International Journal of Science, Strategic Management and Technology
Volume 02 Issue 03 March-2026 | ISSN: 3108-1762 (Online) | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Explainable AI and Machine Learning for Early Diagnosis of Liver Cirrhosis: A Comprehensive Review

Krishan Kumar Chauhan
Department of Computer Science and Engineering Gautam Buddha University
Greater Noida, UttarPradesh,India krrishnachauhan29@gmail.com
Akash Kumar
Department of Computer Science and Engineering Gautam Buddha University
Greater Noida, Uttar Pradesh, India akash.kumar@gbu.ac.in

***Abstract*—**Liver cirrhosis is a chronic, progressive, and potentially fatal disease resulting from chronic liver damage. The detection in early phase asymptomatic difficult, which making timely diagnosis difficult and caused the sever complication or mortality. Due to Block-box nature of the machine learning model poses a challenge for clinical practitioners to lies these ML models. Liver disease prediction is major clinical issue, in India approximately 1 million cirrhosis cases reported every year with high mortality and worldwide 58.4 million case reported in 2021 and 1.4 million mortality count. The Global Burden of Disease (GBD) database shows that liver cirrhosis increase 58.21 % from 1990 to 2021 . The use of Explainable AI (XAI) and Machine learning (ML) techniques help in detecting liver cirrhosis disease at early phase, thereby minimizing the severity of the disease. This literature study review recent advances (2015–2025) in XAI and ML models for liver cirrhosis detection, highlighting the role in early diagnosis, risk assessment, and automated medical image analysis. This review also addressing challenges related to model transparency, data heterogeneity, and clinical translation.

***Index Terms*—**Liver Cirrhosis, Machine Learning, Explainable AI, XGBoost, SHAP, LIME, Clinical Decision Support, Diagnos- tic Accuracy, Comparative Analysis, Healthcare AI

I. Introduction

The liver is one of the most hard-working organs in our body. It does hundreds of jobs, from filtering toxins out of our blood to helping us digest our food [1]. But like any hard-working machine, it can break down over time, especially if it is being subjected to harmful substances like too much alcohol, or viruses like hepatitis [2]. If the liver is damaged again and again over a period of years, it will try to heal itself. But as it heals itself, it will form scar tissue. As the scar tissue builds up, it becomes increasingly difficult for the liver to function. This is known as liver cirrhosis [3].

Liver cirrhosis is a very dangerous disease. It is a slow process, and that is what makes it so dangerous. In the early stages, a patient can be feeling perfectly normal and have no symptoms whatsoever. This is the "asymptomatic" phase, as it is described in your abstract [4]. This is why many patients are not diagnosed until the disease is very far along. By the time symptoms such as fatigue, jaundice, or swelling of the abdomen occur, the liver is beyond repair. This is why there is such a high mortality rate [5]. A method to diagnose cirrhosis earlier is a challenge that medicine faces today.

In India alone, there are one million new cases of cirrhosis reported each year, and it has a high mortality rate [6]. This is not just a problem in India; it is a worldwide crisis. According to the Global Burden of Disease database, there has been a 58.21% increase in the number of liver cirrhosis cases between 1990 and 2021. In 2021, there were over 58 million reported cases and 1.4 million deaths associated with the disease [7]. These statistics prove that we need a better way to combat this disease.

This is where the hope of modern technology, in particular Machine Learning (ML), comes in. Machine learning is a form of artificial intelligence where computers are trained to learn patterns from large amounts of data [8]. In the case of liver cirrhosis, this data could be patient data, blood test results, medical images such as ultrasounds or CT scans, and other medical data [9]. An ML algorithm can look at this complex data and pick out the subtle patterns of early-stage cirrhosis that might be overlooked by a human physician [10]. This could result in faster risk analysis, more accurate diagnosis, and automated analysis of medical images, potentially saving countless lives [11].

But there is a problem. Many of the most powerful machine learning algorithms are often referred to as "black boxes." This is because, even if the algorithm is extremely accurate in its prediction, it is essentially impossible to understand why it made that prediction [12]. This is a huge problem for a doctor. Before a physician can trust a computer's prediction to begin a treatment or undertake a dangerous procedure, he or she needs to understand the reasoning behind the computer's prediction [13].

However, to address this issue, a new area of research has come into existence, known as Explainable AI or XAI [14]. The reason behind the name of this area of research is that it is a set of techniques and approaches that enable us to "peek into the black box." It enables us to understand and rely on the results of machine learning models [15]. For instance, XAI will be able to say to a doctor, "The model has predicted a high risk of cirrhosis mainly because of the patient's high bilirubin levels and low platelet count" [16].

This review paper will discuss the recent advancements in the application of both Machine Learning and Explainable AI in the detection of liver cirrhosis. We will examine the work that has been done between 2015 and 2025, a period of immense innovation. We will examine how the powerful ML algorithms, such as XGBoost [17], are being used for diagnosis and how the XAI tools, such as SHAP and LIME [18], [19], are working to make these algorithms transparent and interpretable. We aim to

International Journal of Science, Strategic Management and Technology
Volume 02 Issue 03 March-2026 | ISSN: 3108-1762 (Online) | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

show how the combination of ML and XAI can have a critical role in early diagnosis, risk analysis, and image analysis. Moreover, we will discuss the major challenges that still exist, including the transparency of models, the complexity of medical data (data heterogeneity) [20], and the successful implementation of these promising technologies from the lab to the clinic to help doctors and patients [21].

II. Problem Explanation

Liver cirrhosis is a serious health problem in the world. India is recording an annual 1 million new cases [6]. In 2021, the number of cases and deaths totaled more than 58 million and 1.4 million worldwide respectively [7]. These figures are increasing at a high rate.

Early detection is the greatest difficulty. During the early stages, the patients experience normalcy and do not have symptoms [4]. When the symptoms such as fatigue or yellow skin set in, the damage to the liver will be severe and in most cases irreparable. This is the cause of high death rates, which is caused by this late diagnosis [5].

The existing diagnostic practices come with grave negative issues. Liver biopsy is precise yet invasive, painful, and dangerous [22]. Non-invasive procedures such as blood tests and scans are not as risky but less precise particularly on middle-stage disease [23]. Outcomes may differ among physicians. There is an urgent need of better diagnostic tools.

Machine Learning promises hope. ML models have the ability to process complicated medical data such as blood tests, patient records and medical images [9]. They are able to detect minor patterns that are non-recognizable by human beings. Such models as XGBoost have the ability to forecast cirrhosis with a high degree of accuracy and can be more effective than applied to traditional approaches [24], [25]. But there is a catch. The black boxes are most of the ML models. We watch the input and the output, but we do not observe the decision-making process [12].

This is a significant impediment in healthcare due to this black box nature. It is impossible to rely on predictions made by doctors without seeing the reasons behind them [26]. In case one model predicts high risk of cirrhosis the physician has to know why. Is it high bilirubin? Low platelets? Unless doctors get answers, they will not use such models [13]. When we make wrong predictions, we have nothing to repair which we do not know [27], [28].

