AUIQ Technical Engineering Science

Manuscript 1056

AI-Driven Stratified Modeling for Early Liver Disease Detection: A Comparative Study of Ensemble and Conventional Machine Learning Classifiers

Ghadeer Murtadha Ali

Ali Aqeel Hadi

Mustafa Abdulkareem Abbas

Abdullah Alkarar Mohammad

Ahmed Fadhil Abdulhussein

AUIQ TECHNICAL ENGINEERING SCIENCE 2026;3:76–96

Scan the QR to view the full-text article on the journal website

ORIGINAL STUDY

AI-Driven Stratified Modeling for Early Liver Disease Detection: A Comparative Study of Ensemble and Conventional Machine Learning Classifiers

Ghadeer Murtadha Ali *, Ali Aqeel Hadi, Mustafa Abdulkareem Abbas, Abdullah Alkarar Mohammad, Ahmed Fadhil Abdulhussein

Hilla, Iraq

ABSTRACT

Background: Early prediction of liver disease remains challenging in routine clinical diagnostics due to the multifactorial nature of hepatic dysfunction and the limited discriminative capacity of conventional laboratory-only assessments.

Objective: This study aims to develop and rigorously evaluate a robust machine learning framework for binary liver disease classification, emphasizing predictive stability, diagnostic balance (sensitivity–specificity), and statistical reproducibility across repeated experiments.

Methodology: A structured dataset of 1,700 records with 11 features representing demographic, behavioral, genetic, and clinical determinants was used to train and compare five supervised models: CatBoost, AdaBoost, Random Forest, Support Vector Machine (SVM), and Decision Tree. Performance was assessed under two stratified train–test partitions (60:40 and 70:30). Each experimental configuration was independently repeated 20 times, and mean ± standard deviation was reported for Accuracy, Precision, Recall, F1-score, Specificity, and ROC-AUC to quantify reliability and cross-run robustness. Confusion matrices and ROC curves were used for complementary diagnostic interpretation.

Results: Ensemble learners consistently outperformed non-ensemble counterparts across both partitions. CatBoost exhibited the strongest and most stable performance, achieving approximately 0.88 Accuracy and 0.95 ROC-AUC, alongside the lowest false negative rate (~0.11). The narrow standard deviations across the 20-run protocol indicate high reproducibility and reduced sensitivity to random sampling effects, supporting the generalization strength of the proposed evaluation design.

Conclusion & Contribution: The study provides a statistically grounded, reproducible ML evaluation framework for early liver disease prediction and demonstrates that optimized ensemble learning, particularly CatBoost, can enhance diagnostic accuracy while reducing clinically critical misclassification. These findings support the feasibility of AI-assisted screening pipelines and establish a methodological foundation for future translation into non-invasive decision-support systems in preventive hepatology.

Keywords: Liver disease prediction, Machine learning, Ensemble learning, Biomedical data analysis, Clinical decision support

Received 2 December 2025; revised 22 January 2026; accepted 23 January 2026.
Available online 31 January 2026

* Corresponding author.
E-mail addresses: ghadeermurtadha2001@gmail.com (G. Murtadha Ali), alialmosay@gmail.com (A. A. Hadi), c1387651@gmail.com (M. A. Abbas), albaiati1998@gmail.com (A. A. Mohammad), ahmedfadil490@gmail.com (A. F. Abdulhussein).

https://doi.org/10.70645/3078-3437.1056

1. Introduction

Liver disease is a substantial morbidity and mortality contributor globally. Recent estimates at the global level estimate approximately two million deaths related to the liver per year when cirrhosis and liver cancer are combined [1, 2]. Thus, it underlines the fact that liver disease is a significant component of premature mortality in high-resource and low-resource settings. The World Health Organization reported that viral hepatitis accounted for approximately 1.3 million deaths annually, thus highlighting a need of early detection and prevention as well [3]. Late diagnosis has an association with irreversible damage to the organ as well as failure and increased total health care costs and utilization, hence making early diagnosis imperative [4]. Biochemical LFTs and imaging methods such as ultrasound or MRI provide markedly vital information within practical limitations due to vulnerability on expert-dependent manual interpretation, additional cost and availability limitations, and varying accuracy in different settings [5–7].

In recent years, machine learning (ML) has become one of the most effective clinical data analysis tools due to its ability to model complex nonlinear relationships between biochemical indicators, demographic factors, and disease outcomes. Numerous studies demonstrated that ML methods could attain at least as good or even better results compared to traditional statistical baselines under relevant validation [8, 9]. Several have applied ML for liver disease prediction on well-established clinical datasets- which include the Indian Liver Patient Dataset (ILPD) from UCI Machine Learning Repository [10]. Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forests (RF) are the algorithms that have been commonly implemented for use and offer substantially varying yet occasionally relatively powerful diagnostic performance. Gradient boosting, RF, XGBoost ensembles typically outperform others on tabular clinical data [11–13]. Deep learning methods have also recorded notable results in tasks related to predictions on the liver, with over 90% accuracies across various works and studies [14, 15]. These include Deep Neural Networks (DNNs) on ILPD and combined Convolutional Neural Network/ Fully Connected Network (CNN/FCN) architectures for acute liver failure, however, accuracy, interpretability, and external validity across different patient populations remain a principal issue [12, 13].

The prospective Clinical Decision Support Systems that ML can offer is optimizing workflows and enables earlier, more objective assessment, as the disease of the liver progresses. Therefore, predictive performance on transparency and clinical plausibility should be optimized in models intended for high-stakes decision-making [16, 17]. The challenges of class imbalance, overfitting, and inappropriate feature selection constrain generalizability and hence practical application. Therefore, standardized preprocessing together with a very rigorous validation process that includes an external test and transparent reporting according to emerging guidelines such as TRIPOD + AI are prerequisites. Consequently, there is also an pressing need for a harmonized pre-processing-evaluation environment to conduct a systematic comparative comparison between different machine-learning algorithms to determine which are clinically relevant frameworks for the prediction of liver diseases [18].

Therefore, the purpose of this study is to develop and rigorously evaluate a machine-learning framework for early liver disease classification using a structured clinical dataset, with emphasis on diagnostic reliability and generalization under repeated experimentation.

The novelty of this work lies in (i) conducting a systematic comparison of multiple supervised learners, including modern ensemble boosting (CatBoost and AdaBoost) alongside classical baselines, (ii) evaluating performance under two stratified train–test partitions (60:40 and 70:30), and (iii) reporting mean ± standard deviation across 20 independent runs using multiple clinically relevant metrics to quantify stability and reduce split-dependent bias; an aspect often underreported in prior liver disease prediction studies.

Despite the substantial progress reported across recent studies, several unresolved methodological limitations continue to restrict the clinical reliability and comparability of current liver disease prediction models. First, performance reporting remains highly split-dependent and protocol-sensitive, with wide variability in reported accuracies across works using similar datasets, suggesting that results are often influenced by inconsistent preprocessing choices and insufficiently standardized validation procedures. Second, many studies emphasize single-run or single-split performance, which limits confidence in model stability and reproducibility under different sampling conditions. Third, interpretability is frequently treated as an auxiliary add-on rather than an integrated diagnostic requirement, despite the need for clinically plausible explanations in high-stakes screening contexts. Accordingly, the main problem addressed in this study is the lack of a harmonized,

Fig. 1. A robust and reproducible machine-learning pipeline for liver disease prediction.

reproducible evaluation setting that enables reliable comparison of competing ML models while quantifying stability and diagnostic balance. To resolve this, the proposed approach implements a consistent preprocessing-and-evaluation workflow, conducts a systematic comparison of classical and ensemble learners, and introduces repeated experimentation across two stratified train–test partitions with mean ± standard deviation reporting over multiple clinically relevant metrics; thereby reducing split bias and strengthening the interpretability-readiness of liver disease prediction models.

2. Methodology

This section details the complete experimental framework adopted for developing and evaluating supervised machine-learning classifiers aimed at predicting liver-disease status. The methodological workflow encompassed dataset acquisition, exploratory data analysis, data preprocessing, and modeling pipeline configuration, all executed in a reproducible Python environment to ensure analytical transparency and clinical interpretability. As illustrated in Fig. 1, the end-to-end workflow proceeds

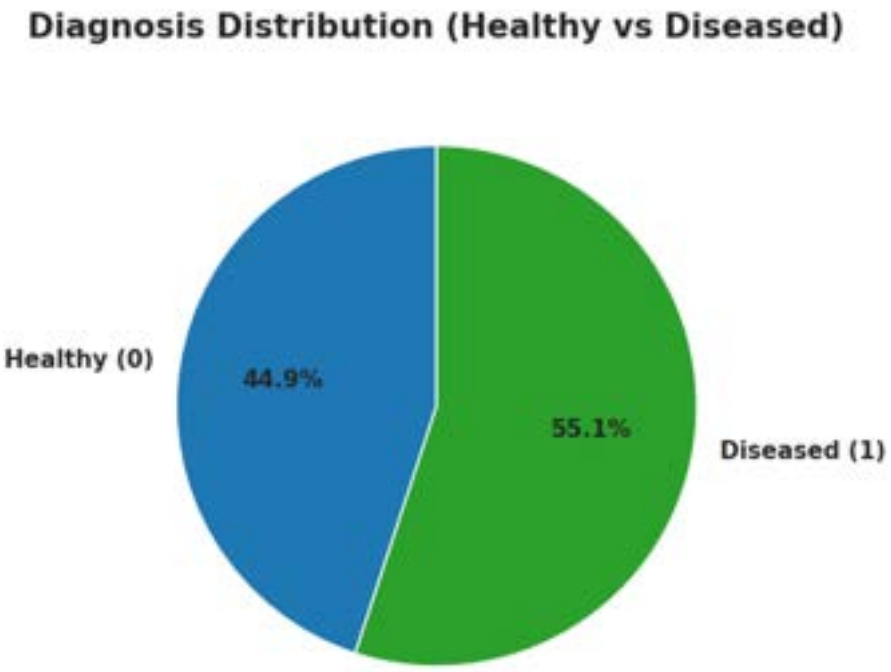

Fig. 2. Diagnosis distribution among participants (Healthy vs Diseased).

from dataset acquisition and exploratory data analysis to preprocessing and robustification, followed by a stratified experimental design (60:40 and 70:30), comparative model training (DT, SVM, RF, AdaBoost, and CatBoost), and repeated evaluation (20 independent runs) using stability-aware reporting (mean ± standard deviation) across multiple clinically relevant metrics.

2.1. Exploratory data analysis (EDA)

The exploratory data analysis (EDA) demonstrates that the dataset is statistically coherent and suitable for supervised classification, while also providing clinically interpretable signals to guide model design. Fig. 2 presents the target-label distribution, showing that liver disease is present in 55.1% of cases and absent in 44.9%. This matters because near-balanced classes reduce the risk of inflated accuracy due to majority-class dominance and enable standard metrics (e.g., ROC-AUC, F1-score) to reflect true discriminative ability. The insight is that only limited rebalancing pressure is expected, allowing algorithmic performance to primarily reflect learning capacity rather than class-correction effects.

