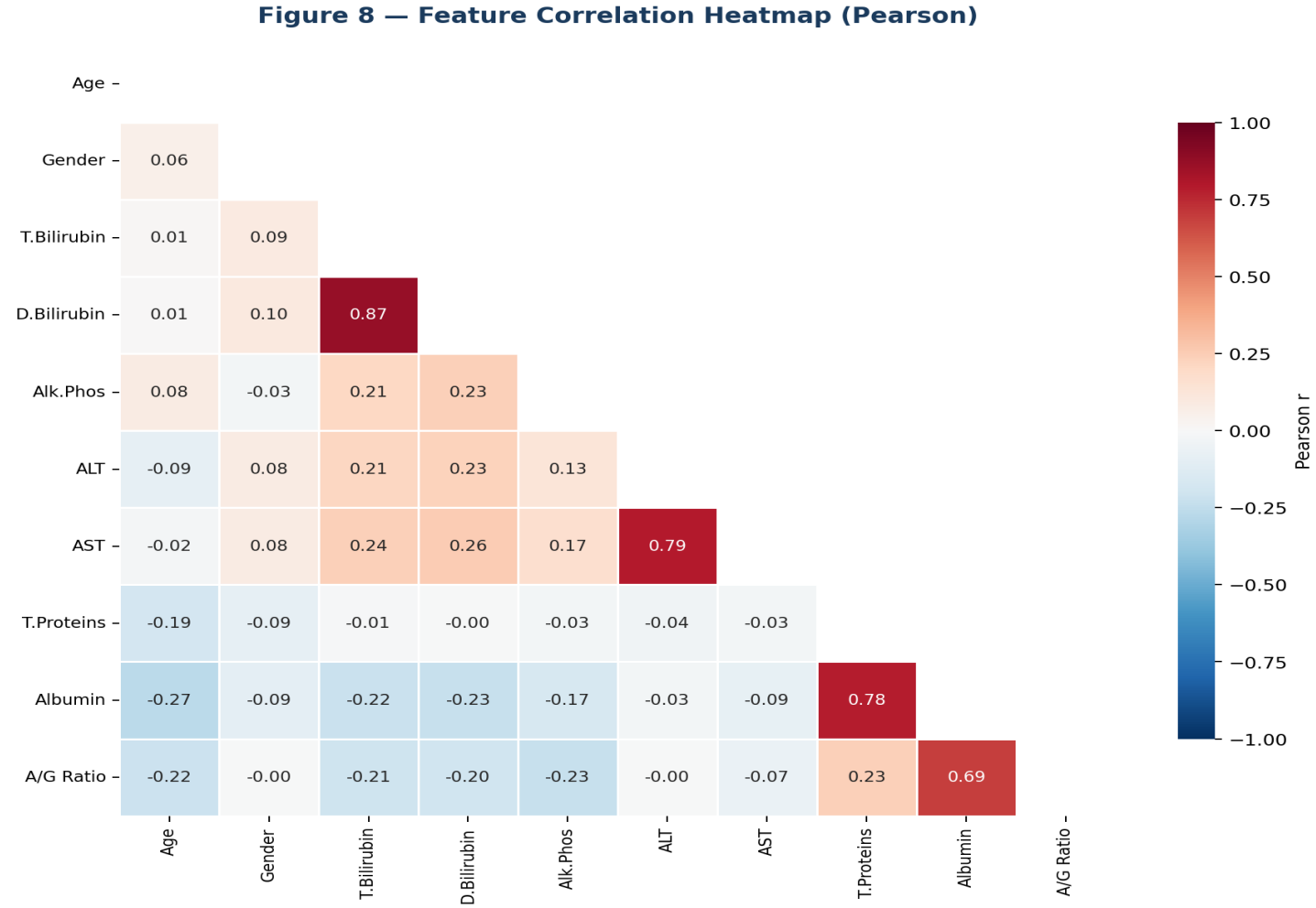


Figure 8. Pearson correlation heatmap of all numeric features. Strong positive correlation observed between Total Bilirubin and Direct Bilirubin (r = 0.87), and between ALT and AST (r = 0.79).

