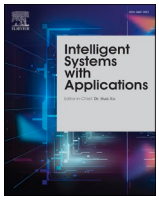
Intelligent Systems with Applications 28 (2025) 200573



Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications



Liver cirrhosis prediction: The employment of the machine learning-based approaches

Genjuan Ma, Yan Li *

College of Media Technology, Communication University of China Nanjing, Nanjing, Jiangsu 211172, China

ARTICLE INFO

Keywords:
Liver disease recognition
Quadratic discriminant analysis (qda)
Healthcare
Machine learning models
early detection

ABSTRACT

Early detection of liver cirrhosis remains problematic due to its asymptomatic onset and the inherent class imbalance in clinical data. This study conducts a comprehensive evaluation of machine learning models for predicting cirrhosis stages, with a focus on addressing these challenges. An approach employing Quadratic Discriminant Analysis (QDA) is benchmarked against seven other models, including powerful ensembles like Stacking and HistGradientBoosting, on a clinical dataset. Methodologies such as SMOTE oversampling, stratified data splitting, and class-specific covariance estimation were implemented to manage data complexity. The results demonstrate that a Stacking ensemble achieves the highest overall predictive performance with a micro-AUC of 0.80. The proposed QDA method also proves to be a highly effective and competitive model, achieving a robust AUC of 0.76 and outperforming several specialized imbalance-learning algorithms. Crucially, QDA offers this strong performance with exceptional computational efficiency. These findings show that while complex ensembles can yield top-tier accuracy, QDA's capacity to model non-linear feature associations makes it a powerful and practical choice for the diagnosis of cirrhosis.

1. Introduction

Liver disease is one of the major health burdens and a leading cause of morbidity and mortality (Assegie et al., 2022), (Sivasangari et al., 2020). The liver plays a key role in detoxification processes, metabolism, and the synthesis of important proteins. It is very prone to many disorders, especially hepatitis, cirrhosis, and liver cancer (de Mattos et al., 2021). Liver cirrhosis generally progresses asymptomatically and therefore needs timely recognition for its effective control and treatment (Huang et al., 2020), (Choi & Oh, 2024). It is of great importance to detect liver diseases at an early stage to allow timely intervention that can impede disease advancement and avoid serious complications (Dutta et al., 2022), (Ghazal et al., 2022). Early-stage detection allows clinicians to treat patients under preventive measures, which may also include a personalized treatment regimen, thus improving the prognosis of quality of life in a patient significantly (Singal & Mathurin, 2021), (Anuradha et al., 2022), (Nowduri & Subramanyam, 2024). Therefore, the early identification of liver disease could reduce the economic burden resulting from extensive treatments or hospitalization caused by a progressed stage of this disease. Consequently, the enhancement of diagnostic accuracy and the identification of effective early recognition frameworks have become two major needs in liver disease management.

Previous studies reviewed different frameworks for the recognition of liver diseases, focusing on both traditional diagnostic frameworks and advanced technological solutions. Traditional frameworks of liver biopsies and imaging are mostly effective but invasive and usually require significant expertise (Anuradha et al., 2022), (Shaheamlung et al., 2020). Recent developments involved the introduction of non-invasive frameworks and computational frameworks, biomarkers, and analysis of images. ML has recently evolved as a promising tool in this area that offers various new opportunities for improving diagnostic accuracy through the analysis of complex patterns present in medical data (Kumar et al., 2018), (Aghamohammadi et al., 2014), (Castaneda et al., 2015). This trend is evident across numerous medical fields, with recent studies successfully applying machine and deep learning techniques for the automated diagnosis of conditions ranging from schizophrenia and tinnitus using EEG/fMRI data (R.A. Lashaki et al., 2025), (Raeisi et al., 2024) to the optimized classification of dental implants from imaging (Z. Raeisi et al., 2025). Despite this progress, however, several challenges persist regarding how to best optimize these frameworks for broad clinical use.

ML has transformed many healthcare domains into a strong tool for

* Corresponding author.
E-mail address: chunyanqing_860821@163.com (Y. Li).

https://doi.org/10.1016/j.iswa.2025.200573
Received 14 October 2024; Received in revised form 14 July 2025; Accepted 17 August 2025
Available online 2 September 2025
2667-3053/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

data analysis and anticipation (Ahn et al., 2021), (Aghamohammadi et al., 2018), (Kumar & Mahajan 2020). Considering liver disease, ML frameworks will analyze huge volumes of patient data that include clinical records and imaging outcomes to determine patterns and predict disease outcomes. Frameworks such as categorization, regression, and clustering have been employed to improve liver disease recognition and prognosis. Because of this, ML-based frameworks give more accurate forecasts and treatment plans since the identification can be done through patterns that might be missed in traditional frameworks (Islam et al., 2024), (Topcu et al., 2024). ML embedded in medical data mining and analysis has the capacity for significant enhancement in efficacy regarding detecting and managing liver diseases.

While much progress has been achieved in the area of ML-based liver disease recognition, further investigation is called for to overcome some issues that still face this domain for improved accuracy (Dritsas & Trigka, 2023), (Lu et al., 2024). The current frameworks may not represent the complexities of liver diseases, and it is urgent to study and refine frameworks that could enhance performance. The objective of the paper is, therefore, to fill these gaps with a high-accuracy predictive framework in detecting liver diseases through ML. This study tries not only to improve previously available frameworks but also to enrich the literature with more reliable and exact tools for early recognition.

While existing models (e.g., SVM, Random Forest) for liver disease prediction work reasonably well, these models perform less well under intricate class distributions seen in cirrhosis data, where steps of disease progression display nonlinear correlations. For instance, linear decision boundaries, as implemented by Logistic Regression, cannot explain subtle correlations between biomarkers (e.g., Albumin, Bilirubin) and cirrhosis grades. QDA serves this niche by estimating class-specific covariance matrices, enabling adaptive decision boundaries tolerant of heterogeneous data patterns. Moreover, previous studies (Topcu et al., 2024), (Kalaiselvi et al., 2021) do not present rigorous comparisons across ensemble architectures (Balanced Bagging, RUSBoost) addressing class imbalance, an inevitable problem across clinical data. The presented work innovates by proposing a QDA-based approach to liver cirrhosis staging, coupling it with stratified splitting, retaining class distributions, and offering comparative analysis across recent models to discover context-specific benefits.

In the current study, an algorithm based on QDA has been developed that meets all the requirements of being the most accurate liver disease recognition model. This study intends to be concerned with creating and evaluating ML-based predictors as Bagging and Voting Classifiers, on the medical records database. These models are trained, validated, and tested to measure their respective performances and accuracies, where the proposed QDA method showed its strength in solving the research gap by providing accurate and efficient outcomes. The novelty here lies in the application of the QDA specifically to liver disease prediction, focusing on the complex cirrhosis phases. Compared to other studies, which predominantly looked at ensemble strategies or other architectures of machine learning, this work demonstrates that QDA is efficient in handling non-linear distributions of classes while retaining high categorization accuracy, necessary for complex patterns found in medical data.

The research contributions of the research:

1) A QDA-based framework is developed for the recognition of liver disease.
2) Analyzing the ML-based frameworks and investigating the ML models on liver disease recognition.
3) Extensive experiments are conducted by providing performance analysis, visual representation, and comparison outcomes that establish the strength and efficacy of the suggested framework.

The remainder of this paper is organized as follows: Section 2 reviews related work and research gaps. Section 3 details the methodology, including data preprocessing and QDA implementation. Section 4 presents experimental results and comparisons. Section 5 discusses limitations and clinical implications. Section 6 concludes the study and outlines future directions.

2. Related works

This section reviews related works. Moreover, an analysis of these studies is presented to address the advantages and weaknesses of the papers. Based on this analysis, the research issue is addressed for this study. This paper (Wu et al., 2019) reviews the application of ML frameworks for the anticipation of fatty liver disease and presents ways of improving the accuracy of the anticipation by various frameworks. First, several categorization frameworks, such as the Support Vector Machines (SVM), the Random Forest, and the Gradient Boosting, are applied to a patient medical record database. In this regard, the investigation shows that the highest value of accuracy and robustness in the case of predicting fatty liver disease is presented by the ensemble frameworks, above all, the Random Forest. Key findings of the research: feature selection and preprocessing of the data contribute much to the improved performance of a model. Limitations include single-database reliance, which might affect generalization in other populations. From the implication of the result, it can be seen that ML has an outstanding role in enhancing the early recognition and management of fatty liver disease.

This study (Kalaiselvi et al., 2021) presents several ML frameworks to predict liver disease, which are represented as a database of clinical and biochemical features. The methodology consists of training and testing frameworks, including logistic regression, decision trees, and neural networks. From this, Neural Networks developed the best performance in the anticipation of liver disease. These are important messages related to the robustness of frameworks of deep learning for correctly handling complex patterns from medical data. The shortcomings comprise issues of overfitting or requirements for more and larger databases for confirmation of the models. Advanced frameworks of ML provide promising solutions for the improvement of liver disease prediction.