Fig. 3 shows the sex distribution (49.6% male, 50.4% female), indicating a highly balanced demographic composition. This matters because imbalanced demographic representation can bias model learning and compromise fairness and generalization. The insight is that performance differences are less likely to be driven by sex-based sampling imbalance, supporting more equitable generalization across sexes.

Fig. 4 reports the distributions of continuous predictors (Age, BMI, Alcohol Consumption, Physical Activity, and Liver Function Test), spanning clinically plausible ranges (e.g., Age 20–80, BMI 15–40 kg/m^2, LFT 20–100). This matters because well-covered ranges and the absence of extreme skewness or sparsity reduce instability in training and improve the reliability of both linear-margin models (e.g., SVM) and nonlinear ensemble learners. The insight is that the dataset provides sufficient variability to learn clinically meaningful boundaries without being overly dominated by outliers or degenerate feature distributions.

Fig. 5 compares key continuous variables across healthy and diseased groups via boxplots, highlighting visible between-group separation patterns, particularly for liver-related biomarkers. This matters because group separation in EDA is an early indicator of feature discriminativeness and supports the feasibility of classification. The insight is that biomarker-centered variables (notably liver function indicators) exhibit diagnostic contrast, suggesting that models capable of capturing nonlinear interactions (e.g., boosting ensembles) may exploit these separations more effectively than simple rule-based splits.

Fig. 6 visualizes class-conditional distributions of binary risk factors (e.g., Smoking, Genetic Risk,

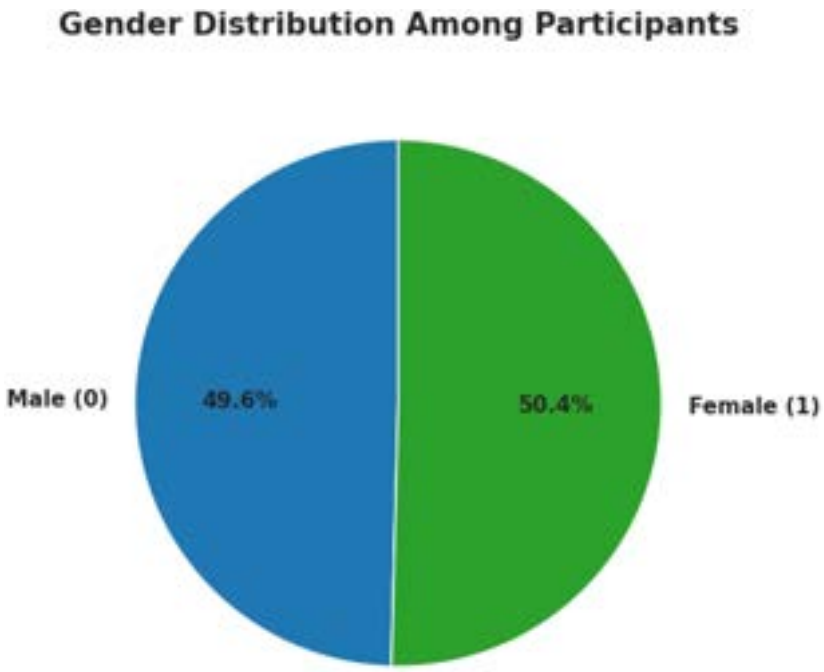

Fig. 3. Gender composition of the study cohort.

Diabetes, Hypertension), showing that positive risk states occur in meaningful but non-dominant proportions (e.g., Smoking 30.4%, Genetic Risk 42%, Diabetes 14.7%, Hypertension 15.5%). This matters because extremely rare binary predictors can be statistically unstable, whereas overly prevalent ones can trivialize prediction. The insight is that these binary variables provide informative clinical variance that can complement continuous biomarkers, supporting a mixed-feature modeling strategy.

Finally, Fig. 7 presents the Spearman correlation heatmap, where several predictors show monotonic association with the target diagnosis, including Liver Function Test ($\rho = 0.36$) and Alcohol Consumption ($\rho = 0.35$), alongside smaller but meaningful correlations for Smoking ($\rho = 0.20$), Gender ($\rho = 0.19$), and BMI ($\rho = 0.17$). This matters because correlation structure helps verify that predictors carry signal while also revealing whether multicollinearity could distort model interpretation or stability. The insight is twofold: (i) multiple clinically plausible variables contribute predictive signal, and (ii) inter-feature correlations are generally low (mostly $|\rho| < 0.10$), suggesting minimal multicollinearity and supporting stable learning across models. Overall, these EDA findings justify proceeding to supervised modeling with confidence that the dataset is balanced, well-behaved, and diagnostically informative.

2.2. Data preprocessing

2.2.1. Dataset loading and initial exploration

The dataset utilized in this study, titled Predict Liver Disease (available at Kaggle) [19], was obtained under the Attribution 4.0 International (CC BY) license, ensuring unrestricted academic use with proper citation. The dataset has been uploaded successfully for analysis and contains 1,700 cases (rows) and 11 features (columns) describing demographic, behavioral, genetic, and clinical variables related to liver health. The individual case profile described by continuous and categorical predictors is attached in Table 1: Dataset Description and Variable Characteristics. Initial examination revealed primary attributes relating to the liver such as results of medical tests on the function of the liver, personal habits such as alcohol consumption and smoking condition other disease conditions e.g., diabetes and hypertension. Such a broad framework creates multidimensional perspectives in understanding determinants of liver disease. Before proceeding into the next steps of normalization, encoding, and handling outliers, a verification was conducted regarding datatype consistency, completeness, and code accuracy of the variables. Diagnosis, being a binary target variable, served as the dependent feature so that an structured supervised classification structure could be developed for predicting the presence of liver disease. Exploratory Data Analysis and Machine Learning

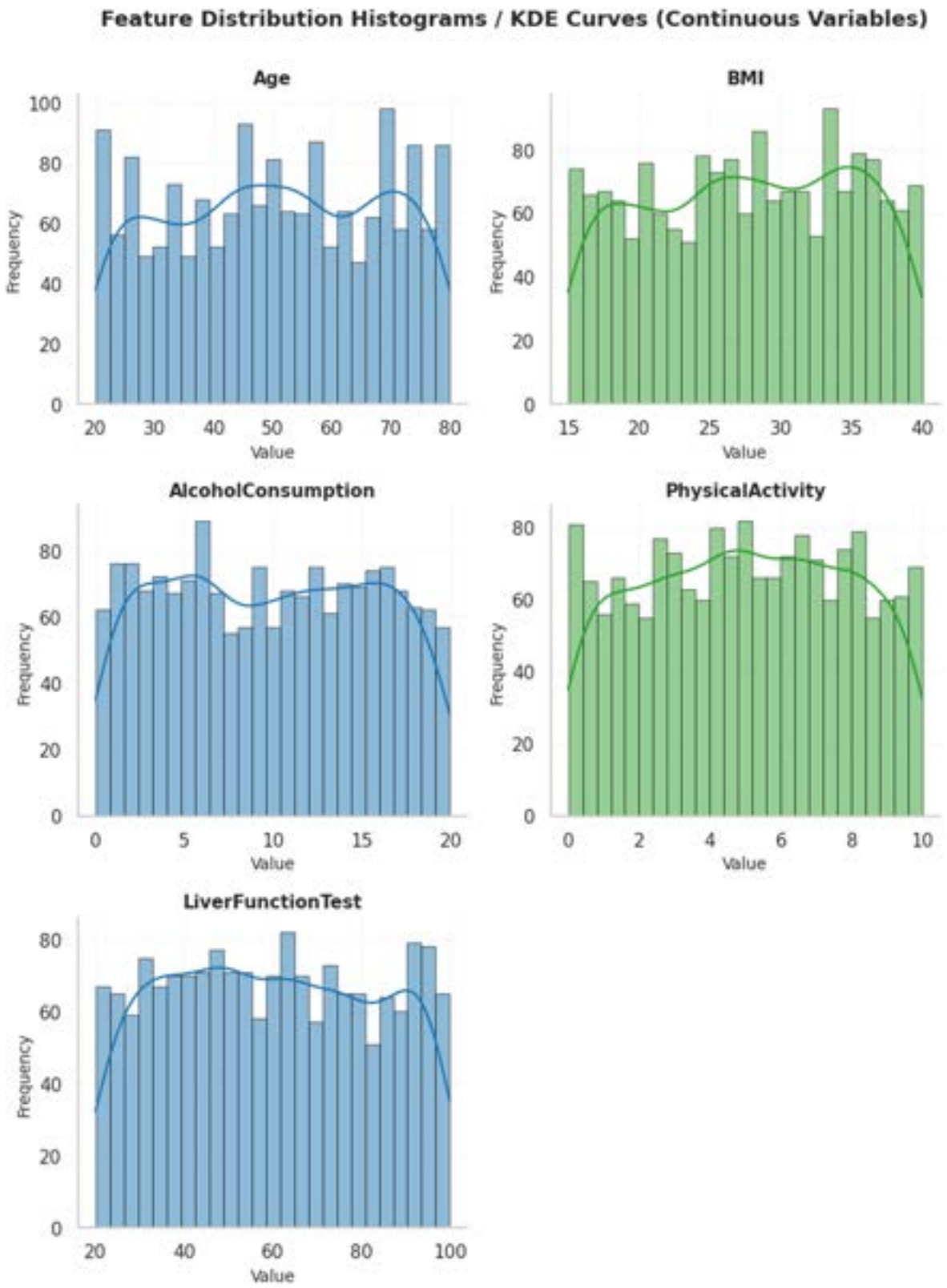

Fig. 4. Feature distribution histograms and KDE curves for continuous variables.