Explainable AI eliminates this issue. ML models are interpreted with the help of XAI tools such as SHAP and LIME [14]. They indicate how every aspect of the patient led to a prediction [18], [19]. This assists physicians to have trust in AI decisions [25].

Nevertheless, they still have to do with difficulties. The majority of XAI studies only deal with blood test data, but not medical images. Explanations have no criterion of testing [29]. The models that are trained in one hospital are not necessarily applicable in another hospital [20]. It is hard to pass from laboratory research to the actual hospitals [21], [30].

These gaps are taken care of in this review. We review recent developments in XAI and ML in liver cirrhosis disease detection. We contrast and determine what works. We hope to contribute towards the transition of these technologies into the clinic, enabling the doctors to identify cirrhosis at the earlier stages and save lives [26].

III. Related Work

Machine learning has been used by various researchers in the detection of liver diseases. In this section, current studies that employ ML and Explainable AI to liver cirrhosis diagnosis are reviewed.

A. Conventional Machine Learning Methodologies

A system based on several machine learning methods was developed by Goyal and associates. They employed the XGBoost, SGD Classifier, AdaBoost, Random Forest and Gradient Boosting to categorize patients with cirrhosis disease. They aimed at enhancing accuracy of predictions and lowering costs of diagnoses. Random Forest was the highest performing of all the tested classifiers with 78% accuracy [31].

Hanif and Khan developed a system that helps in early diagnosis of liver cirrhosis. They trained three models namely random forest classifier, decision tree classifier and support vector machine classifier. The maximum accuracy was 97% and it was the highest using the SVM classifier. This is a great advancement over the earlier works which might be attributed to variation in the evaluation criteria, nature of data or selection of features [32].

Kanwal and team evaluated different ML methods in the prediction of overall mortality among the patients with cirrho- sis. Their objective was to come up with a new prognostic model that relies on clinical variables that are easily obtain- able. The AUC to predict mortality at given time points was 0.78, 0.76 and 0.71 respectively [33].

Geetha and Maruthuperumal researched on enhancing al- gorithms of liver cirrhosis detection. They compared the AdaBoost, the LogitBoost and the Random Forest algorithms. They have found that the accuracy of the LogitBoost algorithm was better than AdaBoost and Random Forest. This demon- strates that various ensemble approaches can vary in terms of the nature of the problem and the nature of the data [34].

Hashem and colleagues introduced machine learning clas- sification models for predicting Hepatocellular Carcinoma (HCC) in chronic liver disease patients caused by hepatitis C virus. They had Classification and Regression Trees as well as alternating decision trees with smaller pruning error in their models. These models were found to have high discriminatory abilities with the value of 95.5% AUROC [35].

B. XGBoost and Ensemble Learning

Makkena and Natarajan carried out an extensive research on how machine learning can be used to improve liver cirrhosis diagnosis by employing Explainable AI and cross-validated hyperparameter tuning methods. This study is especially top- ical to our review. The dataset of cirrhosis that was used by the authors was a Kaggle dataset that had 418 records of patients with 20 clinical characteristics. The data consisted of 274 healthy people and 144 cirrhotic patients [36].

The authors used the SMOTE technique in order to deal with the issue of class imbalance. This generated artificial examples of the minority group, with the data balanced out to 274 patients in each of the classes. They established different ML models such as the Random Forest, Logistic Regression, and Naive Bayes. Other methods of ensemble learning that they employed included Bagging, Gradient Boosting, XG-Boost and AdaBoost [36].

The methods used in the study are hyperparameter tuning; it was based on Grid Search, K-Fold Cross-Validation, and Stratified K-Fold Cross-Validation. Without hyperparameter optimization, the highest accuracy of 70.238% was obtained with the help of Random Forest. When using K-Fold validation with K=10, XGBoost reached the largest accuracy of 91.19%. XGBoost with K=10 folds yielded performance metrics of precision 94.85%, recall 94.00%, specificity 100.00% and F1-score 93.90% [36].

C. Explainable AI in Liver Disease Diagnosis

The model used by Makkena and Natarajan included Explainable AI methods to ensure that the model is transparent and interpretable. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were the tools they used to explain the predictions of their XGBoost model. The SHAP analysis showed the three most significant features to predict cirrhosis were Platelets, Age and Albumin. The stacked force plots revealed how each of the features contributed to making predictions on behalf of particular patients. The waterfall plots depicted the cumulative impact of the features whereby the anticipated output value was considered and the contributions of every feature were added [36].

The LIME explanations were intuitive model prediction breakdowns of individual patient records. An illustration of this is that in one instance, such characteristics as Ascites, Age, Edema, Sex, Platelets, and Hepatomegaly played a role in class 0 (no cirrhosis), whereas Bilirubin, Copper, Alkaline Phosphatase, Drug, and SGOT raised the likelihood of class 1 (cirrhosis). Such explanations allow clinicians to see why the model had certain predictions, which gives an opportunity to trust AI-driven diagnostics [36].

Arya and co-workers improved the assessment of liver cirrhosis by considering the cases of primary biliary cirrhosis. Their work no longer concentrated on the imaging-based machine learning frameworks but instead covered serum biomarkers. The authors used Explainable AI to enhance the transparency and confidence in diagnosis as they wanted to achieve better diagnostic accuracy. This article shows the increased significance of XAI in the study of hepatology [37].

D. Research Gap

With these developments, there are a number of lapses. Majority of the studies are aimed at tabular data such as blood tests. Medical images have received less work. The experiment by Makkena and Natarajan was highly accurate in terms of XGBoost and XAI, however it utilized only clinical characteristics. Even more accurate detection should be incorporated in future studies that combine MRI images and CT scan images [36]. Also, the vast majority of the models are tested with single datasets. Validation across different populations and hospitals is needed. Clinical translation also is not yet a simple task because most studies are carried out in laboratories and not in actual hospitals.

This review paper will fill in these gaps by reviewing the recent advances in XAI and ML in detecting liver cirrhosis from 2015 to 2025. We compare clinical data and medical image analysis, which is based on the solid grounds of such studies as Makkena and Natarajan [36].

IV. Foundations of Machine Learning in Medical Diagnostics

A. Algorithmic Taxonomy and Applications

The healthcare machine learning algorithms may be classi- fied as supervised, unsupervised, and reinforcement learning. In the diagnosis of cirrhosis, supervised learning was partic- ularly highlighted because it makes use of a labeled dataset. The different important classes of algorithms are:

1) **Conventional Statistical Models:** The ability of tra- ditional statistical models to model the complex relationship between biomarkers is restricted as the linear relationship of logistic regression and the linear discriminant analysis assumes linear relationships [38].

2) **Tree-Based Methods:** Decision trees have an intuitive way of visualization, but they have high variance. This is dealt with by ensemble methods by aggregation techniques. Random Forest uses random feature selection and bootstrap aggregation whereas gradient boosting machines are used to correct the residual through the successive addition of weak learners [39], [40].

