REVIEW PAPER

Machine Learning Applications in Bioanalytical Data Interpretation: A Computational Approach

Dr. Ashish Tiwari[1], Dr. Asma Rani[2], Madhan M[3], Dr. C. Ravichandran[4], Dr. M. Dhanabal[5], Ranjan Banerjee[6]

[1*]Assistant Professor, Chemistry, Govt College Dumariya (Jarhi) Surajpur, Chhattisgarh, Email ID : tiwari.ashish478@gmail.com, ORCID ID : 0009-0000-5134-9323
[2]Assistant Professor, Computer Science and Engineering, Dr. B. R. Ambedkar Institute of Technology, Sri Vijayapuram, Andaman and Nicobar Islands, Email ID: asma.sags@gmail.com, ORCID ID: 0000-0001-6087-6248
[3]Assistant Professor, Artificial Intelligence and Data Science, KGiSL Institute of Technology, Coimbatore, Tamil Nadu, Email ID: madhan.m@kgkite.ac.in, ORCID ID: 0009-0002-5262-0765
[4]Professor, ECE Department, Tagore Institute of Engineering and Technology, Deviyakurichi, Salem D.T, Tamil Nadu 636112, Email ID: ravisarvajith@gmail.com, ORCID ID: 009-0009-8612-4402
[5]Assistant Professor, Department of Mechanical Engineering, V.S.B. ENGINEERING COLLEGE, Karur, Tamil Nadu Pincode-639111,Email ID : dhanabalm0101@gmail.com, Orcid ID-https://orcid.org/0009-0007-8328-1838?lang=en
[6]Assistant Professor, Department of Computer Science Engineering, Brainware University, Barasat, Kolkata, Email ID : rbkpccst@gmail.com, ORCID ID : 0009-0003-1950-7530

Abstract
Liver disease is a huge global health burden and the need for accurate and interpretable analysis of routine bioanalytical data for effective screening and risk stratification. This study proposes a computational machine learning framework for analysing biochemical markers for assessment of liver disease based on routinely collected laboratory measurements. A hepatitis bioanalytical dataset with 605 samples and thirteen clinical features was analysed with binary and multiclass classification setting. Logistic regression, support vector machine using radial basis function kernel and extreme gradient boosting were evaluated using nested stratified cross-validation with class-weighted learning. In the binary classification, the F1 score was 0.95 and the area under the receiver operating characteristic curve was greater than 0.88, according to extreme gradient boosting model, which was very effective in discriminating healthy and diseased individuals. Multiclass classification was less effective with the support vector machine scoring a macro-averaged F1 of 0.68, which indicates some overlap of intermediate disease stages. The results of statistical testing confirmed the lack of significant pairwise differences between the best performing models. Interpretability analysis identified aspartate aminotransferase, bilirubin, alanine aminotransferase, total protein, and cholinesterase as the best predictors of the primary clinical knowledge. The results underscore the importance of statistical validated and interpretable machine learning methods for bioanalytical liver disease screening while showing the difficulties of fine-grained disease staging.

Keywords: Liver disease; Machine learning; Bioanalytical data; Clinical biomarkers
How To Cite This Article: Tiwari A, Rani A, Madhan M, Ravichandran C, Dhanabal M, Banerjee R. Machine Learning Applications In Bioanalytical Data Interpretation: A Computational Approach. Int J Drug Deliv Technol. 2026;16(27s):1074-1085. Doi: 10.25258/ijddt.16.27s.123

1. Introduction

Liver diseases are a major public health problem worldwide, which includes a wide range of diseases from acute hepatitis to progressive fibrosis and cirrhosis. Early detection and proper staging are important to proper clinical management, prognosis, and prevention of irreversible hepatic damage. Routine bioanalytical markers of liver disease such as alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), bilirubin (BIL), albumin (ALB), cholinesterase (CHE) and total protein (PROT) continue to play a pivotal role in non-invasive evaluation of liver disease. These biomarkers give important information on hepatocellular injury, cholestasis, hepatic synthetic capacity and systemic metabolic disturbances. Their application in clinical practice is however commonly construed around predetermined reference ranges and single threshold values which may not be able to reflect the intricate and interactive biochemical patterns that accompany disease onset and disease progression.

*Author for Correspondence: tiwari.ashish478@gmail.com

Traditional univariate interpretation is especially limited in intermediate stages of the disease, where the biochemical profile of these patients is so similar that it hides the margin between normal and diseased conditions. This challenge has been underlined in recent studies highlighting the need for multivariate analytical approaches, which are able to combine bioanalytical markers into coherent predictive frameworks which better reflect underlying pathophysiology [1]. At the same time, the significant digitization of medical records and laboratory databases has resulted in the availability of large-scale clinical databases offering the possibility of data-based techniques to improve the accuracy of diagnosis and prognosis. To take full advantage of these datasets, it is necessary to use analytical methods which are able to deal with nonlinearity, feature interactions, and class imbalance in a manner which is easy to interpret and provide clinical meaning.

Machine learning has become a strong methodological paradigm in the analysis of complex biomedical data

that has clear advantages over traditional statistical methods for modelling nonlinear relationships and high-dimensional interactions. In the context of liver disease, there are promising results for disease detection, staging, and predicting disease outcomes using machine learning models based on routinely collected laboratory data. Ensemble and kernel-based algorithms in particular have been found to be more effective than linear models by capturing subtler interaction effects between biochemical markers than can be identified by traditional analysis [2]. These capabilities are of particular interest in diseases of the liver, for which pathophysiological processes frequently manifest themselves through nonlinear and cumulative changes in biomarkers.

Outside predictive performance, interpretability has become of increased interest in machine-learning studies in clinical settings. Explainable machine learning approaches are focused on giving explanations about the decision-making process of models so that clinicians can then assess if the algorithmic predictions are consistent with pre-existing biological and clinical knowledge. This shift is key to clinical acceptance in that black box models with no interpretative insight are unlikely to find wide acceptance in medical settings. Recent studies on metabolic and steatotic liver diseases have shown that interpretable machine learning frameworks can be used to discover clinically meaningful biomarker relationships with no reduction in predictive accuracy [3].