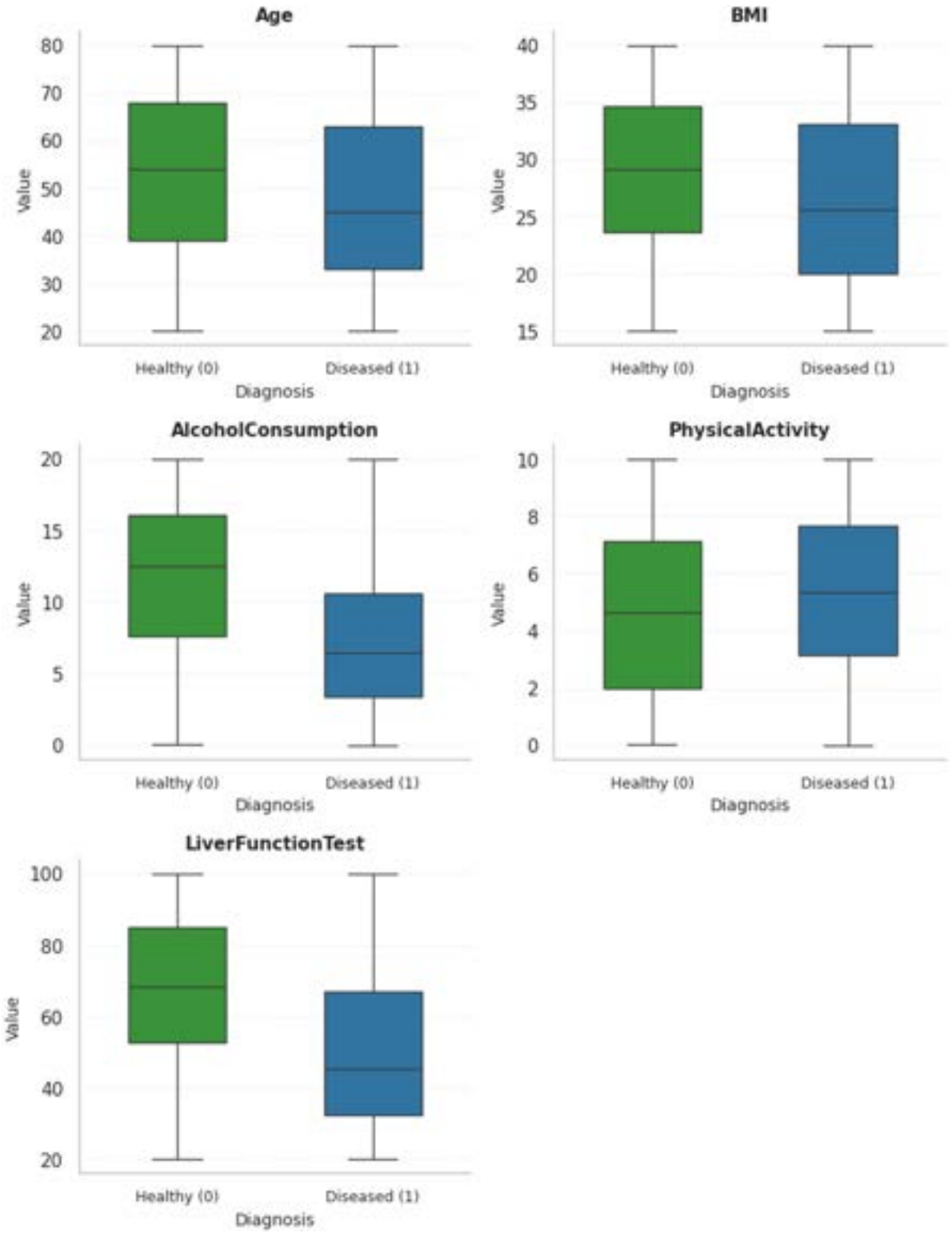

Fig. 5. Comparative box plots of continuous features by diagnostic group.

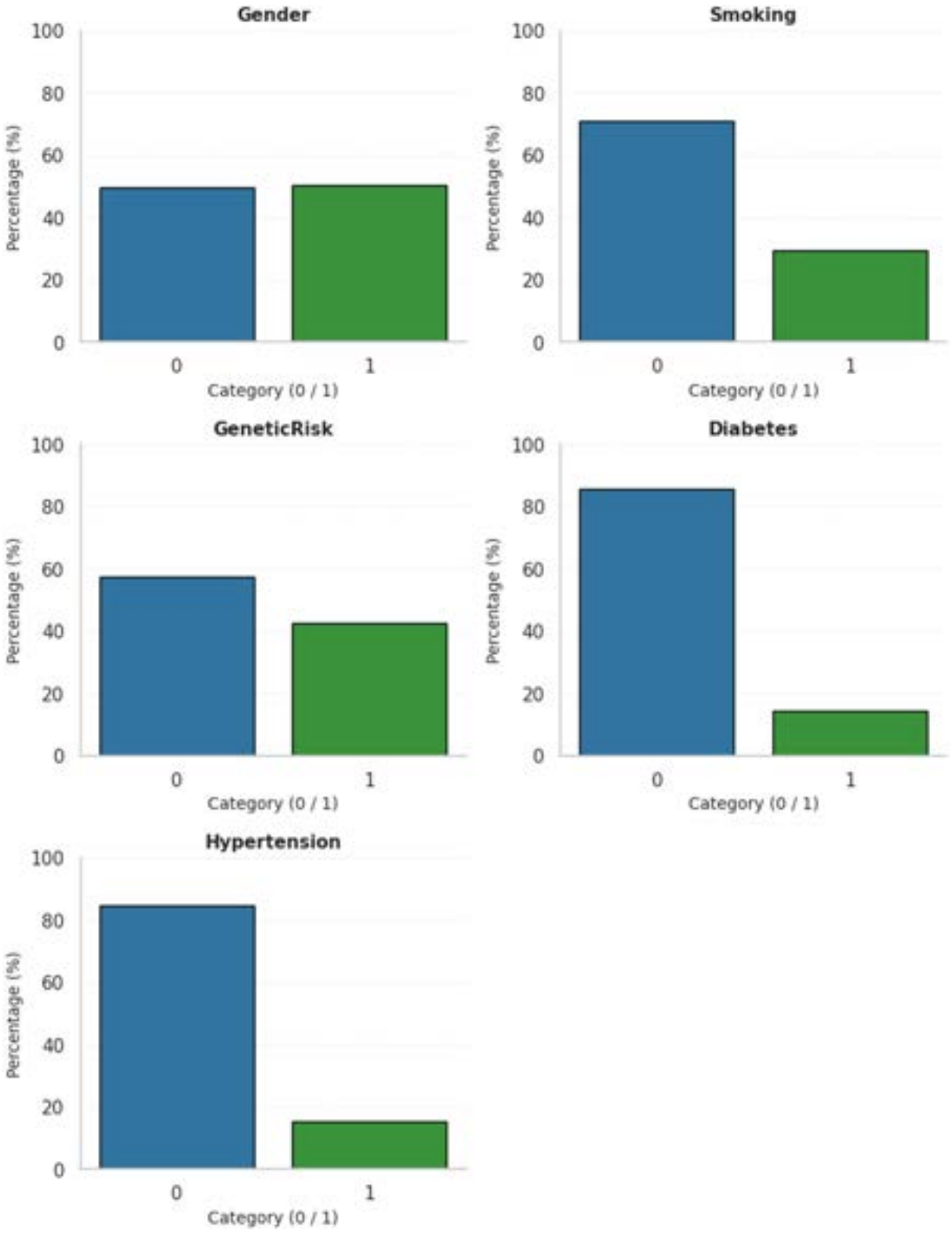

Fig. 6. Distribution of categorical risk factors in the study population.

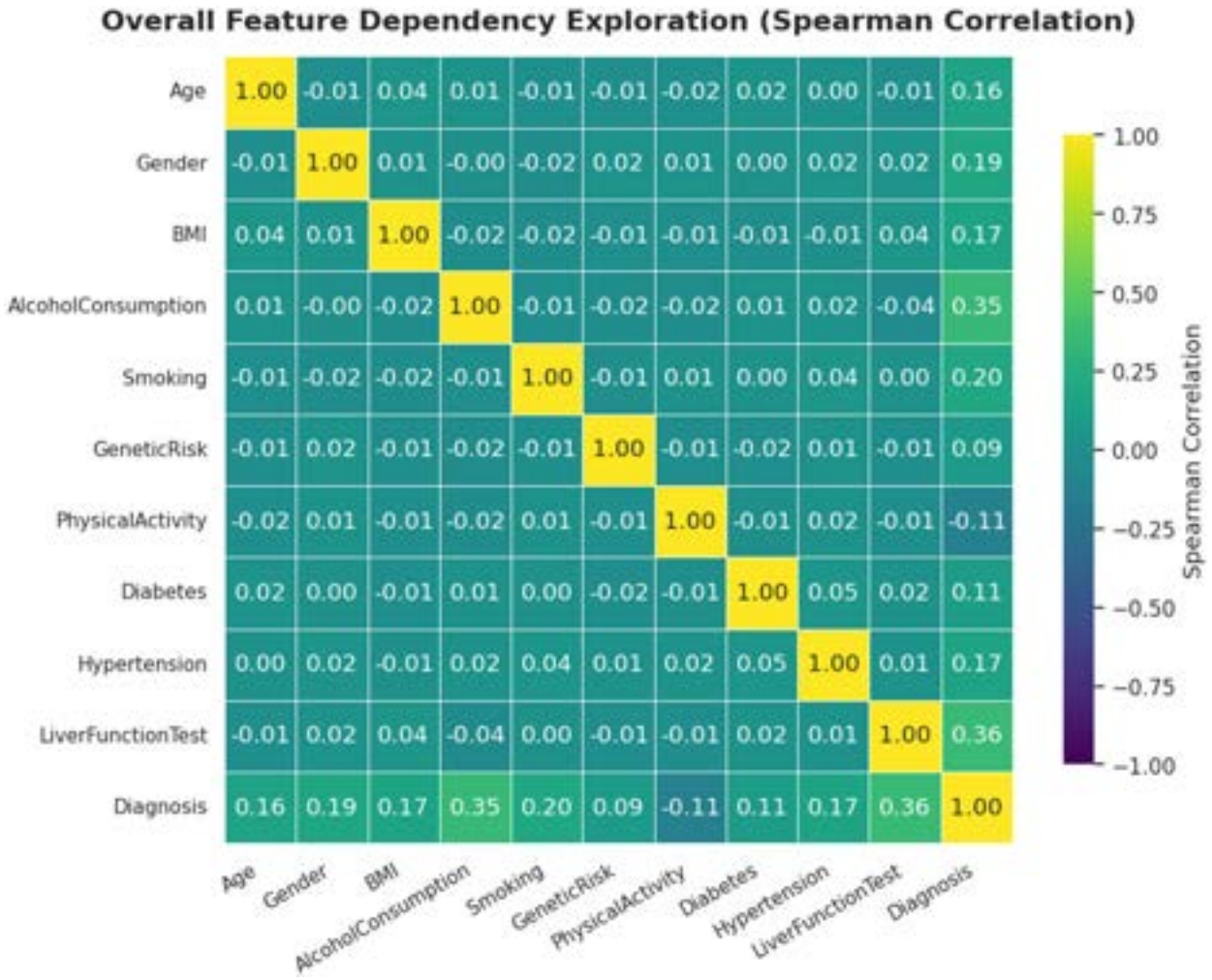

Fig. 7. Spearman correlation matrix of all clinical and lifestyle features.

Model Development reveal that this dataset is notably well-constructed.

