

Artificial Intelligence for Liver Disease Prediction: A Systematic Review and an Explainable Hybrid Ensemble with Automated Clinical Feature Engineering

Sukhpreet Singh

Sukhpreet.gku@gmail.com

Guru Kashi University

Research Article

Keywords: liver disease, ILPD, machine learning, systematic review, feature engineering, feature selection, stacking ensemble, explainable artificial intelligence, SHAP, nested cross-validation

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11363770/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Liver disease prediction using routine clinical biomarkers is a common medical artificial-intelligence problem, there are significant variations in reported performance on the Indian Liver Patient Dataset (ILPD) and these differences stem from variations in preprocessing, feature selection, class balancing, validation and leakage control. This study integrates a DOI verified systematic evidence review, a reproducible experimental benchmark and an explainable hybrid ensemble. Structured data studies published between 2020 and September 2026 that include a DOI were synthesized, especially for the current focus of the research on ILPD. All 583 ILPD records (of 416 liver-disease and 167 non-disease cases) were used in the experimental component. The data-quality analysis, followed by false-discovery-rate (FDR) corrected group testing, repeated stratified cross-validation and by duplicate-sensitivity analysis were followed by nested 5×3-fold model development. Four complementary learners extra trees, LightGBM, Gaussian Naive Bayes and logistic regression were stacked with leakage safe out-of-fold stacking and sigmoid calibration. An automated clinical feature-engineering module expanded the 10 original predictors to a 24-feature candidate pool using clinically interpretable ratios, composite biomarkers, and log transformations; mutual-information feature selection was tuned strictly inside training folds. In the systematic review, DOI-verified studies reported headline accuracies ranging from approximately 71% to 99.6%, but evaluation protocols were highly heterogeneous. In our leakage-aware experiment, automated feature engineering improved ROC-AUC from 0.744 to 0.761 and PR-AUC from 0.893 to 0.897. Under a balanced-accuracy threshold, the engineered model achieved 64.8% accuracy, 71.2% balanced accuracy, 56.3% sensitivity, 86.2% specificity, and 0.761 ROC-AUC. Under a training-only sensitivity-constrained threshold, it achieved 69.5% accuracy, 78.1% sensitivity, 47.9% specificity, F1 = 0.785, and MCC = 0.258. Age, ALT, AST, indirect bilirubin, bilirubin burden, ALT + AST, and log-transformed bilirubin/aminotransferases were selected in every nested-CV model-fit. The findings show that automated clinical feature engineering improves discrimination and screening trade-offs, while also demonstrating that very high headline accuracy on small tabular medical datasets should be interpreted in the context of validation design and leakage control.

1. Introduction

Liver diseases are an important contributor to morbidity and mortality and are a significant burden of disease worldwide. They present with viral hepatitis, alcohol-associated diseases, metabolic dysfunction-associated steatotic liver disease, cirrhosis and hepatocellular carcinoma. Identifying the risk of chronic liver disorder in early stages is an important diagnostic-support objective as many chronic liver disorders are clinically silent until advanced stages and identification of those is an important element to identify in early stages using routinely available clinical and biochemical variables [1]. In hepatology, AI and machine learning (ML) techniques have increasingly applied to combine laboratory values, imaging, pathology and longitudinal clinical data to integrate these different modalities. Structured clinical data, such as clinical chemistry measures, is appealing because models can capture the nonlinear relationships between bilirubin, aminotransferases, alkaline phosphatase, serum proteins and demographic factors that may be hard to capture using a single conventional marker [2]. But model performance cannot be the only factor for clinical credibility, calibration, sensitivity/specificity balance, explanation, design of the model validation, and transportability are also crucial.

In previous studies on the use of computers for the diagnosis of liver disease, there has been great methodological variation in datasets, definitions of the disease, in preprocessing procedures, and in model families [3]. This is particularly apparent in the Indian Liver Patient Dataset (ILPD) benchmarks, which have been used in hundreds of classification experiments, and are both small and imbalanced. The accuracies of the studies have ranged from the low 70s, up to 95% and some reports come close to 100%. This dispersion poses a major methodological question: to what extent is the observed performance gap due to actual gains in predictive performance and to what extent is it due to methodological artifacts like upsampling prior to split, single hold-out partitions, repeated use of the same test data, duplicate rows or feature selection done prior to validation. Both contributions of the present study aim at providing an answer to that question. First, it is a DOI-available systematic review of recent studies on structured-data liver prediction,