This paper (Ghazal et al., 2022) proposes an intelligent machine-learning model for the early recognition of liver disease. The framework includes feature engineering and categorization frameworks like SVM and Random Forest to construct the predictive model. This study establishes that the hybrid framework strengthens early recognition of liver disease due to subtle patterns that may not be picked by any other framework. The model can identify cases of early-stage diseases with high precision. However, this study recognizes limitations in the size and diversity of the database that could affect the general applicability of this model. The result is intended to show that ML can improve early recognition and intervention.

This research (Gupta et al., 2022) focuses on the performance assessment of some ML categorization frameworks for liver disease forecast: Decision Trees, Naive Bayes, and Ensemble Frameworks. This methodology is to train these models on the database containing various clinical features and compare the outcomes. Ensemble Frameworks, especially Bagging, deliver the highest accuracy and stability related to liver disease anticipation. This paper has brought into focus the issues of combining multiple models to gain improvements in the anticipation performance. Limitations include that this paper is restricted to using only one type of database, and extensive hyperparameter tuning has not been considered. The outcomes of this study show that the anticipation and management of liver diseases could be much enhanced by the use of ML categorization frameworks.

Another framework for the recognition of liver disease is presented in this paper (Md et al., 2023), using improved preprocessing combined with ensemble ML frameworks. Advanced data preprocessing frameworks include feature scaling and imputation, while some of the corresponding frameworks are Random Forest and Gradient Boosting. This study has shown that improved preprocessing significantly improves model performance, especially regarding accuracy and precision. Salient

points related to the work are the efficiency of preprocessing for handling missing data and outliers. Limitations pertain to the computational cost of the preprocessing steps and the requirements of extensive parameter tuning. The result has shown that an effective preprocessing technique coupled with ensemble frameworks can improve the recognition of liver disease.

In this paper (Behera et al., 2023). A hybrid algorithm is presented by incorporating Modified Particle Swarm Optimization (PSO) with the widely used SVM for predicting heart and liver diseases. The methodology includes an application of PSO for feature selection and categorization by SVM. This study revealed that the hybrid framework improves predictive accuracy through better feature selection and enhancements in model performance. The key highlights are that the algorithm can support multiple disease anticipations simultaneously. The limitations are in terms of the complexity of the hybrid model and extensive computational resource requirements. The research outcome shows that a hybrid framework provides robust and accurate anticipation for both heart and liver diseases.

This study (Modhugu & Ponnusamy, 2024) performs comparative analysis for liver disease anticipation using the SVM, Logistic Regression, and Decision Tree frameworks. The methodology encompasses the training and testing of these frameworks on a medical database and further comparing their performance based on accuracy, precision, and recall. The findings indicate that SVM always outperforms the other frameworks in predictive accuracy and also handles complex data patterns more effectively. Salient features of the study are a detailed performance comparison that establishes SVM as the most effective algorithm. Some of these limitations are model overfitting and databases that are more heterogeneous. The outcome of this study identifies the selection of SVM to predict liver disease; however, more research is required to establish these findings across a range of databases.

In the paper (Montazeri & Montazeri, 2014), the authors analyze different frameworks for ML in trying to improve the accuracy of diagnosing liver disease. This process also involves utilizing a set of frameworks- Decision Trees, Random Forests, and SVMs-on a database constituted by the clinical and biochemical data of patients. In this paper, all these models have been applied, and their performances are compared based on accuracy, precision, and recall. The best model performance is determined with feature selection and hyperparameter tuning. It became evident from the outcomes that Random Forests and SVMs provide the highest accuracy, thus proving their strength in handling complex and imbalanced databases, a common characteristic of sets related to liver disease diagnosis. Key points to be included in the discussion are how ensemble frameworks can be used to boost the reliability of anticipations and how the identification of critical features significantly affects diagnostic accuracy. The limitations of the study include a reliance on a single dataset that may not allow the generalization of findings across other populations. It further fails to explore model interpretability, an important fact for clinical adoption. This result of the research underlines the seriously improved identification of liver diseases using ML frameworks. Further exploration with diverse databases and model interpretability in focus might improve the clinical applicability and usefulness. Table 1 presents an analysis of each paper and addresses the proposed method in the paper, its advantages, and weaknesses.

Table 1
Analysis of the previous studies.

Ref	Method	Advantages	Weaknesses
(Kumar & Mahajan 2020)	SVM, Random Forest, Gradient Boosting on medical records	Ensemble methods, especially Random Forest, provide high accuracy and robustness. Feature selection and preprocessing improve performance.	Single-database reliance may affect generalization.
(Islam et al., 2024)	Logistic Regression, Decision Trees, Neural Networks on clinical and biochemical features	Neural Networks perform best, handling complex patterns in medical data effectively.	Overfitting risks and the need for large datasets for model validation.
(Dutta et al., 2022)	Feature engineering + SVM, Random Forest for predictive modeling	The hybrid framework enhances early recognition of liver disease with high precision.	Limited database size and diversity may affect general applicability.
(Topcu et al., 2024)	Decision Trees, Naive Bayes, and Ensemble Methods (Bagging) on clinical features	Ensemble methods improve accuracy and stability in disease anticipation.	Single database use and lack of extensive hyperparameter tuning.
(Dritsas & Trigka, 2023)	Advanced preprocessing + Ensemble ML (Random Forest, Gradient Boosting)	Improved preprocessing significantly enhances accuracy and precision. Efficient in handling missing data and outliers.	High computational cost and extensive parameter tuning requirements.
(Lu et al., 2024)	Hybrid model: Modified Particle Swarm Optimization (PSO) + SVM for feature selection and classification	The hybrid framework improves predictive accuracy and supports multiple disease anticipations.	Increased model complexity and high computational resource needs.
(Wu et al., 2019)	Comparative analysis: SVM, Logistic Regression, Decision Tree on medical database	SVM outperforms others in accuracy and effectively handles complex data patterns.	Risk of overfitting and database heterogeneity issues.
(Kalaiselvi et al., 2021)	Decision Trees, Random Forest, SVMs with feature selection and hyperparameter tuning	Random Forest and SVMs provide the highest accuracy, handling imbalanced datasets effectively.	Single database reliance affects generalization; lacks model interpretability for clinical adoption.

3. Methodology

The work is done on ML-based liver disease anticipation, focusing on the anticipation of stages of liver cirrhosis. In this work, a QDA algorithm is developed for categorization and compared with various other models regarding its performance. The QDA estimates per-class covariance matrices by cirrhosis stage, so adaptive decision boundaries that account for nonlinearity over biomarker combinations (i.e., Platelet Count vs. Bilirubin) are possible. QDA does not, unlike LDA, presuppose homogeneous covariance, but heterogeneous class distributions, as one normally has in medical data. Stage 4 cirrhosis, for example, has increased variance over Albumin values over the other stages, modeled by QDA via per-class covariance estimation. In addition, the efficacy of the models in this study is analyzed, drawing on key metrics: accuracy, precision, recall, and F1-score. The subsequent sections discuss in detail the research methodology concerning data preparation, model training, and performance assessment. Fig. 1 presents the workflow of the QDA-based liver cirrhosis prediction framework.

3.1. Data preparation and encoding

The foundation of this study is a publicly available clinical dataset.[1]

[1] https://www.kaggle.com/code/varshaakshir/liver-cirrhosis-prediction/input

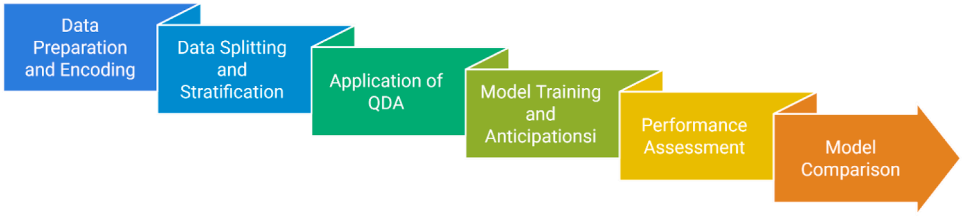


Fig. 1. Workflow of the QDA-based liver cirrhosis prediction framework.

This dataset includes approximately 312 records from patients, with each record consisting of 18 clinical parameters that include demographics, biochemical markers (e.g., Albumin, Bilirubin), and measurements indicative of liver health. As illustrated in Fig. 2, the dataset exhibits a significant class imbalance, with early stages (Stage 0 and 1) being severely underrepresented compared to the more advanced stages (Stage 2, 3, and 4). This imbalance is a critical challenge, as a naive classification model could achieve high accuracy by simply predicting the majority classes, thereby failing to identify patients in the crucial early stages of the disease.

To prepare the data for robust machine learning analysis and create a generalizable model, a systematic preprocessing pipeline was implemented. First, to handle missing values, k-NN imputation with $k = 5$ was employed. This method estimates a missing value based on the values of its most similar data points, effectively preserving the natural relationships within the data.

Second, to mitigate the adverse effects of class imbalance, a two-pronged strategy was adopted. The data was partitioned using a stratified train-test split to ensure the original proportion of each cirrhosis stage was maintained in both the training and testing sets. Concurrently, the training data was augmented using SMOTE to create a more balanced dataset for model training.