2.2.2. Missing data assessment and completeness verification

Missing data was extensively analyzed to check the structural and analytical soundness of the Predict Liver Disease dataset before it underwent further stages of preprocessing. In a structured diagnostic approach implemented in Python, every variable has been assessed by calculating their absolute and relative missing counts as provided in the Missing Values Report. From this report, it was observed that there are no missing values among the 11 features; therefore, this dataset is complete and ready for further statistical analysis. The lack of null entries removes the need for any imputation technique like Median or mode replacement that would keep the original statistical distribution of the dataset and avoid introducing artificial bias. This is the stage at which validation of data completeness becomes very important so that subsequent processes of feature scaling, encoding, and modeling become robust. The dataset was found to be completely populated and consistent as well as analytically sound sufficiently for exploratory data analysis and the machine learning-based classification exercise on liver disease risk.

2.2.3. Duplicate record detection and uniqueness validation

To keep the dataset intact and minimize biases relating to redundancy, a comprehensive analysis of duplicate records on the Predict Liver Disease dataset was performed. Verification was performed through a comprehensive row-wise duplication check using Python's pandas library to ensure there are no replicated entries that might skew the statistical distribution or inject spuriousness in model accuracy by supporting repeated patterns. Diagnostic results demonstrated that no duplicate rows exist within the

Table 1. Dataset description and variable characteristics.

Variable	Type	Measurement/ Coding Scheme	Range/ Categories	Clinical or Analytical Significance
Age	Continuous (Numerical)	Recorded in years	20–80	Represents participant age; a major determinant in metabolic and hepatic function variability. Used for stratified risk assessment.
Gender	Categorical (Binary)	0 = Male, 1 = Female	{0, 1}	Captures sex-specific physiological and hormonal differences influencing liver enzyme activity and disease susceptibility.
BMI (Body Mass Index)	Continuous (Numerical)	Calculated as *weight (kg)/height^2 (m^2)*	15–40	Indicates nutritional and metabolic status; higher BMI is often correlated with hepatic steatosis and non-alcoholic fatty liver disease risk.
Alcohol Consumption	Continuous (Numerical)	Units of alcohol per week	0–20	Quantifies hepatotoxic exposure; higher consumption is associated with elevated liver enzyme levels and chronic liver disease.
Smoking	Categorical (Binary)	0 = No, 1 = Yes	{0, 1}	Reflects exposure to oxidative stress and toxins that may exacerbate hepatic inflammation and fibrosis.
Genetic Risk	Categorical (Ordinal)	0 = Low, 1 = Medium, 2 = High	{0, 1, 2}	Represents hereditary predisposition to liver dysfunction; allows gradient-based analysis of inherited susceptibility.
Physical Activity	Continuous (Numerical)	Average hours per week	0–10	Indicates overall activity level; inversely related to obesity and insulin resistance, which are risk factors for liver disease.
Diabetes	Categorical (Binary)	0 = No, 1 = Yes	{0, 1}	Identifies presence of metabolic dysregulation; frequently co-occurs with hepatic steatosis and non-alcoholic liver disease.
Hypertension	Categorical (Binary)	0 = No, 1 = Yes	{0, 1}	Reflects cardiovascular comorbidity; may indicate systemic inflammation and altered hepatic perfusion patterns.
Liver Function Test (LFT)	Continuous (Numerical)	Composite biochemical measure (e.g., ALT, AST, ALP, bilirubin)	20–100	Serves as a direct quantitative marker of hepatic health; elevated values suggest hepatocellular injury or dysfunction.
Diagnosis	Categorical (Binary; Target Variable)	0 = No liver disease, 1 = Liver disease	{0, 1}	Dependent variable used for supervised learning classification; defines case vs. control grouping.

dataset since the count of duplicated observations is equal to zero. This finding obviates any doubt about structural uniqueness and originality of the dataset, thereby asserting that every record is for a different participant profile. No duplicate records facilitate keep the observation independent, generalization of machine learning models, and avoids model overfitting while training. This leads to an inference that the dataset is complete and unique, hence satisfying one of the principal requirements for high-quality data preprocessing in predictive modeling and clinical analytics.

2.2.4. Outlier detection and IQR-based winsorization

Outlier treatment was conducted to reduce extreme numeric values which, if not addressed, would adversely affect statistical relationships and introduce bias into model learning. The method used here is Winsorization based on the Interquartile Range (IQR) for all continuous variables in the Predict Liver Disease dataset to identify and address extreme readings. Under this methodology, outliers are defined as values greater than 1.5 times the IQR below the first quartile (Q1) or above the third quartile (Q3). Rather than excluding these observations, Winsorization replaces them with values at appropriate boundary limits, maintaining the original data size and ensuring data quality. After performing this operation, all 1,700 records were retained, demonstrating its effectiveness in attenuating outliers without data loss. This adjustment ensured a more stable statistical distribution and improved model resilience by reducing excessive influence from improbable or abnormal values.

Table 2. Skewness analysis of continuous numeric features.

Feature	Skewness	Interpretation
Age	–0.041	Approximately Symmetrical
BMI	–0.072	Approximately Symmetrical
Alcohol Consumption	0.018	Approximately Symmetrical
Physical Activity	–0.023	Approximately Symmetrical
Liver Function Test	0.040	Approximately Symmetrical

2.2.5. Skewness assessment of continuous variables

A distributional symmetry evaluation of all continuous numeric features in the Predict Liver Disease dataset was performed to determine the appropriateness of parametric modeling and statistical balancing before normalization. Skewness for five quantitative attributes, Age, BMI, Alcohol Consumption, Physical Activity, and Liver Function Test, was computed using the Fisher-Pearson standardized moment coefficient. According to Table 2, "Continuous Numeric Features Skewness Analysis," all variables have skewness within the range of –0.072 to 0.040, indicating approximate symmetry. This suggests that the data are evenly distributed around their means, reducing the need for logarithmic or Box-Cox transformations. These variables can therefore be used in subsequent statistical modeling and machine learning without distortion due to asymmetric distributions, supporting accurate and stable parameter estimation. Overall, these results confirm the quantitative quality of the dataset and its suitability for feature scaling and model development.

2.2.6. Target variable distribution and class balance verification

An evaluation of the diagnosis target variable provided insight into class balance for the Predict Liver Disease dataset and helped determine whether resampling techniques would be required prior to training the model. The distribution analysis results showed 764 records coded 0 (no liver disease) and 936 records coded 1 (has liver disease), translating to proportions of 44.94% and 55.06%, respectively. The comparatively balanced scenario confirmed a binary classification structure having reasonably balanced classes; thus, no synthetic oversampling or undersampling steps like SMOTE or random resampling would be required. maintain the natural class proportions helps maintain the representativeness of the dataset intact and reduces any artificial variance that may undermine against model generalization. Balanced class distribution makes sure that learning algorithms find relevant discriminative patterns between healthy and diseased cases, favoring neither as it does not lead to an overreliance on one dominant class, hence supporting fair and robust model evaluation in future predictive analyses.

2.2.7. Feature standardization and scale normalization

The standard scaler () transformation method is used to ensure numerical stability as well as make different heterogeneous features comparable with one another. This rescales continuous variables by centering each feature, yielding it attains a Mean of zero and makes adjustments in its variance to one; that is, all measures are transformed into a standardized normal distribution. It will mitigate such differences in magnitudes particularly between the clinical and behavioral indicators, for example, body mass index and alcohol consumption as well as results from liver function tests. The StandardScaler () method will ensure that no large magnitude variable will dominate the model optimization process even for those algorithms where the sensitivity of the feature scaling is high. All continuous features are harmonized within a consistent scale framework which enables convergence more readily to achieve while improving coefficient interpretability and ensuring a balanced contribution of all predictors throughout the learning process. In this respect, preprocessing creates uniform data space that prevents biased parameter estimation and robust pattern recognition in subsequent analyses.

2.2.8. Modeling pipeline configuration and performance evaluation framework

A modeling framework was developed to evaluate, in a systematic manner, the classification performance of liver disease within a structured and reproducible experimental design. Five supervised learning classifiers, CatBoost, AdaBoost, Random Forest, SVM, and Decision Tree, were applied to a fully preprocessed and standardized dataset. Before model training, z-score normalization for all continuous variables using StandardScaler() was applied to harmonize feature distributions and remove scale disparities, ensuring consistent optimization across algorithms. To maintain proportional representation of both diagnostic categories, stratified sampling was used for data partitioning. Two train–test configurations, 60:40 and 70:30, were applied to assess model stability under different levels of training data availability. Each configuration was executed across twenty independent stratified runs, and performance statistics were aggregated as mean ± standard deviation to minimize stochastic variability and improve experimental reliability. A complete set of confusion-matrix-derived measures was used, including True Positives, False Positives, True Negatives, and False Negatives. Higher-level indicators computed from these elements included Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, False Positive Rate, False Negative Rate, and Receiver Operating Characteristic area under the curve (ROC-AUC).

Together, these diagnostics rigorously assessed discrimination ability, generalization performance, and misclassification behavior to support model comparison and optimal selection. The mathematical formulations of all reported metrics are presented in the section below to ensure transparency and reproducibility.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [20]$$

$$Precision = \frac{TP}{TP + FP} \quad [20]$$

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \quad [20]$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [20]$$

$$Specificity = \frac{TN}{TN + FP} \quad [21]$$

$$ROC - AUC = \int_0^1 TPR\,(FPR)\ d\,(FPR) \quad [22]$$

3. Results and discussion

CatBoost, AdaBoost, Random Forest, Support Vector Machine (SVM), and Decision Tree algorithms were exhaustively assessed regarding their performance in the prediction task, namely classifying liver disease status from a standardized dataset. Z-score normalization was applied to ensure feature standardization of all continuous predictors, as models such as SVM, which rely on margin optimization, tend to perform more effectively when feature magnitudes are comparable. Experiments were conducted using two stratified splits, 60:40 and 70:30, to test training adequacy and generalization capability. Each configuration was executed in 20 independent stratified runs, with all indicators reported as mean ± standard deviation to support robustness and reproducibility. Numerical results are presented in Table 4, "Evaluation Metrics for Stratified 70:30 and 60:40 Splits (20 runs, mean ± std)," and Table 4, "Confusion Matrix Metrics for Stratified 70:30 and 60:40 Splits (20 runs, mean ± std)." Graphical results are shown in Figs. 8 to 13.

The quantitative analysis provided in Table 3 shows that all ensemble-based models were able to outperform their classical SVM and Decision Tree counterparts. CatBoost has been able to achieve the highest mean scores with an accuracy of 0.8787 ± 0.0117 and ROC-AUC of 0.9357 ± 0.0071 under the 60:40 split, marginally increasing to values of 0.8816 ± 0.0120 and 0.9378 ± 0.0095 under a split of 70:30, respectively, for accuracy and ROC-AUC. Results obtained for AdaBoost and Random Forest are comparable close; accuracies lie between approximately 0.87–0.88 and AUCs near 0.93 while SVM shows moderate performance with around .91 for AUC, Decision Tree finally having the lowest capability to discriminate with only approximately 0.80

Table 3. Comparison of machine-learning approaches for liver disease prediction.

