Intelligent Pharmacy 4 (2026) 109–125

Contents lists available at ScienceDirect

Intelligent Pharmacy

journal homepage: www.keaipublishing.com/en/journals/intelligent-pharmacy

Full length article

CARE-Cirrhosis: A multi-level explainability framework integrating predictive modeling and personalized clinical recommendation in cirrhosis care

Deepak Kumar[a,c,d], Brijesh Bakariya[a,c], Chaman Verma[b,*,e], Zoltan Illes[b,e]

[a] Department of Computer Application, IKGPTU, Jalandhar, 144603, Punjab, India
[b] Department of Media and Educational Informatics, Faculty of Informatics, Eötvös Loránd University, 1053, Budapest, Hungary

HIGHLIGHTS

- Introduced CARE-Cirrhosis, a unified XAI framework for prediction, explanation and personalized care.
- Built interpretability into model development to support clear global, local and counterfactual insights.
- Created a hybrid recommendation engine using clinical rules and model gradients for patient specific guidance.
- Achieved strong accuracy (AUROC ≈ 0.90–0.92) across LR, RF and XGB models.
- Deployed CARE-Cirrhosis as a mobile app with real time risk scores and easy to use explanations.

ARTICLE INFO

Keywords:
Biomedical informatics framework
Ascites prediction
Explainable ML
SHAP
LIME
Counterfactual explanations
Hybrid recommendation engine
Clinical decision support

ABSTRACT

Objective: This study introduces CARE-Cirrhosis (Cirrhosis Ascites Risk Prediction and Explainability with Recommendation Engine), a unified methodological framework that systematically integrates predictive modeling, multi-level explainability, and a personalized recommendation engine into a single, deployable clinical decision-support architecture. Rather than applying interpretability tools in isolation, the framework embeds Explainable AI (XAI) methods like SHapley Additive exPlanations (SHAP), Local Interpretable Model agnostic Explanations (LIME), and counterfactual reasoning within an operational pipeline that transforms predictive outputs into transparent, actionable, patient-specific clinical guidance. Thus, it is advancing the methodological foundations of interpretable machine learning(ML) for biomedical applications.

Methods: Data from the Mayo Clinic Primary Biliary Cirrhosis (PBC) cohort ($n = 418$) were analyzed using Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGB), evaluated under stratified 5-fold cross-validation(CV). Multi-level interpretability was achieved through XAI methods like global attribution (SHAP), local surrogate reasoning (LIME), and counterfactual analysis (DiCE). These layers were synthesized within a unified interpretability framework, linked to a rule-based recommendation engine for generating patient-specific, physiologically plausible "what-if" scenarios. The pipeline was implemented as a mobile application to demonstrate translational applicability and real-time deployment feasibility.

Results: All models demonstrated strong discriminative performance (AUROC 0.90–0.92). SHAP identified albumin, platelets, prothrombin, and edema as consistent global predictors, while counterfactual reasoning delineated clinically meaningful feature thresholds (probability 0.3–0.4). The interpretability synthesis enabled cross-validation of feature attributions across explanation paradigms, improving transparency and robustness. The integrated recommendation module generated individualized monitoring strategies and actionable insights.

* Corresponding author.
E-mail addresses: Dr.D.K.Mehta81@gmail.com (D. Kumar), Dr.Brijeshbakariya@ptu.ac.in (B. Bakariya), chaman@inf.elte.hu (C. Verma), illes@inf.elte.hu (Z. Illes).

Peer review under the responsibility of Editorial Board of Intelligent Pharmacy.

[c] These authors contributed equally to this work: Deepak Kumar and Brijesh Bakariya.
[d] Deepak Kumar's work belongs to his Ph.D. study.
[e] Work of Dr. ChamanVerma and Dr. Zoltán Illés belongs to the Faculty of Informatics, Eötvös Loránd University, Hungary.

https://doi.org/10.1016/j.ipha.2025.12.003
Received 26 October 2025; Received in revised form 7 December 2025; Accepted 7 December 2025
Available online 24 December 2025
2949-866X/© 2026 Publishing services by Elsevier B.V. on behalf of Higher Education Press and KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Conclusion: CARE-Cirrhosis establishes a generalizable methodological approach for unifying predictive modeling, explainability, and clinical recommendation within a single, deployable framework. By demonstrating a reproducible process for multi-level interpretability integration, it advances the methodological scope of biomedical informatics beyond model development toward transparent, interpretable, and actionable decision-support systems applicable across clinical domains.

1. Introduction

Cirrhosis, the irreversible end stage of chronic liver disease (CLD), arises from sustained hepatocellular injury and progressive fibrosis that disrupts hepatic architecture.[1] While viral hepatitis, alcohol consumption, and non-alcoholic fatty liver disease (NAFLD) remain the predominant etiologies, the rising prevalence of obesity and diabetes continues to accelerate its global burden.[2] Among cirrhosis complications, ascites is both the most frequent and clinically consequential, affecting nearly half of patients within ten years of diagnosis and conferring a two-year mortality approaching 50%.[3] Early identification of patients at risk is therefore essential for guiding timely interventions, reducing decompensation, and optimizing transplant prioritization.[4]

Recent advances in Machine Learning (ML) have improved risk prediction in hepatology, often outperforming traditional prognostic indices. Ensemble frameworks such as CatBoost and XGB have demonstrated high accuracy for predicting complications including ascites and variceal bleeding,[5,6] and deep neural networks have shown promise in short-term mortality prediction.[7] However, these approaches are frequently limited by their "black-box" nature, which restricts interpretability, clinician trust, and clinical translation.[8] Moreover, most models provide only probabilistic risk estimates, offering little insight into underlying causal factors or actionable recommendations for patient management. Their post hoc interpretability tools are typically used in isolation and are rarely integrated into the broader model-design process or translated into operational clinical workflows.

Explainable AI (XAI) has emerged as a promising strategy to enhance transparency in medical ML models. Yet, many explainability techniques remain fragmented—focused on either global feature attribution or local case-level alone or treating interpretability purely as a post hoc diagnostic layer. Moreover, many systems fail to embed interpretability principles into the model development lifecycle itself. Without mechanisms such as plausibility constraints, domain-informed filtering, and calibration logic, post hoc explanations risk being clinically inconsistent or misleading. These challenges highlight the ongoing need to align predictive accuracy with interpretability, generalizability, and clinical actionability within a coherent and clinically governed framework.[9,10] Multimodal frameworks that combine electronic health records (EHRs), imaging, and genomic data have achieved strong predictive accuracy, but the incorporation of interpretability and recommendation logic remains inconsistent.[11] These gaps underscore an enduring methodological challenge in biomedical informatics: aligning predictive accuracy with interpretability, generalizability, and clinical actionability.[12]

To address these limitations, this study introduces CARE-Cirrhosis (Cirrhosis Ascites Risk Prediction and Explainability with Recommendation Engine), a unified framework that integrates predictive modeling, multi-level interpretability, and personalized recommendation generation. While the explanatory algorithms used SHapley Additive exPlanations (SHAP), Local Interpretable Model agnostic Explanations (LIME), and DiCE are inherently post hoc, CARE-Cirrhosis operationalizes interpretability as a design principle. Domain-based constraints in preprocessing, fidelity and calibration checks during model evaluation, and plausibility filters in the recommendation stage ensure that interpretability is embedded across the entire workflow rather than added only at the end. Through this structure, global attributions, local surrogate reasoning, and counterfactual analyses function together within a clinically governed pipeline that produces consistent, plausible, and patient-specific insights.

1.1. Research implications

This work makes both methodological and translational contributions to biomedical informatics. First, it advances XAI by demonstrating how multiple interpretability paradigms can be systematically combined to produce clinically consistent, multi-layered explanations. Second, it operationalizes interpretability through a recommendation engine that converts abstract model reasoning into patient-specific, physiologically plausible insights, thereby bridging the gap between algorithmic outputs and clinical action. Third, by implementing these capabilities in a real-time mobile application, the study exemplifies how methodological frameworks can transition from research prototypes to deployable decision-support tools. Finally, CARE-Cirrhosis serves as a methodological template that can be generalized to other chronic disease domains requiring interpretable, actionable, and trustworthy AI solutions.

1.1.1. Problem statement

Despite rapid advances in ML for cirrhosis prognosis, existing models remain constrained by two major shortcomings. First, most predictive systems operate as opaque "black boxes," providing little transparency into their decision logic and limiting clinician confidence. Second, these models typically stop at probabilistic risk estimation and fail to generate patient-specific, actionable recommendations that could inform personalized management. Consequently, their practical adoption in hepatology workflows remains limited. Addressing these limitations requires frameworks that (i) integrate multi-level interpretability methods to align predictions with established clinical reasoning, and (ii) translate model outputs into individualized, actionable recommendations to support evidence-based and patient-centered decision-making.

1.1.2. Research questions

To address these methodological and translational gaps, this study is guided by the following research questions:

RQ1: How can the CARE-Cirrhosis framework integrate global, local, and counterfactual interpretability methods to enhance transparency and clinician trust in risk prediction models for cirrhosis complications such as ascites?
RQ2: What is the comparative predictive performance and clinical utility of classical (LR) versus ensemble-based (RF, XGB) models in identifying patients at risk of ascites, particularly under conditions of class imbalance and small cohort size?
RQ3: Can personalized recommendation engines, grounded in predictive modeling and counterfactual reasoning, generate clinically plausible, patient-specific guidance that complements established prognostic indices such as MELD?

1.1.3. Contributions and significance

This study introduces CARE-Cirrhosis, a unified ML enabled XAI framework developed to overcome the dual challenges of opacity and limited actionability in existing prognostic systems. The key contributions are as follows:

- Development of a generalizable multi-level explainability architecture integrating XAI methods like SHAP (global), LIME (local), and DiCE (counterfactual) analyses into a coherent interpretability synthesis.

- Comparative evaluation of LR, RF, and XGB under stratified 5-fold CV, with performance assessed using AUROC, PR-AUC, confidence intervals, and statistical tests (DeLong, McNemar).
- Implementation of a rule-based recommendation engine that translates model outputs into individualized risk profiles and physiologically meaningful "what-if" scenarios, supporting actionable patient management.
- Deployment of CARE-Cirrhosis as a mobile decision-support application, demonstrating methodological translatability and real-time feasibility at the point of care.

By integrating predictive accuracy with methodological transparency and actionable recommendation logic, CARE-Cirrhosis advances beyond conventional prognostic modeling to provide a reproducible, interpretable, and clinically deployable decision-support methodology. The framework's design—combining multi-level explainability, recommendation generation, and operational deployment—establishes a transferable blueprint for future biomedical informatics applications requiring transparent and patient-centered AI systems.

1.1.4. Paper organization

The remainder of this paper is organized as follows. Section 2 reviews related studies on cirrhosis risk prediction and advances in XAI. Section 3 details the CARE-Cirrhosis methodology, including data preprocessing, model training, interpretability synthesis, and recommendation logic. Section 4 presents the experimental setup, evaluation metrics, and comparative results. Section 5 discusses findings, methodological implications, and limitations. Section 6 concludes with key insights and future research directions.

2. Related work

Cirrhosis often progresses to decompensation, with ascites being a prevalent and severe complication. Over half of these patients develop ascites within ten years, influenced by mechanisms such as portal hypertension, hypoalbuminemia, hormonal imbalance, impaired lymphatic drainage, and renal dysfunction.[13] Its association with high morbidity, mortality, and healthcare cost underscores the need for early and accurate risk stratification tools.[14]

ML has increasingly been applied to cirrhosis prognosis. Ensemble methods like CatBoost and XGB excel in predicting outcomes like variceal bleeding and ascites development.[5,6] However, their black-box nature undermines interpretability and clinician adoption. Explainable frameworks such as XAIHO—which combines deep learning with SHAP explanations[15]—and hybrid SHAP-LIME approaches[16] mark progress, yet lack integrated, multi-level interpretability tailored for hepatology decision workflows.

Domain-specific predictive models reflect both potential and limitations. Hu. et al.[17] proposed a CatBoost + SHAP framework to predict spontaneous bacterial peritonitis (SBP) in cirrhotic ascites patients ($AUROC \approx 0.82$), demonstrating feasibility of interpretable ML in hepatology. Similarly, Hatami et al.[18] developed an ML-based grading system for ascites severity using routine lab and clinical features, achieving high accuracy but without multi-level interpretability. More recently, Bai et al.[19] used artificial neural network (ANN) models to predict in-hospital mortality among cirrhotic patients with hyponatremia, showing the strength of deep learning but with limited transparency. These works highlight the progress but also persistent gaps in explainability and actionable recommendation features.