Finally, all features were transformed into a suitable numerical format. Categorical data (i.e., 'Drug', 'Ascites') were converted into a machine-readable format using one-hot encoding. Numeric data (i.e., 'Bilirubin') were scaled using a standard scaler to have a zero mean and unit variance, a crucial step for models sensitive to feature magnitudes.

Following preprocessing, feature engineering and selection were performed to enhance model performance. Key predictive features, including Bilirubin, Albumin, Platelet Count, and Prothrombin Time, were identified via RFE. To better capture the complex, non-linear patterns inherent in disease progression, polynomial terms (e.g., Bilirubin^2) were engineered from these key features. This step explicitly equips the models to learn the non-linear associations that are crucial for accurately distinguishing between the stages of cirrhosis.

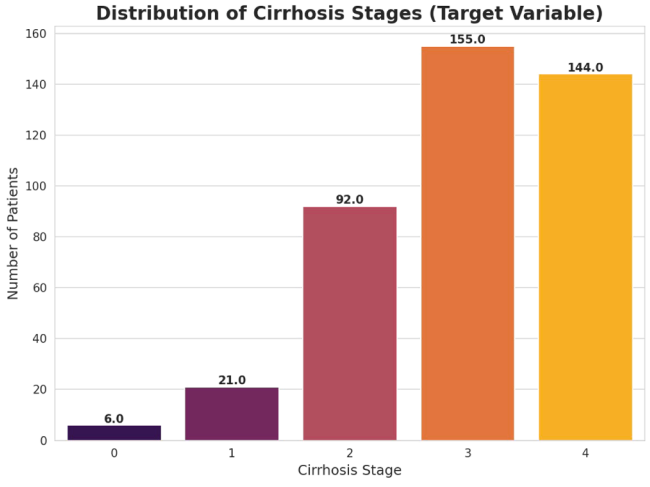


Fig. 2. Distribution of cirrhosis stages in the dataset, highlighting significant class imbalance.

3.2. Exploratory data analysis (EDA)

Following data preprocessing, a comprehensive EDA was conducted to uncover underlying patterns, feature relationships, and class-specific distributions within the cirrhosis dataset. This analysis is crucial for understanding the data's complexity and for informing the selection of an appropriate modeling strategy. Key insights were derived from a correlation matrix and distribution plots, as detailed below.

Fig. 3 provides a correlation matrix of the key numerical features, illustrating the strength and direction of linear relationships between clinical variables.

This correlation matrix offers a quantitative overview of how different biomarkers and clinical measurements relate to one another. The color intensity and shade directly represent the correlation coefficient, with bright yellow tones (e.g., the diagonal) indicating a strong positive correlation (close to $+1.0$) and dark purple/black tones indicating a strong negative correlation (close to -1.0). For instance, the bright block at the intersection of Bilirubin and Copper reveals a strong positive correlation, suggesting that as copper levels accumulate in the liver, bilirubin levels also tend to rise, a known clinical association in cholestatic liver diseases. Similarly, a strong positive correlation is visible between SGOT (a liver enzyme) and Bilirubin. Conversely, the dark block between Albumin and Bilirubin signifies a strong negative correlation; as liver function declines, the production of albumin decreases while bilirubin levels often increase. This detailed map of interdependencies is critical because it demonstrates that the features are not independent. Models that assume feature independence (like Naive Bayes) would be inappropriate, whereas more sophisticated models like QDA, which can inherently model covariance between features, are better suited for this dataset.

To investigate the data's structure from a classification perspective, Fig. 4 presents violin plots that visualize the distribution of four critical biomarkers across each cirrhosis stage.

These violin plots provide a rich and detailed view into how the statistical properties of key features change as liver cirrhosis progresses. Each "violin" is a density plot rotated on its side, showing the probability density of the data at different values; the wider sections represent a higher concentration of patients, while the white dot indicates the median. The plot for Bilirubin clearly shows that for early stages (0 and 1), the values are concentrated near the lower end. As the disease advances to stages 3 and 4, the median value increases significantly, and more importantly, the distribution becomes much wider and more skewed, indicating a higher variance and a greater number of patients with extremely high bilirubin levels. A similar but inverse trend is seen for Albumin, where levels are higher and more tightly distributed in early stages but decrease and spread out as the disease worsens. This visual evidence of heteroscedasticity (i.e., non-constant variance across classes) is the single most important justification for the methodological approach of this paper. It empirically demonstrates that a one-size-fits-all model assuming a common covariance structure for all classes (like Linear Discriminant Analysis) would be fundamentally flawed. Instead, it necessitates a model like QDA, which is specifically designed to handle this complexity by fitting a unique covariance matrix for each class,

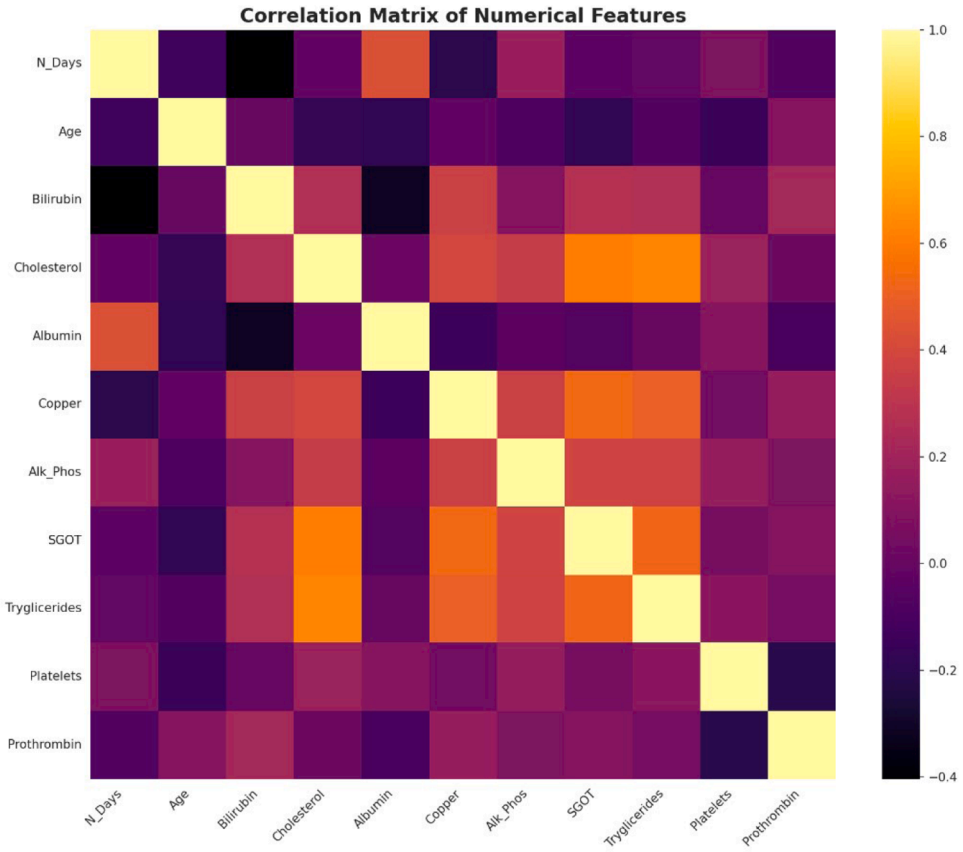


Fig. 3. Correlation matrix visualizing inter-feature relationships.

thereby allowing for more accurate, non-linear classification boundaries.

3.3. QDA

Based on the insights from the Exploratory Data Analysis, particularly the evidence of heterogeneous class distributions, QDA was selected as the primary classification model for this study. QDA is a probabilistic generative model that, unlike its linear counterpart (LDA), does not assume that the covariance matrix is identical for all classes. Instead, it models the probability distribution of each class individually.

In QDA, the posterior probability of a data point x belonging to a class k is calculated using Bayes' theorem, as shown in Eq. (1):

$$P(y=k|x) = \frac{P(x|y=k)\cdot P(y=k)}{P(x)} \tag{1}$$

Here, $P(y=k)$ is the prior probability of class k, and $P(x|y=k)$ is the class-conditional probability density function (the likelihood). QDA assumes that this likelihood follows a multivariate Gaussian (normal) distribution for each class:

$$P(x|y=k) \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Here, μ_k is the mean vector and $\sum_k$ is the covariance matrix for class k. During the training phase, the model estimates a separate mean vector and covariance matrix for each cirrhosis stage from the training data. The formula used to estimate the covariance matrix for each class is detailed in Eq. (2):

$$\Sigma_k = \frac{1}{n_k - 1}\sum_i (x_i - \mu_k)(x_i - \mu_k)^T \tag{2}$$

where n_k is the number of samples in class k.

