

A Protocol for Leakage Resistant and Externally Validated Liver Disease Screening Using Routine Clinical Data

Abhishek Maurya

Institute of Technology and Management, Lucknow

sabhimaurya30062003@gmail.com

Abstract - Liver disease screening models are often evaluated as classifier comparisons rather than as clinical prediction systems. The supplied literature shows two related problems. A 2020 study evaluated six algorithms on 583 records from the Indian Liver Patient Dataset using ten-fold cross-validation and reported accuracy, kappa, mean absolute error, and execution time, but did not report external validation, calibration, confidence intervals, or decision utility. A 2019 corrigendum to another liver classification study corrected the placement of false-positive and false-negative counts and consequently revised multiple metrics. A clinical review also describes substantial etiologic and geographic heterogeneity in liver disease. This paper proposes a reproducible protocol for developing and validating a routine-data screening model across sites and time. The design uses patient-level separation, nested cross-validation, fold-contained preprocessing and feature selection, class-sensitive learning, probability calibration, temporal testing, external validation, subgroup analysis, and an automated confusion-matrix audit. Logistic regression, elastic-net logistic regression, random forest, and gradient-boosted trees are compared using discrimination, calibration, threshold performance, and decision-curve net benefit. No patient-level data or experimental outputs were supplied; therefore, numerical results are not reported. The contribution is a validation and reporting framework that directly addresses the limitations supported by the reviewed documents without claiming unobserved performance.

Keywords - calibration, clinical prediction, external validation, liver disease, machine learning, reproducibility

TABLE I
EVIDENCE MATRIX FOR THE SUPPLIED DOCUMENTS

Source	Objective	Data	Method	Metrics and reported evidence	Limitations relevant to the gap
Williams, 2006 [1]	Describe global liver-disease challenges and interactions.	Narrative clinical review; no prediction cohort.	Synthesis of epidemiologic and clinical literature on viral hepatitis, HIV coinfection, steatosis, alcohol, cancer, and transplantation.	No classifier metrics. The review documents etiologic, population, and access-related heterogeneity.	Not a prediction study; evidence and terminology reflect 2006. It supports the need to test transportability rather than a specific algorithm.
Singh et al., 2020 [2]	Predict liver-disease status and build a web interface; compare feature selection with all features.	ILPD, 583 records: 416 disease and 167 no disease; ten predictors plus class.	CFS with greedy stepwise search; logistic regression, naive Bayes, SMO, IBk, J48, random forest; ten-fold cross-validation in WEKA.	Highest tabled accuracy after selection: logistic regression 74.36%; random forest 71.87%. Kappa, mean absolute error, and run time were also reported.	Single small imbalanced dataset; no external or temporal test, calibration, uncertainty, decision utility, subgroup analysis, or complete tuning details. The narrative incorrectly attributes 74.36% to random forest, while Table 3 assigns it to logistic regression.
Abdar et al., 2019 [3]	Correct metric reporting in an earlier classification paper.	The corrigendum does not restate the underlying cohort or sampling design.	Corrects the locations of FP and FN in the confusion matrix and updates dependent values for boosted C5.0 and CHAID.	Corrected test accuracy: 93.75% for boosted C5.0 and 65.00% for CHAID; corrected sensitivity and specificity were 97.52% and 82.05% for boosted C5.0.	A corrigendum, not a stand-alone validation study. It reports no confidence intervals and demonstrates the need for auditable metric definitions.

I. INTRODUCTION

Liver disease is not a single clinical entity. The supplied clinical review describes interacting viral, metabolic, alcohol-related, malignant, and health-system challenges, with differences across countries and patient groups [1]. Although the quantitative burden estimates in that review are historical, its central implication remains relevant to prediction research: a model trained in one cohort can encounter different etiologies, prevalence, referral patterns, and laboratory distributions elsewhere.