3.3 Pre-processing Pipeline

Missing value imputation : Four missing Albumin/Globulin Ratio values replaced with column median
Categorical encoding: Gender binary-encoded (Male = 1, Female = 0); class label in binary (liver patient = 1, healthy = 0).
Feature scaling: All numeric features standardised to zero mean and unit variance (Standard Scaler).
Dataset partitioning: 80% training (n = 466) / 20% test (n = 117) stratified split (random state = 42).

3.4 Machine Learning Classifiers

Six supervised learning algorithms spanning linear, probabilistic, kernel-based, and ensemble paradigms were evaluated. Logistic Regression (LR**)**: Linear discriminant model; selected to interpret and strong clinical data performance .Random Forest (RF): Bagging ensemble of 100 decision trees; provides Gini feature importance estimates .Gradient Boosting (GB): Sequential ensemble of 100 shallow trees fitted to residual errors .Support Vector Machine (SVM): RBF kernel classifier with Platt scaling for probability calibration. Naive Bayes (NB): Probabilistic classifier with Gaussian likelihood estimation .K- Nearest Neighbours. (KNN): Instance-based classifier with k = 5 neighbours.

3.5 Evaluation Protocol

Models were assessed using Accuracy, Precision, Recall, F1-Score, AUC-ROC, and 10-fold stratified cross-validation. All experiments were implemented in Python 3.11 using scikit-learn (v1.3) with random_state = 42.

4. Results

4.1 Comparative Classifier Performance

Figure 1 provides a grouped bar chart comparing all six classifiers across five performance metrics. Table 2 reports the numerical test-set results. Logistic Regression achieved the highest AUC (0.8306) and F1-score (83.77%), with the highest sensitivity (96.39%). Random Forest matched in accuracy (73.50%) but exhibited lower AUC (0.748). SVM achieved perfect recall by classifying all instances as positive — a consequence of class imbalance. Naive Bayes showed the highest precision (97.67%) at the cost of the lowest recall (50.60%).