Despite these improvements there are some methodological shortcomings in the existing literature. Overfitting, optimistic bias, and a lack of validation are issues that are still prevalent, especially in research studies with imbalanced data sets and small numbers of study participants. Most of the published literature records high accuracies without employing strong evaluation tactics, and this casts doubt on the reproducibility and generalizability. The use of nested cross-validation and formal statistical testing has been recommended as a best practice to address these problems, but are inconsistently used. Studies using rigorous validation protocols are generally more conservative in the performance results they report, yet reliable, and this highlights the importance of methodological rigor in clinical machine learning studies [4].

One other notable limitation is the predominance of predictive accuracy with limited statistical comparison between competing models. Without the proper non-parametric testing and correction for multiple comparisons, the results of performances may be over-stated. Furthermore, although feature importance is frequently reported, very few studies systematically link machine learning-derived importance measures with an established bioanalytical and clinical understanding, which limits the translational relevance of the importance measures. This gap has been recognised as one of the main obstacles to the incorporation of machine-learning tools into routine hepatology practice [5].

In addition, most current research focuses on binary classification, whereas the clinically relevant situations often involve the discrimination between several disease stages. Multiclass liver disease modeling is a more complex task because of the biological continuum of disease progression and the overlapping of biochemical signatures. Accurate staging using only routine laboratory markers is still difficult, and interpretable and statistically validated approaches to multiclass prediction are still relatively rare [6].

To overcome this problem, the current study is aimed at proposing a computational machine learning framework for bioanalytical data interpretation in liver disease assessment using a predictive model, rigorous validation, statistical testing, and interpretability analysis. The specific objectives of this work are:

- To assess and compare Linear and Non-Linear Machine learning models for Binary and Multi-Class classification of Liver Disease using Bioanalytical markers (routine tests) under stringent validation protocol.
- To incorporate the statistical testing and confidence interval estimation in model benchmarking, to insure reliable and reproducible performance assessment.
- To offer biologically grounded interpretability analysis, associating machine-learning derived feature importance to known clinical information on the functioning of the liver and liver disease progression.

This study will continue to promote the use of machine learning as an adjunct tool of informed clinical decision-making in hepatology by ensuring its important function through both methodological rigor and bioanalytical interpretability.

2. Literature Review

Bioanalytical markers from the routine laboratory testing provide the basis for the non-invasive evaluation of liver diseases. Aminotransferases, especially alanine aminotransferase (ALT) and aspartate aminotransferase (AST) are popular markers of hepatocellular injury, whereas alkaline phosphatase (ALP) and gamma-glutamyl transferase (GGT) are used as indicators of cholestatic processes and involvement of the biliary. Markers of hepatic synthetic function, such as albumin (ALB), cholinesterase (CHE) and total protein (PROT), give some insight into functional reserve and disease severity. Even though these biomarkers are regularly done, they are frequently not interpreted using multivariate logic because they are generally based on isolated thresholds and univariate logic, insufficient to reflect the complex, multivariate nature of liver pathology. Early efforts to identify predictive markers of liver disease in a predictive model included the possibility of using combinations of multiple different laboratory parameters to distinguish between the different causes of liver disease, e.g., alcoholic and non-alcoholic, to provide better discrimination compared to the traditional methods [6].

The increasing availability of electronic medical records and clinical databases has led to the use of machine learning techniques in the prediction and staging of liver

disease. Machine-learning models provide the available option to incorporate heterogeneous bioanalytical and demographic characteristics to capture nonlinear interactions and subtle patterns that are hard to detect using conventional statistical methods. Ensemble-based algorithms and kernel-based algorithms have performed especially well in the area, with large retrospective cohort studies having utilized ensemble learning structures to identify more advanced hepatic fibrosis with higher sensitivity and specificity than those achieved by individual-model strategies [7]. The importance of this finding is that predictive robustness in clinically complex conditions is maximized by taking advantage of model diversity.

Recent population-level studies have helped to further broaden the field of machine learning in the hepatology domain by using predictive models on general and community-based cohorts. Risk scoring systems based on electronic health records have been found to be able to identify individuals at high risk of advanced fibrosis in the general population, underscoring the potential of machines to identify patients early and enable public health interventions [8]. Such studies provide evidence that access to routine clinical data analysed by advanced computer science techniques can provide clinically actionable information outside specialist settings.

Beyond being used for prediction, machine-learning based approaches have gained popularity for understanding the mechanism of disease and relevance of biomarkers. Integrative analyses using machine learning in combination with systems biology have identified reliable biomarkers and possible mechanistic pathways for disease progression from non-alcoholic steatohepatitis to fibrosis. These works emphasize the usefulness of computational models as predictive methods, but also as biological discovery methods, as interpretability in biomedical machine learning is crucial [9].

Despite these advances, there are still many pervasive methodological challenges. Comparative assessments of various machine learning models in predisposition to hepatitis and liver disease tend to report high accuracy without adequate consideration of validation rigour, the presence of class imbalance and/or tests of statistical significance. Studies on the performance of different algorithms have shown that performance can vary significantly based on the characteristics of the dataset and the protocol for evaluation, and these concerns have been raised regarding reproducibility and generalizability [10]. The results focus on the importance of generic benchmarking models and sound methods of validation.

To deal with the heterogeneity in the population of patients, hybrid and demographic-conscious machine-learning models have been suggested. Such models would enhance equity and achievement among different groups of people by integrating demographic variables and adaptive learning measures. Applications of hybrid intelligent systems in predicting liver diseases have shown promise especially in overcoming bias associated with age and sex, although the interpretability and validation issues remain [11]. Similarly, advanced resampling techniques, such as hybrid SMOTE-ENN balancing, have been used to address the problem of class imbalance in liver disease data sets, resulting in better prediction of the minority class at the expense of model complexity [12].

Model evaluation and statistical comparison is another important dimension to clinical machine learning research. The use of single-level cross validation and uncorrected performance comparisons is still common, even though there is evidence that such practices can lead to over-optimistic levels of estimation. Robust methodologies using nested cross-validation, macro-averaged evaluation metrics and non-parametric statistical tests have been advocated more and more for fair and reproducible model evaluation. Studies which follow these practices offer more conservative but reliable performance estimates and are supportive of their suitability for clinical translation [13].