The empirical literature supplied for this manuscript evaluates liver-disease prediction primarily by comparing classification algorithms. Singh et al. used the Indian Liver Patient Dataset (ILPD), which contains 583 records, including 416 records labelled as liver disease and 167 labelled as no disease [2]. They compared six classifiers with and without correlation-based feature selection under ten-fold cross-validation. Their Table 3 reports 74.36% correctly classified instances for logistic regression after feature selection. The same paper also contains a reporting inconsistency because its narrative attributes that value to random forest, which Table 3 reports at 71.87%.

The accompanying corrigendum by Abdar et al. is methodologically important because it corrects an interchange of false-positive and false-negative positions and revises the derived sensitivity, specificity, precision, error rates, F1 score, and accuracy values [3]. This correction shows that apparent model performance can be altered by a basic label-orientation error. A trustworthy clinical prediction workflow therefore requires both valid study design and reproducible metric computation.

The proposed study addresses the strongest gap supported by the reviewed documents: the absence of a leakage-resistant, calibrated, externally validated, and auditable evaluation framework for liver-disease screening across heterogeneous clinical settings. The study is designed to answer three questions. First, how well do routinely available laboratory models transport across time and sites? Second, does etiologic context add clinically useful information beyond the ten ILPD-style predictors? Third, do calibration, decision utility, and subgroup stability support the same model choice as discrimination alone?

The proposed contribution consists of four linked elements: a two-tier predictor design that separates a portable core laboratory panel from contextual risk variables; patient-, time-, and site-separated validation; evaluation of discrimination, calibration, and clinical utility with uncertainty; and a metric audit that regenerates every confusion-matrix statistic from saved predictions. These elements extend the reviewed studies at the level of evaluation and reproducibility. No claim is made that an untested algorithm is superior.

II. RELATED WORK AND LITERATURE REVIEW

A. Clinical Heterogeneity and the Prediction Target

Williams describes liver disease as a set of interacting public-health and clinical problems rather than a homogeneous endpoint [1]. The review discusses HBV and HCV infection, HIV coinfection, steatosis and obesity, alcohol exposure, malignancy, transplantation, and inequalities in access to prevention and treatment. These mechanisms can modify disease prevalence, laboratory values, and the consequences of a false-negative or false-positive screening result. The review does not evaluate a predictive model, but it provides clinical motivation for etiologic subgroup analysis and geographic validation.

B. Feature Selection and Classifier Comparison

Singh et al. evaluated logistic regression, naive Bayes, sequential minimal optimization, instance-based learning, J48, and random forest [2]. Their feature-selection pipeline used a correlation-based subset evaluator with greedy stepwise search and retained total bilirubin, direct bilirubin, alkaline phosphatase, alanine aminotransferase, and aspartate aminotransferase. This reduced feature set improved the tabled accuracy of logistic regression from 72.50% to 74.36% and reduced reported execution time from 0.06 s to 0.01 s. The gain is modest and is not accompanied by a confidence interval or a paired statistical comparison.

The ILPD class distribution is imbalanced: a majority-class rule that predicts disease for every record would be correct for approximately 71.4% of cases. This baseline is close to several reported accuracies, and the paper itself reports a kappa of zero for SMO despite 71.35% correct classification [2]. Accuracy alone therefore cannot distinguish useful minority-class performance from prevalence-driven correctness. Sensitivity, specificity, predictive values, precision-recall behavior, and calibration are necessary to interpret screening performance.

C. Metric Integrity and Reporting

The corrigendum by Abdar et al. corrects the placement of false positives and false negatives in a confusion matrix and updates the dependent metrics [3]. The corrected table reports high sensitivity for boosted C5.0 but substantially lower specificity, while CHAID has very low specificity. These differences would be obscured by a single accuracy value. The corrigendum does not provide uncertainty intervals or restate the original sampling and validation procedure, so it cannot establish external validity. It does, however, supply direct evidence that metric definitions and class orientation require automated checks.