focusing on ILPD and being explicit about validation and class-balancing options and their results, as well as the results of the underlying model. Second, it creates an explainable and reproducible leakage-aware benchmark and a hybrid ensemble with the preprocessing, the automatic feature engineering in clinical domains, the feature selection, and the model tuning, all performed on the training data, as well as the stacking, the calibration, and the threshold selection. Instead of an arbitrarily picked target accuracy, discrimination, calibration, and class-specific performance and robustness are evaluated. The novel parts of the experiment are an automatic clinical feature-engineering layer in which interpretable clinical variables (such as indirect bilirubin, direct-to-total bilirubin ratio, De Ritis ratio, estimated globulin, albumin-to-total-protein ratio, composite bilirubin/enzyme burdens, and log-transformed biochemical markers) are extracted and an automatic mutual-information selector is used to select the subset for processing within nested cross validation. These features are not derived from raw imaging/signal but are algorithmically derived from existing pre-set laboratory biomarkers. The architecture thus allows automatic feature engineering and automatic leakage-safe feature selection, automatic calibration of stacking, and automatic post-hoc explanation.

2. Systematic Evidence Review

2.1. Review design and reporting approach

The evidence component was designed in accordance with PRISMA 2020 guidelines for the transparent eligibility, screening and synthesis of evidence [4]. The review period was from 1st January 2020 to 13th September 2026. The objective was to provide a context for the experimental benchmark and not to pool effect sizes, thus necessitating a narrative systematic synthesis approach; given that the eligible studies varied greatly in data/phenomenological aspects (i.e. datasets, resampling, data splits, feature-selection procedures, optimization strategies, metrics reported), a formal meta-analysis approach was considered inappropriate.

2.2. Search strategy and eligibility

The DOI-resolvable bibliographic and publisher-indexed materials (PubMed/PMC, Elsevier/ScienceDirect, Springer/Nature, Wiley, Frontiers, MDPI, PLOS and other DOI-resolvable journal websites) were searched. We used the following combinations of core terms: ("liver disease" OR "liver disorder" OR hepatic) AND ("machine learning" OR "artificial intelligence" OR ensemble OR "feature selection" OR "explainable AI") AND (prediction OR diagnosis OR classification) and also the following combinations: ("liver disease" OR "liver disorder" OR hepatic) AND ("machine learning" OR "artificial intelligence" OR ensemble OR "feature selection" OR "explainable AI") AND (prediction OR diagnosis OR classification) AND (ILPD OR "Indian Liver Patient Dataset"). In this manuscript only references with verified DOIs were kept as requested.

3. Systematic Review Findings

3.1. Evolution of ILPD and structured-data liver prediction

It was already evident from initial studies during the window of review that there was substantial variation in performance. Fathi et al. showed the maximum accuracy of 90.9% for SVM-based liver classification with feature reduction strategy by using ILPD [5]. Singh et al. have also compared the feature-selection and classification methods, and got the accuracy of 74.36% when using Random Forest, which shows that the accuracy can be significantly different with different protocols [6].

Mostafa et al. went on to show that statistical ML pipelines that include imputation, dimensionality reduction and balancing can achieve very high predictive performance e.g., Random Forest 98.14% on their structured liver-disease dataset [7]. The SLT based on Integrated Projection (IP) was introduced by Amin et al. and achieved a high rate of

accuracy of 88.10% [8]. Dritsas and Trigka used an 80.1% accuracy rate in their experiments as a voting classifier using supervised ML with class-balancing, and used multiple performance measures instead of just a score [9].

Alauthman et al. only specifically mentioned tabular-data generation and class balancing for ILPD and achieved accuracy of about 99.6% for KNN with augmentation and SMOTE-based processing [10]. This type of results shows the best possible performance to be obtained with an aggressive synthetic data workflow, but also highlights the need to think very carefully about the sequence of the resampling and data partitioning when comparing studies.

Subsequent research was more recent, but continued to differ. Rehman et al. presented 76.56% Accuracy for MaLLiDD framework for ILPD [11]. Ganie and Pramanik have done a comparison of boosting algorithm with normalization, upsampling and tuning hyper parameters; Gradient Boosting achieved 98.29% accuracy on ILPD [12]. Islam et al. have proposed the use of Extra Trees together with Tree-Structured Parzen Estimator optimization, upsampling, normalization and feature selection and achieved 95.8% accuracy with 10-fold cross-validation [13].