The key advantage of this approach, defined by Eq. (2), lies in the per-class covariance matrix, $\sum_k$. This allows QDA to capture the unique spread and feature correlations within each cirrhosis stage—such as the increased variance in Bilirubin in later stages—and form quadratic decision boundaries. These curved boundaries are significantly more flexible than the rigid, linear boundaries of LDA, making QDA exceptionally well-suited for separating classes in complex medical datasets where non-linear relationships are common. While this flexibility comes at the cost of estimating more parameters, it is a necessary trade-off to accurately model the underlying data structure identified in the EDA.

3.4. Experimental setup

To rigorously evaluate the effectiveness of Quadratic Discriminant Analysis for liver cirrhosis prediction, a comprehensive comparative analysis was conducted. QDA's performance was benchmarked against a suite of other machine learning models, providing context and highlighting its specific strengths. The experimental design, model parameters, and evaluation metrics are detailed in the following subsections.

3.4.1. Model selection

The comparative analysis included a total of eight distinct machine learning models, chosen to represent a diverse range of algorithmic families. The proposed model, QDA, was selected for its theoretical suitability in handling the non-linear decision boundaries suggested by the data's heterogeneous class distributions. This model was compared against several powerful ensemble methods, including HistGradientBoosting, a high-performance gradient boosting algorithm, and Stacking, a meta-classifier that combines predictions from multiple base models to enhance overall accuracy. To specifically address the challenge of class imbalance, the analysis also incorporated specialized algorithms such as RUSBoost, Easy Ensemble, and Balanced Bagging,

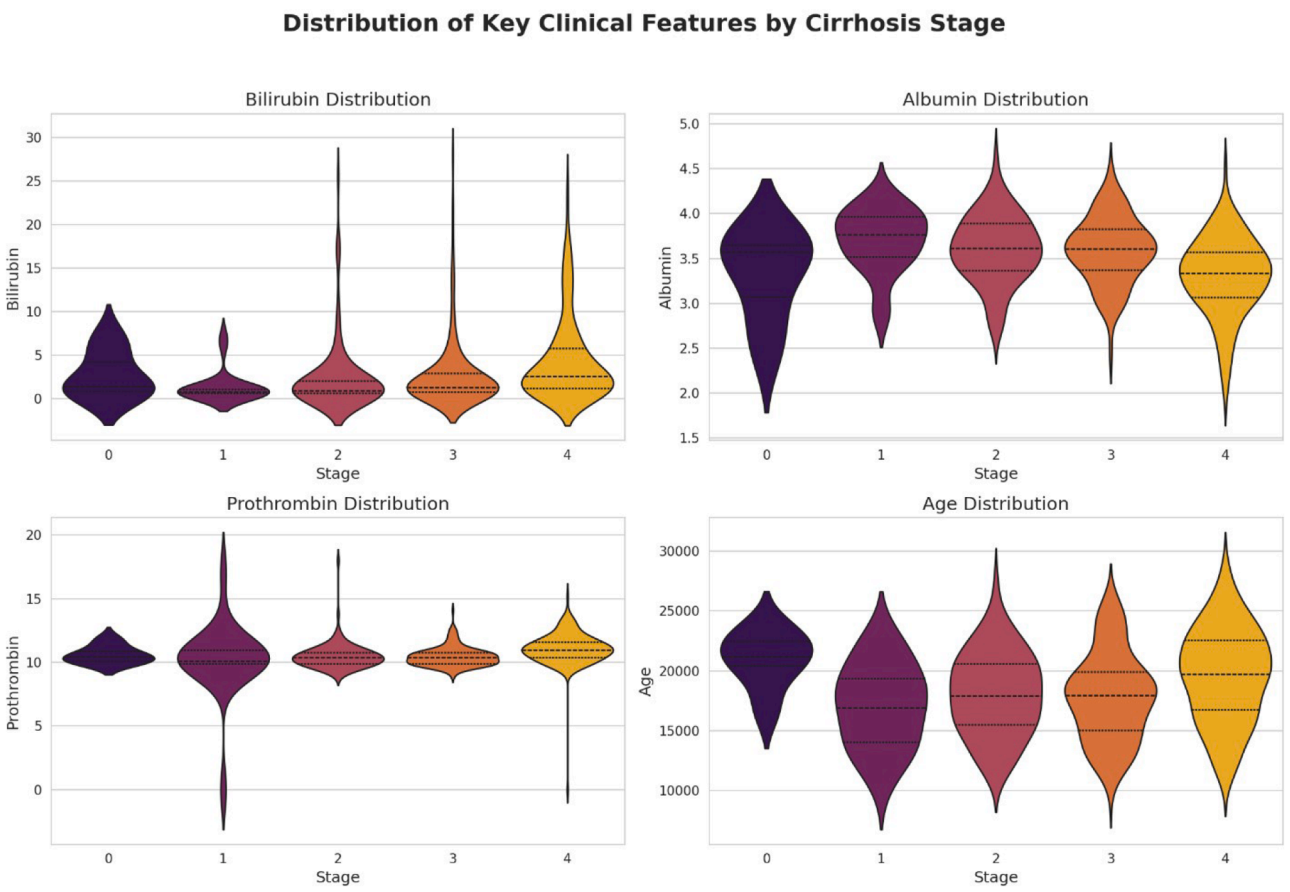


Fig. 4. Violin plots showing biomarker distributions across cirrhosis stages.

which integrate data sampling techniques directly into the ensemble learning process. Finally, a non-parametric probabilistic model, the Gaussian Process classifier, was included to assess a different approach to generating predictive probabilities. This carefully curated selection of models ensures a robust evaluation, pitting QDA against both standard high-performance classifiers and algorithms explicitly designed for imbalanced data.

3.4.2. Implementation and hyperparameters

All models were implemented within a Python environment, primarily using the scikit-learn and imbalanced-learn libraries. For the sake of reproducibility, the specific hyperparameter settings for the proposed QDA model are outlined here. The QDA algorithm's priors parameter was configured to match the class frequencies observed in the training data, a crucial step that allows the model to inherently account for the class imbalance during its probabilistic calculations. To prevent overfitting and enhance the model's ability to generalize to unseen data, a regularization parameter (reg_param) of 0.01 was applied. This technique adds a small constant to the diagonal of the covariance matrix, which stabilizes its inversion, a particularly important consideration for classes with a limited number of samples. Furthermore, the covariance matrices were estimated using Ledoit-Wolf shrinkage, a method that produces a more robust and well-conditioned estimate, thereby further improving the model's stability and reliability.

3.4.3. Performance evaluation metrics

The performance of each trained model was thoroughly assessed using a standard set of classification metrics to provide a multi-faceted and holistic view of its effectiveness. The overall proportion of correctly classified instances was measured by Accuracy. The model's ability to avoid making false positive predictions was quantified by Precision, also known as Positive Predictive Value. The capacity to identify all true positive cases, which is critical for minimizing missed diagnoses, was measured by Recall, which is also referred to as Sensitivity or the True Positive Rate. To balance the trade-off between Precision and Recall, the F1-Score, their harmonic mean, was calculated. Beyond single-threshold metrics, the model's overall ability to discriminate between classes across all possible thresholds was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Finally, the Area Under the Precision-Recall Curve (AUC-PR) was also computed, as it provides a particularly informative assessment of performance on imbalanced datasets. Together, these metrics offer a comprehensive evaluation of each model's diagnostic capabilities.

4. Experimental results and comparisons

This section presents the performance outcomes of the eight machine learning models evaluated on the liver cirrhosis prediction task. The analysis begins with an overview of the consolidated performance metrics and overall discriminative power, followed by a detailed per-class breakdown and an analysis of model errors.

4.1. Overall model performance

To provide a comprehensive a-glance summary of the models' capabilities, the key performance metrics—Accuracy, Precision, Recall, and F1-Score—were consolidated into a single performance heatmap, as shown in Fig. 5. For a deeper visual comparison of the models' overall discriminative power across all classification thresholds, the multi-class ROC and PR curves were also plotted.