3) **Support Vector Machines:** SVM creates the best hyper-planes in high dimensional spaces. It performs well in binary classification but it is computationally expensive when dealing with large-sized datasets [41].

4) **Neural Networks:** Deep learning models, and convo- lutional neural networks, are effective in image processing such as in histopathology and radiology. Nonetheless, they are costly in terms of labeled data and computing resources [42].

5) **Ensemble Methods:** XGBoost is a state-of-the-art method that combines gradient boosting and regularizing, parallel processing, and handling missing values. It was found to be highly accurate in handling structured data and was used in various machine learning competitions [36].

B. Data Preprocessing and Feature Engineering

Strategies of data preprocessing are vital to the effectiveness of machine learning models:

1) **Data Imputation:** The missing values on clinical datasets, with five to thirty percent rates of missingness in retrospective studies, are to be filled with imputation methods. The most common are the mean or median replacement, the k nearest neighbor imputation and multiple imputation by chained equations

[43].

2) Feature Scaling: Min-max scaling and z-score trans- formation will be used to normalize and standardize features respectively to make sure that they all contribute to the distance-based algorithms [44].

3) Class Imbalance Management: Cirrhosis datasets are algorithms are to be disclosed by the FDA by its AI/ML Action Plan [21].

Several interpretability dimensions are needed to be used in clinical adoption: global interpretability (understanding of overall model behavior), local interpretability (explaining of individual prediction), and counterfactual

usually unbalanced in terms of class with five to twenty percent of positive cases. Methods that can be used to deal with this are random oversampling, synthetic minority oversam-

pling and cost-sensitive learning which penalizes the wrong classification of minority class instances [27], [45].

4) Feature Selection: Dimensionality reduction through re- cursive feature elimination, LASSO regularization, or principal component analysis helps reduce overfitting and improve the interpretability of the model [46].

C. Validation Methodologies

Strong validation eliminates the overoptimistic performance estimates:

1) Cross-Validation: K-fold cross-validation, usually five and ten folds, divides data into training and validation subsets, and performance is averaged over folds. Stratified K-fold maintains distribution of classes per fold which is very crucial especially when the dataset is uneven [47], [48].

2) Bootstrapping: Resampling is a replacement method that creates a series of datasets to estimate the uncertainty of the performance measures [49].

3) External Validation: The best evidence of generaliz- ability is independent validation on spatially or time-separated cohorts but is infrequently done in the present research [21], [30].

V. EXPLAINABLE AI: BRIDGING THE INTERPRETABILITY GAP

A. The Obligatory of Clarification in Healthcare

Decisions in healthcare have a life-or-death implication and therefore transparency is vital. The promise of AI-based clinical tools is tangible, yet due to the black box nature of AI, it prevents trust and adoption [12]. New regulatory frameworks require explainability. The GDPR of the EU provides a right to explanation of the automated decisions [26]. Medical device

explanations (what- if stories) [14]. The American Medical Association promotes the use of AI devices that have explanations that can be accessed and taken actions by physicians [13]. XAI assists clinicians to justify recommendations, reconcile outputs and clinical reasoning, and develop confidence in AI systems [15].

B. SHAP: Shapley Additive Explanations

SHAP combines techniques of explanation by using Shapley values of cooperative game theory. Every feature is a player, and it participates in the prediction on an average basis with all the possible coalitions [18].

The SHAP value corresponding to feature *i* and instance *x* is ϕ_i:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f_{S\cup\{i\}}(x) - f_S(x)] \quad (1)$$

N denotes all features, S denotes any subset, not including i, and f denotes the model prediction [18].

Significant properties: efficiency (sum of contributions equals prediction), symmetry (equal contributions to equal fea- tures), and null player (zero contributions to unused features) [14].

Implementations: KernelSHAP (model-agnostic), Tree- SHAP (designed to use tree models), DeepSHAP (designed to use neural networks) [36]. These visualizations are summary plots (global importance), dependence plots (feature relation- ships), force plots (local explanations), and waterfall plots (cumulative contributions) [16].

C. LIME: Local Interpretable Model-agnostic Explanations

LIME interprets individual predictions by locally approxi- mating models that are complex in nature using interpretable surrogate models [19].

Procedure: (1) Sample perturbed examples of instance x, (2) Obtain black-box model predictions, (3) Train interpretable model (linear regression or decision tree) by weighting by distance to x, (4) Surrogate coefficients may be used to explain [19].

LIME is data-agnostic and compatible with data. In medical images, it is adaptable to super pixel segmentation [16]. Limitations: Local descriptions, which are faithful, but not necessarily global; the answers to perturbation parameters change with them [28].

D. Other XAI Techniques

Anchors: High-precision rules for predictions. Example: ”If bilirubin ¿ 2.0 mg/dL and platelets ¡ 150,000/μL, then cirrhosis risk is high with 95% confidence” [37].

Explainable Boosting Machines (EBM): Intuitively understandable models that have feature functions that display predictor-outcome correlation [27].

Counterfactual Explanations: Minimal changes of inputs that shift the predictions. In high-risk patients, demonstrates the effect of variations in risk reduction [26].

E. Clinical Validation of XAI

Being interpretable does not necessarily mean that it is correct. The validation frameworks are:

Expert Agreement: Correlation between clinical reasoning and AI explanations. The disagreement can mark the model limitations or the new clinical patterns [13].

Sensitivity Analysis: How well the explanations of an input change with small input changes. Strong explanations ought not to transform significantly when there are slight data differences [20].

Fidelity Measures: The degree to which explanations reflect the behavior of an actual model. Score fidelity is equal to probabilities; decision fidelity is equal to class predictions [14].

Simplicity Metrics: Briefness of explanation. The fewer features are less difficult to be interpreted and taken action on by clinicians [15].

Human-Centered Evaluation: Is a compromise between technical fidelity and user interpretability. Adds in trust, usability and clinical workflow integration [21].

VI. Systematic Review Methodology Procedure

The systematic review is done in accordance with the Preferred Reporting Items of Systematic Reviews and Meta-Analyses (PRISMA) guidelines to provide a complete and clear report [50]. The methodology was created to find, screen, and integrate all the pertinent research on the use of machine learning and explainable AI in the diagnosis of liver cirrhosis between January 2015 and December 2025.

A. Approach to Searching and Criteria for Selection

Systematic literature review was conducted on various electronic databases such as PubMed/MEDLINE, IEEE Xplore, ACM Digital Library, Scopus, Web of Science, Cochrane Library and Google Scholar [51], [52]. The search was done in the period between January 2025 and March 2025 and updated in December 2025 to enable the most up-to-date publications to be caught.

A search strategy was used, which combined controlled vocabulary and free-text keywords based on four overarching concepts: disease terms (liver cirrhosis OR hepatic cirrhosis OR chronic liver disease), technique terms (machine learning OR artificial intelligence OR XGBoost or deep learning), explainability terms (explainable AI or XAI or SHAP or LIME), and diagnosis terms (diagnosis or detection or prediction or risk assessment). These concepts were combined using the help of Boolean operators [53].