Ganji et al. employed the forward selection and fine-tuned LightGBM model in 2025, achieving an accuracy of 82.12% for ILPD which was higher than the previous 76.21% [14]. In order to enhance the accuracy, Rani et al. evaluated their proposed hybrid configurations of recursive feature elimination, SMOTE-ENN and ensemble learning and obtained a maximum accuracy of 93.2% [15]. El Atifi et al. performed optimization and achieved eight best results by using ensemble classifiers with Indian Liver Patient Records dataset which got 85.17% accuracy on Random Forest [16].

The 2026 literature is especially helpful in making the methodological comparisons. Mohapatra et al. systematically quantified the preprocessing effects and found that the accuracy of Random Forest after preprocessing is around 74% [17] and found that there were four missing albumin/globulin-ratio values and duplicate handling in the benchmark of 583 records. Saraf et al. employed the techniques of demographical stratification, recursive feature elimination, resampling, Bayesian optimization and dynamic hybridization with reported 94.2% accuracy on ILPD [18]. Eray explicitly criticized resampling before splitting and single-split evaluation, implemented hybrid feature selection in a leakage-free design and demonstrated that significant improvements in clinically relevant parameters were better demonstrated via values of specificity and ROC-AUC (0.00-0.33 vs. 0.61–0.92 across classifiers) than by those of accuracy [19].

Leite et al. tested 30 repetitions of 10-fold cross validation, and found that the conventional 30 tree ensembles mostly did not exceed 0.68–0.70 accuracy, while the exploratory TabNet benchmark achieved 0.7307 accuracy and 0.9424 recall [20]. Similarly, Mithu et al. applied interpretable modeling and SHAP to ILPD and found that modest accuracy is also frequently achieved for tasks that involve interpretability as well as traditional validation [21]. Table 1 summarizes the selected studies, datasets, ML approaches, validation protocols, reported accuracy, and key methodological notes. Figure 1 illustrates the reported liver-disease classification accuracy compares these results with the clinical threshold used in this study.

Table 1
Comparison of recent structured-data liver-disease prediction studies

Study	Year	Dataset	Approach	Validation	Headline accuracy	Protocol note
Fathi et al. [5]	2020	ILPD/BUPA	SVM + feature selection	10-fold CV	90.9%	Feature reduction; high SVM result
Singh et al. [6]	2020	ILPD	RF + feature selection	Comparative validation	74.36%	Much lower result on same benchmark
Mostafa et al. [7]	2021	Structured liver dataset	RF; MICE/PCA/SMOTE	Train/test evaluation	98.14%	Strong preprocessing/balancing effect
Amin et al. [8]	2023	ILPD	Projection-based feature extraction + ML	Comparative evaluation	88.10%	Feature-extraction emphasis
Dritsas & Trigka [9]	2023	Structured liver dataset	SMOTE + voting ensemble	10-fold CV	80.1%	Multiple metrics reported
Alauthman et al. [10]	2023	ILPD	Data generation/SMOTE + KNN	Cross-validation	99.6%	Synthetic augmentation; very high headline accuracy
Rehman et al. [11]	2024	ILPD	MaLLiDD	Comparative evaluation	76.56%	Specificity 68.32%; F1 78.48%
Ganie & Pramanik [12]	2024	ILPD	GB + tuning + upsampling	k-fold CV	98.29%	Normalization and upsampling
Islam et al. [13]	2024	ILPD	Extra Trees + TPE	10-fold CV	95.8%	Upsampling + normalization + feature selection
Ganji et al. [14]	2025	ILPD	Forward selection + LGBM	80/20 + 10-fold CV	82.12%	Optimized from 76.21%
Rani et al. [15]	2025	ILPD	RFE + SMOTE-ENN + ensemble	80/20-based pipeline	93.2% max	Hybrid balancing/ensemble configuration
El Atifi et al. [16]	2025	Indian Liver Patient Records	Optimized RF	GridSearchCV	85.17%	Ensemble optimization
Mohapatra et al. [17]	2026	ILPR/ILPD	Preprocessing + RF	Repeated stratified 5-fold ablation	~ 74%	Shows preprocessing sensitivity
Saraf et al. [18]	2026	ILPD	Demographic hybrid + Bayesian tuning	Segmented hybrid evaluation	94.2%	RFE + resampling + demographic segmentation
Eray [19]	2026	ILPD/BUPA	Leakage-free HFS + SMOTE	10 independent splits + Wilcoxon	Not accuracy-centered	Specificity 0.61–0.92; explicitly leakage-aware
Leite et al. [20]	2026	ILPD	TabNet benchmark	30 × 10-fold CV	73.07%	Recall 94.24%; rigorous repeated protocol