The consolidated metrics in Fig. 5 reveal a complex performance

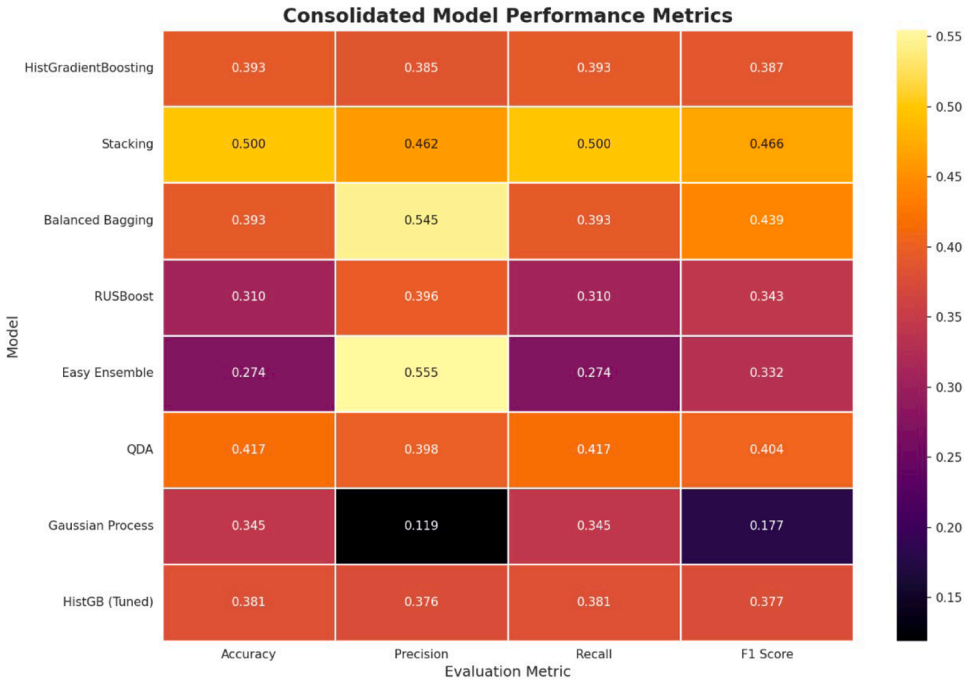


Fig. 5. Heatmap of consolidated performance metrics for all evaluated models.

landscape. The Stacking model achieved the highest accuracy (0.500) and a competitive F1-score (0.466), indicating a strong and balanced overall performance. In contrast, the Gaussian Process model exhibited high precision but suffered from extremely low recall (0.119), resulting in the lowest F1-score and suggesting it is a highly conservative model that misses the vast majority of positive cases. The proposed QDA model demonstrated a balanced profile, achieving a strong F1-score of 0.404, outperforming several specialized imbalance-learning algorithms like RUSBoost and Easy Ensemble. This initial view suggests that while Stacking is the top performer overall, QDA offers a robust and well-balanced alternative.

Fig. 6 presents the multi-class ROC curves, which plot the True Positive Rate against the False Positive Rate to measure a model's ability to distinguish between classes. The AUC serves as a single-number summary of this ability.

As shown in Fig. 4, the Stacking model again demonstrates superior performance, achieving the highest Micro-AUC score of 0.797, indicating it has the strongest capability to discriminate between the different stages of cirrhosis. The HistGradientBoosting model (Micro-AUC = 0.771) and the proposed QDA model (Micro-AUC = 0.755) also

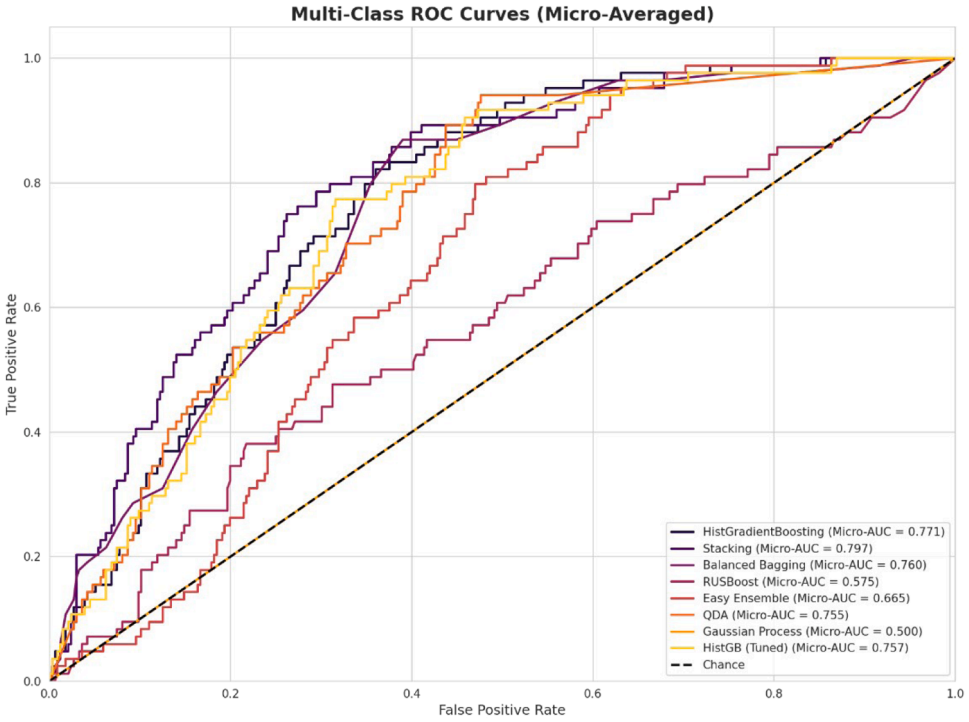


Fig. 6. Micro-averaged ROC curves comparing the overall classification performance.

show strong discriminative power, with their curves positioned well above the random chance line (dashed line, AUC = 0.5). In contrast, the Gaussian Process model's curve is very close to the chance line (Micro-AUC = 0.500), confirming its poor classification performance observed in Fig. 6. These ROC curves validate that the ensemble and QDA approaches are effective at the classification task.

Given the significant class imbalance in the dataset, the PR curves shown in Fig. 7 offer a more informative perspective on model performance, as they focus on the trade-off between identifying true positives and minimizing false positives.

The PR curves in Fig. 7 reinforce the findings from the ROC analysis. The Stacking model achieves the highest Micro-Averaged Precision (Micro-AP) of 0.450, demonstrating the best balance between precision and recall across all thresholds. The HistGradientBoosting and Balanced Bagging models also perform well, with Micro-AP scores over 0.400. The proposed QDA model (Micro-AP = 0.379) shows competent performance, again outperforming several other models. The curves clearly illustrate the difficulty of the task; even the best models operate far below a perfect score (AP = 1.0), which is typical for complex, imbalanced multi-class problems. The "No Skill" line represents the performance of a random classifier, and all serious models perform significantly better.

4.2. Per-Class performance analysis

While overall metrics provide a useful summary, a per-class analysis is essential for understanding how models handle the different stages of liver cirrhosis, particularly the underrepresented early stages. This subsection examines the precision, recall (sensitivity), and F1-score for each class individually.

Fig. 8 presents a bar chart comparing the per-class precision of each model. Precision measures the accuracy of positive predictions for a given stage, indicating how trustworthy a prediction is when the model identifies a patient as belonging to that stage.

The precision analysis in Fig. 8 highlights a significant challenge posed by the class imbalance. For the most underrepresented classes, Stage 0 and Stage 1, most models exhibit very low precision, struggling to make reliable positive predictions due to the scarcity of training examples. However, for the majority classes, performance is much stronger. For Stage 3, the Stacking and QDA models demonstrate high precision, suggesting that when they predict this stage, they are often correct. For Stage 4, the HistGradientBoosting and Stacking models lead in precision. This indicates that while identifying early-stage cirrhosis is difficult for all models, the top-performing ensemble methods can reliably classify the more advanced, well-represented stages of the disease.

Next, Fig. 9 compares the per-class recall of the models. Recall is arguably the most critical metric in a medical diagnostic context, as it measures the model's ability to successfully identify all true cases of a specific disease stage, thereby minimizing false negatives.

The recall results in Fig. 9 show that the models are highly effective at identifying Stage 0 cases, with several models achieving perfect recall. This is likely because the few examples of this class are highly distinct. For Stage 1, however, recall drops significantly, indicating difficulty in distinguishing this stage. The most crucial finding is for Stage 3, where the Stacking model achieves an exceptionally high recall of approximately 0.77, demonstrating a superior ability to find nearly all patients belonging to this class. The QDA model also performs competitively for the later stages. This high recall for the advanced stages is a desirable trait, as it ensures that patients with severe disease are not overlooked.

Finally, Fig. 10 displays the per-class F1-score, which provides a balanced measure by combining both precision and recall. A high F1-score indicates that a model has both high precision and high recall for a specific class.

The F1-score analysis in Fig. 10 synthesizes the previous findings. The Stacking model emerges as the most consistent top performer, achieving the highest F1-scores for the three most prevalent and clinically significant stages (Stage 2, 3, and 4). This confirms that its high recall for these stages does not come at the cost of extremely low precision, making it a well-balanced and robust classifier. The HistGradientBoosting and QDA models also show strong, balanced performance on the majority classes, reinforcing their suitability for this task. The low F1-scores for Stage 1 across all models underscore it as the most challenging class to predict accurately.