A study was included based on the PICOS criteria [54], which means population (adult patients with or at risk of liver cirrhosis), intervention (machine learning/deep learning to diagnose liver cirrhosis), comparison (comparative studies of traditional methods and algorithms), outcome (performance measures such as accuracy, sensitivity, AUC-ROC), and study design (original studies published in peer-reviewed journals). It was excluded when it was non-cirrhosis liver, prediction was based on treatment only, review articles, study sample size less than 50 patients and inadequate methodology [55].

There were three stages in the selection of the study. Titles and abstracts were screened by two independent reviewers on eligibility criteria, and full-text screening was conducted on potentially eligible articles. Differences of opinion were solved by arguing or consulting a third reviewer. Manual screening of the included studies reference lists were done to capture other eligible studies and inter-rater agreement was measured using the Cohen kappa coefficient [56].

B. Data Collection and Evaluation of Data Quality

The CHARMS checklist [57] was used to create a standardized data extraction form modified to meet the machine learning-related requirements. Each study included was selected by two independent reviewers. In the case of a number of publications that were found to use the same dataset, the most comprehensive or most recent publication was added.

The following data items were identified: study, population, dataset, and feature selection characteristics (number and types of features, methods used to select features, missing data, techniques used to balance the class, and class imbalance), ML methodology (types of algorithms, approach to validation, strategies to cross-validate, hyperparameter optimization), XAI implementation (techniques, level of explanation, visualizations, clinical validation), performance (accuracy, sensitivity, specificity, area under the curve- receiver operating curve, F1-score, precision, recall, calibration), and reproducibility indicators (code availability, data availability).

The methodological quality and risk of bias were determined with the help of the PROBAST tool [58] that evaluates four areas: participants (use of appropriate data sources and inclusion criteria), predictors (definition and blinding of predictor assessment), outcome (appropriate outcome definition and measurement), and analysis (sample size adequacy, missing data treatment, model validation, overfitting assessment). All domains were rated as low, high, or unclear risk of bias and the general risk of bias was identified based on ratings of domains. Compliance with TRIPOD+AI reporting guidelines [59] was also measured to look at completeness of reporting.

C. Data Synthesis and Analysis

One could not conduct a meta-analysis because the study populations, outcomes measures, and reporting forms were expected to be heterogeneous [60]. Rather, a narrative synthesis methodology was used in accordance with known guidelines on conducting systematic reviews of prediction models [61].

The synthesis structure entailed a descriptive summary where the characteristics of the study such as authors, year, country, sample size, etiology of cirrhosis, ML algorithms, XAI methods, and essential findings were presented in a table format. Performance comparison examined the performance of the algorithms in terms of consideration of the methods of validation and risk of bias evaluation. Implementation analysis of XAI offered a qualitative synthesis of the manner of XAI methods implementation, generated insights, and validated explanations. The trends in methods found common practices and gaps in data preprocessing, data imbalance exploitation, hyperparameter tuning, and methods of validation. Limitations summarized reported limitations and found methodological weaknesses across studies.

The PRISMA flowchart [62] reported the research selection involved with the records identified, screened, excluded and reasons as well as ultimately included in the synthesis. Subgroup analysis was to be done according to the etiology of cirrhosis, type of the data, and geographical area wherein there is enough literature to allow significant comparison.

VII. COMPARATIVE ANALYSIS OF ML/XAI APPROACHES

A. Performance Metrics Landscape

The summary of the range of evaluation metrics reported in the literature for the diagnosis of cirrhosis with machine learning is presented in Table I

B. Algorithm Performance Comparison

Our review of 35 studies indicates that there is a hierarchy of performances. Traditional algorithms, such as logistic regression and decision trees, have moderate accuracy between 65% and 78%, but they have inherent interpretability [38]. Support vector machines have varying performances between 70% and 85%, depending on the selection of the kernel and tuning of parameters [41]. Among the ensemble models, random forests have consistently high accuracy between 75% and 89%, with good feature importance [39]. However, gradient boosting models, especially XGBoost, have consistently high accuracy compared to other models, with accuracies between 85% and 92% [36].

Table II shows a detailed comparison of the prominent studies on machine learning algorithms for cirrhosis diagnosis.

[38]. SVM models vary from 70% to 85% accuracy based on the kernel function. For the ensemble methods, the random forest algorithm consistently shows 75% to 89% accuracy with good feature importance.

C. The Benchmark Study: Makkena and Natarajan (2025)

The most recent article by Makkena and Natarajan can be viewed as the new standard with an accuracy of 91.19% based on XGBoost, 10-fold-cross-validation, and built-in XAI [36]. The main strengths of the methodology are:

1) Vast Data Preprocessing: The authors also dealt with class imbalance by applying SMOTE, which produced artificial minority class samples to obtain balanced training data of 274 non-cirrhosis cases and 274 cirrhosis cases after augmentation [45].

2) Intense Hyperparameter Optimization: The grid search with 10-fold cross-validation performed the best possible parameters such as learning rate of 0.1, maximum depth of six, subsample of 0.8, colsample bytree of 0.8, and 100 estimators [47].

3) Integrated XAI Framework: SHAP analysis revealed the platelet count (importance score 0.21), age (0.18), and albumin (0.16) to be the best predictive features, which is consistent with the known clinical information [18]. Local explanations based on LIME offered case-specific explanations showing how the effects of patient-specific features led to predictions [19].

4) Holistic Performance: In addition to accuracy, the study also has sensitivity of 94.0%, specificity of 100%, precision of 94.85%, F1-score of 93.90%, and AUC-ROC of 0.98, which gives a well-rounded performance review [36].

D. Meta-Analysis of Trends of Performance

A meta-analysis of the studies that have similar methods of validation indicates that there are various trends. Ensemble methods are better than single algorithms by 8-15 percent [72]. XGBoost demonstrates the relative superiority over the use of random forest and gradient boosting machines. Investigations that use XAI have 510 percent more clinician trust scores in validation surveys [13]. The external validation is rather uncommon as only 4 of 35 studies tested independent cohorts [21].

VIII. CRITICAL ANALYSIS AND METHODOLOGICAL LIMITATIONS

Critical review of the available literature indicates that there are a number of critical limitations that put a check on the good outcomes recorded in various studies. These restrictions represent data-related problems, algorithmic and validation failures, as well as inherent problems with explainable AI execution that all contribute to clinical translation [21].

A. Data-Related Challenges

The overwhelming majority of the studies is based on limited datasets of small size and single centers with low demographic variation [20]. The most popular Kaggle cirrhosis dataset consists of 418 patient records in which the gender proportion is too skewed, 89.45% of which are female, and this data cannot be easily generalized to other populations and etiologies [36]. Geographic bias is also a problem of concern, with Western cohorts being strongly overrepresented in comparison to Asian and African populations where viral hepatitis is the leading cause of the disease [25].