Study	Year	Dataset	Approach	Validation	Headline accuracy	Protocol note
Mithu et al. [21]	2026	ILPD	Interpretable LR + SHAP	Train/test modeling	~ 71%	Interpretability-focused

A consistent methodical trend identified in the review is that larger upsampling, synthetic generation, better optimization, demographic segmentation or hold-out workflows typically lead to the highest accuracy while repeated or designed explicitly to prevent leakage led to more moderate accuracy. It is not a refutation of the quality of the studies but rather reveals that it is not possible to compare the accuracy values without also comparing the time of resampling, whether feature selection was nested or not, whether the same or duplicate observations were crossing partitions or not, the method of tuning the hyperparameters, and the metric optimized.

4. Materials and Methods

4.1. Dataset and variables

Indian Liver Patient Dataset (ILPD) from UCI Machine Learning Repository [22] was used for experimental analysis. There are 583 records, 10 predictors and a binary selector that marks the presence of liver-disease. Of the 416 records identified as liver disease, 167 were identified as non-disease and 441 were male with 142 records for females. There were four values of the albumin/globulin ratio that were blank. The raw file was kept in its original form and a research layer was attached with the identification of explicit records and binary target mapping (1 = liver disease; 0 = no liver disease). Table 2 illustrates the clinical and demographic variables used for liver-disease prediction, including their data types and clinical interpretation. Figure 2 illustrates the class distribution of the ILPD dataset, showing the number of patients with and without liver disease.

Table 2
ILPD predictors used in the experimental analysis

Variable	Type	Role/clinical interpretation
Age	Numeric	Demographic predictor
Gender	Categorical	Demographic predictor
Total bilirubin	Numeric	Overall bilirubin burden
Direct bilirubin	Numeric	Conjugated bilirubin
Alkaline phosphatase (ALP)	Numeric	Cholestatic/hepatobiliary enzyme
ALT/SGPT	Numeric	Hepatocellular injury marker
AST/SGOT	Numeric	Hepatocellular/tissue injury marker
Total protein	Numeric	Circulating protein concentration
Albumin	Numeric	Liver synthetic-function proxy
Albumin/globulin ratio	Numeric	Relative albumin-to-globulin balance
Liver disease	Binary target	1 = disease, 0 = no disease

4.2. Data-quality assessment and exploratory statistics

The master dataset was not globally normalized, imputed, resampled, or feature-selected. Missing values were retained until model fitting so that imputation parameters could be learned from training folds only. Exact duplicate analysis identified 13 excess duplicate rows; therefore, the 583-record source was maintained for comparability with published ILPD studies and a 570-record exact-deduplicated cohort was retained for sensitivity analysis.

Because bilirubin, ALP, ALT, and AST were markedly right-skewed, disease-group comparisons were summarized using median and interquartile range and tested with two-sided Mann-Whitney U tests. Benjamini-Hochberg false-discovery-rate (FDR) correction was applied across the nine numeric predictors. Spearman correlation was used to characterize monotonic inter-feature relationships. Table 3 illustrates the differences in biomarker distributions between liver-disease and non-disease groups using median [IQR] values and FDR-corrected statistical tests. Figure 3 illustrates the Spearman correlation matrix among the clinical and biochemical variables in the ILPD dataset.

Table 3
Disease versus non-disease biomarker comparison (median [IQR]) with FDR-corrected tests