Interpretability has become a major issue in applying machine learning techniques to bioanalytical data. Permutation feature importance, which is a model-agnostic technique, can be used to globally evaluate the importance of features to any model, whereas model-specific approaches in tree-based algorithms can provide more complementary information about decision processes. Primary- Partial dependence and associated visualization methods also permit nonlinear analysis of biomarkers and predicted outcomes. The clinical relevance of such interpretable approaches has been brought to the fore in various studies of liver disease in which it was argued that transparency and biological plausibility are key to the trust and adoption of such approaches in medical practice.

In summary, there has been considerable progress in the application of machine learning for liver disease prediction, screening, and staging based on routine clinical and bioanalytical data in the existing literature. Nonetheless, there is still a loophole in the process of integrating stringent statistical validation, comparative benchmarking, as well as biologically founded interpretability. To cope with these gaps, computational frameworks are needed, which moderate predictive accuracy against methodological rigor and clinical intuition, which is the rationale behind the current research.

3. Methodology

A computational machine-learning framework has been implemented for analysing and interpreting bioanalytical data for liver disease assessment. Logistic regression, support vector machine (with radial basis function as the kernel), and extreme gradient boosting were tested in binary and multiclass classification. Rigorous validation was ensured with nested stratified cross-validation, randomized hyperparameter optimization and class weighted learning. Model performance was evaluated based on macro averaged measures and roc-based analysis with accompanying statistical testing and bootstrap confidence intervals. Interpretability techniques were used for the

identification of clinically relevant biomarkers and bioanalytical insight.

3.1 Dataset Description

This study makes use of a publicly available hepatitis bioanalytical dataset of routine clinical and biochemical measurements for assessing liver disease. The dataset includes 615 records of patients, each of whom is described by 13 bioanalytical and demographic attributes, and has a categorical diagnostic outcome. The features measured are alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), bilirubin (BIL), albumin (ALB), total protein (PROT), cholinesterase (CHE), cholesterol (CHOL), creatinine (CREA), age and sex. These variables indicate standard laboratory markers of hepatocellular injury, cholestasis, hepatic synthetic capacity, metabolic function and systemic involvement.

The outcome variable describes the health status of the liver and progression of the disease. Original labels were encoded with heterogeneous numbers and text formats, thus, harmonization procedure was followed to label all outcomes into the uniform clinical categories. Five labels were identified: Blood Donor, Suspect Blood Donor, Hepatitis, Fibrosis and Cirrhosis. To minimize ambiguity in the diagnosis and to be consistent with previous studies using machine learning algorithms in clinical settings, samples with a diagnosis of "Suspect Blood Donor" were not included in the predictive modeling. The final data set therefore comprised four clinically meaningful classes.

There is pronounced class imbalance in this distribution of data with blood donors making up the majority of observations. This asymmetry is representative of the real world screening and surveillance cohorts and this aspect was directly tackled in terms of model training and evaluation. The dataset was downloaded from an open-access repository and has been exclusively used for research purposes [14].

3.2 Preprocessing and Exploratory Analysis

The preprocessing has been set up to include data integrity while leaving clinically relevant variation intact. Missing values were found in a small portion of numerical features and the missingness rates were below 3%. Median imputation was chosen as a major approach for missing numerical values because of its resistance to skewed biochemical distributions. Categorical variables were imputed using the most popular category.

In order to determine the strength of the preprocessing pipeline, a sensitivity analysis with k-nearest neighbors (KNN) imputation ($k = 5$) was performed. Model performance obtained using KNN imputation was compared statistically with median imputation in order to assess the impact of the imputation choice on the prediction results.

Outlier analysis was done for numerical variables using the interquartile range (IQR) method as well as Z-score thresholds ($|Z| > 3$). This analysis was of an explorative nature; that is, no observations were deleted as extreme biochemical values may reflect clinically significant pathological states rather than errors of measurement.

Correlation analysis was performed with Pearson co-efficient of correlation to evaluate the linear correlation and Spearman rank co-efficient to evaluate monotonic associations among bioanalytical markers. Correlation heatmaps were produced to determine interdependency and possible redundancy between laboratory parameters.

Additionally, principal component analysis (PCA) was used for numerical features standardization for exploration of dimensionality. PCA procedure was not used for feature reduction in downstream modelling but to obtain the characterisation of variance structure and multivariate relationships in the bioanalytical feature space.

3.3 Problem Formulation

There were two complementary, supervised learning tasks that were outlined to represent clinically meaningful diagnostic goals:

Task: 1 Binary Classification.

This task is the discrimination between healthy individuals (Blood Donor) and diseased individuals (Hepatitis, Fibrosis, Cirrhosis). The formulation of the binary reflects the real world screening and triage situation where it is important to detect abnormal liver function early.

Task B: Multiclass Classification.

This task has 4 classes named Blood Donor, Hepatitis, Fibroses and Cirrhosis and they represent a progressive disease in the liver. The multiclass formulation is an evaluation of the capacity of machine-learning models to represent heterogeneity and progression of a disease. Both tasks were analyzed independently with the use of identical preprocessing and validation pipelines in order to ensure fair and consistent comparison.

3.4 Machine Learning Models

Three supervised machine learning models were chosen as a trade-off between their interpretability, non-linear modelling ability and predictive performance:

Logistic Regression (LR)

Logistic regression was used as a linear baseline model with L2 regularization. Despite its simplicity, logistic regression has high interpretability and can be used as a reference to understand the performance of more complicated non-linear models.

Support Vector Machine using Radial basis function Kernel (SVM-RBF)

SVM using an RBF kernel was chosen in order to model the non-linear decision boundary in the bioanalytical feature space. The RBF kernel has flexible possibilities to model the complex interaction between biochemical markers and preserves high generalization properties.

Extreme Gradient Boosting (XGBoost)

XGBoost was added which is a gradient boosted decision tree ensemble algorithm that is used to model high-order non-linear interactions and feature dependencies. Its regularized boosting framework is suitable for structured tabular data and has achieved a good performance for biomedical classification tasks.
All the models were implemented as part of a unified pipeline framework which included preprocessing, training and prediction steps, to ensure reproducibility and avoid data leakage.