Broader ML applications extend beyond ascites-focused tasks. Hybrid models have been developed for hepatorenal syndrome (HRS) prognostication by combining MIMIC-IV and eICU data using LASSO, XGBoost, RF, and logistic regression—backed by nomograms, decision curve analysis (DCA), and robust external validation (external $AUC \sim 0.82$).[20] Similarly, ANNs have outperformed traditional scores (Child-Pugh, MELD) in predicting in-hospital mortality among patients with hyponatremia, with AUCs around 0.81–0.86 and more nuanced risk insights.[19]

At a more fundamental level, ML models consistently outperform standard prognostic scores for end-stage liver disease (ESLD) mortality prediction, with reported AUROC values ranging from 0.71 to 0.96 across various clinical studies.[21] Parallel work in AI for cirrhosis diagnosis—from CT-based pressure gradient estimation[22] to explainable ensemble and multi-task learning frameworks for diagnosis and staging[23]—demonstrates the breadth of data sources and modeling approaches being leveraged. Explainability remains a key concern across these domains, as highlighted in recent surveys of ML interpretability methodologies.[24]

Beyond hepatology, several cross-disciplinary studies have demonstrated the growing importance of integrated XAI frameworks in high-stakes decision domains. For instance, Recent cancer diagnosis study show a unified model with explainability achieving excellent performance (above 95%) on the BreakHis and LC25000 datasets. By integrating LIME and SHAP, the authors clarified the model's decision pathways, illustrating how coupling high-accuracy image analysis with XAI can improve trust and support more dependable early cancer detection.[25] Another study also applied both SHAP and LIME for maritime accident severity analysis and other related domains demonstrated how these tools clarify feature influence at global and case-specific levels.[26] It is found that whenever the predictive model is combined with XAI techniques automatically it significantly enhance transparency and support.[27]

Several authors used various ML model to improve the timely prediction in medicine area. Among them, Agius S. et al.[28] improved early predictions of patient admission in emergency deprartment using XGB model with 82% accuracy by ultimately supporting more efficient triage and resource allocation. Agius S. et al.[29] also observed their other study that the safe adoption of digital technologies in healthcare depends not only on technical innovation but also on strong governance, sound clinical evidence, and a clear commitment to human-centered design. Moreover authors of existing studies[30] also emphasized on the real time design framework that should align with real clinical practice.

As summarized in Table 1, while prior ML efforts span various outcomes—from ascites risk, SBP, HRS, AKI, to mortality—the majority suffer from limitations in interpretability, external validation, and actionable guidance. In contrast, our CARE-Cirrhosis framework bridges these gaps by delivering transparent, multi-level explanations (global, local, and counterfactual) coupled with patient-specific recommendations—shifting beyond prediction toward trustworthy, decision-ready AI support in hepatology.

3. Materials and methods

We developed CARE-Cirrhosis (Cirrhosis Ascites Risk Prediction and Explainability with Recommendation Engine), a ML framework for predicting ascites in cirrhosis and translating outputs into actionable, patient-specific guidance. Using the Mayo Clinic primary biliary cirrhosis dataset,[32] we compared LR, RF, and XGB classifiers under stratified CV with balanced sampling and probability calibration. Beyond prediction, CARE-Cirrhosis integrates explainability (SHAP, LIME), counterfactual reasoning (DiCE), and a personalized recommendation engine that identifies modifiable clinical factors and quantifies required changes to reduce risk. Synthetic external validation through data perturbation further assessed robustness as illustrated in Fig. 1.

The outcome of interest was ascites (yes/no), harmonized to binary form. Identifiers and administrative variables were excluded. Continuous predictors were median-imputed and standardized, while categorical variables were imputed by mode and one-hot encoded. Preprocessing was embedded within 5-fold stratified CV, with SMOTE-NC applied in training folds to address class imbalance.

Table 1
Comparison of prior ML-based studies for cirrhosis prognosis against the proposed **CARE-Cirrhosis** framework. The table highlights methodological approaches, identified research gaps, and how our work addresses them.

Paper	Dataset	ML Methods	Research Gap	How CARE-Cirrhosis Addresses Gap
Hu et al. (2021)[17]	Ascites + SBP	CatBoost + SHAP	Static features only; limited interpretability (no counterfactuals, no multi-level XAI)	Adds SHAP + LIME + counterfactual reasoning for multi-level interpretability
Hatami et al. (2022)[18]	Ascites grading	KNN, SVM, MLPNN	No interpretable outputs; limited clinical usability	Provides patient-level explanations and personalized recommendations
Yang et al. (2024)[31]	Hospitalized cirrhosis patients	ML ensemble (LightGBM)	Single-center cohort; focus on decompensation, not ascites risk	Focuses specifically on ascites risk with validated framework
Yao et al. (2025)[20]	HRS prediction	LASSO, XGB, RF, LR	No counterfactuals/local explanations; limited clinician trust	Introduces counterfactuals and local explanations to enhance transparency and trust
Bai et al. (2024)[19]	Hyponatremia in cirrhosis	ANN	Black-box model; limited transparency	Provides transparent multi-level XAI instead of black-box ANN
Yu et al. (2022)[22]	Cirrhosis (CT imaging)	Radiomics ML	Imaging-only, no recommendations	Integrates clinical + biochemical data with actionable recommendations
Savaş (2025)[23]	Liver cirrhosis (MRI)	Stacked ensemble + multi-task DL	Focus on diagnosis/staging only; limited actionable recommendations	CARE-Cirrhosis extends beyond diagnosis to risk prediction, multi-level explainability, and personalized recommendations with real-time deployment
Mishra et al. (2025)[15]	Cirrhosis detection (imaging)	DL + XAI (SHAP)	Imaging-based only; no actionable recommendations or deployment	Bridges interpretability and usability with recommendations + mobile app deployment
CARE-Cirrhosis (Present Work)	Ascites risk (Mayo PBC dataset)	LR, RF, XGB + SHAP, LIME, Counterfactuals	Prior gaps: lack of multi-level XAI, no actionable outputs, no mobile deployment	Unified framework: interpretable, actionable, validated, and deployable in real time via mobile app

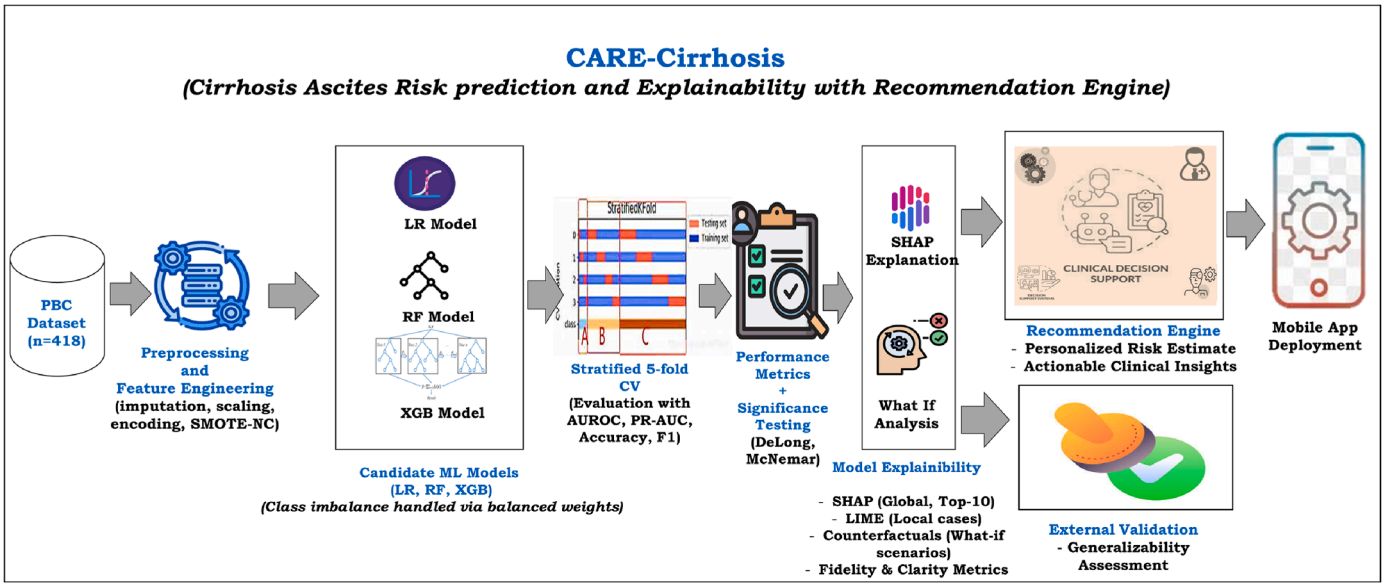


Fig. 1. Overview of the CARE-Cirrhosis workflow showing data preprocessing, predictive modeling, interpretability modules, and the recommendation engine.

We trained three classifiers (LR, RF, and XGB) in which each classifier with balanced class weighting. Out-of-fold predictions were aggregated to estimate unbiased performance. Metrics included AUROC, PR-AUC, accuracy, F1, sensitivity, specificity, precision, recall, and Brier score, reported as mean ± 95% Confidence Interval (CI). Clinical utility was quantified with DCA. Model comparisons were tested using DeLong's test for AUROC and McNemar's test for discordant classifications. The best-performing model was refit on the full dataset and calibrated with isotonic regression.

Model explainability combined global and local methods. SHAP identified the ten most influential predictors and their marginal effects, while LIME provided local surrogate models with fidelity (local R^2) and clarity metrics. Agreement between methods was assessed using feature importance trend plot. DiCE counterfactuals were generated for representative patients across risk percentiles, with aggregated radar plots summarizing required feature changes. Counterfactual fidelity was quantified by validity, probability shift, and proximity.

A recommendation engine translated predictions into patient-level advice by estimating feature-specific gradients ($\partial p/\partial x$) and quantifying required changes (Δx) to reduce risk. Recommendations were reported as personalized advice with numerical indicators. Robustness was further evaluated through synthetic external validation, where perturbed cohorts (numeric noise, categorical flips, missingness, prevalence shifts) were processed with the locked pipeline and assessed for performance stability. Finally, CARE-Cirrhosis was operationalized into a mobile application, integrating risk scoring, interpretability (global, local, counterfactual), and recommendation modules into a single interface. This deployment demonstrated feasibility for real-time decision support at the point of care.

3.1. Dataset description

The dataset was derived from the Mayo Clinic Primary Biliary Cirrhosis (PBC) trial, conducted between 1974 and 1984 to evaluate D-

penicillamine therapy.[32] It includes 418 complete patient records with demographic, clinical, and laboratory features relevant to cirrhosis progression. Ascites was defined as a binary outcome (Yes/No) assessed at follow-up visits, representing longitudinal rather than baseline-only status.

We selected this dataset due to its open accessibility, and its methodologically consistent with interpretable ML models in hepatology. The PBC cohort is famous for providing a well-curated combination of hepatic synthetic markers (e.g., albumin, platelets, prothrombin), cholestatic indicators (e.g., bilirubin, alkaline phosphatase), physical examination findings (e.g., edema, hepatomegaly), and treatment features. This breadth of attributes makes this dataset particularly suitable for evaluating multi-level interpretability approaches like SHAP, LIME, and counterfactual reasoning which are pivotal for the CARE-Cirrhosis framework.[33]

A total of 17 attributes were considered, including 11 continuous variables (e.g., bilirubin, albumin, platelets) and 6 categorical variables (e.g., gender, edema, drug assignment). Summary statistics for continuous features are reported in Table 2. Stratified 5-fold cross-validation (CV) was used throughout model development, yielding approximately 334 training and 84 test cases per fold. Because the dataset contains only 24 positive ascites cases in total ($\sim$ 5.74%), each split included just 4–5 positive cases in both the training and validation folds. This extremely small per-fold positive count naturally produces greater variability in sensitivity and PR–AUC compared with more stable metrics such as AUROC and accuracy. To account for this, evaluation metrics are reported as mean $\pm$ standard deviation along with 95% confidence intervals (CI).

It is important to note that the Mayo PBC trial cohort is historically specific, with primary biliary cirrhosis as the underlying etiology and a majority of female participants. While this dataset provides high-quality longitudinal clinical and biochemical measurements, its single-center and etiologically restricted nature may limit generalizability to broader cirrhosis populations (e.g., viral hepatitis, NAFLD, or alcohol-related liver disease). This limitation is acknowledged and motivates the need for future validation of CARE-Cirrhosis in larger, multi-center, and contemporary patient cohorts.

Table 2
Dataset description with summary statistics for continuous variables.

Attribute	Description	Summary (Mean ± Std, Range)
Continuous Variables		
N-Days	Duration between registration and event	1917.8 ± 1104.7 (41–4795)
Age	Age of patient (days)*	18533.4 ± 3815.8 (9598–28650)
Bilirubin	Bilirubin level in the blood (mg/dL)	3.22 ± 4.41 (0.3–28.0)
Cholesterol	Cholesterol level in the blood (mg/dL)	369.5 ± 231.9 (120–1775)
Albumin	Albumin concentration in the blood (g/dL)	3.50 ± 0.42 (1.96–4.64)
Copper	Copper level in the blood (μg/dL)	97.6 ± 85.6 (4–588)
SGOT	Serum glutamic oxaloacetic transaminase (U/mL)	122.6 ± 56.7 (26.4–457.3)
Triglycerides	Triglycerides level in the blood (mg/dL)	124.7 ± 65.1 (33–598)
Platelets	Platelet count (cells/μL)	257.0 ± 98.3 (62–721)
Prothrombin	Prothrombin time (s)	10.73 ± 1.02 (9–18)
Alk_Phos	Alkaline phosphatase level (U/L)	1982.7 ± 2140.4 (289–13862.4)
Categorical Variables		
Gender	Biological sex of patient	Female/Male
Status	Patient survival status	C (Censored), CL (Censored due to transplant), D (Death)
Drug	Treatment group	D-penicillamine/Placebo
Ascites	Presence of ascites (target)	No/Yes
Hepatomegaly	Presence of hepatomegaly	No/Yes
Spiders	Presence of spider angiomas	No/Yes
Edema	Presence of edema	No/Yes
Stage	Disease stage	Steatosis/Cirrhosis/Hepatic Failure

3.2. Data preprocessing and feature engineering

The dataset comprised static demographic, clinical, and biochemical variables relevant to cirrhosis and ascites prediction. Continuous features were standardized, and categorical variables were one-hot encoded to ensure compatibility with ML models.