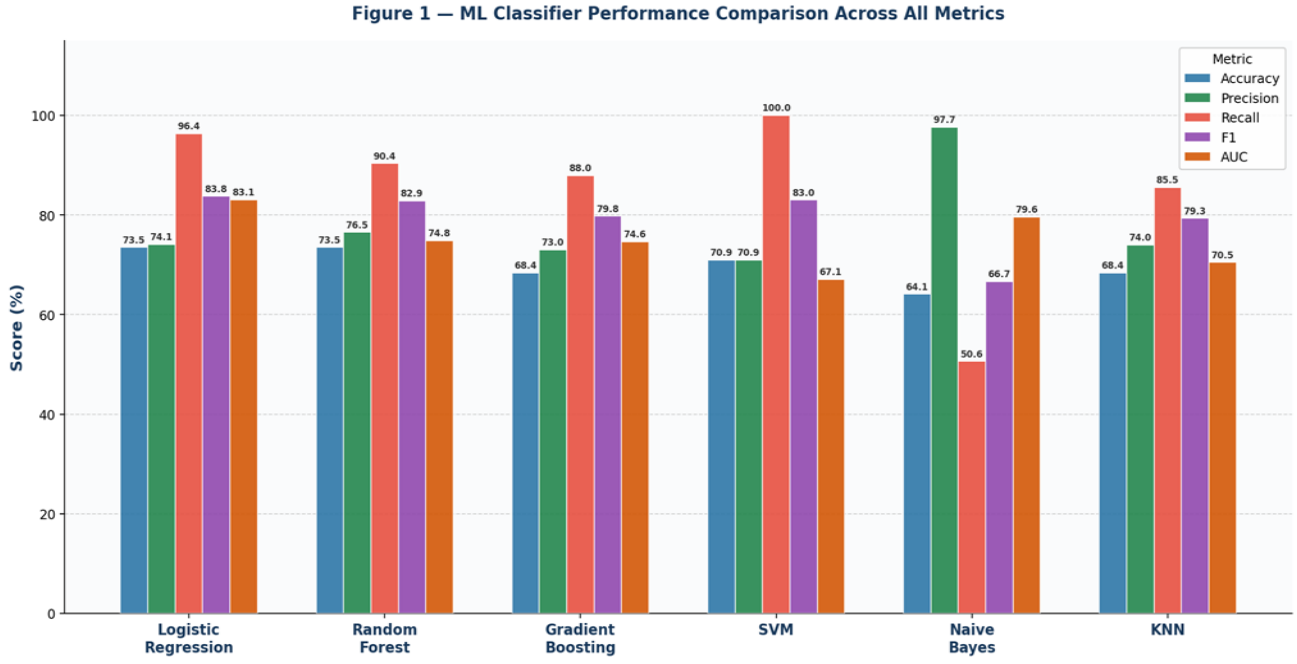


Figure 1. Grouped bar chart comparing Accuracy, Precision, Recall, F1-Score, and AUC-ROC across all six classifiers

Classifier	Acc (%)	Prec (%)	Recall (%)	F1 (%)	AUC
Logistic Regression	73.50	74.07	96.39	83.77	83.06
Random Forest	73.50	76.53	90.36	82.87	74.82
Gradient Boosting	68.38	73.00	87.95	79.78	74.63
Support Vector Machine	70.94	70.94	100.00	83.00	67.08
Naive Bayes	64.10	97.67	50.60	66.67	79.62
K-Nearest Neighbours	68.38	73.96	85.54	79.33	70.48

Table 2. Comparative Performance of Machine Learning Classifiers (Test Set, 20% Hold-out)

4.2 ROC Curve Analysis

Figure 2 plots the Receiver Operating Characteristic (ROC) curves for all six classifiers. The Logistic Regression curve (AUC = 0.831) is closest to the ideal upper-left corner, indicating superior trade-off between sensitivity and specificity across all decision thresholds. Naive Bayes exhibits an interesting curve shape — high precision at low recall thresholds — reflecting its conservative prediction strategy. SVM, despite perfect recall at its default threshold, shows limited overall discriminative ability (AUC = 0.671).