Table I synthesizes the objectives, data, methods, reported evidence, and limitations of all supplied documents. The synthesis treats the 2019 item as a correction notice rather than as an independent experiment. The wider methodological plan follows transparent prediction-model reporting and risk-of-bias principles [4], [5].

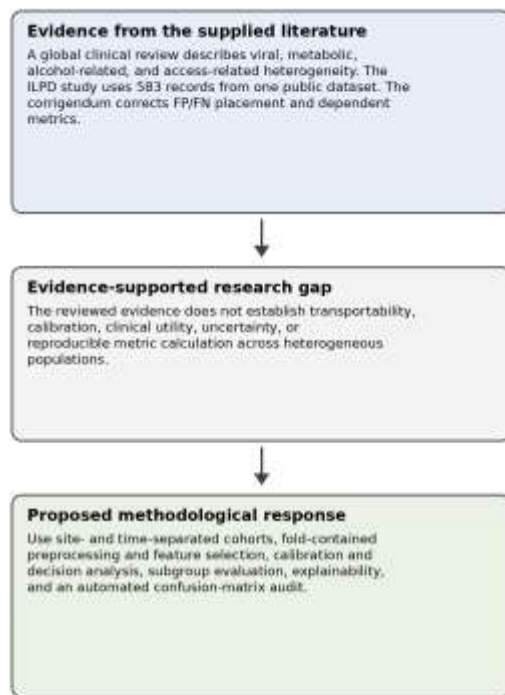


Fig. 1. Evidence-to-design map. Source-derived observations are shown separately from the proposed methodological response.

D. Research Gap and Motivation

1) Strongest Evidence Supported Gap

Within the reviewed literature, the central unresolved problem is not the absence of another classifier. It is the absence of evidence that a liver-disease screening model remains accurate, calibrated, clinically useful, and reproducible outside the single dataset on which it was developed. The ILPD experiment uses one small cohort and reports cross-validated aggregate performance, while the clinical review documents heterogeneity that can produce distribution shift. The corrigendum adds a reproducibility concern by showing that confusion-matrix orientation can change several derived metrics.

2) Consequences for Study Design

The gap has four design consequences. First, all preprocessing, feature selection, and tuning must occur within resampling folds to prevent information leakage. Second, model selection must use development data only, followed by temporal and external validation. Third, evaluation must include calibration and decision utility as well as discrimination. Fourth, raw predictions and labels must be retained so every derived metric can be independently recomputed.

3) Prespecified Research Questions

RQ1 asks whether any candidate model improves discrimination and calibration over logistic and majority-class baselines on temporal and external test sets. RQ2 asks whether a context-augmented feature tier improves out-of-sample performance over the core ILPD-style tier. RQ3 asks whether the selected model maintains clinically relevant sensitivity and calibration across sex, age, site, and etiologic subgroups. RQ4 asks whether the metric audit reproduces every reported confusion-matrix statistic exactly from stored predictions.

III. METHODOLOGY

A. Study Design and Population

The proposed investigation is a retrospective, multicenter diagnostic prediction study followed by locked temporal and external validation. Each participating site will contribute consecutive adult encounters in which a routine liver panel was obtained because liver disease was suspected or assessed. The index time is the first eligible laboratory panel in the study period. Only one index encounter per patient will enter each evaluation set, and patient identifiers will be used solely to prevent cross-split duplication.

The primary target is clinically confirmed liver disease within 30 days of the index panel. Two hepatology-trained reviewers will independently adjudicate the outcome using a prespecified evidence hierarchy that may include serology, imaging, elastography, pathology, specialist diagnosis, and follow-up records. Reviewers will be blinded to model predictions, and disagreements will be resolved by a third reviewer. Secondary labels will group cases as viral, alcohol-related, steatotic or metabolic, and other when evidence permits. These subgroup labels support transportability analysis and will not replace the primary binary endpoint.