Variable	Liver disease	No liver disease	FDR-adjusted p	Significant
Age	46.00 [34.00, 58.00]	40.00 [28.00, 55.00]	0.002	Yes
Total Bilirubin	1.40 [0.80, 3.62]	0.80 [0.70, 1.10]	1.030e-12	Yes
Direct Bilirubin	0.50 [0.20, 1.80]	0.20 [0.20, 0.35]	2.229e-12	Yes
Alkaline Phosphatase	229.00 [186.00, 315.25]	186.00 [161.50, 213.00]	7.825e-11	Yes
ALT / SGPT	41.00 [25.00, 76.50]	27.00 [20.00, 37.50]	5.249e-12	Yes
AST / SGOT	52.50 [29.75, 108.75]	29.00 [21.00, 43.50]	8.289e-13	Yes
Total Protein	6.55 [5.70, 7.20]	6.60 [5.90, 7.30]	0.437	No
Albumin	3.00 [2.50, 3.62]	3.40 [2.90, 4.00]	7.158e-05	Yes
Albumin/Globulin Ratio	0.90 [0.70, 1.10]	1.00 [0.90, 1.20]	8.718e-06	Yes

The automated feature engineering was added as a deterministic transformer within each of the ML pipelines. It was applied after the median imputation for the folds and before the feature selection; therefore, outer test folds did not affect the imputation and/or the model-selection process. The intent was to make physiologically interpretable relationships explicit, and preserve the original variables. Table 4 illustrates the engineered biochemical features and their corresponding formulas, highlighting derived measures of bilirubin, enzyme activity, protein balance, and nonlinear transformations.

Table 4
Automatically generated clinical features

Engineered feature	Formula	Interpretation
Indirect bilirubin	Total bilirubin – Direct bilirubin	Unconjugated bilirubin component
Direct/Total bilirubin ratio	Direct / Total bilirubin	Relative conjugated fraction
AST/ALT (De Ritis) ratio	AST / ALT	Relative aminotransferase pattern
Estimated globulin	Total protein – Albumin	Derived globulin concentration
Albumin/Total-protein ratio	Albumin / Total protein	Synthetic-protein balance
Bilirubin burden	Total + Direct bilirubin	Composite bilirubin magnitude
Enzyme sum	ALT + AST	Composite aminotransferase burden
Enzyme geometric mean	$\sqrt{\text{ALT} \times \text{AST}}$	Multiplicative enzyme burden
Log biochemical features	$\log_{10}(\text{TB}), \log_{10}(\text{DB}), \log_{10}(\text{ALP}), \log_{10}(\text{ALT}), \log_{10}(\text{AST})$	Skewness-reducing nonlinear representation
Recomputed A/G ratio	Albumin / (Total protein – Albumin)	Consistency/alternative ratio representation

These transformations with the 10 original predictors resulted in a 24-feature candidate pool. The manual fixing of feature selection was replaced by automatic fixing. The ranking based on mutual information was included in the training pipeline, and the number of candidates in the subsets was determined by inner cross-validation. This is because dimensionality reduction should be integrated into the model building process and not be carried out as a preprocessing step before developing the model with the entire data set [23].

4.3. Hybrid ensemble, calibration, and threshold policies

Four intentionally complementary learners, namely Extra Trees, LightGBM, Gaussian Naive Bayes and logistic regression, were proposed. Tree-based, probabilistic, and linear learning are different kinds of errors made by learners of different types, and repeated learners of one type might not be able to provide as good a second stage representation as learners of other types. This is useful for ensemble diversity. The outer loop was cross-validation over training folds for which candidate hyperparameters and numbers of features were optimized in the inner loop, which was a 3-fold stratified cross-validation. The tuned base learners then predicted inner out-of-fold probabilities and these were used as input for a class-balanced logistic meta-learner. The probabilities at the meta-level were sigmoided. Figure 4 illustrates the proposed liver-disease prediction pipeline, integrating feature engineering, MI-based selection, ensemble learning, probability calibration, threshold optimization, and SHAP-based interpretation.

Two threshold policies were assessed. The threshold chosen with the balanced-accuracy policy was the one that achieved the best (balanced) accuracy with training-only out-of-fold predictions. The policy used in the screening scenario had thresholds that had a training sensitivity of 0.80 or higher and was applied in such a way as to maximize specificity (balanced accuracy and MCC were used as tie-breakers). The chosen threshold was applied to the outer test fold (untouched) as is. Thus, there was no optimization for the outer-fold performance on test labels.

4.4. Validation, baseline benchmarking, and metrics

Repeated stratified 10-fold cross validation was done three times (30 test folds per condition) for stage-3 baselines. In each training fold we fitted median imputation and scaling. To avoid this, the sensitivity analysis by SMOTENC was

performed for some models after the training-fold process. Thus, the sensitivity analysis using SMOTENC was carried out for some models after training-fold process given that conventional SMOTE can lead to optimistic bias when applied before partitioning [25]. The proposed model comparison was the nested stratified cross-validation (CV) approach consisting of five outer folds and three inner folds. There were no outer tests observed that were used for imputation, feature engineering, feature ranking, subset-size selection, hyperparameter tuning, stacking, calibration, or threshold selection.