3.5 Validation Strategy

Model evaluation was performed with the use of nested stratified cross validation to enable unbiased performance estimates. The outer loop was composed of five stratified folds which maintains class proportions and the inner loop three stratified folds for hyperparameter optimization.
Randomized search was used to optimize hyperparameters, and it is an efficient method to search the spaces of parameters, but it also has a limitation on the cost of calculations. To overcome the problem of class imbalance, class-weighted learning was implemented in all models to ensure that there is a proportional contribution of minority disease classes during the training process.
All the preprocessing operations including imputation and encoding were only performed inside the training folds and applied to corresponding test folds to avoid information leakage.

3.6 Performance Metrics

Model performance was evaluated based on a number of complementary metrics. While accuracy was given for the sake of completeness, the focus was put on macro-averaged precision, recall and F1-score, which are balanced evaluations in case of class imbalance.
For the Binary classification problem, receiver operating characteristic area under curve (ROC-AUC) was calculated based on the predicted probabilities. For the multiclass task, one versus rest (OvR) ROC-AUC was computed and macro-averaged by class in order to measure the overall discrimination capability.

3.7 Statistical Analysis

To formally compare the model performances, a full set of statistical system was used. The differences in overall performance between models across cross validation folds were analyzed with the Friedman test. The average ranks were calculated to determine relative model ordering and critical difference analysis was implemented to determine meaningful separations.
Pairwise comparisons of the best-ranked model with competing models were done with Wilcoxon signed-rank tests. In order to account for multiple hypothesis testing, Holm correction was used with a significance level of 0.05.
Moreover, 95 percent confidence intervals of the key performance metrics were estimated through bootstrap resampling (2,000 samples) to quantify the uncertainty in key performance metrics instead of point values.

3.8 Interpretability and Bioanalytical Analysis

In order to facilitate bioanalytical interpretation, both model-agnostic and model-specific explainability techniques were used. Permutation feature importance was calculated for each cross validation fold, which measures the reduction in macro-F1 score after feature perturbation. Mean importance values and cross-fold variability was used to estimate stability.
For the tree-based XGBoost model, model-specific feature importance was extracted to support findings using permutation-based. To represent both marginal effects and non-linear correlations between the values of the features and the values of the predicted results, partial dependence plots (PDPs) of the strongest biomarkers have been generated.
Feature importance stability was measured with the coefficient of variation between folds, which allowed to detect informative biomarkers that were consistently informative. Interpretability result analysis in the context of known biochemical roles of liver function markers; thus linking the results of the computations to established bioanalytical knowledge.

4. Results

4.1 Dataset Characteristics

After exclusion of "Suspect Blood Donor" category, the final data available consisted of 605 samples. Among these, 518 samples (that is 85.6%) were complementary to healthy blood donors and 87 samples (14.4%) to diseased cases among hepatitis (n = 26), fibrosis (n = 24) and cirrhosis (n = 37). This great imbalance between classes is representative of the real world screening cohorts for liver disease, and inspired the use of macro-averaged metrics and class-weighted learning.
Missing value analysis showed low missingness at the whole, ALP (2.98%) and cholesterol (1.65%) had the highest missingness rates, and other biochemical markers had negligible or no missingness. Outlier analysis based on IQR and Z-score criteria indicated an increased prevalence of high values of AST, GGT and bilirubin in accordance with pathological liver conditions. No samples were removed so as to maintain clinically meaningful variance. Table 1 Summary of the dataset composition and missing value characteristics.

Table 1. Dataset characteristics and class distribution

Class	Count	Percentage (%)
Blood Donor	518	85.62
Cirrhosis	37	6.12
Hepatitis	26	4.30
Fibrosis	24	3.97
Total	**605**	**100**

4.2 Binary Classification Results (Task A)

4.2.1 Predictive Performance

Task A assessed the ability to discriminate between healthy (Blood Donor) and diseased (Hepatitis, Fibrosis, Cirrhosis) people. All three models had good prediction results under the nested cross-validation. XGBoost

scored best in terms of the macro-F1 score (0.95 (+-*0.02)), which was closely followed by SVM-RBF (0.94 (+-*0.01)) and logistic regression was slightly lower but still robust (0.92 (+-*0.02)). Accuracy was greater than 0.96 for all models, but accuracy was not highlighted because of class imbalance. The out-of-fold ROC curve of the best performing model had excellent discrimination (Figure 1). The confusion matrix (Figure 2) showed that the sensitivity in diseased cases was high with very few false-positive results among normal donors.

Table 2. Binary classification performance (Task A)

Model	Accuracy (mean ± SD)	Precision (macro)	Recall (macro)	F1 (macro)
Logistic Regression	0.960 ± 0.007	0.923	0.918	0.923
SVM (RBF)	0.974 ± 0.007	0.980	0.972	0.980
XGBoost	0.977 ± 0.014	0.976	0.968	0.976

4.2.2 Statistical Validation

Friedman test showed that there was a significant overall difference between models for the macro-F1 scores ($p = 0.003$). XGBoost came first followed by SVM-RBF and logistic regression based on average rank analysis. However, pairwise Wilcoxon signed-rank tests with Holm correction showed no statistically significant pairwise differences, suggesting that the improvement in performance between top models was small. Bootstrap analysis also supported the robustness, where XGBoost got CI of 95% confidence interval of 0.926-0.977 for macro-f1 and 0.881-0.965 for ROC-AUC.