Study	Dataset/Task	Algorithms Evaluated	Best Model Reported	Best Accuracy (%)/AUC-ROC
Ghosh et al. (2021) [23]	Chronic liver disease prediction (public clinical dataset)	LR, RF, XGBoost, SVM, AdaBoost, KNN, DT	Random Forest	83.70/–
Dritsas & Trigka (2023) [24]	Liver disease risk prediction (ILPD)	Multiple ML + Voting ensemble	Voting classifier	80.10/0.884
Wu et al. (2019) [25]	Fatty liver disease (clinical cohort, 577 patients)	RF, NB, ANN, LR	Random Forest	87.48 / 0.925
Ghazal et al. (2022) [26]	Early liver disease prediction	Multiple ML models	Proposed ML model	88.40 / –
Md et al. (2023) [27]	Liver disease detection (ILPD)	GB, XGBoost, Bagging, RF, Extra Trees, Stacking	Extra Trees	91.82 / –
El Atifi et al. (2025) [28]	Liver disease prediction (public dataset)	RF, AdaBoost, Gradient Boosting	Tuned Random Forest	85.17 / –
Moturi et al. (2023) [29]	Liver disease prediction (ILPD)	KNN, DT, Extra Trees, LR, RF	Extra Trees	92.50 / –
Lakumarapu et al. (2024) [30]	Liver disease prediction	LR, KNN, DT, RF	RF / LR	≈ 74.0 / –
Modhugu & Ponnusamy (2024) [31]	Liver disease prediction (Kaggle, large-scale)	SVM, LR, DT	SVM	85.0 / –
Nararrya et al. (2025) [32]	Liver disease prediction + feature importance	RF, LR, XGBoost, GB, SVC, LightGBM, CatBoost	CatBoost	92.18 / –
Soares et al. [33]	Liver disease classification (Kaggle)	SVM (IQR outlier handling)	SVM	84.74 / 0.933
This study	Liver disease prediction (Kaggle, 1,700 records, 11 features)	DT, SVM, RF, AdaBoost, CatBoost	CatBoost	≈ 88.0 / ≈ 0.94–0.96

Table 4. Evaluation metrics for stratified 70:30 and 60:40 Splits (20 runs, mean ± std).

Model	Split	Accuracy (mean ± std)	Precision (mean ± std)	Recall (Sensitivity) (mean ± std)	F1-score (mean ± std)	Specificity (mean ± std)	ROC-AUC (mean ± std)
CatBoost	60:40	0.8787 ± 0.0117	0.8928 ± 0.0148	0.8861 ± 0.0176	0.8893 ± 0.0109	0.8696 ± 0.0207	0.9357 ± 0.0071
AdaBoost	60:40	0.8732 ± 0.0145	0.8874 ± 0.0177	0.8817 ± 0.0200	0.8843 ± 0.0134	0.8627 ± 0.0247	0.9334 ± 0.0077
Random Forest	60:40	0.8691 ± 0.0119	0.8852 ± 0.0115	0.8758 ± 0.0191	0.8803 ± 0.0115	0.8609 ± 0.0159	0.9295 ± 0.0090
SVM	60:40	0.8319 ± 0.0105	0.8401 ± 0.0132	0.8580 ± 0.0168	0.8488 ± 0.0096	0.8000 ± 0.0206	0.9061 ± 0.0073
Decision Tree	60:40	0.8036 ± 0.0152	0.8246 ± 0.0161	0.8170 ± 0.0186	0.8206 ± 0.0142	0.7873 ± 0.0222	0.8021 ± 0.0154
CatBoost	70:30	0.8816 ± 0.0120	0.8980 ± 0.0174	0.8863 ± 0.0224	0.8918 ± 0.0114	0.8758 ± 0.0247	0.9378 ± 0.0095
AdaBoost	70:30	0.8727 ± 0.0130	0.8870 ± 0.0182	0.8820 ± 0.0223	0.8842 ± 0.0121	0.8614 ± 0.0272	0.9353 ± 0.0097
Random Forest	70:30	0.8750 ± 0.0116	0.8909 ± 0.0150	0.8815 ± 0.0220	0.8859 ± 0.0113	0.8670 ± 0.0214	0.9322 ± 0.0104
SVM	70:30	0.8373 ± 0.0187	0.8442 ± 0.0207	0.8648 ± 0.0256	0.8541 ± 0.0172	0.8035 ± 0.0308	0.9093 ± 0.0114
Decision Tree	70:30	0.8047 ± 0.0207	0.8242 ± 0.0220	0.8212 ± 0.0265	0.8224 ± 0.0194	0.7845 ± 0.0313	0.8028 ± 0.0210

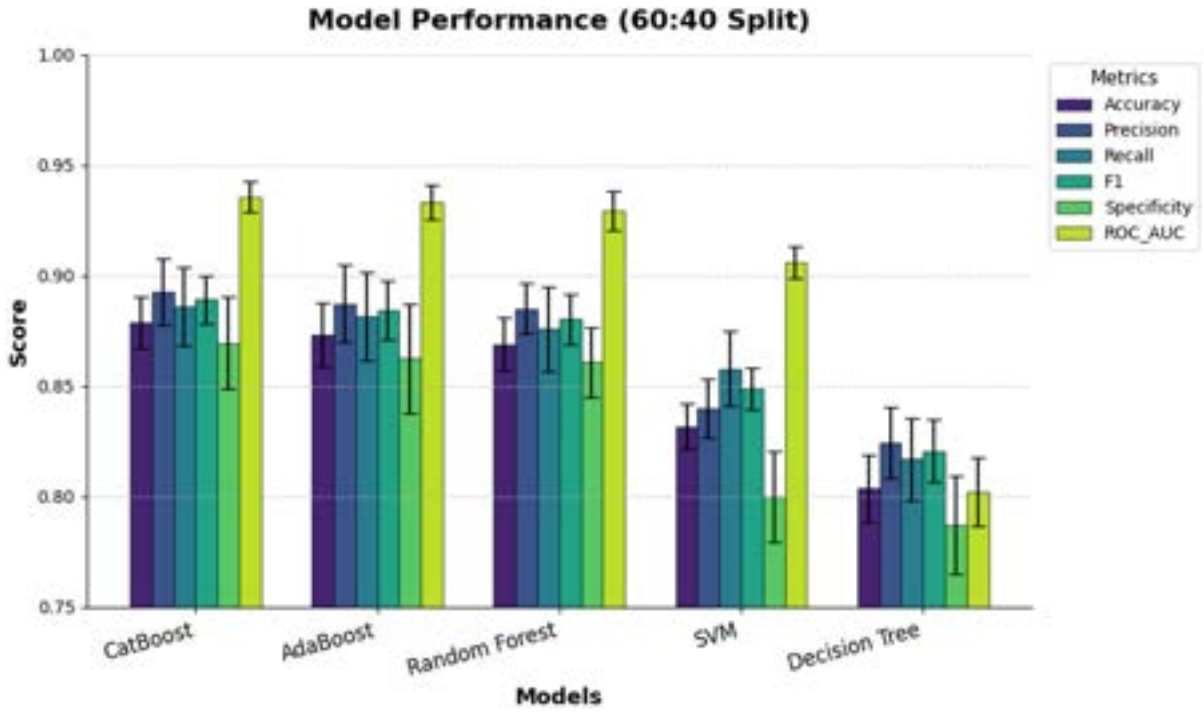

Fig. 8. Model performance (60:40 Split).

for AUC. All results indicate a very high level of consistency since standard deviation is quite minimal. Minor improvements for accuracy, recall, and AUC when using the 70:30 split indicate improved generalization of the model with an increased portion of the training subset. This fact has visual support in Figs. 8 and 9, where the bar-plot distributions reveal the dominance of CatBoost, AdaBoost, and Random Forest across all evaluation measures.

Further insights are created from the diagnostic metrics that emanate from the confusion matrix, as summarized in Table 5. The components of the confusion matrix-True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) have all been averaged to come up with some basic but important diagnostic indicators that include a False Positive Rate (FPR) and a False Negative Rate (FNR). CatBoost possesses the lowest FNR values among evaluated algorithms (0.1139 ± 0.0176 at 60:40 and 0.1137 ± 0.0224 at 70:30) and the lowest FPR values (0.1304 ± 0.0207 and 0.1242 ± 0.0247), respectively-all this puts sensitivity above others toward discovering diseased cases while keeping precision in minimizing false alert among healthy individuals. AdaBoost and Random Forest have produced the same diagnostic reliability, while SVM and Decision Tree showed higher error rates (FPR about 0.20–0.21; FNR about 0.14–0.18) that attested a less favorable balance between sensitivity and specificity, hence marking ensemble methods clinically meaningful since they are robust in detecting diseases with minimum misclassification.

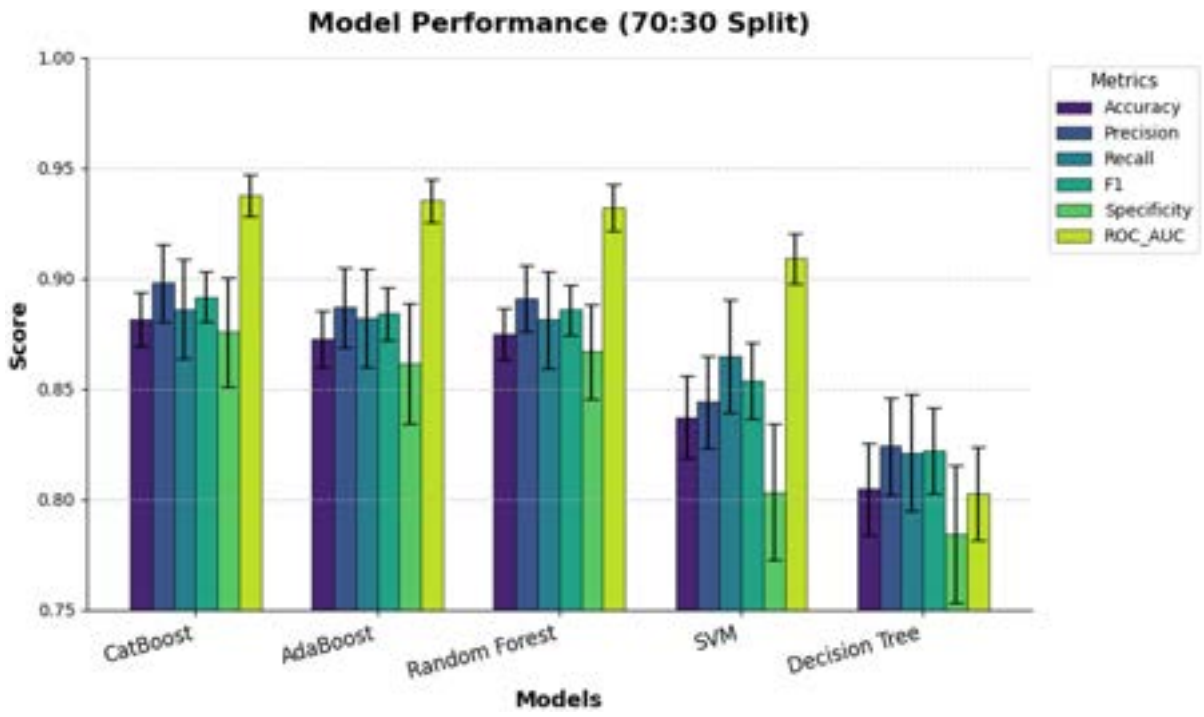

Fig. 9. Model performance (70:30 Split).