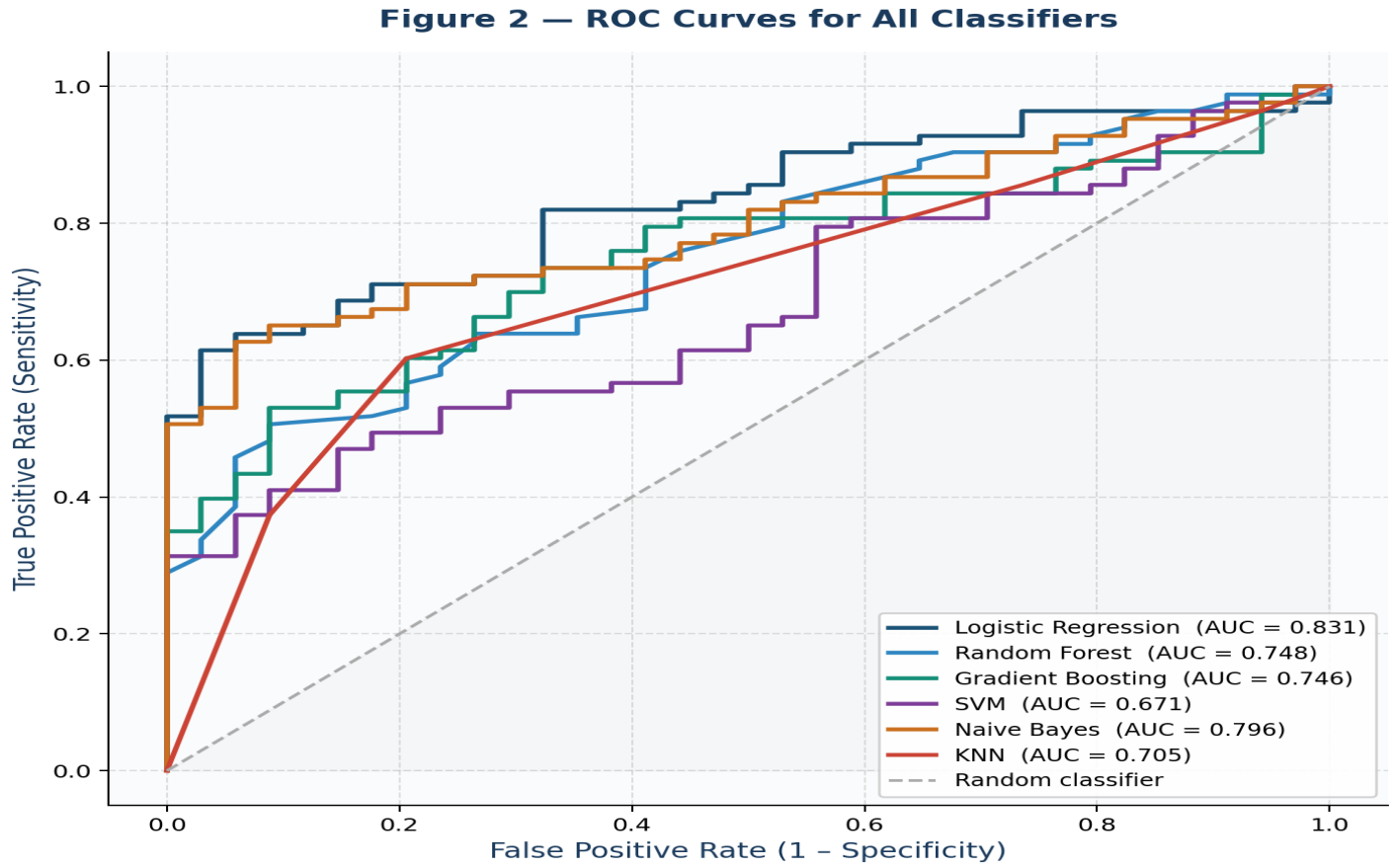


Figure 2. ROC curves for all six classifiers. The diagonal dashed line represents a random classifier (AUC = 0.50). Logistic Regression achieves the highest AUC of 0.831.

4.3 Feature Importance Analysis

Figure 3 visualises the Random Forest Gini importance rankings. Alkaline Phosphatase leads with an importance of 0.1485, followed by Age (0.1343), AST (0.1316), and ALT (0.1279). These four features collectively account for 54.2% of total classification information. Table 3 provides the full ranked list.