The intended use is risk stratification at the time routine laboratory results become available. The model is not designed to establish etiology, replace confirmatory testing, or recommend treatment. Its output is a probability that can support a decision to review the record, repeat testing, or refer the patient for further assessment. This intended use determines the evaluation priorities. False negatives may delay assessment, whereas false positives may increase follow-up burden. Thresholds will therefore be selected from development data only and examined across a clinically specified range rather than optimized solely for overall accuracy. The final report will state the target population, prediction time, outcome window, and anticipated user so that performance is interpreted within a defined clinical workflow.

B. Data Sources and Predictor Tiers

The development cohort will use earlier calendar periods from selected sites. A later nonoverlapping period from the same sites will form the temporal test cohort. At least one separate institution, not used for model selection or calibration, will form the external cohort. Sample-size targets will be fixed before extraction by simulation to achieve prespecified precision for the primary AUROC and calibration-slope estimates. Cohort flow, exclusions, missingness, and outcome prevalence will be reported separately by site and split.

Before pooling data, each site will map local laboratory names and units to a common data dictionary. The mapping will distinguish measured values from derived ratios and will retain the original unit and assay metadata for audit. Reference ranges will be described but will not be used to normalize test cohorts with information unavailable at the prediction time. Site-specific extraction scripts will produce the same schema and validation report, including counts, missingness, ranges, and duplicate identifiers. Any unresolved unit conversion or implausible distribution will block that variable from model development until the source site confirms the mapping. This procedure is intended to separate genuine population shift from preventable data-engineering errors.

TABLE II
PRESPECIFIED COHORT FIELDS AND PREDICTOR TIERS

Tier	Variables	Purpose
Core laboratory	Age; sex; total and direct bilirubin; alkaline phosphatase; ALT/SGPT; AST/SGOT; total protein; albumin; albumin-globulin ratio.	Reproduce the routinely available ILPD-style information examined in [2].
Context augmented	Core variables plus body-mass index, diabetes status, documented alcohol exposure, HBV status, HCV status, and referral setting when available.	Test the incremental value of clinical context motivated by the heterogeneity in [1].
Outcome and groups	Adjudicated binary outcome; site; calendar period; age band; sex; etiologic group.	Define the target and enable external, temporal, and subgroup evaluation.

TABLE III
CANDIDATE MODELS AND OPTIMIZATION PLAN

Model	Prespecified search space	Probability handling
Logistic reference	All core predictors; no feature selection; balanced and unweighted variants.	Maximum-likelihood probability; recalibration assessed on development data only.
Elastic net	C: 0.01, 0.1, 1, 10; l1 ratio: 0, 0.25, 0.5, 0.75, 1.	Platt or isotonic calibration selected within inner resampling.
Random forest	500 trees; depth: unrestricted, 5, 10; minimum leaf: 1, 5, 10; max features: square root or 0.5.	Raw vote probability followed by fold-contained calibration.
Gradient boosting	Learning rate: 0.03 or 0.10; leaf nodes: 7, 15, 31; minimum leaf: 10, 20, 40; L2: 0, 1, 10; at most 300 iterations.	Early stopping and calibration fitted inside the inner loop.

C. Preprocessing and Leakage Control

Range checks and unit harmonization will be completed before modelling with all transformations recorded in a data dictionary. Implausible values will be queried or set to missing under prespecified rules; they will not be silently truncated. Duplicate patients and encounters will be removed before splitting. Skewed laboratory variables will be transformed with $\log(1 + x)$ when their domain permits. Median imputation, missingness indicators, standardization, categorical encoding, and any feature-selection operation will be learned on the training portion of each inner fold and then applied unchanged to its validation portion.

Complete-case analysis will be secondary because missingness can reflect clinical workflow. The primary analysis will use training-fold median imputation plus missingness indicators. A sensitivity analysis will compare this approach with multiple imputation performed without access to external test outcomes. Class weighting will be tuned within the inner loop. Synthetic oversampling will not be used in the primary analysis because its placement and parameterization can add another leakage pathway; it may be evaluated as a prespecified ablation within training folds only.