The accuracy, balanced accuracy, sensitivity, specificity, precision, F1 score, receiver-operating-characteristic area under the curve (ROC-AUC), precision-recall AUC (PR-AUC), Matthew's correlation coefficient (MCC) and Brier score were used to summarize performance. The emphasis was placed on the balanced accuracy and the MCC since disease is the majority class. A repeated rows sensitivity analysis was performed to see if baseline performance was materially inflated by repeated rows.

4.5. Explainability

There were two levels of evaluation of explainability. The first feature was the feature-selection stability, which calculated the number of times that each of the original or engineered features was selected in the 20 outer-fold base-model fits (five folds $\times$ four learners). Second, a tuned LightGBM base learner was used for providing post-hoc feature-attribution estimates using SHAP. SHAP values link local additive explanations with the importance of a tree ensemble [26]. The results presented were interpreted using the SHAP analysis but were NOT used as a stand-alone validation result. The extremely randomized trees were kept as a base learner as they can provide high variance reduction and good nonlinear discrimination abilities on tabular data, due to their split construction based on the random feature [27].

5. Results

5.1. Data characteristics and statistical profile

The patients in the cohort were predominantly liver disease (416/583, 71.4%). After FDR correction, eight of the nine numeric predictors were found to be significantly different between the diagnostic groups: age, total bilirubin, direct bilirubin, ALP, ALT, AST, albumin, and albumin/globulin ratio. The only numeric predictor that was not significant (FDR adjusted $p = 0.437$) was total protein. There was very strong correlation between the total/direct bilirubin and between ALT/AST. These findings encouraged automated selection, not only because the researchers had no reason to believe that all the original variables contained the same amount of information, but also because they were optimists about what they could learn from the selected variables. Table 5 illustrates the comparative performance of baseline machine-learning models using accuracy, balanced accuracy, sensitivity, specificity, ROC-AUC, and MCC.

5.2. Leakage-aware baseline benchmark

Table 5
Stage-3 baseline performance on the original 583-record dataset

Model	Accuracy	Balanced acc.	Sensitivity	Specificity	ROC-AUC	MCC
Logistic Regression	0.723	0.576	0.921	0.231	0.750	0.210
KNN	0.673	0.569	0.813	0.325	0.657	0.149
Gaussian NB	0.560	0.679	0.402	0.956	0.726	0.354
Decision Tree	0.666	0.604	0.750	0.458	0.604	0.204
Random Forest	0.716	0.600	0.873	0.327	0.757	0.234
SVM-RBF	0.715	0.502	1.000	0.004	0.666	0.014
Extra Trees	0.735	0.622	0.886	0.357	0.779	0.284
XGBoost	0.690	0.572	0.848	0.296	0.740	0.164
LightGBM	0.709	0.611	0.841	0.381	0.751	0.244
CatBoost	0.714	0.564	0.914	0.213	0.750	0.179

Extra Trees produced the highest raw baseline accuracy (0.735) and ROC-AUC (0.779), but specificity was only 0.357. Gaussian Naive Bayes produced the highest balanced accuracy (0.679) and specificity (0.956), but sensitivity was only 0.402. The most striking imbalance pathology was SVM-RBF: accuracy was approximately 0.715 and sensitivity was 1.000, while specificity was only 0.004. Thus, a model could appear acceptable by raw accuracy while classifying almost every patient as diseased. This result independently mirrors the majority-class collapse described in recent leakage-aware ILPD work [19]. Table 6 illustrates the impact of feature engineering and decision-threshold selection on the proposed model's classification and calibration performance.