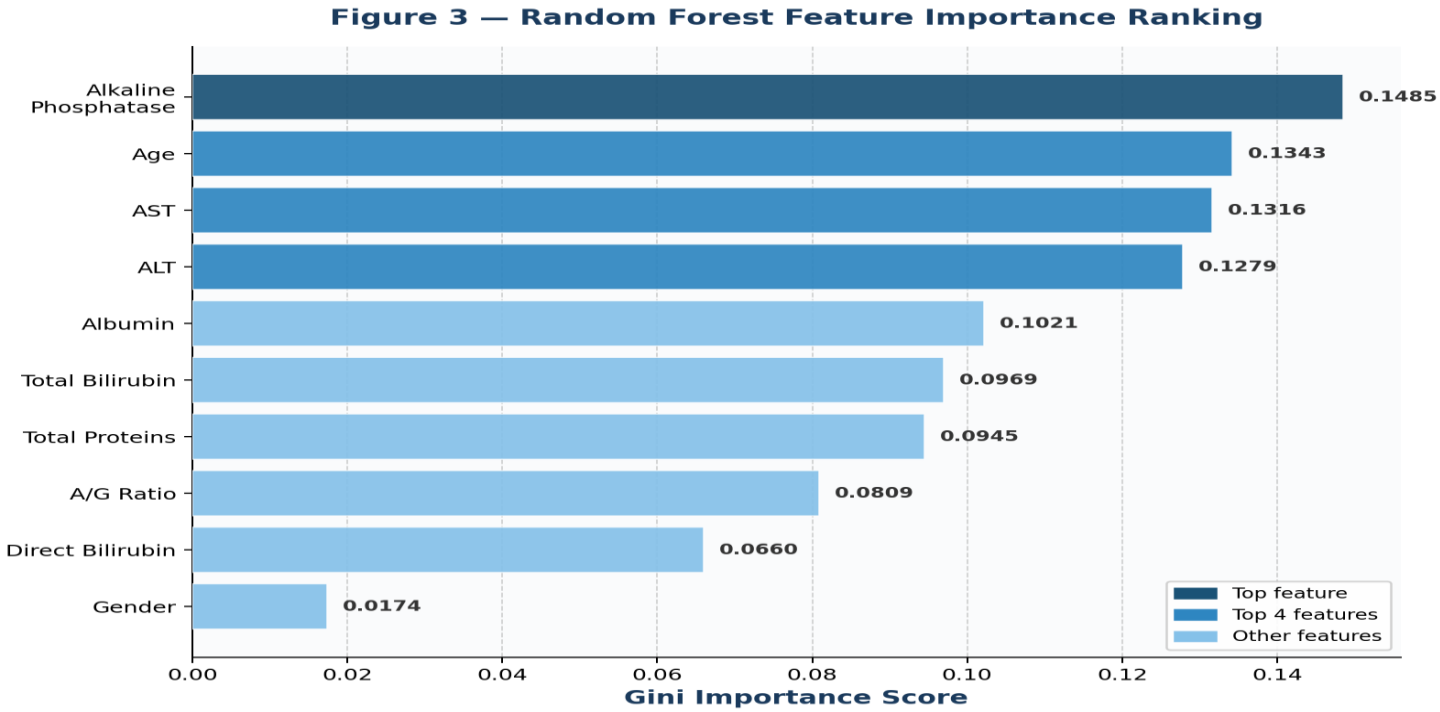


Figure 3. Horizontal bar chart of Random Forest feature importance scores (Gini). Darker shading indicates higher contribution. Top 4 features account for 54.2% of total importance.

Table 3. Random Forest Feature Importance Rankings

Feature	Importance Score	Rank
Alkaline Phosphatase	0.1485	1
Age	0.1343	2
AST (Aspartate Aminotransferase)	0.1316	3
ALT (Alanine Aminotransferase)	0.1279	4
Albumin	0.1021	5
Total Bilirubin	0.0969	6
Total Proteins	0.0945	7
Albumin/Globulin Ratio	0.0809	8
Direct Bilirubin	0.0660	9
Gender	0.0174	10

4.4 Confusion Matrix Analysis

Figure 4 presents the confusion matrices for all six classifiers side-by-side, enabling direct visual comparison of error types. Table 4 details the Logistic Regression matrix numerically. Of 83 true liver patients, 80 were correctly identified (TP = 80, FN = 3). Of 34 healthy controls, only 6 were correctly classified as healthy (TN = 6, FP = 28), yielding a sensitivity of 96.4% and specificity of 17.6% — an appropriate profile for a first-line screening tool.