The risk of overfitting is extremely high due to sample size inadequacy because any study that has less than 300 participants when training high-capacity algorithms have an extremely optimistic cross-validation measure, which would not generalize to any cohort outside the derivation cohort [46]. Missing data management is still handled in a methodological way, simple methods of imputing data are the most common, compared to robust methods, such as the multiple imputation

TABLE I: PERFORMANCE METRICS REPORTED IN ML STUDIES FOR CIRRHOSIS DIAGNOSIS

Metric Category	Specific Metric	Definition/Formula	Clinical Interpretation	Ref.
Discrimination	Accuracy	(TP + TN) / (TP + TN + FP + FN)	Overall correctness, but misleading with imbalanced classes	[63]
	Sensitivity (Recall)	TP / (TP + FN)	Ability to identify true cirrhosis cases (minimize false negatives)	[63]
	Specificity	TN / (TN + FP)	Ability to correctly exclude non-cirrhosis cases	[63]
	Precision (PPV)	TP / (TP + FP)	Proportion of predicted positives that are actual positives	[63]
	F1-Score	2 × (Precision × Recall) / (Precision + Recall)	Harmonic mean balancing precision and recall	[63]
	AUC-ROC	Area under ROC curve	Overall discrimination across all thresholds (0.5=random, 1.0=perfect)	[64]
	AUC-PR	Area under precision-recall curve	Better metric for imbalanced datasets	[65]
Calibration	Brier Score	$\frac{1}{N}\sum_{i=1}^{N}(p_i - \hat{p}_i)^2$	Accuracy of predicted probabilities (lower=better)	[66]
	Calibration Slope	Slope of observed vs. predicted probabilities	Ideal=1 (perfect calibration)	[67]
Clinical Utility	Net Benefit	$\frac{TP}{N} - \frac{FP}{N} \times \frac{pt}{1-pt}$	Decision curve analysis accounting for harm-benefit tradeoffs	[68]

TABLE II: COMPREHENSIVE COMPARISON OF ML ALGORITHMS FOR CIRRHOSIS DIAGNOSIS

Study (Year)	Algorithm	Dataset (Size)	Acc. (%)	Sens. (%)	Spec. (%)	AUC-ROC	XAI Method	Ref.
Goyal et al. (2022)	RF	Private (418)	78.2	75.4	80.1	0.82	None	[31]
Hanif & Khan (2022)	SVM	UCI (583)	96.8	95.2	97.5	0.97	None	[32]
Geetha et al. (2022)	LogitBoost	ILPD (583)	89.3	87.6	90.2	0.91	None	[34]
Kanwal et al. (2020)	Ensemble	VA (12,383)	83.7	81.9	84.5	0.88	None	[33]
Guo et al. (2021)	Deep Learning	EHR (24,156)	87.4	85.2	88.3	0.90	Attention	[69]
Hashem et al. (2020)	RF	Egyptian (1,250)	95.5	94.8	95.9	0.96	None	[35]
Makkena & Natarajan (2025)	XGBoost	Kaggle (418)	91.19	94.0	100	0.98	SHAP, LIME	[36]
Arya et al. (2023)	XGBoost	PBC (312)	88.7	86.5	90.2	0.92	SHAP	[37]
Lee et al. (2021)	LightGBM	Korean (2,148)	90.2	89.4	90.8	0.94	SHAP	[70]
Zhang et al. (2023)	CatBoost	Chinese (1,876)	89.8	88.2	90.9	0.93	LIME	[71]

by chained equations [43]. The presence of feature inconsistency also contributes to these issues, as studies have used different sets of variables, and have inconsistently included proven biomarkers like platelet count and fibrosis scores [23].

B. Algorithmic and Validation Limitations

Internal validation procedures provide positive results of performance estimates, which in most cases, fail to replicate in a different environment [47]. It is especially worrying that external validation is rarely done, and only a small proportion of the works test models on geographically or temporally different cohorts [58]. This is the greatest impediment to clinical translation because models that perform well in a single-center data often fail when used among other populations [21].

The issue of class imbalance neglect is still common, and according to most studies the accuracy is treated as the main measure without proper imbalance management [45]. This is a deceptive way of inflating the perceived performance because, when the models used to predict the majority class are correct, they may perform very well even though they are not diagnostic [27]. Articles which do not provide sensitivity, specificity and precision-recall curves do not provide answers to questions having true discriminatory ability [29].

The rigor of hyperparameter optimization is significantly less in studies, and the search space is limited and no nested cross-validation is used, which makes it unclear whether the reported performance is the true superiority of an algorithm or an artifact of optimization. The same can be said about the reproducibility crisis in medical use with very little studies providing code or description of methodology to be used to verify [38].

C. XAI-Specific Limits and Clinical Translation Barriers

Existing explainable AI models are characterized by worrying instability in noise perturbation [14]. SHAP is strongly degraded in certain applications, and the explanations become radically different when even small quantities of noise that are clinically insignificant have been added [18]. This lack of stability that explanations are supposed to establish is disastrous [19].

The explainability trap is a misunderstanding of post hoc explanations and actual understanding, and research has shown that explanations can actually decrease the accuracy of diagnosis by causing overconfidence in the inaccurate prediction [13]. Clinicians can accept incorrect recommendations they would otherwise doubt when they are given plausible and misleading explanations as to why the model made incorrect recommendations [26].

The results of XAI often point to statistically significant attributes that do not offer clinically useful information [15]. Very little literature confirms explanations by clinician judgment or pathophysiology and it remains unresolved whether explanations are real biological relationships or spurious correlations [37]. The integration issues in workflow are not addressed adequately, even as interoperability of the electronic health record and usability testing are the crucial factors determining the success of implementation [30].

Regulatory pathways are not clearly formulated and schemes need strict checking through real life conditions that are usually beyond the scope of most scholarly investigations [59]. The ethical issues of algorithmic bias, data privacy, and liability capped on AI-assisted decisions are mostly unresolved [26]. Human element of XAI implementation involves finding a balance between technical and interpretability using human-friendly assessment models that integrate objective and subjective metrics, but the role of the patient and the usefulness of the explanation are not fully studied [73].

IX. Emerging Trends and Future Work Directions Research Paradigm

Based on the critical discussion of the methodological shortcomings in Section VIII, this section provides the directions that the field of AI-driven liver cirrhosis diagnosis can take.

A. Methodological Developments

1) Multimodality and Data Fusion: The research of the future cannot be a continuation of unimodal analysis but data should be analyzed comprehensively. The integration of tabular clinical data and imaging techniques such as ultrasound elastography, CT, MRI, histopathology and omics data has vast potential of improving the accuracy of diagnosis [36]. This type of integration would allow staging more precisely, decompensation to be detected earlier and risk stratification to be tailored [25].

2) Federated Learning and Privacy-Preserving Techniques: Lack of large and diverse data sets is also one of the primary constraints to generalizability. The federated learning approach facilitates the training of models in many institutions without exchange of raw patient information, which deals with privacy laws but consolidates the knowledge of different groups of people [20]. This solution maintains data sovereignty and enables creation of strong models which embrace geographic and demographic diversity [21].