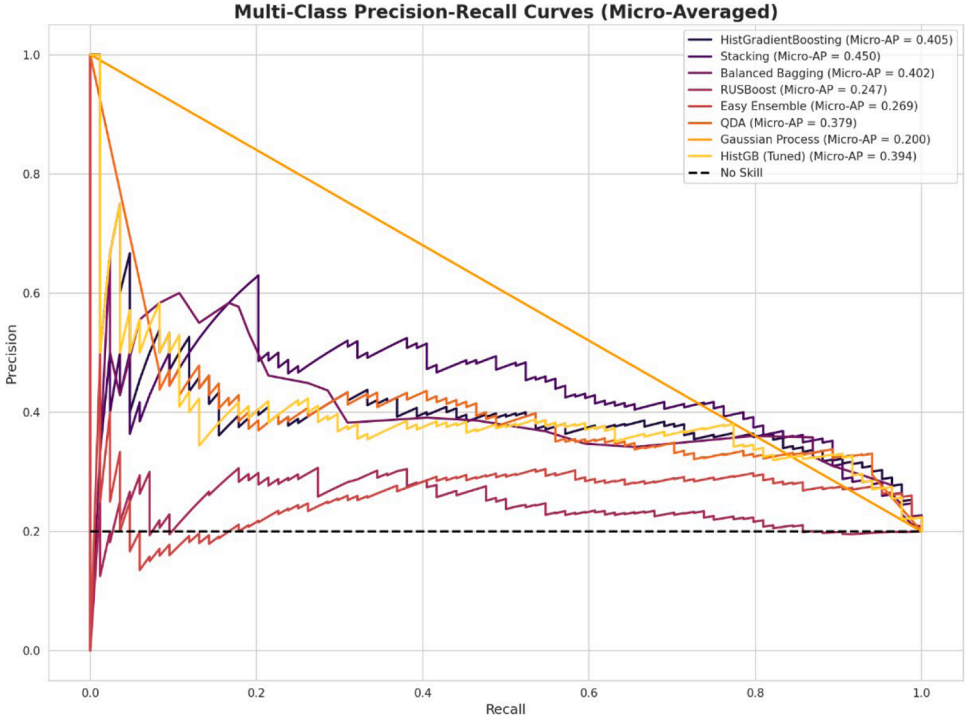


Fig. 7. Micro-averaged Precision-Recall curves, which are highly informative for imbalanced datasets.

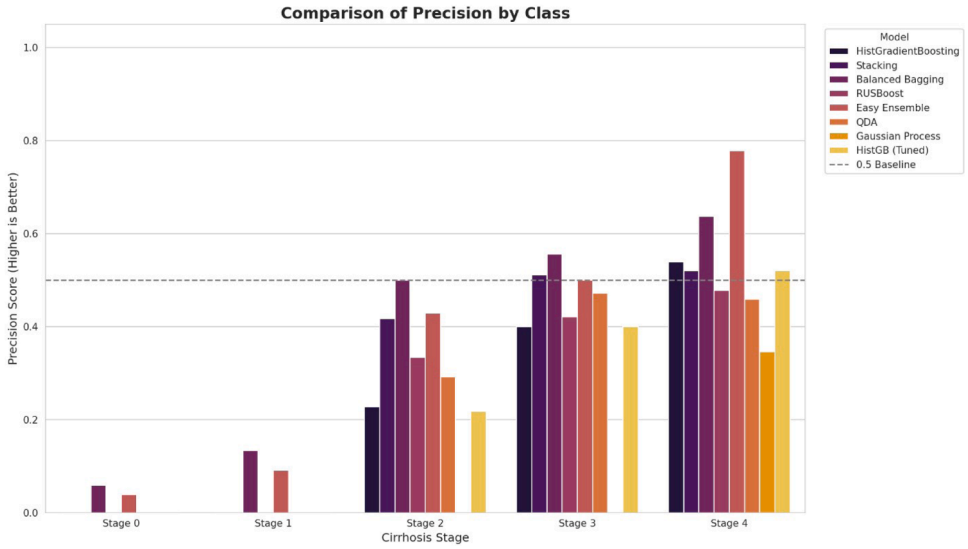


Fig. 8. Per-class precision scores for each model across cirrhosis stages.

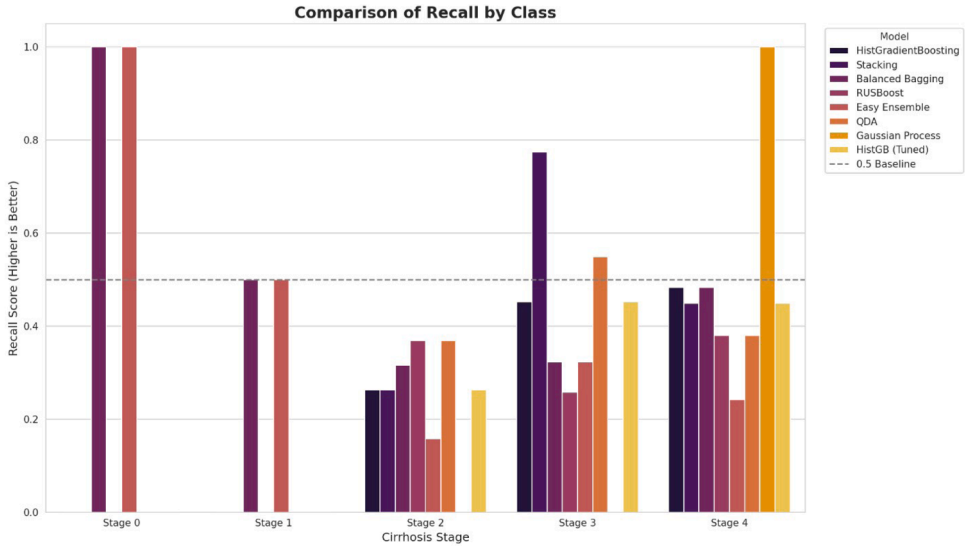


Fig. 9. Per-class recall scores for each model.

4.3. Error analysis and model efficiency

To gain a deeper understanding of the models' predictive behavior, an error analysis was performed to identify the specific types of mistakes made during classification. Additionally, the computational efficiency of each model was evaluated.

Fig. 11 presents the confusion matrix for the best-performing model, Stacking. The matrix provides a detailed breakdown of correct and incorrect predictions for each cirrhosis stage, with the diagonal elements representing correct classifications.

The confusion matrix reveals a critical pattern in the model's errors. The majority of misclassifications occur between adjacent stages. For example, a significant number of true Stage 4 patients are misclassified as Stage 3, and true Stage 2 patients are misclassified as Stage 3. This type of error, while not ideal, is less severe in a clinical context than mistaking an early-stage case for an advanced one, or vice versa. The matrix also confirms the difficulty in predicting the minority classes (Stage 0 and 1), which have very few true instances in the test set.

To compare the error patterns across all models, Fig. 12 visualizes the count of FP and FN for each class. A False Positive occurs when a model incorrectly predicts a stage (e.g., calling a Stage 2 patient "Stage 3″), while a False Negative represents a missed diagnosis for that stage (e.g., failing to identify a Stage 3 patient).

The error analysis in Fig. 12 shows that for the clinically crucial Stage 3 and Stage 4, the top-performing models like Stacking and HistGradientBoosting manage to keep the number of False Negatives relatively low, which is vital for ensuring patients with advanced disease are not overlooked. The proposed QDA model also demonstrates a competitive error profile in these later stages. The high number of False Positives for Stage 3 across many models corresponds with the confusion matrix finding, where patients from other stages are frequently misclassified into this central, majority class.

Finally, model efficiency was assessed by plotting the training time against the weighted F1-score, as shown in Fig. 13. This analysis is important for considering the practical deployment of models, where both predictive power and computational cost are relevant.

The efficiency plot in Fig. 13 illustrates a clear trade-off between performance and speed. The QDA model is by far the most computationally efficient, with a training time of less than a second while still delivering a strong F1-score. In contrast, the top-performing Stacking

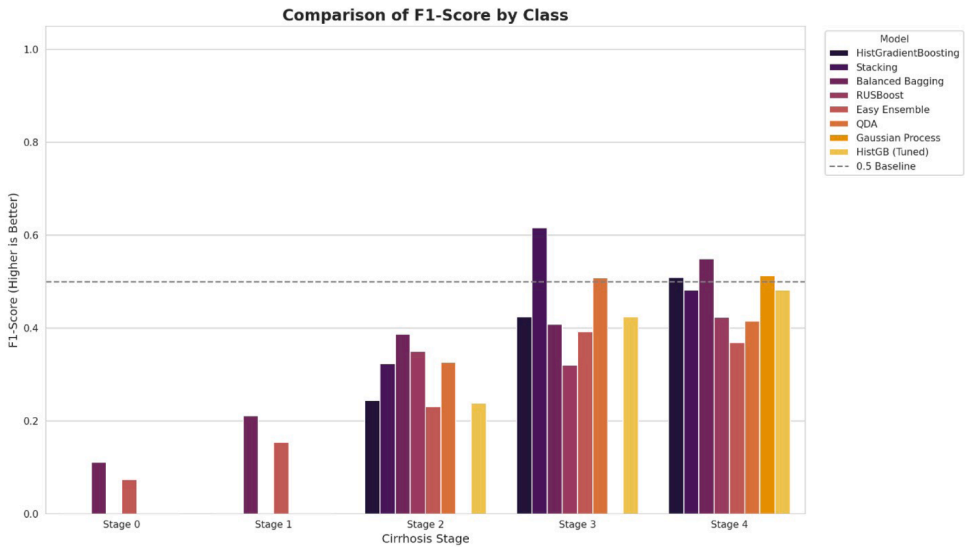


Fig. 10. Per-class F1-scores, balancing precision and recall for each stage.