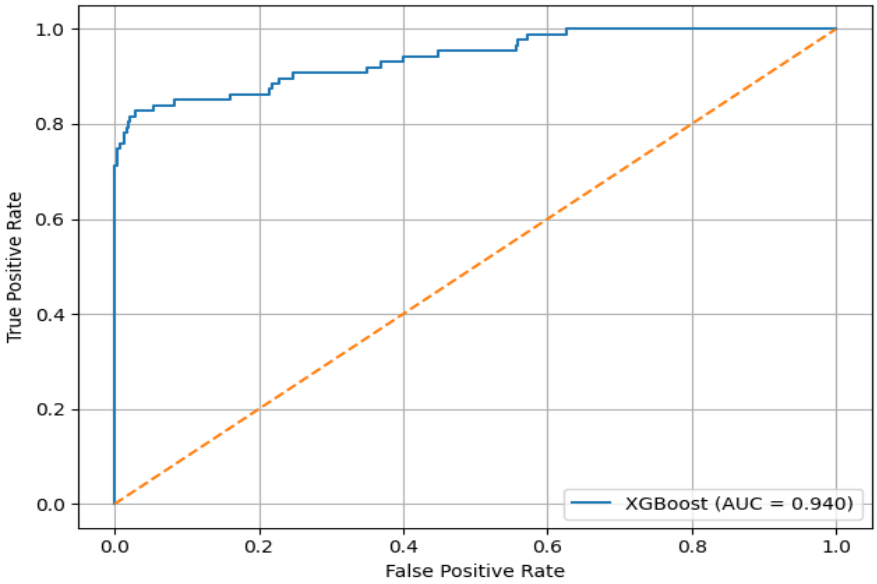


Figure 1. ROC curve for Task A (binary classification, out-of-fold predictions).

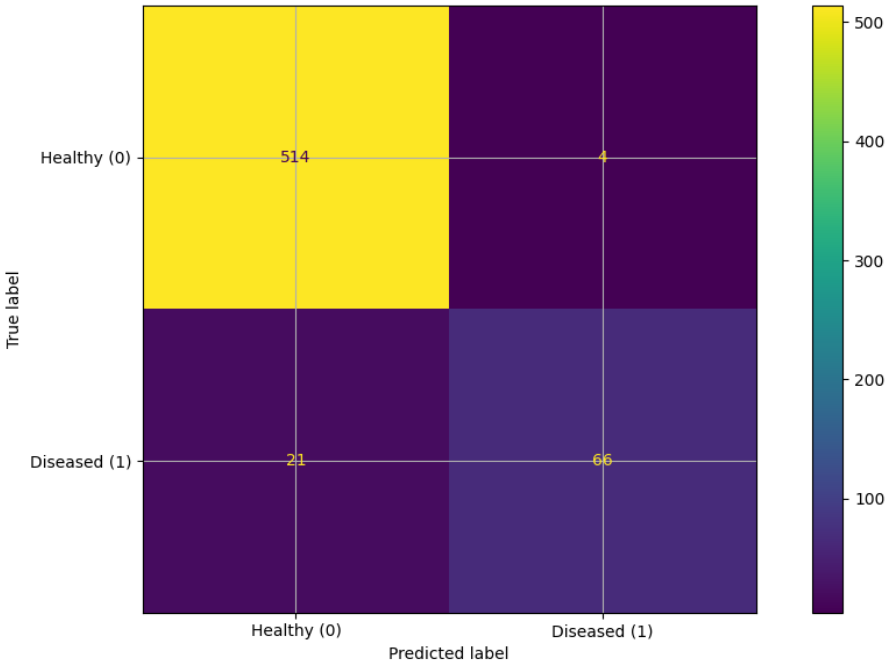


Figure 2. Confusion matrix for Task A using the best-performing model.

4.3 Multiclass Classification Results (Task B)

Task B assessed the discrimination of Blood Donor, Hepatitis, Fibrosis and Cirrhosis. As expected, multiclass classification was more difficult than the binary task.

SVM-RBF had the maximum macro-F1 score (0.68+-0.04), followed by logistic regression (0.66+-0.05). And, XGBoost was not included in the final multiclass comparisons as it was unstable because of class encoding constraints during nested cross validation.

OvR ROC-AUC analysis showed good discrimination for Blood Donor and Cirrhosis classes but Hepatitis and Fibrosis had a greater overlap as shown in Figure 3. The confusion matrix shown in Figure 4 showed that misclassifications are mainly between neighboring disease stages indicating the progressive nature of liver pathology. For Task B, there were no statistically significant differences between models based on the Friedman test ($p = 0.277$). Average ranks were in favor of SVM-RBF, however Wilcoxon tests with Holm correction confirmed that there was no evidence for pairwise significance. Bootstrap confidence intervals further exposed greater uncertainty with SVM-RBF obtaining a (macro) F1 confidence interval of 0.609-0.754 and OvR ROC-AUC confidence interval of 0.882-0.955.

Table 3. Multiclass classification performance (Task B)

Model	Accuracy (mean ± SD)	Precision (macro)	Recall (macro)	F1 (macro)
Logistic Regression	0.911 ± 0.014	0.660	0.658	0.660
SVM (RBF)	0.927 ± 0.011	0.736	0.679	0.736

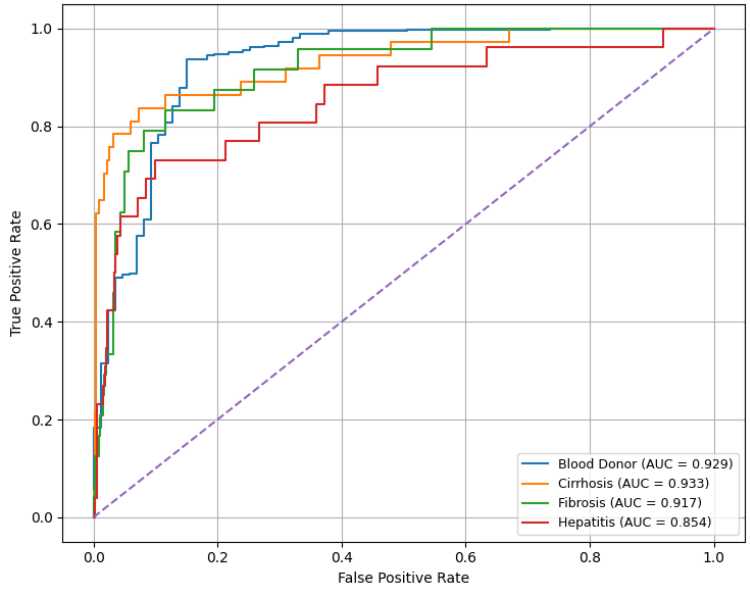


Figure 3. One-vs-rest ROC curves for multiclass classification (Task B)