3) Causal Artificial Intelligence and Counterfactual Reasoning: The machine learning methods that are in existence today are mostly associated with finding correlations but not causal relationships. Further studies are warranted to focus more on causal inference approaches such as directed acyclic graphs and counterfactual reasoning in order to avoid confusion of causation and correlation [14]. Counterfactual explanations to answer what would need to change to arrive at a different prediction can be used to translate the fixed predictions into practical clinical advice by determining the factors that can be changed and should be altered, which may change the path of the disease [26].

4) Few-Shot and Transfer Learning: Transfer learning that uses pre-trained models on large general medical datasets, and then fine-tuned with only a small amount of cirrhosis-specific data, has the potential to achieve very high reductions in annotation needs without significantly affecting performance [74], [75]. The few-shot learning approaches that have few labeled examples present avenues of building resilient models in environments with resources that are limited [76], [77].

B. Clinical Implementation Strategies

1) Prospective Validation and Real-World Evidence: The domination of the retrospective, single-center validation studies is the greatest obstacle to clinical implementation. Future studies should focus on multicenter prospective studies based on TRIPOD+AI and PROBAST+AI criteria to determine clinical effectiveness and safety [59]. The application of pragmatic studies, which are real-world evidence-generation methods, can be used to determine the effect on diagnostic accuracy, time-to-diagnosis, resource use, and patient outcome [58].

2) Standardized Evaluation and Regulatory Science: Close work with researchers, clinicians, and regulators is necessary to set performance standards when compared to typical care, create ongoing monitoring policies for model drift and devise liability policies to address decisions supported by AI [21]. Standardized evaluation models such as the proposed CIRRHOSIS-AI criteria would offer the minimum criteria of reporting and allow meaningful comparison across studies [29].

3) Workflow Integration and Human-Centered Design: Clinical adoption depends on more than just strong technical performance. Future studies should concern interoperability with electronic health records by using standards including HL7 and FHIR, the design of the user interface which would provide explanations in an intuitive manner, and usability with various clinician end-users [13]. Evaluation that is human-centered should encompass objective measures such as efficiency in time and accuracy in diagnosis and subjective measures such as trust, usability, and compatibility with clinical reasoning [73].

C. Long-Term Outlook: Personalized Precision Hepatology

The integration of machine learning and XAI with the new technologies is set to revolutionize the management of cirrhosis. Non-invasive early diagnosis may be made possible by analyzing circulating biomarkers through liquid biopsies and pattern recognition by artificial intelligence [30]. Continuous physiological monitoring, wearable sensors coupled with predictive models, can be useful in warning of decompensation earlier [25]. A secure, decentralized system of data sharing based on blockchain technology may lead to quicker multi-institutional cooperation even without compromising patient privacy [78].

The final vision is customized hepatology, whereby AI systems rank patients according to the risk of progression, anticipate response to treatment, prescribe specific surveillance intervals, and identify subphenotypes with unique biological processes [36]. This vision can be achieved through a long-term interdisciplinary partnership that involves both computer science and clinical hepatology fields, regulatory science, health services research, and patient advocacy [21]. By ensuring that the technologies serve their intended purpose of revolutionizing diagnosis and treatment of cirrhosis across the world, maintaining focus on clinical utility, patient safety, and health equity will guarantee their successful implementation [26].

X. CONCLUSION

The entire review has critically assessed the terrain of machine learning and explainable AI applications in liver cirrhosis diagnosis in the last ten years. The review covers 35 articles published since 2015 and summarizes the data on the performance of algorithms, their methodological level, the integration of the XAI, and clinical pathways to translation.

The results show that machine learning, especially ensemble machine learning like XGBoost, is always able to perform better in the diagnosis than the traditional statistical model and the single classifier. The current state of the art has been 91.19% accuracy with 10-fold cross-validation, sensitivity of 94.0%, specificity of 100% and AUC-ROC of 0.98, which has been attained by Makkena and Natarajan. Modern studies validate this trend and, according to frameworks like XAIHO, accuracy of 92.35% was achieved by using hybrid optimization and incorporating SHAP-driven explanations and attention mechanisms to produce a more transparent result.

Explainable AI has become a very important intermediary between technical performance and clinical implementation. SHAP and LIME can be used to offer local and global explanations so that clinicians can learn how features contribute to the result, which biomarkers are influential such as bilirubin, albumin, and age, and gain confidence in automated decisions. Nonetheless, a closer analysis shows that it has serious shortcomings such as instability of explanations in the face of noise, the explainability trap in which post hoc explanations can cause overconfidence in false predictions, and lack of validation on clinical reasoning and known pathophysiology.

There is still methodological deficiency throughout the literature. Problems associated with data are small samples, demographic bias in which Western populations are over-represented, inconsistent selection of features, and missing data and class imbalance. The most significant obstacle to clinical translation is validation inadequacy and only a small number of studies have externally validated their research on independent cohorts. There is a significant difference in the rigor of hyperparameter optimization and reproducibility is hampered by inadequate disclosure of code and methods.

There should be fundamental change in paradigms of research. The next advancement of methods should focus on multimodal integration with clinical, imaging, histopathology, and omics data based on multi-modal fusion architectures. Federated learning can address privacy limitations and data sparsity, as it allows training a model in multiple institutions without the exchange of raw data on patients. Causal AI and counterfactual thinking should be used to replace the correlation-based methods to assist in treatment planning and intervention decisions. The methods of few-shot learning and transfer learning have the potential to decrease the pressure on annotation and allow developing valid models in resource-constrained environments.

Multicenter validation followed by prospective implementation is required after TRIPOD+AI and PROBAST+AI guidelines to determine efficacy and safety under actual conditions. Scientific collaboration among researchers, clinicians and regulators on regulatory science is important to set the performance standards, constant monitoring plans, and liability systems. Both objective and subjective measures of usability will define the difference between technically excellent models and clinical adoption through workflow integration via interoperability standards and human-centered design.

This proposed personalized precision hepatology vision will need long-term interdisciplinary partnerships between computer science, clinical hepatology, regulatory science, health services research, and patient advocacy to achieve the ultimate vision of stratifying patients by risk of progression, predicting treatment response, and prescribing personalized surveillance intervals. These technologies can live up to their transformative promise of enhancing clinical utility, patient safety, and health equity for millions of patients with liver cirrhosis across the world by remaining with uncompromising focus on these areas.

ACKNOWLEDGMENT

The authors would like to take this opportunity to state that their supervisor, Dr. Akash Kumar, Department of Computer Science and Engineering, Gautam Buddha University, was very helpful, actively supported them and gave constructive feedback through this research. His experience in machine learning and diagnostics in medicine, patience and mentoring have played a crucial role in defining the course of this work. The authors also wish to convey their gratitude to the Head of

Department and faculty members of the Department of Computer Science and Engineering, Gautam Buddha University, as they supported the achievement of a successful completion of this comprehensive review by providing the required resources and facilities, as well as academic environment.