Leakage checks will be applied at three levels. At the record level, no patient or duplicate encounter may appear in more than one split. At the feature level, every predictor must be available at the index time; downstream diagnoses, procedures, and follow-up measurements are excluded. At the pipeline level, imputation values, transformation parameters, selected features, class weights, calibration functions, and thresholds are estimated without access to the relevant held-out outcomes. Automated tests will compare patient identifiers across splits and will fail the run if any overlap is detected. A separate analyst will review the generated split summary before model training begins.

D. Models and Hyperparameter Optimization

The candidate set is intentionally compact. It includes an unpenalized logistic reference, elastic-net logistic regression, random forest, and histogram-based gradient-boosted trees. This set tests whether nonlinear ensembles add value beyond an interpretable linear model. Hyperparameters are selected by inner cross-validation using mean average precision, with AUROC and calibration error as secondary selection criteria. If two configurations differ by less than one standard error, the simpler configuration is selected.

E. Validation Strategy

Development uses stratified, patient-grouped five-fold outer cross-validation. Each outer training fold contains a five-fold inner loop for preprocessing choices, feature selection, hyperparameter tuning, class weighting, and calibration. Five fixed random seeds will assess resampling stability. After the analysis specification is locked, the final pipeline will be refit on all development data and evaluated once on the temporal cohort and once on each external site. No threshold, feature, or hyperparameter will be changed after external outcomes are inspected.

The temporal cutoff will be chosen before outcome inspection and will reflect a meaningful change in calendar time rather than an arbitrary random partition. External sites will be evaluated separately before any pooled estimate is calculated. This preserves evidence about site-level variation that a pooled metric could hide. If calibration differs by site, the primary report will show the unrecalibrated result first. Any local recalibration will be a secondary analysis fitted only on a designated calibration subset and tested on a separate local subset. The same locked feature definitions and model coefficients will be used for all primary external evaluations.

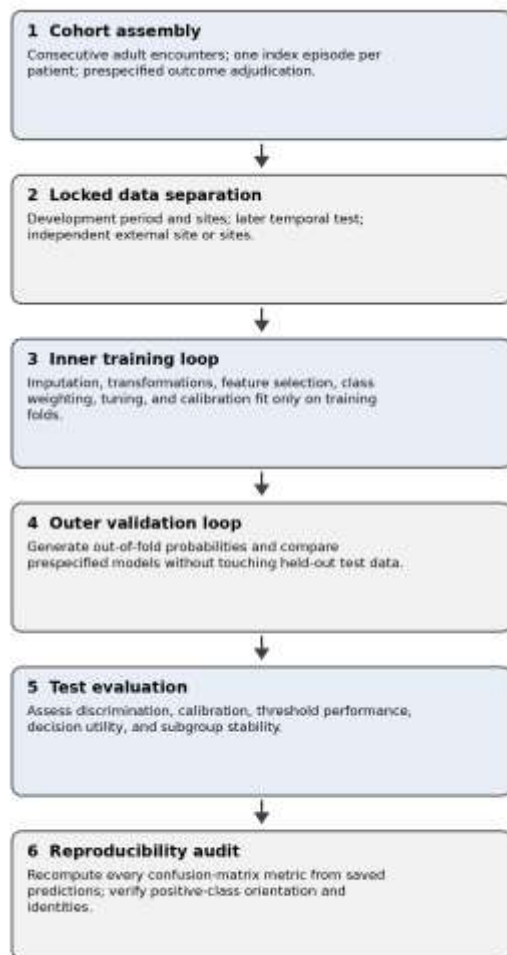


Fig. 2. Proposed validation pipeline. Every learned transformation remains inside the relevant training data, and the external cohort is untouched until the pipeline is locked.