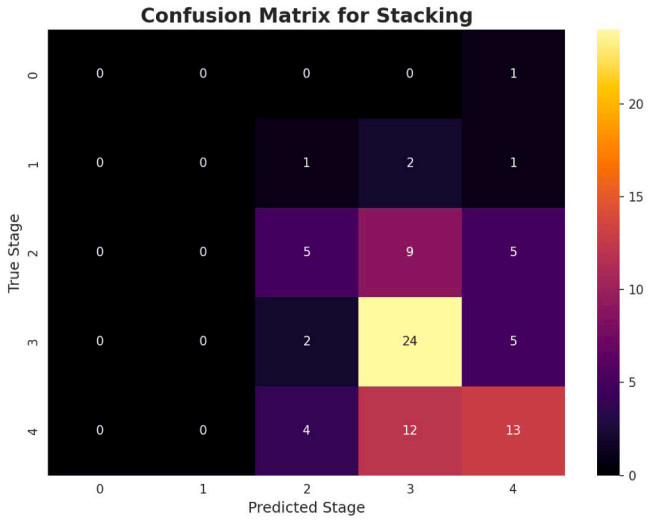


Fig. 11. Confusion matrix for the Stacking model, detailing per-class predictions.

model requires significantly more training time due to its nature as a meta-ensemble. This suggests that in scenarios where rapid model training or retraining is required, QDA presents a highly compelling alternative, offering a robust balance of predictive accuracy and computational efficiency.

5. Discussion

The experimental results of this study provide a multi-faceted view of the challenges and opportunities in using machine learning for liver cirrhosis prediction. The analysis highlights that while no single model excelled on every metric, a clear hierarchy of performance emerged, with the Stacking ensemble demonstrating the most robust overall predictive power and the proposed QDA model offering a highly efficient and well-balanced alternative.

The overall performance analysis established the Stacking model as the top performer, achieving the highest accuracy (0.500), Micro-AUC (0.797), and Micro-AP (0.450). This finding aligns with a growing body of research demonstrating the power of sophisticated computational models in medical diagnostics, where machine learning is increasingly used for the automatic classification of complex neurological conditions like autism and trigeminal neuralgia (R.A. Lashaki et al., 2025), (Khadem-Reza et al., 2025), Its success can be attributed to its ensemble nature, which combines the strengths of multiple base classifiers to create a more powerful and generalizable meta-model. This approach effectively mitigated the weaknesses of individual models, leading to superior discrimination between cirrhosis stages. In contrast, the Gaussian Process model's performance was no better than random chance, likely due to its kernel-based approach struggling with the high dimensionality and complex, non-linear structures within the clinical data. The proposed QDA model performed competitively, with an AUC of 0.755, positioning it firmly among the effective models and significantly outperforming several specialized imbalance-learning algorithms like RUSBoost.

The per-class analysis offered deeper insights into the practical utility of these models. This detailed examination of model performance on specific biomarkers is crucial, as the identification of reliable, predictive features is a central goal in computational medicine, mirroring efforts to find EEG microstate biomarkers for conditions like schizophrenia (Z. Raeisi et al., 2025), The results underscored the profound difficulty of predicting the earliest stages of cirrhosis (Stage 0 and 1) due to their extreme scarcityThe results underscored the profound difficulty of predicting the earliest stages of cirrhosis (Stage 0 and 1) due to their extreme scarcity in the dataset, leading to low precision across all models. However, for the more prevalent and clinically advanced stages, a clearer picture emerged. The Stacking model's high recall for Stage 3 (approx. 0.77) is a particularly significant finding, as it indicates a strong ability to correctly identify patients with advanced disease, which is paramount for ensuring timely clinical intervention. The F1-score analysis confirmed that Stacking's high recall did not come at the cost of unacceptably low precision, making it a well-balanced and reliable classifier for the most critical disease stages.

Furthermore, the error analysis revealed that the Stacking model's mistakes were often clinically plausible, with most misclassifications occurring between adjacent stages. This suggests the model is learning the continuous nature of disease progression. The efficiency analysis introduced a crucial practical consideration. While Stacking was the most accurate, it was also one of the most computationally expensive. QDA, on the other hand, stood out as exceptionally efficient, delivering a strong F1-score in a fraction of the training time. This makes QDA a highly attractive option for clinical settings where rapid model deployment, frequent retraining on new data, or resource constraints are important factors.

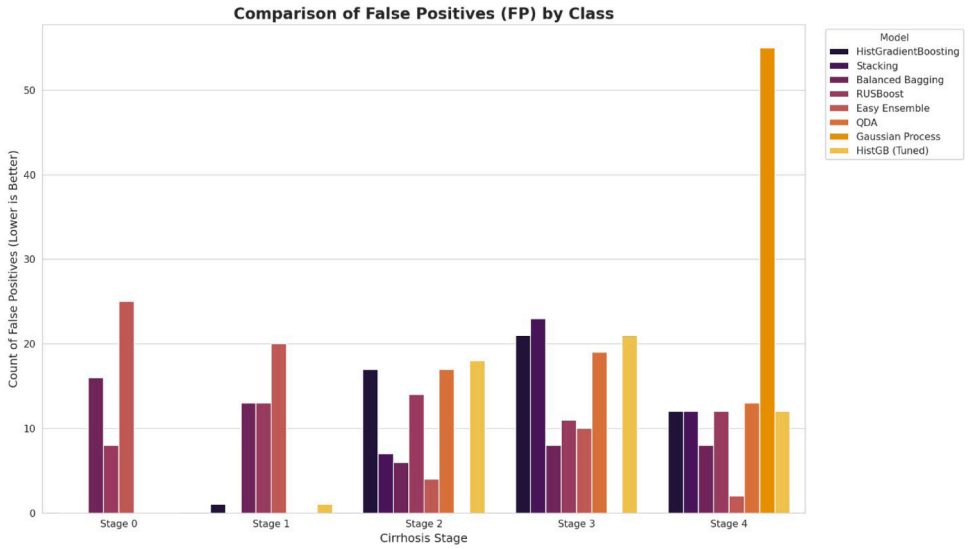


Fig. 12. Comparison of False Positives and False Negatives by class for all models.

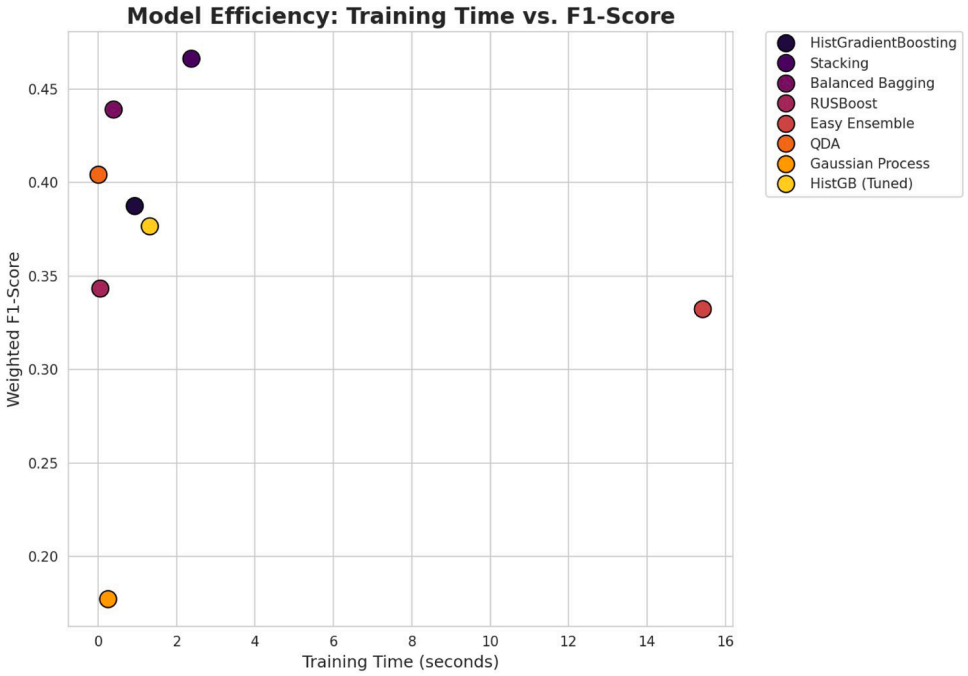


Fig. 13. Model efficiency, comparing training time against weighted F1-score.

Finally, this study has limitations that must be acknowledged. First, the analysis is based on a single, publicly available dataset. The findings' generalizability should be confirmed on larger, more diverse, and multi-center clinical cohorts. Second, while several imbalance-learning techniques were compared, extensive hyperparameter tuning for every model was beyond the scope of this work, and further optimization could potentially alter the performance landscape. Lastly, statistical significance tests were not performed to formally compare model performances. Future work could employ tests like the DeLong test for ROC curves or McNemar's test to statistically validate the observed differences. Addressing these limitations will be crucial for developing models that are not only accurate but also robust and reliable enough for real-world clinical adoption.