The confusion matrix and ROC curve figures visually support these quantitative results. Major diagonal dominance in Figs. 10 and 11, "Confusion Matrices of Best Runs," for the stratified 60:40 and 70:30 splits, respectively, is most evident for CatBoost, AdaBoost, and Random Forest. This reflects strong true-positive and true-negative counts, consistent with the numerical results. The ROC curves in Figs. 12 and 13 further demonstrate separability strength, as the curves lie close to the top-left region of the plots, which is typically associated with high discrimination between positive and negative cases. CatBoost attained the highest area under the curve, with AUC approximately in the range 0.951 to 0.958, closely followed by AdaBoost and Random Forest with AUC values near 0.95. In contrast, SVM and Decision Tree exhibited less pronounced curves, consistent with lower AUC values. The similarity of ROC profiles across both splits is reassuring, as it supports model reliability across partitions and validates the evaluation protocol in statistical terms.

The stability of the performance metrics for twenty random runs demonstrates how well results can be reproduced with this proposed framework. Ensemble models showed low variance in accuracy, recall, and specificity further attesting to their robustness to perturbations in the data. Misclassification analysis revealed that the majority of errors were originating from borderline clinical profiles where patient biomarkers were overlapping between mild disease and healthy ranges. SVM and Decision Tree algorithms primarily misclassified these marginal cases as non-diseased (hence increasing FNR), while AdaBoost sometimes tends to over-prediction (hence increasing FPR). However, CatBoost balanced error control owing to an ordered boosting and feature dependency handling resulting in a symmetrical distribution of errors across classes.

Overall, the experimental findings clearly establish CatBoost as the most reliable and diagnostically accurate for liver disease prediction among the five classifiers tested in addition to being the best-performing model. A high degree of accuracy accompanied by a high ROC-AUC, and low rates of both type I and II errors render it both statistically and clinically credible. The close correspondence between the tabular results as well as the graphical evidence and the stability observed in two splits validates claims of robustness as well as generalizability for this developed modeling framework. Additionally, CatBoost inherently provides feature-importance estimates, and this interpretability component can be incorporated in future work to highlight the most influential clinical predictors and support clinically transparent decision-making.

As summarized in Table 3, several prior studies have reported competitive or higher peak accuracy values on liver disease datasets; however, these results are obtained under heterogeneous datasets, clinical endpoints, and validation protocols, which

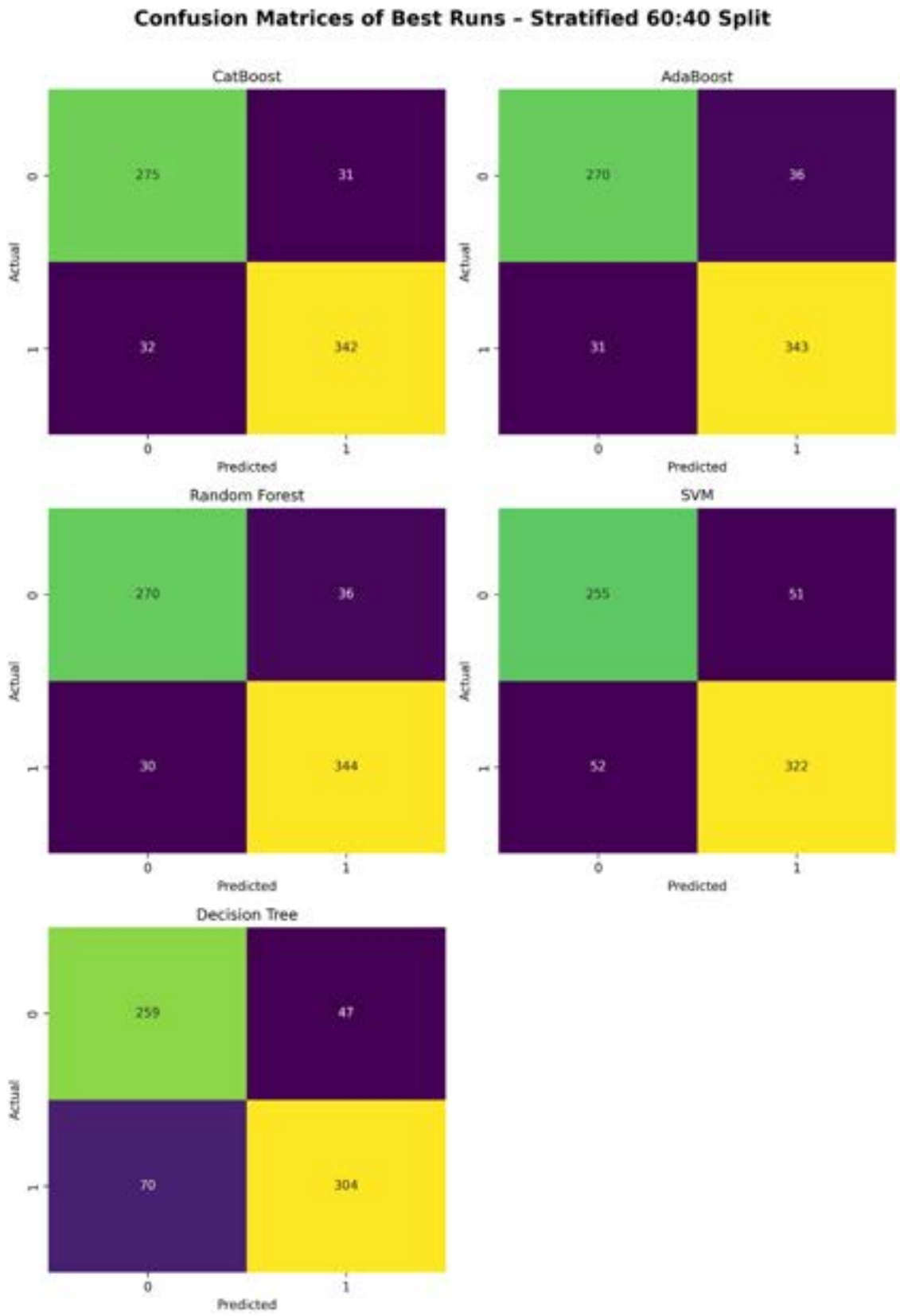

Fig. 10. Confusion matrices of best runs – Stratified 60:40 Split.

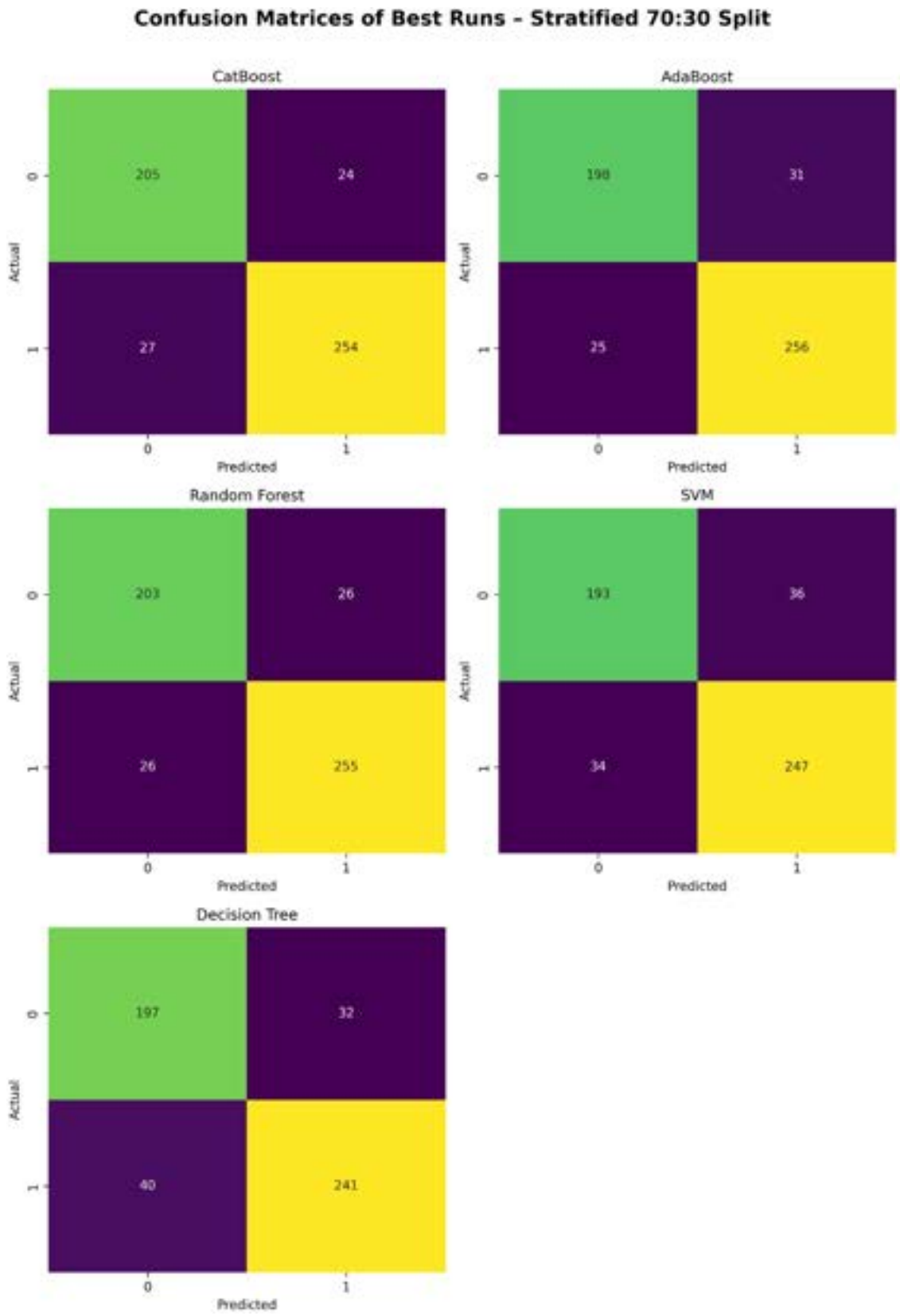

Fig. 11. Confusion matrices of best runs – Stratified 70:30 Split.