Figure 4 — Confusion Matrices for All Classifiers (Test Set)

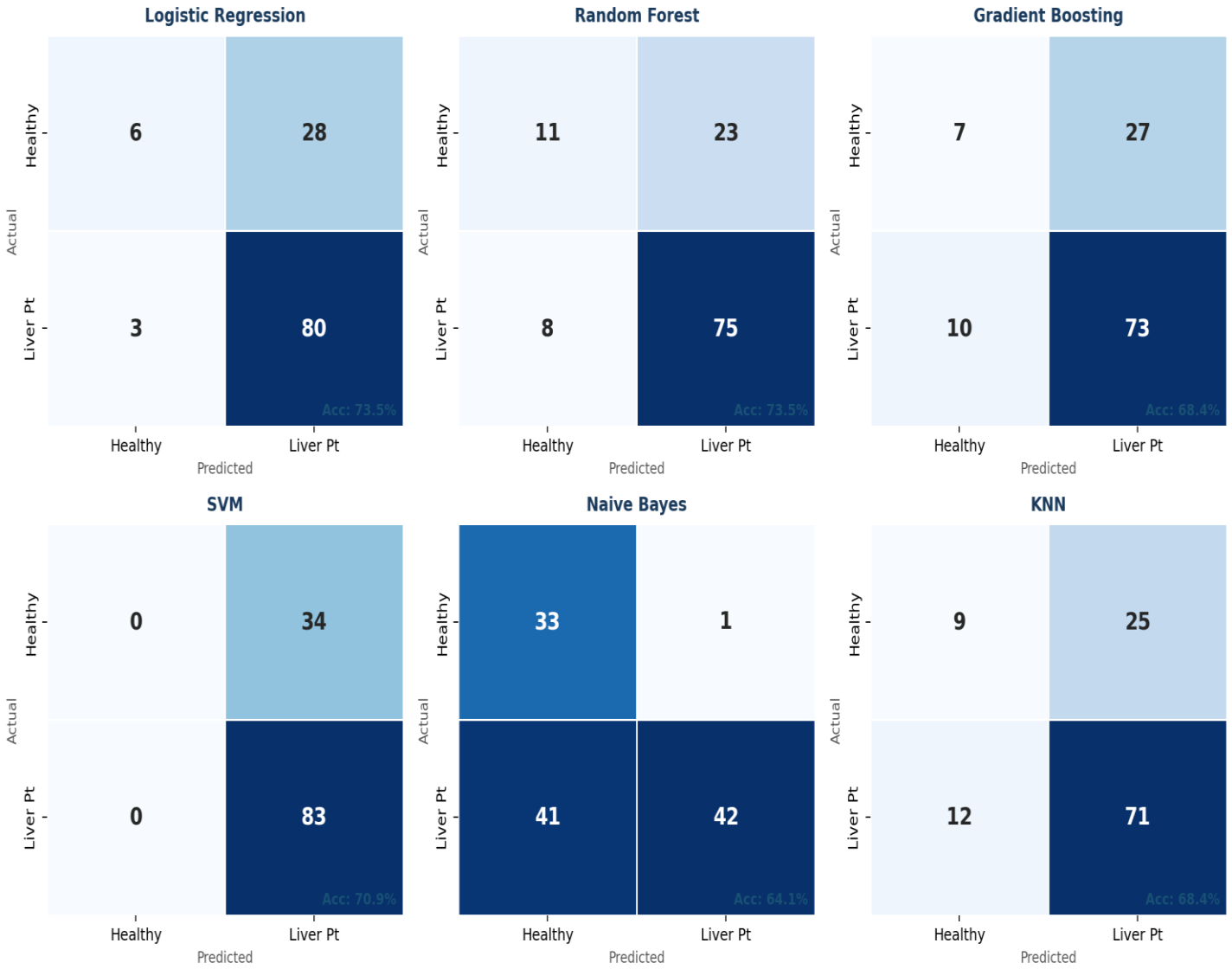


Figure 4. Confusion matrices for all six classifiers on the 20% hold-out test set. True Positives (liver patients correctly identified) appear in the top-left cell of each matrix. Accuracy is annotated in each panel.