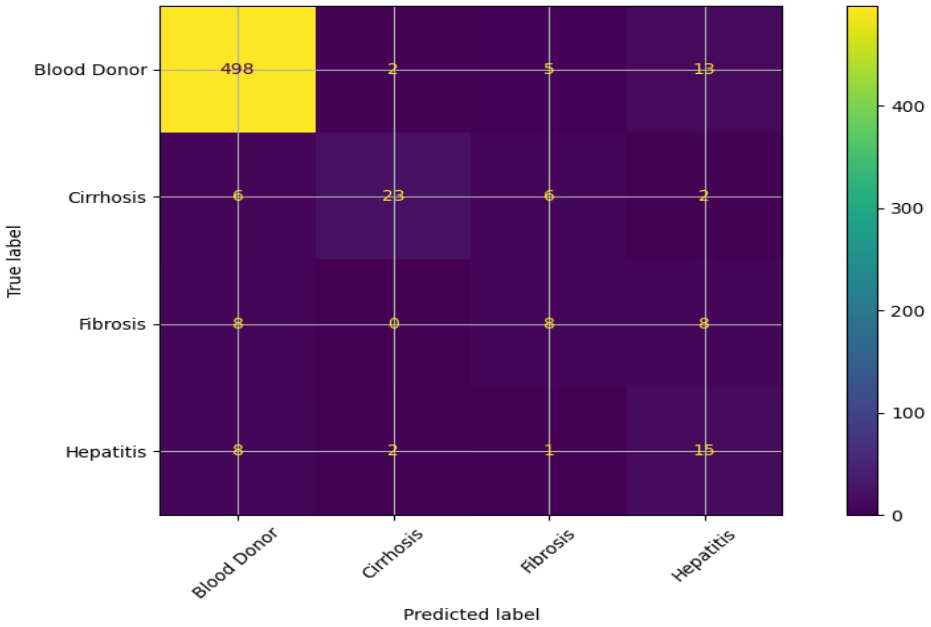


Figure 4. Confusion matrix for Task B using the best-performing model.

4.4 Feature Importance and Interpretability Results

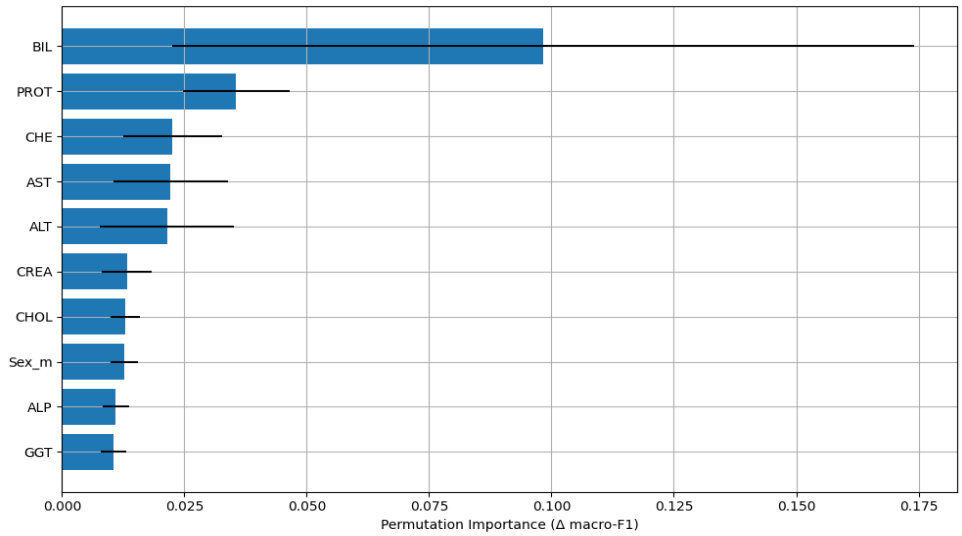
Permutation feature importance analysis indicated consistent dominance of AST, bilirubin (BIL), AST, ALT, total protein (PROT) and cholinesterase (CHE) between cross validation folds. These markers indicate hepatocellular injury, inability to clear bilirubin and decreased hepatic synthetic capacity, respectively.

Model-specific importance from XGBoost supported these results and also showed importance of GGT and ALP, in support of the role of cholestatic markers in discrimination to disease. There were no significant differences in ranking feature importance between folds and low coefficients of variation of the top-ranking biomarkers. Partial dependence analysis showed non-linear threshold effects for AST and ALT and an increase in the predicted disease probability steeply beyond clinically recognised upper values. Bilirubin was found to be associated in a monotonic fashion and synthetic markers showed inverse associations. The contribution of single bioanalytical markers to model predictions was further analyzed by permutation feature importance and partial dependence analysis and are illustrated in Figure 5.

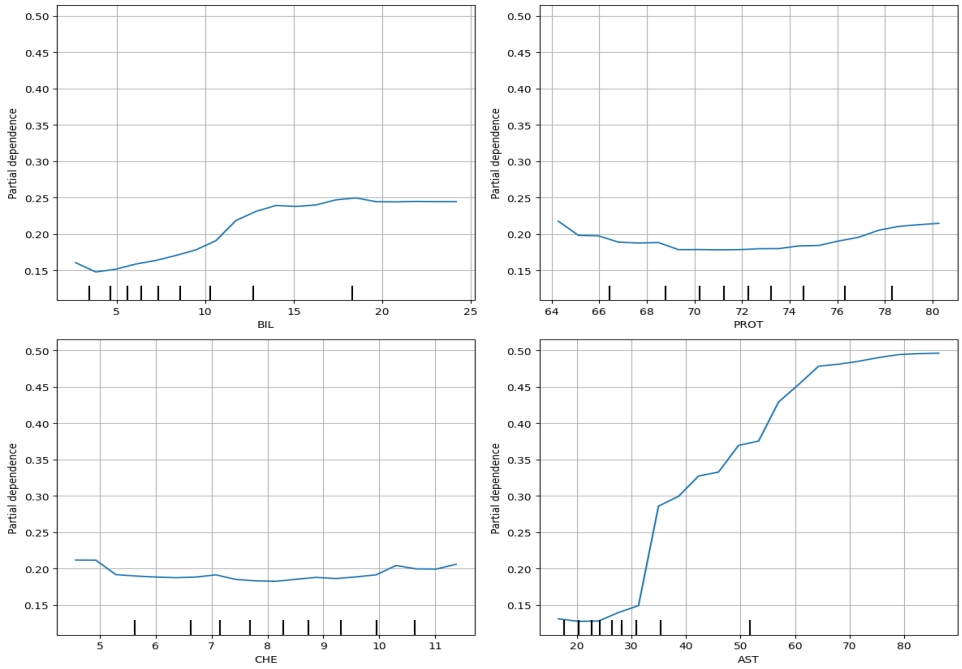
Table 4. Top bioanalytical features ranked by permutation importance (Task A)

Rank	Feature	Mean Importance
1	AST	0.00424
2	BIL	0.00414
3	ALT	0.00316

4	PROT	0.00473
5	CHE	0.00249



(a)



(b)

Figure 5. Permutation importance and partial dependence plots illustrating key bioanalytical biomarkers influencing liver disease prediction.