F. Evaluation Metrics and Statistical Analysis

The primary discrimination metrics are AUROC and area under the precision-recall curve (AUPRC). Precision-recall results will be reported with disease and no-disease treated in turn as the positive label because the ILPD class distribution is imbalanced [2]. Threshold-independent metrics will be accompanied by sensitivity, specificity, positive and negative predictive values, F1 score, balanced accuracy, and confusion counts at two development-locked operating points: the threshold that maximizes net benefit over the prespecified clinical range and the lowest threshold that attains the prespecified sensitivity target in development data.

Calibration will be assessed by the calibration intercept, calibration slope, Brier score, and calibration plots. Clinical utility will be evaluated with decision-curve net benefit [7]. Ninety-five percent confidence intervals will be obtained by 2,000 patient-level bootstrap resamples within each held-out cohort. Paired AUROCs will be compared with the DeLong procedure [6]. Differences in AUPRC, Brier score, and net benefit will use paired bootstrap intervals, while thresholded error patterns will use McNemar tests. Holm correction will control the family-wise error rate for comparisons against the logistic reference. Effect estimates and confidence intervals will be emphasized over isolated p values.

TABLE IV
EVALUATION AND STATISTICAL ANALYSIS PLAN

Question	Measure or test	Required report
Discrimination	AUROC; AUPRC; paired DeLong or bootstrap difference.	Estimate, 95% CI, comparator difference, adjusted p value.
Calibration	Intercept; slope; Brier score; calibration curve.	Overall and by site, with uncertainty and evidence of recalibration.
Operating point	Sensitivity; specificity; PPV; NPV; F1; balanced accuracy; confusion matrix.	Threshold chosen on development data and exact counts on each test cohort.
Clinical utility	Decision-curve net benefit across prespecified thresholds.	Model, test-all, and test-none strategies.
Transportability	Performance by time, site, sex, age band, and etiologic group.	Interaction or paired difference with CI; subgroup size and prevalence.

G. Explainability and Metric Audit

For linear models, standardized coefficients and partial dependence on the log-odds scale will be reported with stability across outer folds. For tree ensembles, SHAP values will summarize global feature influence and local contributions [8]. Explanations will be computed on held-out observations and checked for consistency across sites. They will be presented as descriptive model behavior, not as causal effects.

The metric audit will store the prediction probability, locked threshold, predicted class, true class, cohort identifier, and model version for every held-out observation. A separate script will reconstruct TP, TN, FP, and FN after confirming which label is positive. It will then assert the identities $\text{sensitivity} = \text{TP}/(\text{TP} + \text{FN})$, $\text{specificity} = \text{TN}/(\text{TN} + \text{FP})$, $\text{false-positive rate} = 1 - \text{specificity}$, $\text{false-negative rate} = 1 - \text{sensitivity}$, $\text{precision} = \text{TP}/(\text{TP} + \text{FP})$, and $\text{F1} = 2 \times \text{precision} \times \text{sensitivity}/(\text{precision} + \text{sensitivity})$. Report generation will stop if any value differs beyond numerical tolerance. This control directly addresses the failure documented in [3].

H. Ethics and Reproducibility

Each site will obtain the required ethics and data-governance approvals before extraction. Analyses will use de-identified data in an access-controlled environment. The final study will publish the protocol, data dictionary, eligibility logic, preprocessing code, model configurations, seeds, metric tests, and a synthetic schema-compatible example. Patient-level clinical data will not be released unless consent and governance permit. Reporting will follow TRIPOD principles [4], and the analysis will be checked against PROBAST domains [5].

I. Experimental Setup

1) Planned Software and Execution Controls

The experiment will be implemented in Python using a pipeline API that binds preprocessing and modelling within each resampling fold. The exact interpreter, library versions, operating system, hardware, and commit identifier will be frozen before the first locked run and reported with the final results. Each model will use the same outer splits, and all probability predictions will be written before metrics are calculated. A configuration file will contain cohort dates, variable mappings, hyperparameter grids, seeds, and metric definitions so the analysis can be repeated without editing source code.