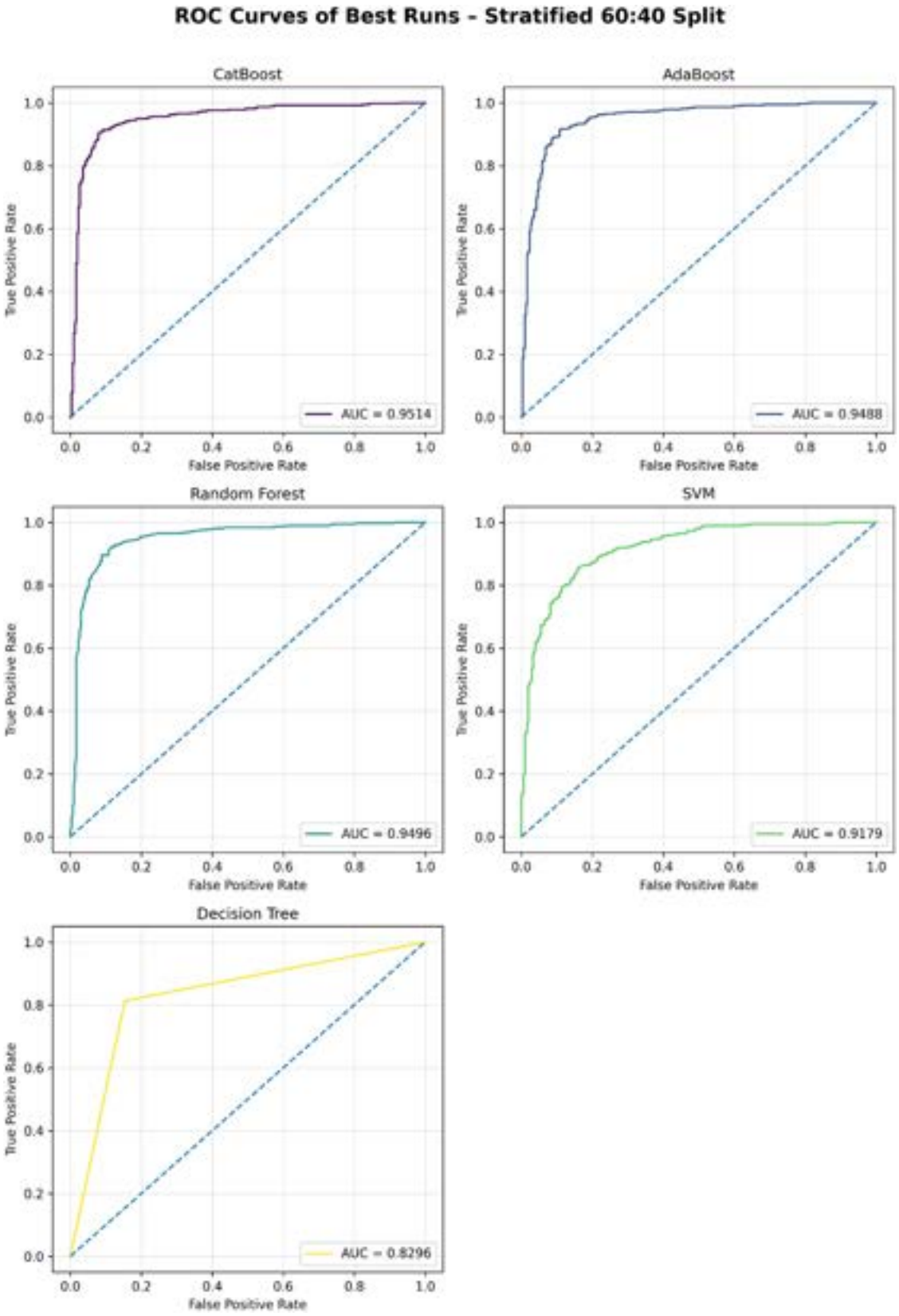

Fig. 12. ROC curves of best runs – Stratified 60:40 Split.

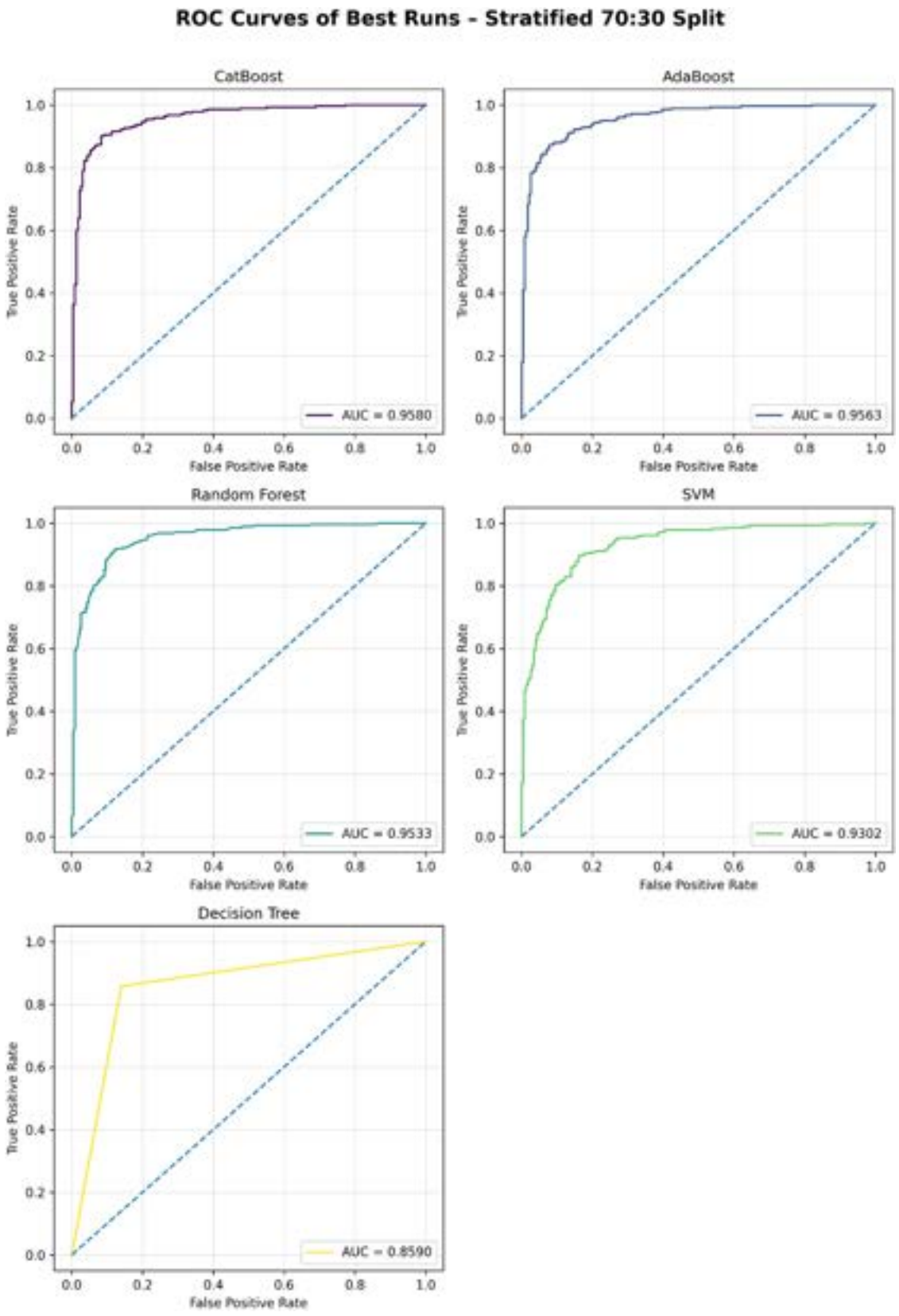

Fig. 13. ROC curves of best runs – Stratified 70:30 Split.

Table 5. Confusion matrix metrics for stratified 70:30 and 60:40 Splits (20 runs, mean ± std).

Model	Split	TN (mean ± std)	FP (mean ± std)	FN (mean ± std)	TP (mean ± std)	FPR (mean ± std)	FNR (mean ± std)
CatBoost	60:40	266.1000 ± 6.3317	39.9000 ± 6.3317	42.6000 ± 6.5757	331.4000 ± 6.5757	0.1304 ± 0.0207	0.1139 ± 0.0176
AdaBoost	60:40	264.0000 ± 7.5631	42.0000 ± 7.5631	44.2500 ± 7.4825	329.7500 ± 7.4825	0.1373 ± 0.0247	0.1183 ± 0.0200
Random Forest	60:40	263.4500 ± 4.8629	42.5500 ± 4.8629	46.4500 ± 7.1587	327.5500 ± 7.1587	0.1391 ± 0.0159	0.1242 ± 0.0191
SVM	60:40	244.8000 ± 6.3135	61.2000 ± 6.3135	53.1000 ± 6.2761	320.9000 ± 6.2761	0.2000 ± 0.0206	0.1420 ± 0.0168
Decision Tree	60:40	240.9000 ± 6.7816	65.1000 ± 6.7816	68.4500 ± 6.9532	305.5500 ± 6.9532	0.2127 ± 0.0222	0.1830 ± 0.0186
CatBoost	70:30	200.5500 ± 5.6611	28.4500 ± 5.6611	31.9500 ± 6.2966	249.0500 ± 6.2966	0.1242 ± 0.0247	0.1137 ± 0.0224
AdaBoost	70:30	197.2500 ± 6.2360	31.7500 ± 6.2360	33.1500 ± 6.2791	247.8500 ± 6.2791	0.1386 ± 0.0272	0.1180 ± 0.0223
Random Forest	70:30	198.5500 ± 4.9038	30.4500 ± 4.9038	33.3000 ± 6.1895	247.7000 ± 6.1895	0.1330 ± 0.0214	0.1185 ± 0.0220
SVM	70:30	184.0000 ± 7.0569	45.0000 ± 7.0569	38.0000 ± 7.1903	243.0000 ± 7.1903	0.1965 ± 0.0308	0.1352 ± 0.0256
Decision Tree	70:30	179.6500 ± 7.1713	49.3500 ± 7.1713	50.2500 ± 7.4557	230.7500 ± 7.4557	0.2155 ± 0.0313	0.1788 ± 0.0265

limits direct numerical comparability. In contrast, the present study emphasizes evaluation robustness and reproducibility, demonstrating consistently strong CatBoost performance across two stratified splits with twenty repeated runs and low variance across all diagnostic metrics. This stability-aware assessment, together with balanced control of false-positive and false-negative errors, provides stronger evidence for real-world screening applicability than single-split or accuracy-only evaluations.