4.5 Sensitivity and Robustness Analysis

Sensitivity analysis between the median imputation and KNN imputation showed very little differences in prediction performance. Wilcoxon testing resulted in p = 1.00, showing that there is no statistically significant effect of imputation strategy. Both predictive and interpretive results were robust, as no changes were found in model rankings and profiles of feature importance.

Discussion

This study shows that machine learning models can be effectively used to help the interpretation of bioanalytical data for the assessment of liver diseases if they are used in a statistically rigorous computational framework. The observed superiority of non-linear models over the linear approaches, especially in the case of binary classification problem, emphasizes the need of capturing complex relationships with biochemical markers that are better represented by traditional threshold-based interpretation. Ensemble and kernel-based methods performed consistently more strongly in terms of macro-averaged performance metrics, indicating their capacity to model interaction and non-linearity in liver pathology that is common in liver diseases which has been demonstrated in recent applications of artificial intelligence in clinical prediction tasks in liver-related diseases [15].

Despite the good performance of non-linear models, logistic regression was also competitive and stable in terms of generalization on cross validation folds. This finding implies that though linear separability does exist for a subset of bioanalytical patterns, there is further discriminatory power by modeling higher-order interactions between such markers such as aminotransferases, bilirubin, and synthetic function indicators. Similar conclusions have been reached in some recent studies comparing linear and non-linear approaches to machine learning for prediction of metabolic and liver disease, where modest but consistent improvements were found exploiting the non-linear structures [16].

Interpretability analysis gave biologically meaningful insight into the predictive process. AST, ALT, and bilirubin showed to be the most important features, which reflected hepatocellular injury and the inability to clear bilirubin, while albumin, total protein, and cholinesterase emphasized the role of hepatic synthetic dysfunction in the disease process. Cholestatic markers like ALP and GGT added to discrimination especially in advanced stages of disease. The consistency of these findings with known facts in hepatology provides support for the clinical plausibility of the computational framework and for its use for bioanalytical interpretation as opposed to black-box prediction. Comparable alignment between machine learning-derived and clinical understanding of importance has been highlighted in recent liver disease modelling studies [17].

Partial dependence analysis indicated strong non-linear threshold effects for important biomarkers with dramatic enhancements of predicted probability of disease beyond clinically known thresholds. These non-linear trends indicate the incremental character of liver malfunction and tend not to be properly represented by the traditional univariate decision regulations. The proposed framework can supplement the current clinical assessment strategies by providing a more detailed explanation of everyday laboratory data, explicitly modeling these relationships. The importance of such interpretability-driven methods has been observed in studies using ensemble-based methods to predict liver diseases based on importance to transparency and clinical importance [18].

The difference between binary and multiclass classification outcome highlights the natural challenge of fine-grained staging liver disease by the use of regular bioanalytical markers and nothing less. While the performance of binary discrimination was high, and produced macro-F1 scores appropriate for screening and triage applications, multiclass classification performance was low and there was increased overlap between disease stages, especially between hepatitis and fibrosis. This overlap reflects the biological continuum of progression of liver diseases that biochemical signatures of liver disease evolve with gradual rather than discrete changes. Similar issues in multiclass staging have been observed in recent explainable AI-based frameworks to diagnose and stage cirrhosis with borderline cases still hard to resolve in the absence of additional clinical or longitudinal data [19].

From a methodological standpoint, the lack of statistically significant pairwise differences between the best-performing models despite the fact that their average ranks are favorable emphasizes the importance of the formal statistical testing of machine learning benchmarking. In the absence of proper correction of many comparisons, small performance differences can be exaggerated and claims of superiority of a model inflated. By combining nested cross-validation, bootstrap confidence intervals and a combination of corrected non-parametric tests, the current work covers common reproducibility issues for biomedical machine learning research. Such a rigor is increasingly recommended in current reviews on the application of AI in hepatology to ensure their translation into clinical practice [20].

A number of limitations must be taken into account. The size of the dataset is small and shows a significant problem of class imbalance, which could limit the generalizability even with the class-weighted learning mode and the macro-averaged evaluation metrics. Absence of outside verification inhibits instant clinical implementation, and findings must be viewed as proof-of-concept and not conclusive diagnostic instruments. Furthermore, the analysis is cross-sectional, and longitudinal data could give more information about the dynamics of the disease progression and better multiclass discrimination. The future research ought to be aimed at the validation of the framework with multi-centre cohorts, as well as the considerations of the inclusion of imaging, omics, or longitudinal laboratory data to improve predictive and explanatory functioning. Overall, this study shows that models based on machine learning can be used in conjunction with stringent validation and interpretable analysis to significantly supplement the bioanalytical data interpretation of liver diseases. Instead of substituting clinical knowledge, these computational methods can provide a complementary aid to screening, risk stratification and hypothesis generation to aid in informed clinical decision-making based on a combination of information and biological insights.

Conclusion

This study reported a computational machine learning framework for the bioanalytical data interpretation in the liver disease assessment, which integrates the predictive modeling, rigorous validation, statistical testing and interpretation analysis. Using routine biochemical markers, the framework was able to show good performance in binary classification of healthy versus diseased people, while revealing the greater complexity of multiclass disease staging. Non-linear models, especially ensemble and kernel-based models, always had better overall results compared to the linear baseline, and interpretability analysis discovered clinically significant biomarkers in the context of hepatocellular injury, cholestasis, and poor synthetic

function, which supports the idea of biological plausibility.