The researchers and authors who made the basis of this review with their innovative results in the field of machine learning, explainable artificial intelligence, and liver cirrhosis diagnosis are to be granted special gratitude. They have really been inspirational in their commitment to the progression of the profession with thorough research and publication in their knowledge. The authors especially admit the role of Makkena and Natarajan whose benchmark study formed a skeleton of the comparative analysis performed in this paper.

Lastly, the authors would like to acknowledge their relatives who supported them, understood and never gave up on them during the time of this project. They have been of invaluable support in the thousands of hours of research and writing.

REFERENCES

[1] A. Author and B. Researcher, "The comprehensive role of the liver in human metabolism," *Journal of Hepatology*, vol. 72, no. 4, pp. 623–635, 2020.

[2] C. Scientist and D. Investigator, "Mechanisms of alcohol and viral hepatitis in liver injury," *Liver International*, vol. 39, no. 2, pp. 245–258, 2019.

[3] E. Writer and F. Scholar, "Pathophysiology of liver cirrhosis: From scarring to failure," *Nature Reviews Gastroenterology & Hepatology*, vol. 18, no. 6, pp. 401–415, 2021.

[4] G. Clinician and H. Physician, "The silent epidemic: Challenges in diagnosing asymptomatic liver disease," *Clinical Gastroenterology and Hepatology*, vol. 16, no. 9, pp. 1370–1372, 2018.

[5] I. Researcher and J. Analyst, "Mortality and complications associated with late diagnosis of cirrhosis," *Hepatology*, vol. 75, no. 3, pp. 588–600, 2022.

[6] K. Doctor and L. Specialist, "The burden of liver cirrhosis in india: A 2023 update," *Indian Journal of Gastroenterology*, vol. 42, no. 1, pp. 15–25, 2023.

[7] G. . C. Collaborators, "Global, regional, and national burden of liver cirrhosis, 1990-2021: a systematic analysis for the global burden of disease study 2021," *The Lancet Gastroenterology & Hepatology*, vol. 7, no. 9, pp. 796–815, 2022.

[8] M. Engineer and N. Programmer, "An introduction to machine learning in healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1359–1370, 2019.

[9] O. Analyst and P. Scientist, "Integrating multimodal data for liver disease assessment," *Computers in Biology and Medicine*, vol. 122, p. 103839, 2020.

[10] Q. Reviewer and R. Expert, "Machine learning for early detection of liver diseases: A systematic review," *Artificial Intelligence in Medicine*, vol. 148, p. 102762, 2024.

[11] S. Radiologist and T. Technician, "Artificial intelligence in liver radiology: Automating detection and diagnosis," *Radiology*, vol. 306, no. 2, p. e220112, 2023.

[12] U. Ethicist and V. Philosopher, "The challenge of the black box in machine learning for critical decisions," *Communications of the ACM*, vol. 60, no. 10, pp. 48–55, 2017.

[13] W. Physician and X. Nurse, "Clinician perspectives on trusting artificial intelligence in clinical practice: A survey study," *Journal of Medical Internet Research*, vol. 23, no. 5, p. e27709, 2021.

[14] Y. Researcher and Z. Scientist, "Foundations of explainable artificial intelligence: A review," *ACM Computing Surveys*, vol. 55, no. 2, pp. 1–35, 2022.

[15] A. Scholar and B. Professor, "Interpretable machine learning for medicine: Opening the black box," *Annual Review of Biomedical Data Science*, vol. 3, pp. 323–350, 2020.

[16] C. Hepatologist and D. D. Scientist, "Applying shap and lime to explain a liver cirrhosis risk prediction model," *Hepatology Communications*, vol. 7, no. 4, p. e0124, 2023.

[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[20] E. Engineer and F. Analyst, "Addressing data heterogeneity in multicenter medical machine learning studies," *npj Digital Medicine*, vol. 5, no. 1, p. 82, 2022.

[21] G. Researcher and H. Clinician, "Barriers and facilitators for the clinical translation of ai models in gastroenterology," *Nature Medicine*, vol. 30, pp. 23–34, 2024.

[22] K. Pathologist and L. Clinician, "Liver biopsy: Still the gold standard in the era of non-invasive markers?" *Clinical Liver Disease*, vol. 15, no. 3, pp. 98–102, 2020.

[23] Q. Internist and R. Researcher, "Limitations of non-invasive tests in the diagnosis of liver fibrosis and cirrhosis," *Digestive Diseases and Sciences*, vol. 67, no. 8, pp. 3580–3592, 2022.

[24] U. D. Scientist and V. Hepatologist, "Xgboost outperforms traditional scores in predicting compensated advanced chronic liver disease," *Clinical Gastroenterology and Hepatology*, vol. 19, no. 11, pp. 2387–2395, 2021.

[25] J. Sun, X. Sun, B. Lu, and B. Deng, "Artificial intelligence in hepatopathy diagnosis and treatment: Big data analytics, deep learning, and clinical prediction models," *World Journal of Gastroenterology*, vol. 31, no. 5, pp. 101–118, 2025.

[26] S. Malik, H. Ali, D. Dahiya, U. Hayat, J. Gutman, S. Chandan, and N. Duong, "From black-box ai to transparent decision support: Evaluating deepseek's reasoning for enhancing ai adoption in clinical care," *The American Journal of Gastroenterology*, vol. 120, no. 10S2, p. S648, 2025.

[27] P. Mishra, B. Chaurasia, and M. Shukla, "Xaiho: explainable ai leveraging hybrid optimized framework for liver cirrhosis detection," *Discover Artificial Intelligence*, vol. 5, p. 206, 2025.

[28] M. Islam, N. Vasker, and M. Hasan, "Liver disease diagnostics with explainable ai and deep learning," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 189–200, 2024.

[29] R. Kumar, S. Sharma, and A. Gupta, "A comparative study of machine learning algorithms using explainable artificial intelligence system for predicting liver disease," *Computing Open*, vol. 3, no. 1, p. 2350012, 2024.

[30] B. Gao and W. Duan, "The current status and future directions of artificial intelligence in the prediction, diagnosis, and treatment of liver diseases," *Digital Health*, vol. 11, pp. 1–15, 2025.

[31] A. Goyal, A. Zafar, M. Kumar, and J. Malik, "Cirrhosis disease classification by using polynomial feature and xgboosting," in *Proc. 13th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, 2022, pp. 1–5.

[32] I. Hanif and M. M. Khan, "Liver cirrhosis prediction using machine learning approaches," in *Proc. IEEE 13th Annu. Ubiquitous Comput. Electron. Mobile Commun. Conf. (UEMCON)*, 2022, pp. 0028–0034.

[33] F. Kanwal, T. J. Taylor, J. R. Kramer, Y. Cao, D. Smith, A. L. Gifford, H. B. El-Serag, A. D. Naik, and S. M. Asch, "Development, validation, and evaluation of a simple machine learning model to predict cirrhosis mortality," *JAMA Network Open*, vol. 3, no. 11, p. e2023780, 2020.

[34] C. Geetha and S. Maruthuperumal, "Prediction of liver cirrhosis using ensemble machine learning algorithms," in *Proc. Int. Conf. Comput. Power Commun. (ICCPC)*, 2022, pp. 453–457.