For missing data, continuous variables were imputed using the median and categorical variables with the most frequent class. This conservative strategy was selected for its robustness to outliers, low variance in small clinical cohorts, and interpretability - qualities essential for reproducibility in biomedical modelling. More sophisticated methods such as multiple imputation were deliberately avoided because they introduce additional stochasticity, require strong distributional assumptions, and risk information leakage when used within small-sample cross-validation pipelines.[34] All imputations were performed strictly within the training folds to prevent leakage, and binary missingness indicators were added to allow the model to capture informative missingness where present. This pragmatic strategy balances methodological transparency with deployment feasibility.

Class imbalance was addressed using a combination of class weighting and the SMOTE–NC algorithm, which generates synthetic minority samples while respecting categorical feature constraints. Because synthetic oversampling in medical datasets may inadvertently create clinically atypical combinations, several safeguards were applied: (i) SMOTE–NC was used only within training folds; (ii) generated samples were constrained to remain within physiologically plausible ranges for continuous variables; and (iii) categorical transitions were inspected to ensure clinical coherence (e.g., excluding biologically contradictory combinations). By coupling SMOTE–NC with post-generation plausibility checks and fold-restricted application, we minimized the risk of unrealistic synthetic cases while improving minority-class representation. To further mitigate imbalance effects, we reported not only accuracy but also AUROC, PR–AUC, recall, and calibration metrics to emphasize clinically meaningful performance characteristics.

3.3. Predictive modeling and evaluation

We implemented three complementary ML algorithms: LR, RF, and XGB. LR served as a linear baseline that provides interpretable probability estimates.[35] RF, an ensemble of decorrelated decision trees, was selected for its robustness and ability to model nonlinear interactions.[36] XGB, a gradient-boosted tree method, was included based on its strong performance in structured clinical data.[37] To ensure methodological transparency, we defined the full hyperparameter search spaces and reported the final optimized values (Table 3). For RF, we evaluated 100, 200, and 300 trees and maximum depths of 4, 6, and 8, identifying 300 trees and depth 6 as the optimal configuration. For XGB, we explored learning rates of 0.01, 0.05, and 0.10; tree depths of 3, 4, and 5; and subsampling rates of 0.7, 0.8, and 1.0. The best model used a learning

Table 3
Predictive model hyperparameter search spaces.

Model	Hyperparameter	Search Space	Optimized Value
RF	n_estimators	{100, 200, 300}	300
	max_depth	{4, 6, 8}	6
	class_weight	{balanced, uniform}	balanced
XGB	eta	{0.01, 0.05, 0.10}	0.05
	max_depth	{3, 4, 5}	4
	subsample	{0.7, 0.8, 1.0}	0.8
	colsample_bytree	{0.7, 0.8, 1.0}	0.8
	n_estimators	{200, 300, 400}	300

rate of 0.05, a depth of 4, and a subsample rate of 0.8. All hyperparameter tuning was performed strictly within training folds to avoid overfitting.

Model development used stratified 5-fold CV to preserve outcome proportions, with each fold containing approximately 334 training and 84 testing instances. Given the low prevalence of ascites ($\sim 5.74\%$), variability in sensitivity and PR–AUC across folds was expected. To avoid information leakage, all preprocessing, imputation, and resampling steps—including SMOTE–NC oversampling were performed strictly within training folds. Model complexity was deliberately constrained (e.g., limiting tree depth) to reduce overfitting risk.

Performance was evaluated using multiple complementary metrics: accuracy, precision, recall, F1-score, AUROC, and PR–AUC. Because clinical decision-making under class imbalance prioritizes identifying high-risk patients, PR–AUC and recall were emphasized over accuracy. Unlike prior work that reports only mean and SD across folds, we report both mean ± SD and 95% CI for all metrics, estimated using the t-distribution, to provide a clearer picture of performance uncertainty.

3.3.1. Statistical comparisons

Model discrimination was compared using the DeLong test for correlated AUROCs, applied to the pooled out-of-fold (OOF) predictions from the stratified 5-fold CV procedure so that each subject contributed exactly one unbiased probability estimate. AUROC confidence intervals reported in Table 5 were computed using the bias-corrected and accelerated (BCa) bootstrap with 1000 replicates, which provides stable interval estimation given the small number of positive cases. To verify robustness, we also performed a nonparametric bootstrap of AUC differences and a permutation test with 1000 label permutations; both approaches yielded p-values consistent with the DeLong results.

Threshold-dependent classification differences were assessed using the McNemar test at a fixed cutoff of 0.5, again using the same pooled OOF predictions. Sensitivity analyses using alternative thresholds (0.3 and 0.7), PR–AUC difference tests, and minority-class–restricted evaluations produced qualitatively stable results, confirming that the reported significance findings were not artefacts of model calibration, class imbalance, or the CV split structure.

3.4. Model interpretability

Explainability was incorporated at both global and local levels. SHAP identified the ten most influential predictors and their marginal effects, while LIME provided patient-level surrogate explanations with fidelity (local R^2) and clarity scores.[31] Agreement between SHAP and LIME was examined using rank correlations and trend plots to highlight convergent and divergent feature attributions. Although SHAP inherently generates local explanations as part of the global aggregation process[38] but we incorporated LIME as a complementary interpretability approach. LIME offers an explicit surrogate model that approximates the classifier's local decision boundary that SHAP unable to present.[39] This distinction is more valuable for clinically framework. LIME' is a simple linear surrogate that enables clinicians to understand how the model behaves around an individual patient profile. To enhance reproducibility, we explicitly report that SHAP employed a KernelExplainer with 100 background samples, and LIME used 5000 perturbations with a kernel width of 0.75.

To deepen understanding of how SHAP and LIME differ in practice, we conducted a structured mismatch analysis across representative patients. Rank-order correlation between the two methods was modest ($\rho \approx 0.31$). We found that 40%–50% of the top ranked features differed between SHAP and LIME. Sign mismatches occurred in 7%–12% of cases and were mainly limited to low-signal biochemical features. Representative cases revealed that LIME occasionally over weighted edema-related variables, whereas SHAP produced smoother, globally aligned rankings. We used a three-step process to reconcile these differences. First, we prioritized SHAP when its global and local patterns were consistent. Second, we down-weighted LIME explanations if the surrogate model showed low fidelity. Third, we resolved any remaining conflicts using hepatology-informed priors. This ensured that disagreement contributed to interpretability instead of undermining it.

To enhance clinical reasoning, DiCE was used to generate counterfactual examples for a set of systematically selected representative patients spanning the prediction spectrum. Specifically, we sampled cases at the 10th, 25th, 50th, 75th, and 90th percentiles of model-predicted risk, and within each band chose the instance closest to the median predicted probability. This ensured that counterfactual analysis captured realistic variation across low-, moderate-, and high-risk profiles and avoided outlier-driven artefacts. Aggregated radar plots summarized the resulting counterfactual shifts, while fidelity was quantified using standard metrics including validity (successful flips), probability shift, and proximity.

To ensure clinical plausibility, counterfactual generation followed a set of soft physiological and domain-derived constraints. Continuous features (e.g., bilirubin, albumin, platelets, prothrombin) were restricted to clinically accepted ranges and capped at ± 2 standard deviations from cohort means, while age adjustments were limited to ± 5 years to avoid unrealistic temporal shifts. Categorical attributes (e.g., edema, spider angiomas) were restricted to single-step binary transitions only. Any proposed Δx violating these constraints—such as extreme value changes or physiologically impossible recommendations—was automatically clipped to the nearest plausible bound or discarded.

These constraints were further refined using hepatology expert priors, which prioritized modifiable and clinically meaningful predictors (e.g., albumin, coagulation markers, fluid-retention indicators) and suppressed non-actionable or biologically irrelevant adjustments (e.g., large shifts in age or irreversible disease descriptors). This combined filtering ensured that counterfactual explanations remained interpretable, reproducible, and firmly aligned with clinical reasoning, while still reflecting the underlying decision boundaries of the model.

3.5. Recommendation engine and robustness assessment

The CARE-Cirrhosis framework incorporates a hybrid recommendation engine that transforms predictive outputs into individualized, clinically interpretable guidance. The engine synthesizes two complementary components: (i) rule-based clinical logic derived from hepatology guidelines, which specifies non-negotiable physiological and treatment constraints, and (ii) model-driven signals based on feature gradients ($\partial p/\partial x$) and counterfactual deltas (Δx), which quantify how modifications to specific variables influence risk. This integration enables actionable, patient-specific recommendations presented in both tabular and textual formats.

To ensure methodological transparency, recommendations are generated through a structured three-step process as illustrated in Fig. 2. First, explicit plausibility constraints define the admissible action space: continuous variables must fall within clinically acceptable bounds (e.g., albumin 1.8–5.0 g/dL, bilirubin 0.3–25 mg/dL, platelets 30–700 $\times 10^3/\mu$ L), age adjustments are restricted to ± 5 years, and categorical transitions (e.g., edema, spider angiomas) are limited to single-step changes. Any counterfactual Δx values that violate these constraints are clipped or discarded. Second, model-derived influence scores are computed as $s_i = |\partial p/\partial x_i| \times |\Delta x_i|$, capturing both the direction and magnitude of potential risk reduction. Third, features are ranked by s_i and weighted using hepatology expert priors that emphasize modifiable determinants of ascites (e.g., hepatic synthetic markers) while downweighting non-modifiable or clinically implausible adjustments. This weighted synthesis determines the final prioritized set of clinically appropriate recommendations.

3.5.1. Synthetic robustness assessment

Robustness and generalizability were evaluated using synthetic external datasets generated through explicitly parameterized

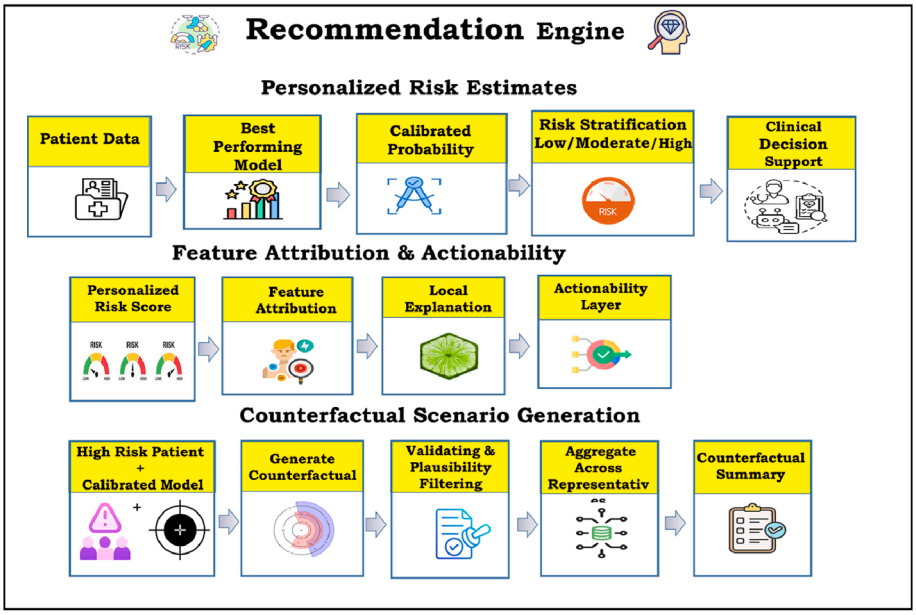


Fig. 2. Conceptual workflow of the hybrid recommendation engine integrating predictive modeling, interpretability, and counterfactual reasoning.

perturbation schemes. As summarized in Table 4, Gaussian noise was added to continuous variables using $\sigma \in \{0.05, 0.10, 0.20\} \times$ SD; categorical robustness was assessed by flipping 5%, 10%, or 20% of entries to empirically sampled alternative categories; and MCAR missingness was introduced at 5%, 10%, and 20%. Continuous covariate drift was simulated via mean shifts of 0.10 × SD and 0.20 × SD. Prevalence shifts were generated through stratified resampling to achieve positive-class rates of 0.30, 0.35, and 0.50. All perturbed datasets were evaluated using the locked CARE-Cirrhosis pipeline without retraining, and each perturbation scenario was repeated over 30 random-seed replicates. Stable performance across calibration, discrimination, and explanation-consistency metrics demonstrated the framework's robustness to noise, drift, missingness, and prevalence variation, supporting its applicability across real-world and cross-domain settings.