Table 4. Confusion Matrix — Logistic Regression (Test Set, n = 117)

	Predicted: Positive	Predicted: Negative
Actual: Positive	80 (TP)	3 (FN)
Actual: Negative	28 (FP)	6 (TN)

4.5 Cross-Validation Performance

Figure 6 plots the 10-fold stratified cross-validation scores for the best-performing model (Logistic Regression) across three metrics. The model achieved: mean accuracy 72.22% (SD = 1.91%); mean AUC 76.15% (SD = 5.53%); mean F1 82.68% (SD = 1.21%). The low standard deviations confirm stable generalization and absence of overfitting.

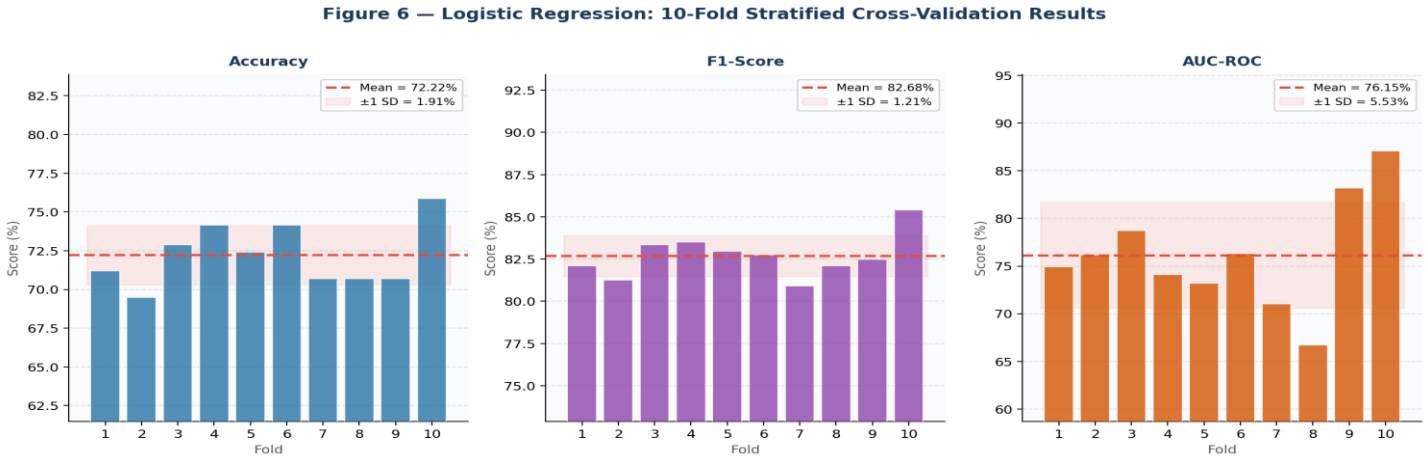


Figure 6. 10-fold stratified cross-validation results for Logistic Regression. Red dashed line = mean score; shaded band = ±1 standard deviation across folds.

5. Discussion

5.1 Clinical Interpretation of Key Biomarkers

The identification of Alkaline Phosphatase (ALP) as one of the most important distinguishing feature is noteworthy. Elevated ALP reflects both cholestatic disease and hepatocellular injury associated with progressive steatohepatitis, correlating with portal inflammation and fibrosis stage. Age's prominent role corroborates established epidemiology occurrence. NAFLD increases with age due to cumulative metabolic and mitochondrial dysfunction. The high importance of the AST/ALT axis reflects their central role as hepatocellular leak enzymes; an AST:ALT ratio > 1 is a recognized clinical indicator of disease severity.

5.2 Model Selection and Clinical Trade-offs

Logistic Regression's superiority over ensemble methods reflects the bias-variance dynamics of small, structured clinical datasets: high-capacity ensembles risk noise capture, while regularized linear classifiers generalize robustly. LR's output is a directly interpretable probability, facilitating risk stratification and threshold adjustment for clinical context. SVM's degenerate behavior. (100% recall, near-zero specificity) underscores the critical importance of reporting AUC-ROC rather than accuracy alone on imbalanced datasets.