6. Conclusion

This study successfully developed and evaluated a machine learning framework for the prediction of liver cirrhosis stages, a task complicated by the asymptomatic onset of the disease and significant data imbalance. Through a comprehensive comparative analysis of eight different models, this work has established a clear performance hierarchy and highlighted the trade-offs between predictive accuracy, per-class reliability, and computational efficiency.

The primary finding is that a Stacking ensemble classifier delivered the most robust and accurate predictions, demonstrating the highest overall AUC and the best-balanced performance on the most clinically significant, advanced stages of cirrhosis. However, the proposed QDA model emerged as a highly compelling alternative. Its performance was competitive with top-tier models, and it proved to be exceptionally

computationally efficient, making it a practical choice for resource-constrained environments or applications requiring rapid model training. The study's EDA visually confirmed the theoretical rationale for QDA's effectiveness by revealing heterogeneous, non-linear data distributions across disease stages, which QDA is uniquely equipped to handle.

Ultimately, this research contributes to the field by providing a rigorous, side-by-side comparison of modern machine learning techniques on a challenging clinical problem. It demonstrates that while complex ensemble methods can achieve superior accuracy, simpler, well-chosen models like QDA can offer a powerful and efficient solution. Future work should focus on validating these findings on larger, multi-center datasets and exploring hybrid models to further enhance predictive performance for the early diagnosis of liver cirrhosis.

Authors' contributions

All authors contributed to the study's conception and design. Data collection, simulation and analysis were performed by" Genjuan Ma and Yan Li". Also, the first draft of the manuscript was written by Genjuan Ma. Yan Li commented on previous versions of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical approval

The research paper has received ethical approval from the institutional review board, ensuring the protection of participants' rights and compliance with the relevant ethical guidelines.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We would like to take this opportunity to acknowledge that there are no individuals or organizations that require acknowledgment for their contributions to this work.

Data availability

The authors do not have permission to share data.

References

Aghamohammadi, A. H., Ang, M. C., Ng, K. W., & Mogharrebi, M. (2014). TRIZ-based innovative solution for wheelchair assistive technology. *Visual informatics international seminar* (pp. 152–157).

Aghamohammadi, A., Ang, M. C., Sundararajan, E. A., Ng, K. W., Mogharrebi, M., & Banihashem, S. Y. (2018). Correction: A parallel spatiotemporal saliency and discriminative online learning method for visual target tracking in aerial videos. *PloS One, 13*(3), Article e0195418.

Ahn, J. C., Connell, A., Simonetto, D. A., Hughes, C., & Shah, V. H. (2021). Application of artificial intelligence for the diagnosis and treatment of liver diseases. *Hepatology (Baltimore, Md.), 73*(6), 2546–2563.

Anuradha, C., Swapna, D., Thati, B., Sree, V. N., & Praveen, S. P. (2022). Diagnosing for liver disease prediction in patients using combined machine learning models. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 889–896). IEEE.

Assegie, T. A., Subhashni, R., Kumar, N. K., Manivannan, J. P., Duraisamy, P., & Engidaye, M. F. (2022). Random forest and support vector machine based hybrid liver disease detection. *Bulletin of Electrical Engineering and Informatics, 11*(3), 1650–1656.

Behera, M. P., Sarangi, A., Mishra, D., & Sarangi, S. K. (2023). A hybrid machine learning algorithm for heart and liver disease prediction using modified particle swarm optimization with support vector machine. *Procedia Computer Science, 218*, 818–827.

Castaneda, C., et al. (2015). Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics, 5*, 1–16.

Choi, Y. S., & Oh, E. (2024). Investigating of machine learning based algorithms for liver cirrhosis prediction. *Advances in Engineering and Intelligence Systems, 003*(01), 115–130. https://doi.org/10.22034/aeis.2024.446087.1177

de Mattos, A. Z., et al. (2021). Current impact of viral hepatitis on liver cancer development: The challenge remains. *WORLD (Oakland, Calif. : 1993) Journal of Gastroenterology, 27*(24), 3556.

Dritsas, E., & Trigka, M. (2023). Supervised machine learning models for liver disease risk prediction. *Computers, 12*(1), 19.

Dutta, K., Chandra, S., & Gourisaria, M. K. (2022). Early-stage detection of liver disease through machine learning algorithms. *Advances in data and information sciences* (pp. 155–166). Springer.

Ghazal, T. M., Rehman, A. U., Saleem, M., Ahmad, M., Ahmad, S., & Mehmood, F. (2022). Intelligent model to predict early liver disease using machine learning technique. In *2022 International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1–5). IEEE.

Gupta, K., Jiwani, N., Afreen, N., & Divyarani, D. (2022). Liver disease prediction using machine learning classification techniques. In *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 221–226). IEEE.

Huang, T., Behary, J., & Zekry, A. (2020). Non-alcoholic fatty liver disease: A review of epidemiology, risk factors, diagnosis and management. *Internal Medicine Journal, 50*(9), 1038–1047.

Islam, R., Sultana, A., & Islam, M. R. (2024). A comprehensive review for chronic disease prediction using machine learning algorithms. *The Journal of Electrical Systems and Information Technology (Elmsford, N.Y.), 11*(1), 27.

Kalaiselvi, R., Meena, K., & Vanitha, V. (2021). Liver disease prediction using machine learning algorithms. In *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)* (pp. 1–6). IEEE.

Khadem-Reza, Z. K., Lashaki, R. A., Shahram, M. A., & Zare, H. (2025). Automatic diagnosis of autism spectrum disorders in children through resting-state functional magnetic resonance imaging with machine vision. *Quantitative Imaging in Medicine and Surgery, 15*(6), 4935.

Kumar, Y., & Mahajan, M. (2020). Recent advancement of machine learning and deep learning in the field of healthcare system. *Computational Intelligence for Machine Learning and Healthcare Informatics, 1*, 77.

Kumar, M. K., Sreedevi, M., & Reddy, Y. (2018). Survey on machine learning algorithms for liver disease diagnosis and prediction. *International Journal of Engineering and Technology (IJET (UAE)), 7*(18), 99–102.

Lashaki, R. A., Raeisi, Z., Razavi, N., Goodarzi, M., & Najafzadeh, H. (2025a). Optimized classification of dental implants using convolutional neural networks and pre-trained models with preprocessed data. *BMC Oral Health, 25*(1), 535.

Lashaki, R. A., Raeisi, Z., Sodagartojgi, A., Abedi Lomer, F., Aghdaei, E., & Najafzadeh, H. (2025b). EEG microstate analysis in trigeminal neuralgia: Identifying potential biomarkers for enhanced diagnostic accuracy. *Acta Neurologica Belgica*, 1–21.

Lu, F., et al. (2024). Artificial intelligence in liver diseases: Recent advances. *Advances in Therapy, 41*(3), 967–990.

Md, A. Q., Kulkarni, S., Joshua, C. J., Vaichole, T., Mohan, S., & Iwendi, C. (2023). Enhanced preprocessing approach using ensemble machine learning algorithms for detecting liver disease. *Biomedicines, 11*(2), 581.

Modhugu, V. R., & Ponnusamy, S. (2024). Comparative analysis of machine learning algorithms for liver disease prediction: SVM, logistic regression, and decision tree. *Asian The Journal of Research; A Journal of Science and Its Applications in Computer Science (New York, N.Y.), 17*(6), 188–201.

M. Montazeri and M. Montazeri, "Machine learning models for predicting the diagnosis of liver disease.," 2014.

Nowduri, S., & Subramanyam, M. M. (2024). Automated seizure detection through EEG analysis and deep learning technique. *The Journal of Artificial Intelligence and System Modelling, 02*(02), 52–61. https://doi.org/10.22034/jaism.2024.458242.1042

Z. Raeisi, R.A. Lashaki, F. Sharafkhani, A. Roshanzamir, and H. Najafzadeh, "Enhanced classification of tinnitus patients using EEG microstates and deep learning techniques," 2024.

Raeisi, Z., Mehrnia, M., Lashaki, R. Ahmadi, & Lomer, F. Abedi (2025a). Enhancing schizophrenia diagnosis through deep learning: A resting-state fMRI approach. *Neural Computing & Applications*, 1–33.

Raeisi, Z., Bashiri, O., EskandariNasab, M., Arshadi, M., Golkarieh, A., & Najafzadeh, H. (2025b). EEG microstate biomarkers for schizophrenia: A novel approach using deep neural networks. *Cognitive Neurodynamics, 19*(1), 1–26.

Shaheamlung, G., Kaur, H., & Kaur, M. (2020). A survey on machine learning techniques for the diagnosis of liver disease. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 337–341). IEEE.

Singal, A. K., & Mathurin, P. (2021). Diagnosis and treatment of alcohol-associated liver disease: A review. *JAMA, 326*(2), 165–176.

Sivasangari, A., Reddy, B. J. K., Kiran, A., & Ajitha, P. (2020). Diagnosis of liver disease using machine learning models. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 627–630). IEEE.

Topcu, A. E., Elbasi, E., & Alzoubi, Y. I. (2024). Machine learning-based analysis and prediction of liver cirrhosis. In *2024 47th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 191–194). IEEE.

Wu, C.-C., et al. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine, 170*, 23–29.