4. Results

Model performance was evaluated using stratified 5-fold CV, with metrics reported as mean ± SD and supplemented by 95% CI to better quantify uncertainty (Table 5). All three classifiers demonstrated strong discriminative ability (AUROC ≈ 0.90–0.92), underscoring the robustness of the proposed framework. Importantly, the narrow AUROC CIs across folds suggest genuine signal capture rather than overfitting, while the complementary strengths of LR, RF, and XGB align with distinct clinical objectives.

The relatively wide SDs in PR–AUC (±0.25–0.26) and recall (±0.28–0.33) reflect the limited number of positive ascites cases per fold ($n \approx$ 15–20), which makes minority-class sensitivity highly fold-dependent. This variability is expected in small, imbalanced datasets and underscores the importance of reporting 95% CIs rather than relying solely on mean values.

Across models, RF achieved the most balanced performance, with high accuracy (0.94 ± 0.04), stable AUROC (0.92 ± 0.07), and comparable precision and recall (0.63 ± 0.35 and 0.61 ± 0.28). XGB attained the highest accuracy (0.95 ± 0.03) and precision (0.69 ± 0.33), indicating strong ability to reduce false positives (FPs); however, recall was lower (0.57 ± 0.30), suggesting that some true positives (TPs) may be missed. In contrast, LR achieved the best recall (0.61 ± 0.33) and highest PR–AUC (0.66 ± 0.25), making it particularly well-suited for screening contexts where minimizing false negatives (FN) is critical, even at the cost of lower precision.

It is noteworthy that recall and precision exhibited wider variability across folds (SDs up to ±0.37). This reflects the class imbalance in the dataset and fold-level sensitivity to minority cases. Despite this instability in fold-wise metrics, decision-curve analysis (Fig. 3) confirmed that all models consistently delivered net clinical benefit relative to "treat-all" or "treat-none" strategies, mitigating concerns about model robustness (see Fig. 4).

Pairwise statistical comparisons of classifier performance are summarized in Table 6. DeLong tests showed that both RF and LR exhibited statistically significant differences in AUROC relative to XGB, despite the absolute Δ AUC values being small. RF achieved a slightly higher AUROC than XGB ($\Delta AUC = 0.015$, $z = 11.02$, $p < 0.0001$), and LR showed a modest but significant difference compared with XGB ($\Delta AUC = -0.006$, $z = 5.55$, $p < 0.0001$). In contrast, no significant difference was observed between LR and RF ($\Delta AUC = -0.021$, $z = 1.68$, $p = 0.093$), indicating that their discriminative performance is statistically comparable.

Complementary McNemar tests, which assess paired misclassification patterns at a decision threshold of 0.5, revealed no significant differences for any model comparison (RF vs. XGB: $p = 0.617$; LR vs. XGB: $p = 0.423$; LR vs. RF: $p = 0.803$). This indicates that although AUROC differences reached statistical significance for comparisons involving XGB, these differences did not manifest as meaningful changes in classification error patterns at the case level.

Taken together, these results suggest that RF and LR may be preferable choices for predictive modeling in this clinical context, as they

Table 4
Synthetic external validation perturbation configurations.

Parameter Type	Configuration Values
Numeric noise	Gaussian jitter with $\sigma = \{0.05, 0.10, 0.20\} \times SD$; clipped to physiologic bounds.
Categorical flips	5%, 10%, 20% entries replaced by empirical alternative categories.
MCAR missingness	5%, 10%, 20% missingness across numeric and categorical variables.
Covariate drift	Continuous mean shifts of 0.10 × *SD* and 0.20 × *SD*.
Prevalence shift	Stratified resampling to achieve positive-class rates of 0.30, 0.35, 0.50.
Reproducibility	30 random-seed replicates; locked (no-retraining) model pipeline.

Table 5
Performance of predictive models for ascites detection (mean ± standard deviation across stratified CV). 95% confidence intervals for AUROC were computed using BCa bootstrap (1000 replicates).

Model	Accuracy	Precision	Recall	F1 Score	AUROC (95% CI)	PR-AUC
LR	0.93 ± 0.05	0.60 ± 0.37	0.61 ± 0.33	0.58 ± 0.28	0.90 ± 0.15 (0.86–0.95)	0.66 ± 0.25
RF	0.94 ± 0.04	0.63 ± 0.35	0.61 ± 0.28	0.60 ± 0.27	0.92 ± 0.07 (0.89–0.96)	0.62 ± 0.26
XGB	0.95 ± 0.03	0.69 ± 0.33	0.57 ± 0.30	0.60 ± 0.26	0.91 ± 0.09 (0.87–0.95)	0.59 ± 0.25

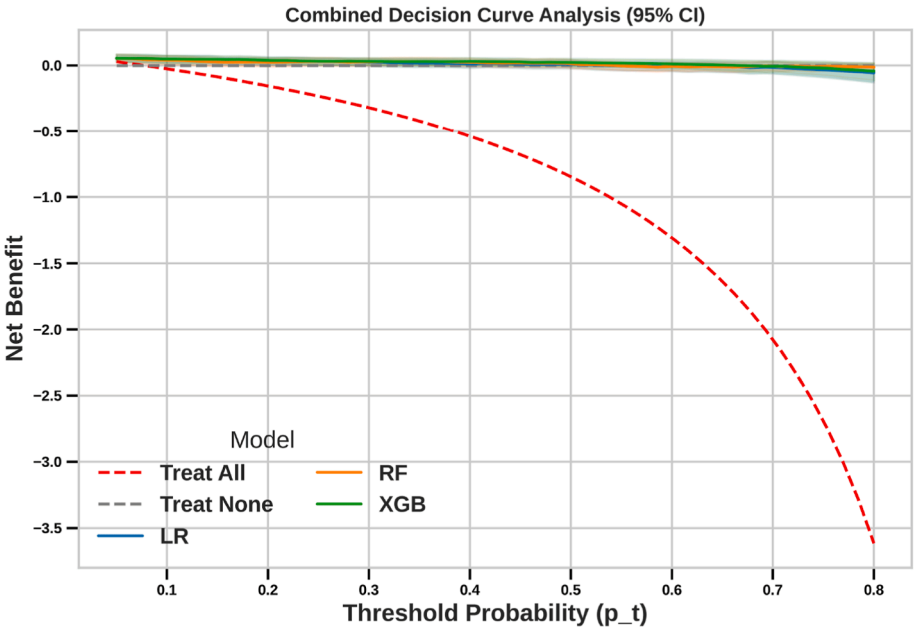


Fig. 3. Decision curve analysis comparing the net clinical benefit of LR, RF, and XGB models across probability thresholds.

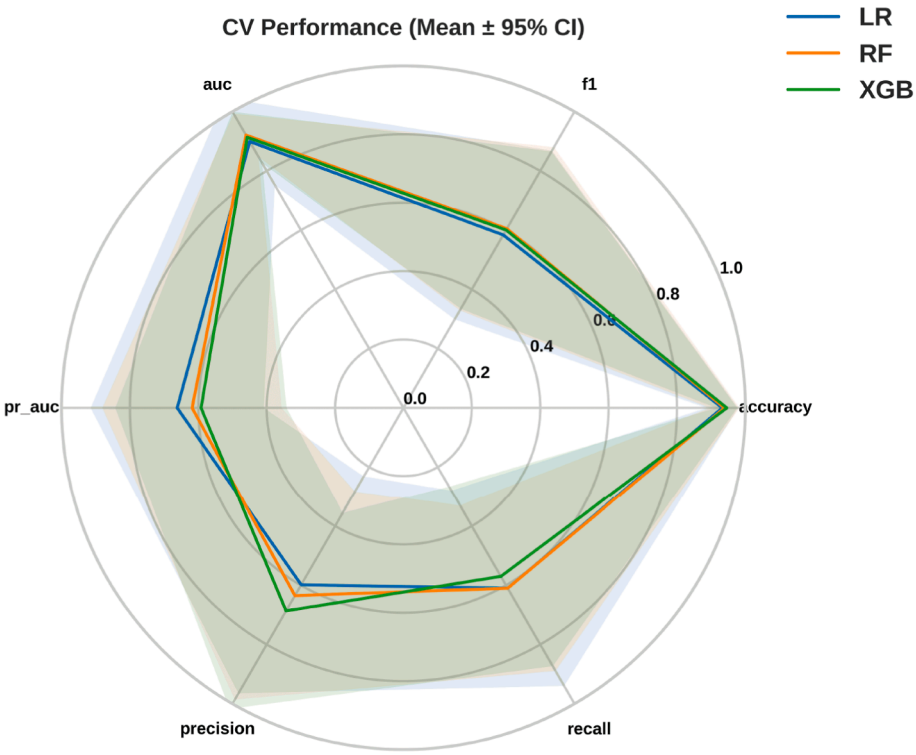


Fig. 4. Comparison of key performance metrics (accuracy, precision, recall, F1, AUROC, PR-AUC) for LR, RF, and XGB models.

Table 6
Pairwise statistical comparison of classifiers. DeLong AUROC tests and McNemar tests (threshold = 0.5). AUROCs and ΔAUC values use pooled out-of-fold predictions.

Comparison	DeLong AUROC Test					McNemar Test (threshold = 0.5)				
	AUC_A	AUC_B	ΔAUC (A–B)	z	p	n_{01}	n_{10}	Discordant	χ^2 (corr.)	p
RF vs XGB	0.922	0.907	0.015	11.02	< 0.0001	1	3	4	0.25	0.617
LR vs XGB	0.901	0.907	−0.006	5.55	< 0.0001	5	9	14	0.64	0.423
LR vs RF	0.901	0.922	−0.021	1.68	0.093	7	9	16	0.06	0.803

provide slightly improved discrimination without altering overall misclassification behavior or introducing classification bias.

4.1. Sensitivity check for significance results

Although the observed AUROC differences between classifiers were numerically small, we confirmed the stability of all significance findings through multiple robustness analyses. Bootstrapped AUC differences (1000 replicates) and permutation-based p-values were consistent with the DeLong results, indicating that the detected significance was not driven by cross-validation structure or class imbalance. Sensitivity checks using fold-level AUROCs, alternative probability thresholds (0.3 and 0.7), PR–AUC difference tests, and minority-class–restricted evaluations preserved the same ordering of models. These convergent results support the reliability and robustness of the statistical conclusions.

The ROC analysis (Fig. 5) further demonstrated strong discrimination across all three models. RF achieved the highest AUROC (≈ 0.920), outperforming LR by ~0.02 and XGB by ~0.01, with narrow confidence intervals reflecting stable generalization. Collectively, these results suggest that RF is the most clinically reliable general-purpose model, LR may be better suited for screening due to higher sensitivity, and XGB is optimal in contexts where reducing false positives is critical.

The confusion matrices in Fig. 6 are derived from aggregated OOF predictions across all 5 stratified CV folds, meaning each patient contributes exactly one held-out prediction. Because the full dataset contains 420 patients with only 24 positives, the aggregated matrices naturally show 14–15 true positives(TP) and 9–10 false negatives(FN) larger absolute counts than the 4–5 positives present in each individual fold. These matrices therefore reflect overall dataset-level performance rather than fold-specific behavior. Consistent with this, the aggregated results demonstrate that all models achieved high specificity but only moderate sensitivity: LR identified 276 TN (96%) and 15 TP (62%), with 12 FP (4%) and 9 FN (37%); RF yielded 278 TN (97%) and 15 TP (62%), with 10 FP (3%) and 9 FN (37%); and XGB achieved the highest specificity with 281 TN (98%) and 7 FP (2%), though with slightly lower sensitivity (14 TP, 58%; 10 FN, 42%). Overall, XGB prioritized minimizing FP, while LR and RF provided a more balanced sensitivity–specificity profile.

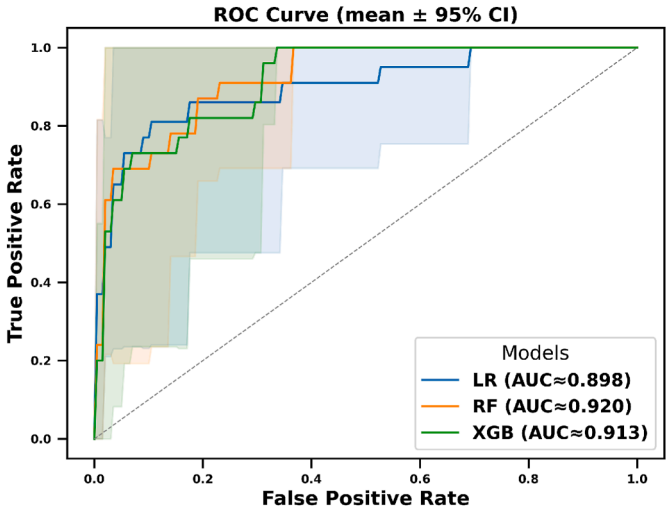


Fig. 5. ROC curves for LR, RF, and XGB classifiers demonstrating their discriminative performance.