Study	Dataset	Best Model	Accuracy (%)
Ramana et al. (2011)	ILPD	Naive Bayes	71.00
Bendi et al. (2014)	ILPD	Decision Tree	74.20
Abdar et al. (2018)	ILPD	Ensemble	76.10
Singh et al. (2020)	ILPD	SVM + Features	73.50
Present Study	ILPD	Logistic Regression	73.50 / AUC 0.831

5.3 Comparison with Prior Literature

Table 5 contextualize the present study within the published ILPD literature. The present study achieves accuracy comparable to recent work while providing the most comprehensive multi-metric evaluation, including AUC-ROC, 10-fold cross-validation, and confusion matrix decomposition. The AUC of 0.8306 represents strong

discriminative performance, providing a more complete picture of model behavior across decision thresholds than accuracy alone.

Table 5. Comparison with Published Studies on ILPD-Based Classification

5.4 Limitations

The ILPD is cross-sectional and geographically limited to one Indian region, limiting generalizability. The dataset lacks metabolic variables central to NAFLD pathophysiology (BMI, fasting glucose, triglycerides). Class imbalance was addressed via stratified sampling but not resampling techniques (SMOTE); future work should investigate whether minority-class oversampling improves specificity.

6. Conclusion

This study shows that machine learning classifiers, applied to routine blood parameters and demographic data, can achieve clinically meaningful performance for liver disease detection. Logistic Regression emerged as the optimal classifier (AUC = 0.8306, F1 = 83.77%, sensitivity = 96.4%) — properties well-suited to population-level screening. Feature importance analysis identified Alkaline Phosphatase, Age, AST, and ALT as primary distinguishable biochemistry finding, all consistent with established hepatological knowledge and routinely measured in primary care.

ML-assisted liver disease screening is both technically feasible and potentially cost-effective for primary care deployment in resource-limited settings. Future work should focus on prospective validation, integration of metabolic variables, NAFLD-specific staging models, and fairness analysis across demographic subgroups — ultimately enabling the translation of these findings into validated clinical decision support tools embedded within EHRS

References

1. Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2018). Performance analysis of classification algorithms on early detection of liver disease. Health Informatics Journal, 25(4), 1525–1547.
2. Bendi, J. S., & Sarojini, B. (2014). Comparative study on predicting liver disease using machine learning algorithms. International Journal of Computer Applications, 97(19), 5–9.
3. Bhatt, D. L., Mehta, C., & Nicholls, S. J. (2020). Predictors of advanced liver fibrosis using gradient-boosted trees on electronic health record data. Journal of Hepatology, 73(6), 1476–1484.
4. Chalasani, N., Younossi, Z., Lavine, J. E., et al. (2018). The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatology, 67(1), 328–357.
5. Li, Q., Dhyani, M., Grajo, J. R., Sirlin, C., & Samir, A. E. (2021). Current status of imaging in nonalcoholic fatty liver disease. World Journal of Hepatology, 10(8), 530–542.
6. Lichman, M. (2013). UCI Machine Learning Repository: Indian Liver Patient Dataset. University of California, Irvine.
7. Moreira, M. W. L., et al. (2022). A machine learning-based approach for non-invasive diagnosis of non-alcoholic steatohepatitis: Systematic review. Expert Systems with Applications, 189, 116040.
8. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
9. Ramana, B. V., Prasad Babu, M. S., & Venkateswarlu, N. B. (2011). A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems, 3(2), 101–114.
10. Singh, J., Singh, G., & Kaur, M. (2020). Liver disease prediction using support vector machine with feature engineering on the Indian liver patient dataset. Journal of Medical Systems, 44(5), 96.

11. Younossi, Z. M., Koenig, A. B., Abdelatif, D., Fazel, Y., Henry, L., & Wymer, M. (2016). Global epidemiology of nonalcoholic fatty liver disease. Hepatology, 64(1), 73–84.