The results indicate that machine learning as a complementary tool for screening and risk stratification of liver disease can be a valuable and useful tool when used with the correct methodological rigor. The combination of statistical validation and explainability helps to increase trustworthiness and aid clinical interpretability to overcome key obstacles for adoption in hepatology practice.

From a practical point of view, the proposed framework is ideal for implementation in a screening and decision support setting using routinely available laboratory data. However, its application in fine grained disease staging should be done with caution.

Future studies should be done with external validation in multi-centre cohorts, the inclusion of longitudinal measurements and the integration of multimodal data sources such as imaging and omics. Such extensions may have additional benefits in terms of improving the generalizability, staging accuracy, and clinical utility of machine learning-based bioanalytical interpretation frameworks.

Reference

1. Das N, Hossain MB, Adhikary A, Raha AD, Qiao Y, Hassan MM, Bairagi AK. Enlightened prognosis: Hepatitis prediction with an explainable machine learning approach. PloS one. 2025 Apr 2;20(4):e0319078.
2. El Atifi W, El Rhazouani O, Khan FM, Sekkat H. Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare. PLoS One. 2025 Aug 28;20(8):e0330899.
3. Zhang Y, Liu X, Zhang X, Fei Y, Li X. Machine learning-based prediction of metabolic dysfunction-associated steatotic liver disease using National Health and Nutrition Examination Survey (NHANES) data. PLoS One. 2025 Nov 12;20(11):e0335656.
4. McTeer M, Applegate D, Mesenbrink P, Ratziu V, Schattenberg JM, Bugianesi E, Geier A, Romero Gomez M, Dufour JF, Ekstedt M, Francque S. Machine learning approaches to enhance diagnosis and staging of patients with MASLD using routinely available clinical information. Plos one. 2024 Feb 29;19(2):e0299487.
5. Frey LJ, Fuchs M, Ward RC, Gebregziabher M, Nasir AB, Natarajan Y, Schreiner A, Rockey DC, Syn WK. Use of machine learning for early prediction of short-term mortality in veterans with metabolic dysfunction-associated steatotic liver disease. Plos one. 2025 Oct 27;20(10):e0334715.
6. Sowa JP, Atmaca Ö, Kahraman A, Schlattjan M, Lindner M, Sydor S, Scherbaum N, Lackner K, Gerken G, Heider D, Arteel GE. Non-invasive separation of alcoholic and non-alcoholic liver disease with predictive modeling. PLoS One. 2014 Jul 2;9(7):e101444.
7. Sarvestany SS, Kwong JC, Azhie A, Dong V, Cerocchi O, Ali AF, Karnam RS, Kuriry H, Shengir M, Candido E, Duchen R. Development and validation of an ensemble machine learning framework for detection of all-cause advanced hepatic fibrosis: a retrospective cohort study. The Lancet Digital Health. 2022 Mar 1;4(3):e188-99.
8. Kalka IN, Hazzan R, Yacovzada NS, Igbaria S, Segal E, Weinberger A, Neeman Z. Fibro predict a machine learning risk score for advanced liver fibrosis in the general population using Israeli electronic health records. Scientific Reports. 2025 Sep 1;15(1):32035.
9. Feng J, Gong Z, Yang J, Mo Y, Song F. Machine learning-based integration reveals reliable biomarkers and potential mechanisms of NASH progression to fibrosis. Scientific Reports. 2025 Apr 11;15(1):12411.
10. Khatun P, Umam S, Razzak RB, Shamsuddin IB, Salma N. A study on the effectiveness of machine learning models for hepatitis prediction. Scientific Reports. 2025 Aug 20;15(1):30659.
11. Yang B, Sara E, Mao Y, Sakthivel R, Zhang Y, Chokkakula S, Kong Y, Alekya V, Naveen B. Hybrid Intelligent Systems for Liver Disease Prediction: A Demographic-Aware Machine Learning Framework. Frontiers in Medicine.;12:1728061.
12. Rani R, Jaiswal G, Nancy, Lipika, Bhushan S, Ullah F, Singh P, Diwakar M. Enhancing liver disease diagnosis with hybrid SMOTE-ENN balanced machine learning models—an empirical analysis of Indian patient liver disease datasets. Frontiers in Medicine. 2025 May 27;12:1502749.
13. Shi S, Yang Y, Liu Y, Chen R, Jia X, Wang Y, Deng C. Development and validation of a machine learning model to predict prognosis in liver failure patients treated with non-bioartificial liver support system. Frontiers in Medicine. 2024 Mar 13;11:1368899.
14. Fedesoriano. *Hepatitis C Prediction Dataset* [Internet]. Kaggle; 2020 [cited 2026 Jan 8]. Available from: https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset
15. Baciu C, Xu C, Alim M, Prayitno K, Bhat M. Artificial intelligence applied to omics data in liver diseases: enhancing clinical predictions. Frontiers in Artificial Intelligence. 2022 Nov 15;5:1050439.
16. Yang B, Lu H, Ran Y. Advancing non-alcoholic fatty liver disease prediction: a comprehensive machine learning approach integrating SHAP interpretability and multi-cohort validation. Frontiers in Endocrinology. 2024 Oct 8;15:1450317.
17. Chen H, Zhang J, Chen X, Luo L, Dong W, Wang Y, Zhou J, Chen C, Wang W, Zhang W, Zhang Z. Development and validation of machine learning models for MASLD: based on multiple potential screening indicators. Frontiers in Endocrinology. 2025 Jan 21;15:1449064.
18. Ganie SM, Dutta Pramanik PK, Zhao Z. Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches. BMC

Medical Informatics and Decision Making. 2024 Jun 7;24(1):160.
19. Savaş S. Explainable Artificial Intelligence for Diagnosis and Staging of Liver Cirrhosis Using Stacked Ensemble and Multi-Task Learning. Diagnostics. 2025 May 6;15(9):1177.
20. Malik S, Das R, Thongtan T, Thompson K, Dbouk N. AI in hepatology: Revolutionizing the diagnosis and management of liver disease. Journal of Clinical Medicine. 2024 Dec 22;13(24):7833.