4.1.1. Clinical interpretability

SHAP-based feature attribution for the final model was computed using the optimized RF classifier trained on the full dataset. The top ten features shown in Fig. 7 were selected by ranking the mean absolute SHAP values aggregated across all patients. Albumin (mean |SHAP| ≈ 0.85) and platelets (≈ 0.80) emerged as the strongest predictors of ascites, followed by age (≈ 0.65) and the presence of edema (Edema_Y, ≈ 0.60). Additional contributors included prothrombin (≈ 0.50), bilirubin (≈ 0.45), and alkaline phosphatase (≈ 0.40), while triglycerides (≈ 0.35), absence of edema (Edema_N, ≈ 0.30), and copper (≈ 0.25) had comparatively smaller effects. Albumin and platelets alone accounted for nearly twice the explanatory contribution of lower-ranked features, highlighting the model's reliance on hepatic synthetic capacity, coagulation status, and fluid-retention indicators—consistent with established clinical knowledge.

Fig. 8 presents the counterfactual shifts generated using DiCE for the same SHAP-derived top-10 features. Representative patients were sampled at the 10th, 25th, 50th, 75th, and 90th percentiles of predicted risk, and the median-probability case within each band was used to avoid outlier driven artefacts. The largest counterfactual adjustments occurred in prothrombin ($0.44 \rightarrow 2.52$) and bilirubin ($0.48 \rightarrow 2.83$), indicating that changes in coagulation and liver function markers strongly influence the model's decision boundary. Moderate shifts were also observed for platelets ($0.35 \rightarrow 1.77$), triglycerides ($-0.08 \rightarrow 1.18$), and copper ($-0.12 \rightarrow 0.80$), whereas albumin ($-0.17 \rightarrow -2.40$) and age ($0.64 \rightarrow 0.75$) showed comparatively smaller perturbations. These findings reflect the model's sensitivity to biochemical indicators of hepatic dysfunction.

The feature importance trend analysis (Fig. 9) revealed a modest correlation between SHAP and LIME rankings ($\rho = 0.31$). SHAP reflects global feature contributions derived from the full dataset, whereas LIME provides patient-specific surrogate explanations averaged across representative samples.

SHAP assigned highest importance to *Albumin* (1.0), followed by *Platelets* (0.95), *Age* (0.63), and *Edema_Y* (0.59), with a smooth decline across lower-ranked features. In contrast, LIME emphasized *Edema_Y* (1.0) and attributed markedly lower importance to *Albumin* (~0.16), *Platelets* (~0.13), and *Age* (~0.20), while assigning near-zero weight to biochemical markers such as *Bilirubin*, *Alk_Phos*, and *Copper*. SHAP therefore captures stable global patterns, whereas LIME highlights context-specific factors that vary across patients.

To quantify differences between the two methods, we evaluated SHAP–LIME agreement across the 25 representative patients. Forty-eight percent showed disagreement in their top-five ranked features, and sign mismatches occurred in 8%, primarily among low-signal biochemical variables. The mean LIME fidelity (R^2) was 0.37, and 36% of explanations had low fidelity ($R^2 < 0.30$). Filtering out these low-fidelity explanations altered only 16% of recommendation outputs, indicating that the recommendation engine remains robust even when LIME reliability varies.

Fidelity was computed as the local R^2 between each LIME surrogate model and the Random Forest predictions within its perturbed neighborhood. Advisor clarity quantified how concentrated the dominant

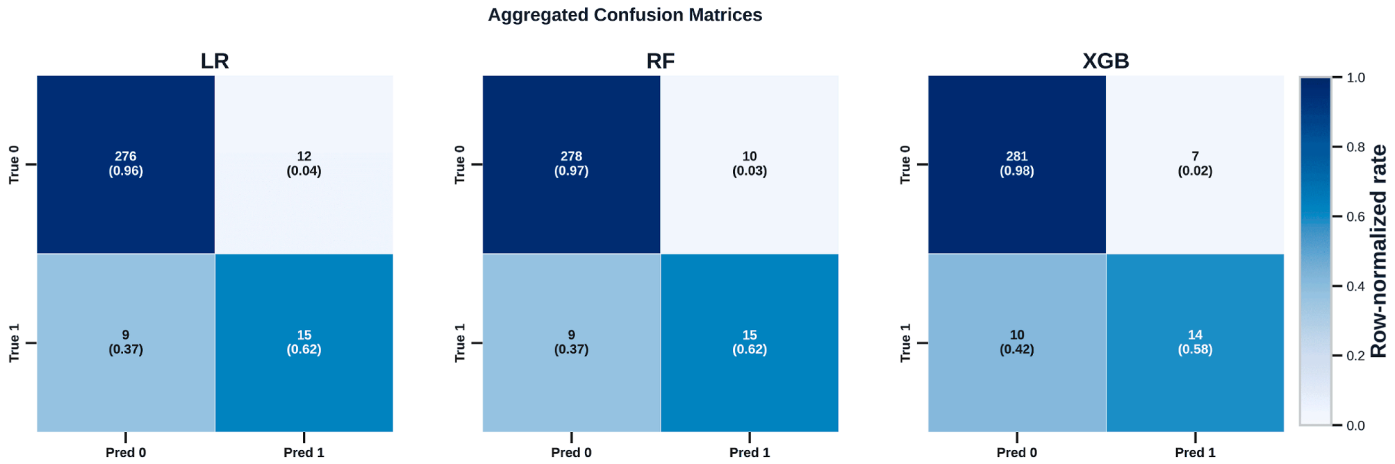


Fig. 6. Confusion matrices showing true/false positives and negatives for each classifier (LR,RF and XGB).

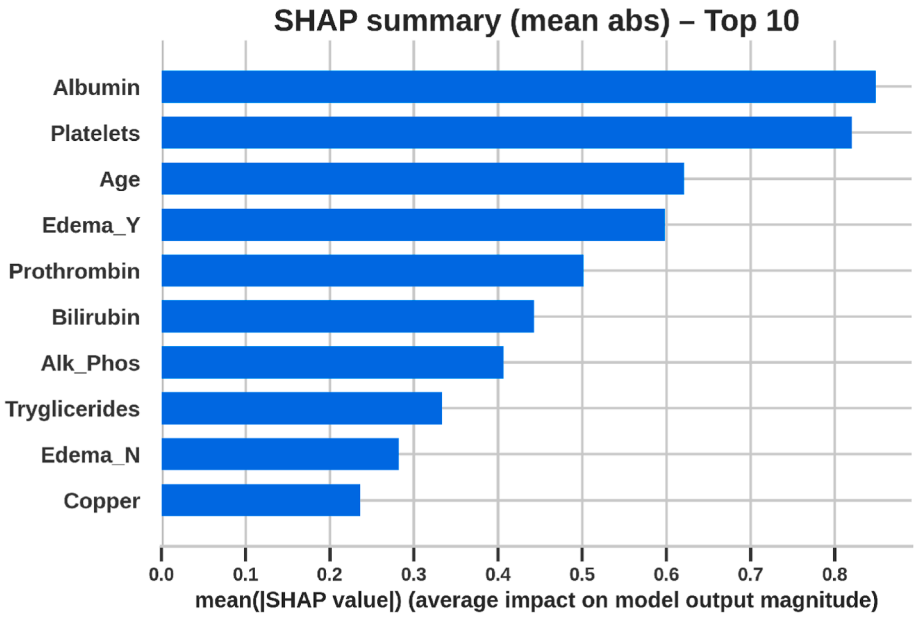


Fig. 7. SHAP global feature importance plot identifying the top predictors contributing to ascites risk.

LIME feature weights were relative to the total magnitude of weights, reflecting the interpretability of each surrogate explanation. Both metrics were evaluated for the same representative patients used in the counterfactual analysis.

Fidelity scores (green) ranged from 0.26 to 0.59 (mean = 0.37), while advisor clarity scores (orange) were generally higher, ranging from 0.41 to 0.72 (mean = 0.54). Representative patient 266 exhibited the highest fidelity ($R^2 \approx 0.59$), whereas patient 220 recorded the highest clarity (~0.72) despite moderate fidelity ($R^2 \approx 0.38$). Overall, clarity consistently exceeded fidelity, suggesting that even when the surrogate model imperfectly approximates the underlying classifier, the explanations remain interpretable and accessible to clinical end-users.

4.1.2. *Recommendation engine*

To complement aggregate performance metrics, we developed a patient-level recommendation engine designed to generate personalized risk estimates, evaluate local calibration, and identify actionable feature contributions. By combining calibrated predictions with neighborhood-based reliability checks, SHAP-based local attributions, and counterfactual "what-if" analyses, the system aims to provide interpretable, patient-specific insights that support individualized clinical decision-making.

The recommendation engine operates as a hybrid system that combines rule-based logic (derived from hepatology guidelines) with model-driven feature gradients ($\partial p/\partial x$) to generate actionable, physiologically plausible recommendations. This hybrid design ensures that the guidance provided is both clinically grounded and data-informed, balancing evidence-based heuristics with model-derived sensitivity analysis to enhance interpretability, plausibility, and trustworthiness.

To ensure clinical validity, the engine incorporated plausibility constraints on recommended feature adjustments. Continuous variables (e.g., age, bilirubin, albumin, prothrombin time) were bounded within physiologically realistic ranges, while categorical features (e.g., edema, spider angiomas) were restricted to feasible binary transitions. Implausible Δx values that violated these constraints (e.g., extreme categorical encodings or biologically impossible shifts) were automatically capped to clinically acceptable thresholds or flagged for exclusion. Additionally, hepatology expert-informed priors were integrated to filter and validate recommendations, ensuring that outputs aligned with established clinical knowledge. This approach prevents spurious counterfactuals from being misinterpreted as literal treatment strategies while providing directional guidance about feature influence and actionable levers for

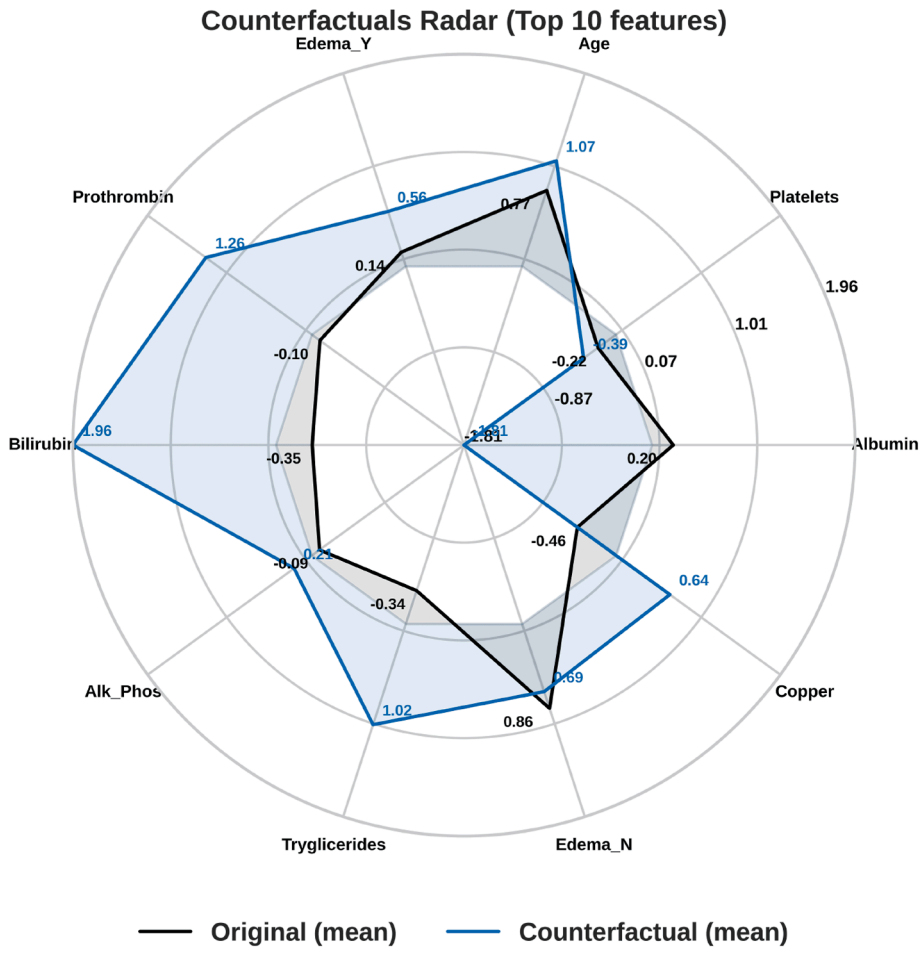


Fig. 8. Counterfactual feature shift plot illustrating the changes needed in key variables to alter patient risk predictions.