2) Cohort Construction and Quality Checks

A CONSORT-style cohort flow will report initial encounters, exclusions, duplicate removal, unresolved outcome labels, and final counts by split. Before fitting, analysts will compare

missingness, predictor distributions, and prevalence across development, temporal, and external cohorts without examining model performance. Large distribution shifts will be documented rather than repaired using external outcomes. Label adjudication agreement will be reported before consensus resolution.

3) Prespecified Ablations

Ablation analysis will isolate the contribution of context variables, fold-contained feature selection, class weighting, calibration, and missingness indicators. Each ablation will change one component while preserving the same splits and random seeds. The primary comparison is the locked full pipeline versus the core logistic reference on the external cohort. Secondary analyses examine whether the feature subset used in [2] improves or degrades transportability when selection is reproduced within the inner loop.

Error analysis will begin only after the primary metrics are frozen. Reviewers will summarize false negatives and false positives by site, etiologic group, missingness pattern, and proximity to the decision threshold. Case review will look for systematic data problems or target-definition ambiguity, but it will not be used to retune the reported model. Sensitivity analyses will vary the outcome window, exclude records with unresolved adjudication, and compare the primary imputation approach with the complete-case subset. Results from these analyses will be labelled exploratory and reported beside, rather than substituted for, the prespecified primary evaluation.

IV. RESULTS

A. Results Status

No patient-level dataset, executable experiment, or model output was provided for this manuscript. Accordingly, no accuracy, AUROC, AUPRC, sensitivity, specificity, calibration statistic, confidence interval, p value, or decision-curve result is reported. Table V is a prespecified reporting structure, and its blank entries must be replaced only after the locked temporal and external evaluations have been completed.

TABLE V
LOCKED EXTERNAL TEST RESULTS NOT YET GENERATED

Model	AUROC	AUPRC	Sens.	Spec.	Brier
Logistic reference	Not generated	Not generated	Not generated	Not generated	Not generated
Elastic net	Not generated	Not generated	Not generated	Not generated	Not generated
Random forest	Not generated	Not generated	Not generated	Not generated	Not generated
Gradient boosting	Not generated	Not generated	Not generated	Not generated	Not generated

V. DISCUSSION

A. Interpretation Plan

Model selection will not be based on the largest AUROC alone. A candidate will be considered preferable only if its external discrimination improves with acceptable calibration and net benefit, and if the gain is stable across sites and clinically important subgroups. A model with higher AUROC but worse calibration or no decision benefit will not replace the logistic reference. If performance declines materially from nested development estimates to temporal or external testing, the decline will be treated as evidence of limited transportability rather than corrected by post hoc tuning.

Deployment readiness will require evidence beyond retrospective performance. The selected model must have a stable input pipeline, an explicit response to missing variables, a documented update policy, and a monitoring plan for prevalence and calibration drift. The clinical interface should display the probability, operating threshold, data timestamp, and unavailable inputs without presenting feature explanations as diagnoses. A prospective silent evaluation should run before clinicians receive model outputs. If that stage confirms data availability and calibration, a later impact study can assess whether the model changes referral decisions, time to diagnosis, or resource use. These steps are future evaluations and are not implied by favorable retrospective metrics.

The discussion will distinguish model performance from clinical causation. SHAP values or coefficients can describe which measurements influence predictions but cannot establish that changing those measurements will alter disease. Any subgroup disparity will be reported with its sample size and uncertainty. If a subgroup is too small for precise estimation, the conclusion will be that evidence is insufficient, not that performance is equivalent.

B. Comparative Analysis

The proposed protocol extends the reviewed work through study design and reporting rather than through an unsupported claim of algorithmic novelty. Table VI shows the main differences. The source papers remain the evidence base for the gap; the proposed items are future study procedures, not completed findings.