4. Conclusion

This study developed and rigorously evaluated a supervised machine-learning framework for early liver disease classification, with emphasis on diagnostic stability, clinically meaningful error control, and reproducibility. Five classifiers (CatBoost, AdaBoost, Random Forest, SVM, and Decision Tree) were assessed on a standardized clinical dataset using two stratified train–test partitions (60:40 and 70:30). To strengthen statistical reliability, each model–split configuration was executed across 20 independent runs, and performance was summarized as mean ± standard deviation using Accuracy, Precision, Recall, F1-score, Specificity, and ROC-AUC, supported by confusion-matrix–derived error rates and ROC-curve analysis.

Across both splits, ensemble learners consistently outperformed non-ensemble baselines, with CatBoost achieving the most stable and best overall performance. CatBoost delivered accuracy of approximately 0.88 and ROC-AUC in the range of ∼0.94–0.96, while maintaining low false-negative and false-positive rates (FNR ∼0.11 and FPR ∼0.12), indicating strong sensitivity for detecting diseased cases without sacrificing specificity. The narrow variance observed across repeated runs further supports the robustness of the evaluation protocol and reduces sensitivity to split-dependent effects, providing a more reliable basis for model selection than single-run reporting. Collectively, these findings support the feasibility of stability-aware, data-driven ensemble modeling as a non-invasive screening approach that can assist clinical decision-making through accurate and reproducible risk assessment.

Future work should validate the proposed framework on external and multi-center clinical cohorts, integrate interpretability through CatBoost feature-importance analysis and post-hoc explanation methods (e.g., SHAP/LIME), and evaluate deployment feasibility within real-time clinical decision-support settings.

Limitations: This study relied on a single publicly available Kaggle dataset, which may not fully represent real-world clinical heterogeneity, and no external clinical validation has yet been performed; therefore, the generalizability of the reported performance should be interpreted cautiously until prospective and multi-center testing is conducted.

Acknowledgment

The authors would like to express their sincere gratitude to the academic staff and faculty members at our university for their valuable guidance, scientific feedback, and continuous support throughout the development of this research. We are especially grateful to our supervisors for their mentorship, constructive recommendations, and rigorous academic oversight, which substantially enhanced the methodological clarity and scholarly quality of the manuscript. We also acknowledge the helpful discussions and feedback provided by colleagues and peers during internal review and manuscript refinement, which contributed to improving the presentation and interpretability of the study.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The study was conducted without external sponsorship, and no third party had

any role in the study design; preprocessing decisions; model development; statistical analysis; interpretation of results; manuscript writing; or the decision to submit the work for publication. Computational experiments were performed using standard academic and personal computing resources available to the authors. Any software tools and libraries used in this study were accessed through publicly available open-source ecosystems.

Author contributions

All authors contributed equally to the conception and design of the study, data curation and preprocessing, methodology development, model implementation and experimentation, statistical analysis, interpretation of findings, preparation of figures and tables, drafting of the manuscript, and critical revision for intellectual content. All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath, "Burden of liver diseases in the world," *J. Hepatol.*, vol. 70, no. 1, pp. 151–171, 2019, doi: 10.1016/j.jhep.2018.09.014.
2. H. Devarbhavi, S. K. Asrani, J. P. Arab, Y. A. Nartey, E. Pose, and P. S. Kamath, "Global burden of liver disease: 2023 update," *J. Hepatol.*, vol. 79, no. 2, pp. 516–537, 2023, doi: 10.1016/j.jhep.2023.03.017.
3. World Health Organization, *Global Hepatitis Report 2024: Action for Access in Low- and Middle-Income Countries*, Geneva, Switzerland: WHO, Apr. 2024, ISBN 978-92-4-009167-2.
4. S. Cheemerla and M. Balakrishnan, "Global epidemiology of chronic liver disease," *Clin. Liver Dis.*, vol. 17, no. 5, pp. 365–370, 2021, doi: 10.1002/cld.1061.
5. V. Lala and D. A. Minter, "Liver function tests," *StatPearls*, StatPearls Publishing, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK482489/.
6. E. Goceri, Z. K. Shah, R. Layman, X. Jiang, and M. N. Gurcan, "Quantification of liver fat: A comprehensive review," *Comput. Biol. Med.*, vol. 71, pp. 174–189, 2016, doi: 10.1016/j.compbiomed.2016.02.013.
7. R. Hernaez, M. Lazo, S. Bonekamp, *et al.*, "Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: A meta-analysis," *Hepatology*, vol. 54, no. 3, pp. 1082–1090, 2011, doi: 10.1002/hep.24452.
8. S. M. Ganie, M. Hussain, A. Altaf, *et al.*, "Improved liver disease prediction from clinical data through ensemble learning and feature optimization," *BMC Med. Inform. Decis. Mak.*, vol. 24, p. 151, 2024, doi: 10.1186/s12911-024-02550-y.
9. N. Salkić, P. Jovanović, M. B. Jaman, *et al.*, "Machine learning for short-term mortality in acute decompensation of liver cirrhosis: Better than MELD score," *Diagnostics*, vol. 14, no. 10, p. 981, 2024, doi: 10.3390/diagnostics14100981.
10. B. Ramana and N. Venkateswarlu, *ILPD (Indian Liver Patient Dataset)* [Dataset], UCI Machine Learning Repository, 2022, doi: 10.24432/C5D02C.
11. S. Dalal, E. M. Onyema, and A. Malik, "Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy," *World J. Gastroenterol.*, vol. 28, no. 46, pp. 6551–6563, 2022, doi: 10.3748/wjg.v28.i46.6551.
12. W. El Atifi, O. El Rhazouani, F. M. Khan, and H. Sekkat, "Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare," *PLoS One*, vol. 20, no. 8, e0330899, 2025, doi: 10.1371/journal.pone.0330899.
13. N. A. Khan, M. F. Bin Hafiz, M. A. Pramanik, S. Hossain, S. Barman, and N. Hossain, "Machine learning and explainable AI for liver disease prediction: An integrated interpretability framework," *Biomed. Mater. Devices*, pp. 1–14, 2025, doi: 10.1007/s44174-025-00414-1.
14. H. Xie, L. Chen, S. Zhang, *et al.*, "A deep learning approach for acute liver failure prediction with combined fully connected and convolutional neural networks," *BMC Med. Inform. Decis. Mak.*, vol. 24, p. 119, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38759076/.
15. G. H. Choi *et al.*, "Development of machine learning-based clinical decision support system for hepatocellular carcinoma," *Sci. Rep.*, vol. 10, no. 1, p. 14855, 2020, doi: 10.1038/s41598-020-71796-z.
16. C. Rudin, "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
17. H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
18. G. S. Collins, P. Dhiman, C. L. Andaur Navarro, *et al.*, "TRIPOD + AI statement: Updated guidance for reporting prediction model studies that use regression or machine learning methods," *BMJ*, vol. 385, e078378, 2024, doi: 10.1136/bmj-2023-078378.
19. R. El Kharoua, "Predict Liver Disease - 1700 Records Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset
20. M. Rashidi, S. Arima, A.C. Stetco, C. Coppola, D. Musarò, M. Greco, and M. Maffia, "Prediction of Parkinson Disease Using Long-Term, Short-Term Acoustic Features Based on Machine Learning," *Brain Sci.*, vol. 15, 739, 2025, https://doi.org/10.3390/brainsci15070739.
21. S. Lahmiri, D.A. Dawson, and A. Shmuel, "Performance of Machine Learning Methods in Diagnosing Parkinson's Disease Based on Dysphonia Measures," *Biomed. Eng. Lett.*, vol. 8, 29–39, 2018, https://doi.org/10.1007/s13534-017-0058-2.
22. V. Škvára, T. Pevný, and V. Šmídl, "Is AUC the best measure for practical comparison of anomaly detectors?," 2023. arXiv preprint arXiv:2305.04754?
23. M. Ghosh, M. M. S. Raihan, M. Raihan, L. Akter, A. K. Bairagi, S. S. Alshamrani, and M. Masud, "A comparative analysis of machine learning algorithms to predict liver disease," *Intelligent Automation & Soft Computing*, vol. 30, no. 3, 2021.
24. E. Dritsas and M. Trigka, "Supervised machine learning models for liver disease risk prediction," *Computers*, vol. 12, no. 1, p. 19, 2023, doi: 10.3390/computers12010019.
25. C.-C. Wu, W.-C. Yeh, W.-D. Hsu, M. M. Islam, P. A. A. Nguyen, T. N. Poly, H.-C. Yang, Y.-C. Li, and Y.-C. J. Li, "Prediction of fatty liver disease using machine learning algorithms," *Computer Methods and Programs in Biomedicine*, vol. 170, pp. 23–29, 2019, doi: 10.1016/j.cmpb.2018.12.032.

26. T. M. Ghazal, A. U. Rehman, M. Saleem, M. Ahmad, S. Ahmad, and F. Mehmood, “Intelligent model to predict early liver disease using machine learning technique,” in *Proc. Int. Conf. Business Analytics for Technology and Security (ICBATS), Dubai, UAE*, 2022, pp. 1–5, doi: 10.1109/ICBATS54253.2022.9758929.
27. A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, “Enhanced preprocessing approach using ensemble machine learning algorithms for detecting liver disease,” *Biomedicines*, vol. 11, no. 2, p. 581, 2023, doi: 10.3390/biomedicines11020581.
28. W. El Atifi, O. El Rhazouani, F. M. Khan, and H. Sekkat, “Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare,” *PLOS ONE*, vol. 20, no. 8, e0330899, 2025, doi: 10.1371/journal.pone.0330899.
29. S. Moturi, J. V. Bolla, M. Anusha, M. M. N. Bhavani, S. Vemuru, S. T. Rao, and S. A. Mallipeddi, “Prediction of liver disease using machine learning algorithms,” in *Data Science and Applications (ICDSA 2023), Singapore: Springer Nature*, 2023, pp. 243–254.
30. S. Lakumarapu, R. Nithyanandhan, V. S. Bhargavi, A. T. P., N. M., and S. S. R., “Machine learning approaches for liver disease prediction: A comparative analysis,” in *Proc. 5th Int. Conf. Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India*, 2024, pp. 159–164, doi: 10.1109/ICESC60852.2024.10689974.
31. V. R. Modhugu and S. Ponnusamy, “Comparative analysis of machine learning algorithms for liver disease prediction: SVM, logistic regression, and decision tree,” *Asian Journal of Research in Computer Science*, vol. 17, no. 6, pp. 188–201, 2024, doi: 10.9734/ajrcos/2024/v17i6467.
32. M. F. R. Nararrya, T. A. Putra, A. A. Madjid, and M. A. Ibrahim, “Feature importance analysis of clinical variables in liver disease prediction using machine learning algorithms,” *Procedia Computer Science*, vol. 269, pp. 953–967, 2025, doi: 10.1016/j.procs.2025.09.038.
33. T. G. Soares, M. Tonggiroh, M. Erkamim, and E. Widarti, “Enhancing liver disease classification using support vector machine with IQR-based outlier handling,” *Jurnal Ilmiah FIFO*, vol. 17, no. 1, pp. 91–101.