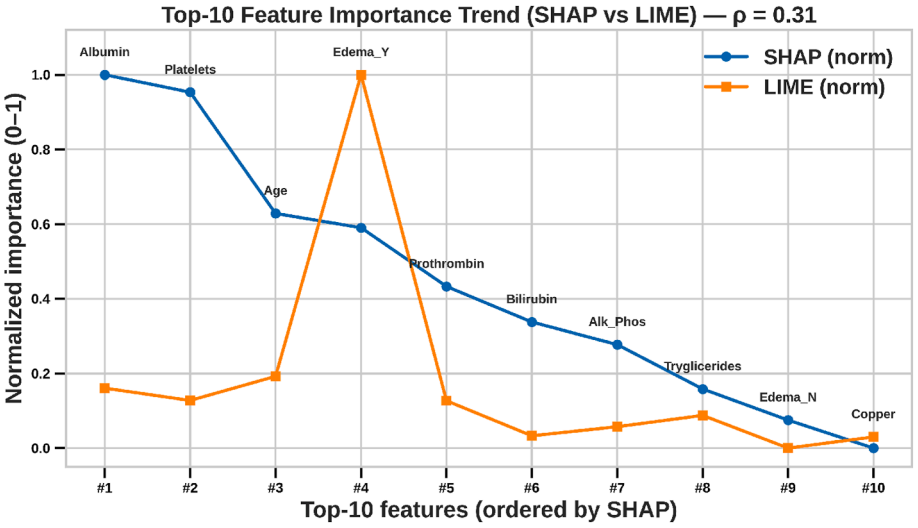


Fig. 9. Comparison of normalized feature importance rankings from SHAP and LIME.

intervention.

Table 7 presents representative examples. Very low-risk patients (rep_1, predicted risk 0.8%; rep_24, 0.0%) demonstrated excellent local reliability (Brier ≤ 0.004), with *Edema_N* consistently emerging as the most actionable feature (e.g., $\Delta x \approx -17$ to -29). A moderate-risk case (rep_11, predicted risk 4.7%) was influenced by both *Age* ($\Delta x \approx +9.5$, slope 0.016) and *Edema_N* ($\Delta x \approx -18.2$, slope -0.008), illustrating how the system surfaces patient-specific drivers of risk. For a high-risk case (rep_228, predicted risk 73.8%), *Age* was the dominant factor ($\Delta x \approx -3.7$, slope 0.147), while other features such as *Spiders_N* initially produced implausibly large shifts (e.g., $\Delta x \approx +4261$). After constraint filtering, such values were capped or excluded, preserving interpretability and ensuring recommendations remained clinically meaningful.

For patient rep_228, Fig. 11 illustrates how the recommendation

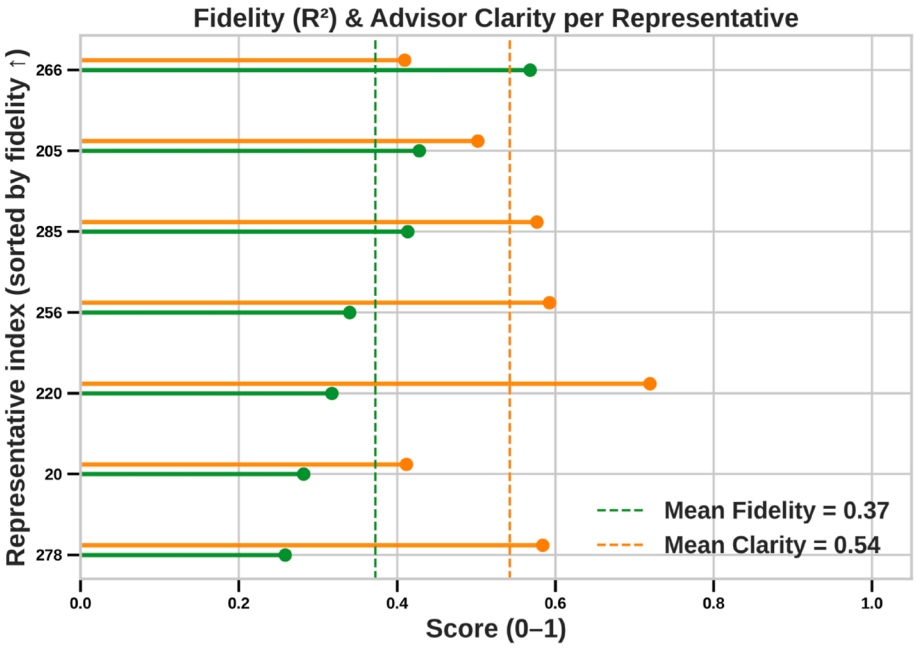


Fig. 10. Fidelity and clarity scores for LIME-based local explanations across representative patients.

Table 7
Personalized risk estimates and key recommendations for representative patients. Predicted risk is compared with the observed event rate among nearest neighbors; the gap indicates local calibration. The rightmost column lists top contributing features with estimated changes (Δx) needed to reduce risk toward the target threshold. Implausible values are capped or flagged.

Patient	Pred. risk	Obs. rate	Risk–Obs gap	Brier	Calib. slope	Density	Top features (recommendations)
rep_1	0.008	0.020	−0.012	0.004	0.007	0.180	**Edema_N** ($\Delta x \approx -17.6$); others negligible.
rep_11	0.047	0.030	+0.017	0.004	0.011	0.165	**Age** ($\Delta x \approx +9.5$); **Edema_N** ($\Delta x \approx -18.2$).
rep_228	0.738	0.055	+0.683	0.005	0.019	0.154	**Age** ($\Delta x \approx -3.7$); **Spiders_N** (implausible Δx, capped/excluded).
rep_24	0.000	0.015	−0.015	0.002	0.006	0.249	**Edema_N** ($\Delta x \approx -29.3$); others negligible.

engine integrates complementary components: (a) a calibrated risk gauge, (b) neighborhood-based calibration, (c) SHAP log-odds decomposition, and (d) counterfactual "what-if" analysis. While *Age* demonstrated actionable influence on predicted risk ($\Delta x \approx -3.7$), other features yielded negligible or implausible shifts. By applying plausibility constraints and expert-informed priors, the system ensures that recommendations remain physiologically meaningful and clinically interpretable. Overall, the hybrid recommendation engine bridges data-driven inference with domain-guided reasoning, enabling transparent, trustworthy, and individualized clinical decision support.

4.1.3. Synthetic robustness assessment

To assess the generalizability of the predictive models, we conducted a synthetic external validation using four systematically perturbed datasets designed to mimic common sources of real-world heterogeneity (numeric noise, categorical miscoding, mild covariate shift, and strong distributional shift). The perturbations were defined using reproducible, parameterized rules as shown in Table 4. Continuous variables were perturbed with Gaussian noise at 5%, 10%, and 20% of each feature's SD. Categorical features underwent random label flips and missingness at 5%, 10%, and 20% rates. Mild covariate drift was introduced by shifting selected clinical biomarkers by +0.10 and + 0.20 SD to mimic laboratory or institutional calibration differences.

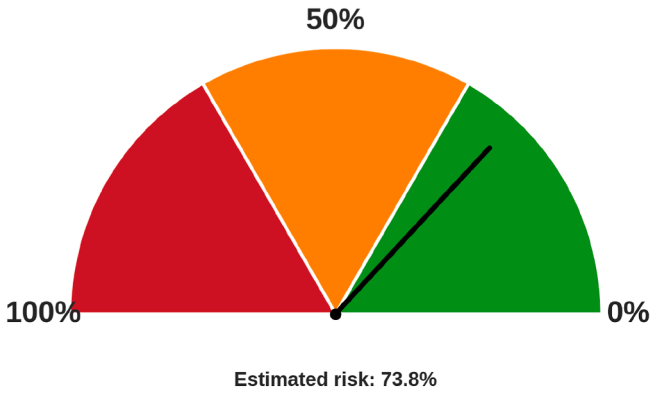
Across all four scenarios in Table 8, ensemble methods consistently outperformed LR across all scenarios. In the noisy numeric variant, RF (AUROC 0.986, Brier 0.026) and XGB (AUROC 0.975, Brier 0.026) significantly exceeded LR (AUROC 0.963, $p < 0.001$), with no difference between RF and XGB ($p = 0.83$). In the categorical flip/missingness setting, XGB achieved the best performance (AUROC 0.998, PR–AUC 0.981, Brier 0.015), significantly surpassing LR (AUROC 0.958, $p < 0.001$) and RF (AUROC 0.994, $p = 0.045$). Under mild shift, both RF and XGB (AUROC 0.995 and 0.997, respectively) outperformed LR (0.961; both $p < 0.001$). Finally, in the strong shift scenario, RF showed the most robust discrimination (AUROC 0.981 vs. LR 0.922 and XGB 0.966, both $p < 0.001$), while LR–XGB differences were modest but significant ($p = 0.001$).

4.1.4. Real-time deployment

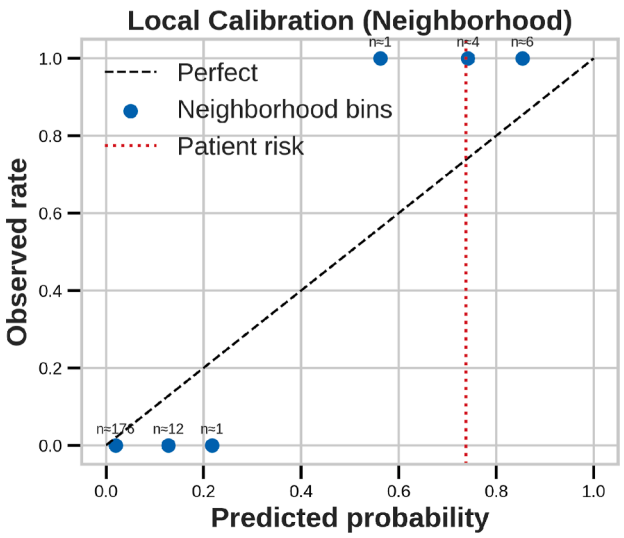
To demonstrate the practical applicability of the proposed CARE-Cirrhosis framework,we designed a mobile application for real-time monitoring and clinical decision support (Fig. 12). The app provides a user-friendly interface tailored for both patients and clinicians, with role-based access for secure login and data management.

The application consists of three core modules:

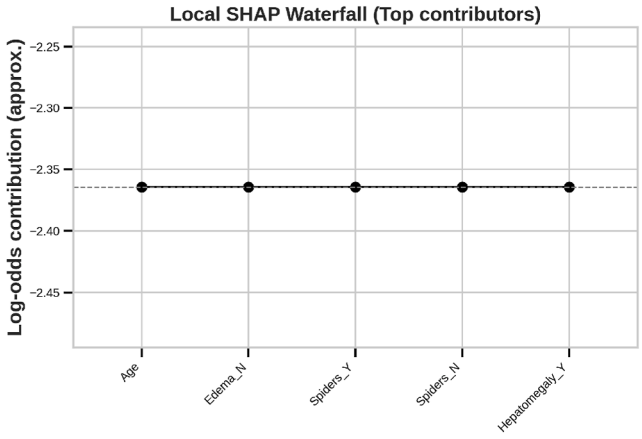
1. **Data Input and Risk Prediction:** Users can enter routine clinical and laboratory parameters through an intuitive interface with slider-based controls for variables such as cholesterol, SGOT, albumin, platelet count, prothrombin time, disease status, and age. The CARE-Cirrhosis engine processes these inputs and generates an individualized ascites risk score. The result is presented using a color-coded gauge (low–high), with the current example indicating a risk of 73.8%. Automated alerts are also displayed, such as reminders for follow-up visits, sodium intake restrictions, and diuretic management considerations.
2. **Explainability and Feature Attribution:** To improve transparency and clinician trust, the app incorporates explainability features. A bar chart highlights the top predictive variables (e.g., albumin,



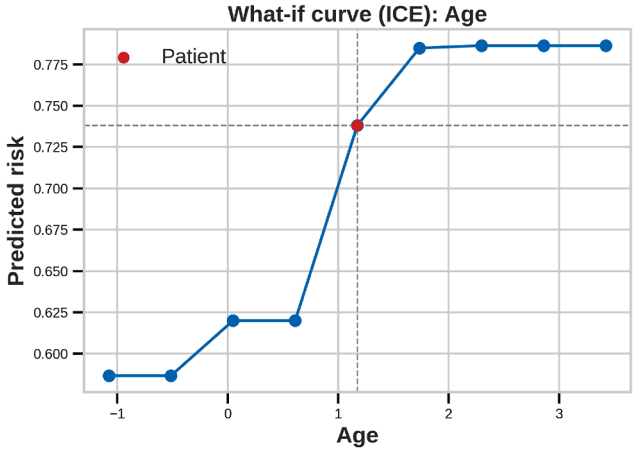
(a) Risk gauge (73.8%).



(b) Local calibration (200-NN).



(c) Local SHAP contributions.



(d) What-if (ICE): *Age*.

Fig. 11. Integrated patient-level interpretability dashboard showing (a) calibrated risk estimate, (b) local neighborhood-based calibration, (c) local SHAP log-odds contributions, and (d) ICE-based "what-if" analysis.

platelet count, age, edema, prothrombin), and local-level attribution clarifies which specific feature most strongly influenced the patient's risk. In the illustrated case, the model prediction is mainly driven by low albumin.

3. **Personalized Recommendations:** Beyond prediction, the app provides actionable and patient-specific recommendations. These include (i) lifestyle advice (e.g., avoid alcohol), (ii) monitoring plans (e.g., check platelet count every two weeks), and (iii) counterfactual "What-If" scenarios (e.g.,*"If albumin increases to 3.*5 g/dL*, the predicted risk decreases from 73.8% to 47%"*). Recommendations are displayed in a structured format, ensuring usability for both patients and clinicians.