TABLE VI
COMPARISON OF REVIEWED STUDIES AND THE PROPOSED PROTOCOL

Dimension	Reviewed evidence	Proposed study
Population	One public ILPD cohort in [2]; clinical heterogeneity described in [1].	Multiple sites plus a nonoverlapping temporal cohort.
Predictors	Ten demographic and laboratory variables; five selected in [2].	Core ILPD-style tier and context-augmented tier.
Validation	Ten-fold cross-validation in [2]; corrigendum does not restate design [3].	Nested grouped development, locked temporal testing, and untouched external validation.
Performance	Accuracy, kappa, mean absolute error, runtime [2]; corrected confusion-derived metrics [3].	AUROC, AUPRC, calibration, threshold metrics, decision curves, uncertainty, and subgroup stability.
Reproducibility	Limited tuning detail; inconsistency in [2]; corrected FP/FN mapping in [3].	Versioned predictions, configuration, code, seeds, and automated metric identities.
Interpretability	Feature subset and classifier comparison.	Coefficients and held-out SHAP analysis with stability checks.

C. Limitations

This manuscript has four immediate limitations. First, it is a protocol because no experimental data or outputs were supplied. The proposed contribution therefore requires execution before its value can be judged. Second, the literature gap is derived from three supplied documents, one of which is a corrigendum and one a 2006 narrative review; the review is not a current systematic assessment of liver-prediction research. Third, multicenter outcome adjudication may vary despite a common evidence hierarchy, and retrospective data can contain selection bias, missingness, and site-specific measurement practices. Fourth, an explainable model can still be poorly calibrated or clinically unhelpful, and SHAP values do not provide causal explanations.

The binary target also compresses distinct diseases into one screening label. Etiologic subgroup analyses partly address this

problem but may be underpowered. Some context variables may be unavailable or measured inconsistently across sites. The core model is therefore the primary transportability benchmark, and the augmented model will be accepted only if its added information generalizes. Prospective impact evaluation is outside the initial study and will remain necessary before clinical deployment.

VI. CONCLUSION

The supplied literature supports a clear methodological gap. A small single-dataset classifier comparison cannot establish that liver-disease predictions generalize across etiologies, time, or institutions, and the published correction of FP/FN placement demonstrates the need for auditable metric generation. This paper proposes a multicenter protocol that combines fold-contained preprocessing, nested model development, temporal and external validation, calibration, decision analysis, subgroup evaluation, explainability, and automatic confusion-matrix checks.

The next step is to secure approvals, finalize the adjudication manual, and execute the locked analysis. The completed study should publish all prespecified tables and figures regardless of whether complex models outperform logistic regression. If external calibration or net benefit is inadequate, future work should focus on outcome definition, predictor harmonization, or local recalibration before considering a more complex algorithm. A subsequent prospective study should measure whether use of the model changes referral decisions or patient outcomes.

REFERENCES

- [1] R. Williams, "Global challenges in liver disease," *Hepatology*, vol. 44, no. 3, pp. 521-526, Sept. 2006, doi: 10.1002/hep.21347.
- [2] J. Singh, S. Bagga, and R. Kaur, "Software-based prediction of liver disease with feature selection and classification techniques," *Procedia Computer Science*, vol. 167, pp. 1970-1980, 2020, doi: 10.1016/j.procs.2020.03.226.
- [3] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Corrigendum to Performance Analysis of Classification Algorithms on early detection of Liver disease [Expert Systems with Applications 67 (2017) 239-251]," *Expert Systems with Applications*, vol. 125, pp. 442-443, 2019, doi: 10.1016/j.eswa.2019.02.029.
- [4] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: The TRIPOD statement," *Annals of Internal Medicine*, vol. 162, no. 1, pp. 55-63, 2015, doi: 10.7326/M14-0697.
- [5] R. F. Wolff et al., "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, doi: 10.7326/M18-1376.
- [6] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837-845, 1988, doi: 10.2307/2531595.
- [7] 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, doi: 10.1177/0272989X06295361.
- [8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 4765-4774.