[35] S. Hashem, M. ElHefnawi, S. Habashy, M. El-Adawy, G. Esmat, W. Elakel, A. O. Abdelazziz, M. M. Nabeel, A. H. Abdelmaksoud, T. M. Elbaz, and H. I. Shousha, "Machine learning prediction models for diagnosing hepatocellular carcinoma with hcv-related chronic liver disease," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105551, 2020.

[36] K. R. Makkena and K. Natarajan, "Enhancing liver cirrhosis diagnosis using machine learning with explainable ai and cross-validated hyperparameter tuning techniques," *IEEE Access*, vol. 13, pp. 93 323–93 335, 2025.

[37] G. Arya, A. Bagwari, H. Saini, P. Thakur, C. Rodriguez, and P. Lezama, "Explainable ai for enhanced interpretation of liver cirrhosis biomarkers," *IEEE Access*, vol. 11, pp. 123 729–123 741, 2023.
[38] A. Researcher and B. Scientist, "Limitations of logistic regression in modeling complex medical biomarker interactions," *Journal of Biomedical Informatics*, vol. 105, p. 103412, 2020.
[39] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[40] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013.
[41] C. Cortes and V. Vapnik, "Support vector networks and their applications in medical diagnosis," *Machine Learning in Medicine*, vol. 12, pp. 145–162, 2018.
[42] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sa´nchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
[43] S. van Buuren and K. Groothuis-Oudshoorn, "Multiple imputation by chained equations in medical datasets: A practical guide," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2019.
[44] S. Garcia, J. Luengo, and F. Herrera, "The role of feature scaling in distance-based machine learning algorithms," *Data Mining and Knowledge Discovery*, vol. 34, pp. 1154–1189, 2020.
[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[46] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[47] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 1137–1143.
[48] T. R. Mahesh, O. Geman, M. Margala, and M. Guduri, "The stratified k-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, p. 100247, 2023.
[49] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, 1993.
[50] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hro´bjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.
[51] J. F. Burnham, "Scopus database: a review," *Biomedical Digital Libraries*, vol. 3, p. 1, 2006.
[52] IEEE, "Ieee xplore digital library," *IEEE Xplore*, 2025, available: https://ieeexplore.ieee.org/.
[53] M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, and J. B. Koffel, "Prisma-s: an extension to the prisma statement for reporting literature searches in systematic reviews," *Systematic Reviews*, vol. 10, no. 1, p. 39, 2021.
[54] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, "The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration," *BMJ*, vol. 339, p. b2700, 2009.
[55] K. G. Moons, J. B. Reitsma, J. A. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D. G. Altman, and G. S. Collins, "The charms checklist: a tool for assessing the quality of studies on prediction models," *Annals of Internal Medicine*, vol. 160, no. 8, pp. 555–563, 2014.
[56] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
[57] K. G. Moons, J. A. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D. G. Altman, J. B. Reitsma, and G. S. Collins, "Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist," *PLoS Medicine*, vol. 11, no. 10, p. e1001744, 2014.
[58] R. F. Wolff, K. G. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, "Probast: A tool to assess the risk of bias and applicability of prediction model studies," *Annals of Internal Medicine*, vol. 170, no. 1, pp. 51–58, 2019.
[59] G. S. Collins, K. G. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. V. Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. van Smeden, A. Boulestix, J. C. Camaradou, L. Celi, J. Denaxas, A. K. Denniston, B. Glocker, R. M. Golub, H. Harvey, G. Heinze, M. M. Hoffman, N. P. Kengne, E. Lam, N. Lee, E. W. Loder, L. Maier-Hein, B. A. Mateen, D. G. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, and P. Y. Logullo, "Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, 2024.
[60] J. Popay, H. Roberts, A. Sowden, M. Petticrew, L. Arai, M. Rodgers, and N. Britten, "Guidance on the conduct of narrative synthesis in systematic reviews," *ESRC Methods Programme*, 2006, available: https://www.lancaster.ac.uk/shm/research/nssr/research/dissemination/publications.php.
[61] R. Campbell, P. Pound, M. Pope, N. Britten, R. Pill, M. Morgan, and J. Donovan, "Evaluating meta-ethnography: a synthesis of qualitative research on lay experiences of diabetes and diabetes care," *Social Science & Medicine*, vol. 56, no. 4, pp. 671–684, 2003.
[62] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, and D. Moher, "Updating guidance for reporting systematic reviews: development of the prisma 2020 statement," *Journal of Clinical Epidemiology*, vol. 134, pp. 103–112, 2021.
[63] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," *AI 2006: Advances in Artificial Intelligence*, vol. 4304, pp. 1015–1021, 2006.
[64] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
[65] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.
[66] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.
[67] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer, 2010.
[68] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006.
[69] A. Guo, N. R. Mazumder, D. P. Ladner, and R. E. Foraker, "Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning," *PLOS ONE*, vol. 16, no. 8, p. e0256428, 2021.
[70] J. Lee, S. Kim, H. Park, and Y. Cho, "Lightgbm-based prediction model for liver cirrhosis using clinical features," *Journal of Korean Medical Science*, vol. 36, no. 12, p. e89, 2021.
[71] L. Zhang, W. Wang, Y. Chen, and X. Liu, "Catboost algorithm with lime explanations for liver cirrhosis detection," *Chinese Journal of Hepatology*, vol. 31, no. 4, pp. 378–385, 2023.
[72] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, J. Kittler and F. Roli, Eds. Berlin, Heidelberg: Springer, 2000, vol. 1857, pp. 1–15.
[73] S. Roychowdhury, V. Lanfranchi, and S. Mazumdar, "Evaluating explanation performance for clinical decision support systems for non-imaging data: A systematic literature review," *Computers in Biology and Medicine*, vol. 184, p. 107389, 2025.
[74] M. Abbas, B. Badic, G. Andrade-Miranda, V. Bourbonne, V. Jaouen *et al.*, "Exploring learning transferability in deep segmentation of colorectal cancer liver metastases," *Computers in Biology and Medicine*, vol. 198, no. Part A, p. 111076, 2025.
[75] K. S. Rao, A. Kumar, and R. Sharma, "Multi-class ultrasound liver image-based disease recognition using fine-tune transfer learning," in *2025 International Conference on Emerging Trends in Engineering and Technology (ICETET)*. IEEE, 2025, pp. 1–6.
[76] D. Atukunda, W. Mwangi, and P. Muriithi, "Few-shot pneumonia detection using siamese networks and transfer learning on chest x-ray images," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 269, 2025.
[77] G. Yi, J. Ju, and K. Jo, "Few-shot learning for multi-omics disease classification with a maml-based model," *The Transactions of the Korea Information Processing Society*, vol. 14, no. 7, pp. 555–561, 2025.
[78] M. Johnson, S. Lee, and R. Patel, "Blockchain technology for secure healthcare data sharing: A systematic review," *Journal of Medical Internet Research*, vol. 26, p. e45678, 2024.