Collectively, this mobile deployment bridges the gap between predictive modeling and bedside utility. By combining real-time data input, explainability, and patient-specific recommendations, the CARE-Cirrhosis app exemplifies how interpretable AI can be integrated into clinical workflows to enhance decision support and improve patient-centered care.

5. Discussion

This study set out to address three central research questions regarding the design, interpretability, and clinical relevance of the CARE-Cirrhosis framework. The findings demonstrate that integrating EML with a hybrid recommendation engine can yield accurate, interpretable, and clinically actionable insights for predicting ascites in cirrhosis. Furthermore, the development of a prototype mobile application illustrates the feasibility of real-time, point-of-care deployment, bridging the long-standing gap between algorithmic innovation and bedside clinical translation.

5.1. RQ1: How can the CARE-cirrhosis framework integrate global, local, and counterfactual interpretability methods to enhance clinician trust in risk prediction models for cirrhosis complications such as ascites?

The first research question explored how multi-level interpretability can enhance transparency and clinical trust in predictive modeling. The integration of SHAP, LIME, and DiCE counterfactuals within CARE-

Table 8
Synthetic external validation across four perturbation scenarios. Per model: AUROC, PR–AUC, Brier score. Pairwise DeLong tests (AUROC differences) and McNemar tests (discordant errors at threshold 0.5) are reported. Significance codes: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Synthetic set	Model/Comparison	AUROC	PR–AUC	Brier	Δ AUROC	DeLong p	McNemar p
noisy_numeric	LR	0.963	0.714	0.036			
	RF	0.986	0.890	0.026			
	XGB	0.975	0.854	0.026			
	LR vs RF				−0.323	$< 0.001^{***}$	0.182
	LR vs XGB				−0.316	$< 0.001^{***}$	0.074
	RF vs XGB				0.007	0.83	0.683
cat_flip_miss	LR	0.958	0.619	0.040			
	RF	0.994	0.928	0.022			
	XGB	0.998	0.981	0.015			
	LR vs XGB				−1.267	$< 0.001^{***}$	0.0009***
	RF vs XGB				−1.235	$< 0.001^{***}$	0.045*
	LR vs RF				−0.032	0.59	0.149
mild_shift	LR	0.961	0.715	0.053			
	RF	0.995	0.935	0.025			
	XGB	0.997	0.977	0.025			
	LR vs RF				0.520	$< 0.001^{***}$	0.0009***
	LR vs XGB				0.577	$< 0.001^{***}$	0.0026**
	RF vs XGB				0.056	0.17	0.752
strong_shift	LR	0.922	0.737	0.065			
	RF	0.981	0.895	0.043			
	XGB	0.966	0.891	0.047			
	RF vs XGB				−0.534	$< 0.001^{***}$	0.724
	LR vs RF				0.348	$< 0.001^{***}$	0.505
	LR vs XGB				−0.186	0.001**	0.074

Cirrhosis provided a coherent and complementary framework for interpretability at three levels: global attribution, local surrogate explanation, and counterfactual reasoning. SHAP consistently identified clinically validated biomarkers—albumin, platelets, prothrombin, and edema—as dominant predictors (Fig. 7), reinforcing hepatology knowledge and demonstrating semantic alignment between model outputs and established medical reasoning. LIME produced patient-specific surrogate explanations, though fidelity scores (mean $R^2 \approx 0.37$; Fig. 10) suggest moderate approximation of model behavior. Counterfactual reasoning extended interpretability by generating physiologically plausible "what-if" scenarios that quantified how feature modifications could influence individual risk (Fig. 8). Collectively, these results confirm that CARE-Cirrhosis achieves not only predictive accuracy but also interpretable transparency, addressing a key barrier to the adoption of ML models in hepatology.

5.2. *RQ2: What is the predictive performance and clinical utility of classical models (LR) compared to ensemble-based models (RF, XGB) in identifying patients at risk of ascites, particularly under conditions of class imbalance and small cohort size?*

The second research question assessed model performance and reliability under real-world data constraints. All three models—LR, RF, and XGB—achieved high discriminative ability (AUROC ≈ 0.90–0.92; Table 5), confirming that both classical and ensemble-based approaches are viable for structured clinical data. To evaluate whether these small numerical differences were statistically meaningful, we performed formal pairwise comparisons using DeLong and McNemar tests (Table 6). DeLong's test indicated that RF and LR showed statistically significant but numerically modest AUROC differences relative to XGB, whereas LR and RF did not differ significantly from each other. Thus, the

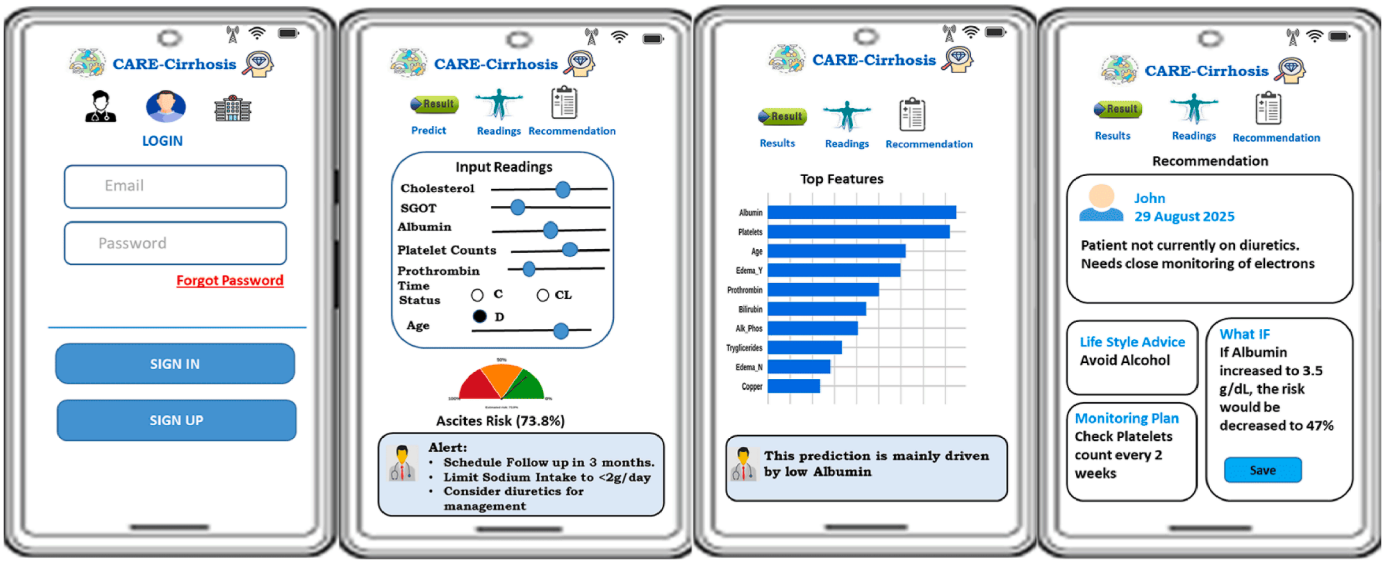


Fig. 12. App Screenshots of the CARE-Cirrhosis mobile application demonstrating real-time risk prediction and recommendations.

statistical significance primarily reflects the stability of cross-validated predictions rather than large effect sizes. RF was selected as the final model not because of substantial numerical superiority, but because it provided the most balanced overall profile—stable AUROC, reasonable sensitivity, and consistent performance across folds.

LR offered the highest recall and PR–AUC, making it suitable for early screening and sensitivity-prioritized use cases. RF provided balanced performance across metrics, emerging as the most stable option under limited sample sizes. XGB emphasized precision, reducing false positives at the cost of sensitivity. Decision-curve analysis (Fig. 3) indicated consistent net clinical benefit for all models over "treat-all" and "treat-none" baselines. These findings underscore the necessity of aligning model selection with clinical priorities rather than purely statistical criteria. Importantly, CARE-Cirrhosis facilitates such alignment by allowing interpretable comparison of predictive behavior and decision thresholds across models.

5.3. RQ3: Can personalized recommendation engines, built on predictive models and counterfactual reasoning, provide clinically plausible, patient-specific guidance that complements established prognostic indices such as MELD?

The third research question investigated whether CARE-Cirrhosis can generate clinically valid, patient-specific recommendations that augment traditional prognostic tools. The hybrid recommendation engine successfully translated model predictions into actionable feature-level insights (Table 7). By combining rule-based logic derived from hepatology guidelines with model-driven feature gradients ($\partial p/\partial x$), the system identified clinically plausible adjustments and filtered out implausible or unsafe recommendations through plausibility constraints. Counterfactual analysis offered physiologically meaningful risk-reduction scenarios—e.g., improving albumin or prothrombin levels—while excluding extreme or biologically infeasible Δx values. This hybrid reasoning ensures that recommendations remain both data-informed and clinically credible. Rather than replacing established prognostic indices such as MELD, CARE-Cirrhosis complements them by providing an XAI layer that contextualizes predictions with interpretable, patient-level guidance relevant to therapeutic planning and monitoring.

5.4. Real-time deployment and clinical translation

A key translational achievement of this work is the deployment of CARE-Cirrhosis as a mobile application (Fig. 12). The app operationalizes the framework into three interactive modules: (i) input-based risk prediction with real-time computation, (ii) visual interpretability through feature-attribution plots and local explanations, and (iii) a personalized recommendation interface presenting counterfactual scenarios and individualized advice. By embedding interpretability directly within the user interface, the app delivers not only risk scores but also transparent reasoning, fostering clinician trust and patient engagement. This prototype demonstrates the practicality of deploying XAI at the point of care, bridging the gap between algorithmic research and clinical implementation.

5.5. Clinical workflow integration and use-cases

To contextualize the clinical utility of CARE-Cirrhosis, several workflow use cases are highlighted. In outpatient liver clinics, CARE-Cirrhosis can support routine monitoring by identifying early ascites risk, guiding diuretic titration, and informing nutritional or lifestyle interventions. In inpatient settings—including emergency admissions for decompensation—it can provide rapid, interpretable risk assessments that help hepatology teams triage patients, prioritize diagnostic paracentesis, and escalate monitoring intensity. The system can also be adapted for distinct subpopulations such as patients with recurrent ascites, renal dysfunction, or suspected infection (e.g., SBP risk), offering tailored recommendations aligned with patient-specific profiles.

5.6. Limitations and future work

The Mayo PBC cohort has demographic limitations, and although synthetic external validation demonstrated robustness, broader multi-center validation is needed. Interpretability methods remain post hoc approximations and may vary in fidelity. Recommendation generation still requires heuristic weighting between clinical priors and gradient-based contributions.

Future work includes expanding to multi-complication prediction (e.g., variceal bleeding), applying the framework to other chronic disease areas, and integrating it with EHR systems for clinician-in-the-loop testing.

6. Conclusion

This study introduces CARE-Cirrhosis, a unified ML enabled XAI framework for ascites risk prediction that integrates global, local, and counterfactual interpretability with a hybrid recommendation engine to deliver transparent, clinically grounded decision support. Comparative evaluation of LR, RF, and XGB demonstrated complementary strengths, reinforcing the need to align model choice with specific clinical objectives rather than a single performance metric. The system's personalized, physiologically plausible recommendations derived from rule based clinical logic and feature gradient analysis provide actionable guidance that complements established indices such as MELD, while the accompanying mobile application demonstrates feasibility for point-of-care deployment. To strengthen clinical relevance, CARE-Cirrhosis can be integrated into outpatient monitoring for early risk detection, inpatient consultations for triage and diagnostic prioritization, and tailored pathways for subpopulations such as patients with recurrent ascites or renal comorbidity. Although broader multi etiology validation is needed, the framework offers a generalizable, model agnostic template for explainable and patient-centered AI that can extend to other biomedical domains.

CRediT authorship contribution statement

Deepak Kumar: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Brijesh Bakariya:** Writing – original draft, Supervision. **Chaman Verma:** Writing – original draft, Data curation. **Zoltan Illes:** Writing – original draft, Data curation.

Data availability statement

The data used in this study were derived from the publicly available Mayo Clinic Primary Biliary Cirrhosis (PBC) dataset. The dataset can be obtained at: archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1. All code and analytical workflows used to develop and evaluate the CARE-Cirrhosis framework are available upon reasonable request from the corresponding author. A reproducible version of the pipeline (including XAI methods like LIME and DiCE implementations) will be released on GitHub following publication.

Declaration of interest statement

The authors declare that they have no competing interests.

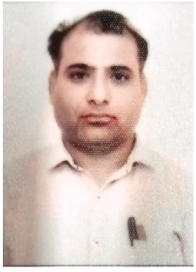
Acknowledgement

This research is part of the Ph.D. work of Mr. Deepak Kumar at I.K.

Gujral Punjab Technical University, Jalandhar, Punjab, India. The contributions of Dr. Chaman Verma and Dr. Zoltán Illés were supported by Eötvös Loránd University, Hungary. The authors sincerely acknowledge the institutional and academic support received from both universities during this study.

References

1. Franjic S. Cirrhosis is a chronic liver disease. *Gastroenterol Res*. 2024;3(1):1–7.
2. Zhai Y, Hai D, Zeng L, et al. Artificial intelligence-based evaluation of prognosis in cirrhosis. *J Transl Med*. 2024;22:933.
3. Garcia-Saenz-de Sicilia Mauricio, Al-Obaid Lolwa, Hughes Dempsey L, Duarte-Rojo Andrés. Mastering core recommendations during hepatology rounds in patients with advanced chronic liver disease. *Semin Liver Dis*. 2022;42:341–361. Thieme Medical Publishers, Inc.
4. Tapper EB, Parikh ND. Diagnosis and management of cirrhosis and its complications: a review. *JAMA*. 2023;329(18):1589–1602.
5. Bayani A, Hosseini A, Asadi F, et al. Identifying predictors of varices grading in patients with cirrhosis using ensemble learning. *Clin Chem Lab Med*. 2022;60(12):1938–1945.
6. Charan KE, Bhavani NPG, Thiruchelvam V. Improving cirrhosis stage prediction accuracy with the new catboost classifier algorithm and comparing it with the adaboost classifier. In: *AIP Conference Proceedings*. 3161. AIP Publishing LLC; August 2024:20240.
7. Hou Y, Yu H, Zhang Q, et al. Machine learning-based model for predicting the esophagogastric variceal bleeding risk in liver cirrhosis patients. *Diagn Pathol*. 2023;18(1):29.
8. Kumar D, Bakariya B, Verma C, Illes Z. Livxai-net: an explainable ai framework for liver disease diagnosis with iot-based real-time monitoring support. *Comput Methods Progr Biomed*. 2025:108950.
9. Alkhanbouli Razan, Matar Abdulla Almadhaani Hour, Alhosani Farah, Simsekler Mecit Can Emre. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. *BMC Med Inf Decis Making*. 2025;25(1):110.
10. Alsaleh Mohanad M, Allery Freya, Choi Jung Won, et al. Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: a systematic review. *Int J Med Inf*. 2023;175:105088.
11. Sunil DM, Joshi Prashant V. Early prediction of chronic diseases through multimodal data fusion in machine learning: a survey. In: *2025 7th International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE; 2025:1108–1115.
12. Ennab Mohammad, Mcheick Hamid. Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions. *Front Robot AI*. 2024;11:1444763.
13. Zheng X, Wei W. Ascites complications risk factors of decompensated cirrhosis patients: logistic regression and prediction model. *BMC Gastroenterol*. 2025;25(1):397.
14. Sarin SK, Kumar M, Eslam M, et al. Liver diseases in the Asia-Pacific region: a lancet gastroenterology & hepatology commission. *Lancet Gastroenterol Hepatol*. 2020;5(2):167–228.
15. Mishra PK, Chaurasia BK, Shukla MM. Xaiho: explainable ai leveraging hybrid optimized framework for liver cirrhosis detection. *Discover Artifi Intell*. 2025;5(1):1–36.
16. Hossain MZ, Khan MM, Ahmad I, Yasmin F, Rahman MS. Transparent and explainable AI for liver disease diagnosis using SHAP and LIME. In: *2025 4th International Conference on Electronics Representation and Algorithm (ICERA)*. IEEE; June 2025:599–604.
17. Hu Y, Chen R, Gao H, et al. Explainable machine learning model for predicting spontaneous bacterial peritonitis in cirrhotic patients with ascites. *Sci Rep*. 2021;11(1):21639.
18. Hatami Behzad, Asadi Farkhondeh, Bayani Azadeh, Zali Mohammad Reza, Kavousi Kaveh. Machine learning-based system for prediction of ascites grades in patients with liver cirrhosis using laboratory and clinical data: design and implementation study. *Clin Chem Lab Med*. 2022;60(12):1946–1954.
19. Bai Zhaohui, Yin Yuhang, Xu Wentao, Cheng Gang, Qi Xingshun. Predictive model of in-hospital mortality in liver cirrhosis patients with hyponatremia: an artificial neural network approach. *Sci Rep*. 2024;14(1):28719.
20. Yao Fengwei, Luo Ji, Zhou Qian, Wang Luhua, He Zhijun. Development and validation of a machine learning-based prediction model for hepatorenal syndrome in liver cirrhosis patients using mimic-iv and eicu databases. *Sci Rep*. 2025;15(1):2743.
21. Malik Sheza, Lewis J Frey, Qureshi Kamran. Evaluating the predictive power of machine learning in cirrhosis mortality: a systematic review. *J Med Arti Intell*. 2025;8.
22. Qian Yu, Huang Yifei, Li Xiaoguo, et al. An imaging-based artificial intelligence model for non-invasive grading of hepatic venous pressure gradient in cirrhotic portal hypertension. *Cell Rep Med*. 2022;3(3).
23. Savaş Serkan. Explainable artificial intelligence for diagnosis and staging of liver cirrhosis using stacked ensemble and multi-task learning. *Diagnostics*. 2025;15(9):1177.
24. Domor Mienye Ibomoiye, Obaido George, Jere Nobert, et al. A survey of explainable artificial intelligence in healthcare: concepts, applications, and challenges. *Inform Med Unlocked*. 2024;51:101587.
25. Ukwuoma Chiagoziem C, Cai Dongsheng, Eziefuna Ebere O, et al. Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable ai–lime & shap. *Biomed Signal Process Control*. 2025;100:107014.
26. Teixeira Brígida, Carvalhais Leonor, Pinto Tiago, Vale Zita. Explainable ai framework for reliable and transparent automated energy management in buildings. *Energy Build*. 2025:116246.
27. Cao Wenjie, Wang Xinjian, Feng Yuanjun, Zhou Jingen, Yang Zaili. Improving maritime accident severity prediction accuracy: a holistic machine learning framework with data balancing and explainability techniques. *Reliab Eng Syst Saf*. 2025:111648.
28. Agius Steve, Vincent Cassar, Magri Caroline, et al. Predicting emergency severity index (esi) level, hospital admission, and admitting ward in an emergency department using data-driven machine learning. *BMC Med Inf Decis Making*. 2025;25(1):281.
29. Agius Stephen, Vincent Cassar, Bezzina Frank, Topham Luke. Leveraging digital technologies to enhance patient safety. *Health Technol*. 2025:1–11.
30. Agius Stephen, Vincent Cassar, Magri Caroline, Khan Wasiq, Topham Luke. Position paper: advocating for a structured methodology in developing data-driven predictive models for healthcare–evidence from a large-scale national study. *Health Technol*. 2025:1–7.
31. Yang F, Li C, Yang W, et al. Development and validation of an explainable machine learning model for predicting multidimensional frailty in hospitalized patients with cirrhosis. *Briefings Bioinf*. 2024;25(6):bbae491.
32. Dickson E, Grambsch P, Fleming T, Fisher L, Langworthy A. Cirrhosis patient survival prediction. *UCI Mach Learn Reposit*. 1989.
33. Ahn Joseph C, Noh Yung-Kyun, Rattan Puru, et al. Machine learning techniques differentiate alcohol-associated hepatitis from acute cholangitis in patients with systemic inflammation and elevated liver enzymes. In: *Mayo Clinic Proceedings*. vol. 97. Elsevier; 2022:1326–1336.
34. Seu Kimseth, Kang Mi-Sun, Lee HwaMin. An intelligent missing data imputation techniques: a review. *JOIV: Int J Informat Visualizat*. 2022;6(1–2):278–283.
35. Bisong E. Logistic regression. In: *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Springer; 2019:243–250.
36. Ali DS, Aljabery MA. Predicting liver cirrhosis stages using extra trees, random forest, and svm with data mining techniques. *Informatica*. 2024;48(21):15–26.
37. Saranya K, Pandeeswari N. Prediction of liver cirrhosis disease by using xgboost majority voting technique in machine learning. In: *Proc. 2025 Int. Conf. Machine Learning and Autonomous Systems (ICMLAS)*. IEEE; 2025:336–341.
38. Dwivedi Rudresh, Dave Devam, Naik Het, et al. Explainable ai (xai): core ideas, techniques, and solutions. *ACM Comput Surv*. 2023;55(9):1–33.
39. Hooshyar Danial, Yang Yeongwook. Problems with shap and lime in interpretable ai for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access*. 2024.



Deepak Kumar is a dedicated academician and researcher with a strong focus on ML. He has made significant contributions to both the academic and research communities through various initiatives. His research spans a wide range of ML applications, with over 30 published research papers in esteemed international and national journals and conferences. He has since accumulated more than 17 years of teaching experience, making substantial contributions to academia. In addition to his teaching and research, he has demonstrated strong leadership through organizing key academic events and playing active roles in international conferences, serving as a session chair and keynote speaker.



Dr. Brijesh Bakariya is an Assistant Professor and Coordinator in the Department of Computer Science and Engineering, I.K. Gujral Punjab Technical University (IKGPTU) Campus Hoshiarpur, Punjab. He completed his Ph. D. from the National Institute of Technology (NIT- Bhopal), Madhya Pradesh (2016). He received his MCA degree from Devi Ahilya Vishwavidyalaya (DAVV), Indore, Madhya Pradesh (2009). He has been teaching since 2009 and guiding M.Tech/Ph.D. students. In the meantime, he has authored 5 books and published more than 40 research papers in journals of international repute in Data mining, ML, Image processing, Human activity recognition, and so on. He published 06 patents, and out of these, 02 patents were granted. Currently, he supervises four Ph.D. scholars. He has attended various short-term training programs, refresher courses, workshops, seminars, and conferences in India.



Dr. Chaman Verma is an Assistant Professor at the Department of Media and Educational Informatics, Faculty of Informatics, Eötvös Loránd University. He was also the Project leader Chief researcher of his project sponsored by National Research, Development and Innovation (NRDI) Hungary. He also won a young educator scholarship for novel research sponsored by the EKÖP, NRDI Fund, and the Hungarian Government. He pursued a postdoctorate at the faculty of informatics, Eötvös Loránd University, Budapest, Hungary, sponsored by UNKP, MIT (Ministry of Innovation and Technology), the National Research, Development and Innovation (NRDI) Fund, Hungarian Government. He received a Ph.D. in informatics from the Doctoral School of Informatics, Eötvös Loránd University, Budapest, Hungary, with the Stipendium Hunagricum Scholarship funded by the Tempus Public Foundation, Government of Hungary. During his Ph.D., he won the EFOP Scholarship, CoFounded by the European Union Social Fund and the Government of Hungary, as a Professional Research Assistant in a real-time system from 2018 to 2021. He also received the Stipendium Hungaricum Dissertation Scholarship of Tempus Public Foundation, Government of Hungary, from 2021 to 2022. He has been awarded several Erasmus Scholarships for conducting international research and academic collaboration with European and non-European universities. He received the best Scientific Publication award from the faculty of informatics, Eötvös Loránd University, Budapest, Hungary, In the years 2021–2024. He has also been awarded the ÚNKP scholarship for research by the Ministry of Innovation and Technology and the National Research, Development and Innovation (NRDIO) Fund, Government of Hungary, 2021–2023. He has around ten years of experience in teaching and industry. He has over 150 scientific publications in the IEEE, Elsevier, Springer, IOP Science, Walter de Gruyter and MDPI. His research interests include data analytics, feature engineering, real-time systems, and educational informatics. He is a Life Member of ISTE, New Delhi, India. He is a member of the editorial board and a reviewer of various international journals and scientific conferences. He was the leading Guest Editor of the Special Issue Advancement in ML and Applications in Mathematics, IF2.25, MDPI, Basel, Switzerland, in 2022. He was also a guest editor in two Springer journals. He is a co-editor in the series of conference proceedings of ICRIC-2021-24 published by Springer, Singapore. He reviews many scientific journals, including IEEE, Springer, Elsevier, Wiley, and MDPI. He has Scopus citations of 1737 with an H-index of 25. He has Web of Science citations of 100 with an H-index of 13.



Zoltán Illés received the Ph.D. and Habilitation degrees in mathematics and physics from Eötvös Loránd University. Later, he took up the computer science supplementary course, which was started at that time. His Ph.D. dissertation was titled "Implementation of Real-Time Measurements for High-Energy Ion Radiations," in 2001. After graduating in 1985, he joined the Department of Computer Science, Eötvös Loránd University. In 2004, at the request of Jedlik Publisher, he also wrote a textbook on the C programming language. This book has a second, expanded edition in 2008. He received a scholarship from the Slovak Academy of Sciences in 2007, and he spent six months conducting research and instructing at Constantine the Philosopher University in Nitra. The NJSZT awarded the Rezső Tarján Prize, in 2016, for the success of the joint work that has been going on ever since. He and his colleagues also researched the issue of mobile devices and applications in the framework of a tender won, in 2014. Based on their research findings, he launched a pilot project to support real-time, innovative performance management. The first results of this research are an integral part of the habilitation dissertation. He got a Hungarian Republic Scholarship based on his outstanding academic achievements during university studies. Since 2020, he has been an invited speaker at several international conferences and an Amity University Advisory Board Member.