5.3. Effect of automated clinical feature engineering

Table 6
Original versus engineered feature sets under two training-only threshold policies

Feature set	Threshold	Acc.	Bal. acc.	Sens.	Spec.	Prec.	F1	ROC-AUC	PR-AUC	MCC	Brier
Original features	Balanced threshold	0.630	0.712	0.519	0.904	0.931	0.667	0.744	0.893	0.391	0.172
Original features	Sensitivity-constrained threshold	0.684	0.612	0.781	0.443	0.778	0.779	0.744	0.893	0.225	0.172
Engineered features	Balanced threshold	0.648	0.712	0.562	0.862	0.911	0.695	0.761	0.897	0.387	0.169
Engineered features	Sensitivity-constrained threshold	0.695	0.630	0.781	0.479	0.789	0.785	0.761	0.897	0.258	0.169

Automated feature engineering improved discrimination under both threshold policies because the underlying probability model is shared. ROC-AUC increased from 0.744 to 0.761 (+ 0.016), PR-AUC from 0.893 to 0.897, and Brier score improved from 0.172 to 0.169. Under the balanced-accuracy threshold, engineered features increased accuracy from 0.630 to 0.648 and sensitivity from 0.519 to 0.563; F1 increased from 0.667 to 0.695, while specificity decreased from 0.904 to 0.862. Balanced accuracy remained essentially unchanged (0.712 in both configurations), indicating a shift toward better disease sensitivity rather than a simple global performance increase.

The clinically constrained policy gave sensitivity of 0.781, an increase in specificity from 0.443 to 0.479, an increase in accuracy from 0.684 to 0.695, an increase of 0.612 to 0.630 in balanced accuracy, an increase of 0.779 to 0.785 in F1, and an increase of 0.225 to 0.258 in MCC. This operating point would be more suited to a discussion of screening versus the balanced-accuracy operating point, as it minimizes missed cases of disease, but is not sufficiently specific for use in the clinical setting without external validation and more data. Figure 5 illustrates the effect of feature engineering on accuracy, balanced accuracy, sensitivity, specificity, and MCC under the sensitivity-constrained threshold. Figure 6 illustrates the out-of-fold ROC curves for the original and engineered feature sets, showing a slight improvement in ROC-AUC from 0.744 to 0.761.

5.4. Automated feature-selection stability

Table 7 illustrates the feature-selection frequency of original and engineered features across the overall analysis and individual ML models.

Table 7
Automated feature-selection stability across 20 nested-CV base-model fits

Feature	Type	Overall	Extra Trees	LightGBM	GNB	LR
Age	Original	100%	100%	100%	100%	100%
ALT / SGPT	Original	100%	100%	100%	100%	100%
AST / SGOT	Original	100%	100%	100%	100%	100%
Indirect Bilirubin	Engineered	100%	100%	100%	100%	100%
Bilirubin Burden (Total+Direct)	Engineered	100%	100%	100%	100%	100%
ALT + AST Enzyme Sum	Engineered	100%	100%	100%	100%	100%
log1p Total Bilirubin	Engineered	100%	100%	100%	100%	100%
log1p ALT	Engineered	100%	100%	100%	100%	100%
log1p AST	Engineered	100%	100%	100%	100%	100%
Total Bilirubin	Original	95%	100%	80%	100%	100%
Direct/Total Bilirubin Ratio	Engineered	95%	100%	80%	100%	100%
Direct Bilirubin	Original	85%	100%	60%	100%	80%
log1p Direct Bilirubin	Engineered	85%	100%	40%	100%	100%
ALT-AST Geometric Mean	Engineered	80%	80%	60%	100%	80%
Recomputed Albumin/Globulin Ratio	Engineered	80%	100%	40%	100%	80%
Albumin/Total Protein Ratio	Engineered	75%	100%	40%	100%	60%
log1p Alkaline Phosphatase	Engineered	75%	80%	40%	100%	80%
Alkaline Phosphatase	Original	70%	80%	40%	80%	80%
Albumin/Globulin Ratio	Original	70%	80%	40%	80%	80%
Albumin	Original	65%	80%	20%	100%	60%
Globulin (Total Protein-Albumin)	Engineered	40%	40%	20%	40%	60%
Gender	Original	25%	20%	20%	40%	20%
Total Protein	Original	20%	20%	20%	20%	20%
AST/ALT Ratio (De Ritis)	Engineered	20%	20%	20%	20%	20%

Nine variables were selected in 100% of the nested-CV base-model fits: Age, ALT, AST, indirect bilirubin, bilirubin burden, ALT + AST enzyme sum, log total bilirubin, log ALT, and log AST. Several of these are engineered representations rather than original ILPD columns, providing direct evidence that the feature-engineering step contributed information consistently recognized by multiple model families. By contrast, Total Protein was selected in only 20% of fits, consistent with its non-significant univariate comparison. Figure 7 illustrates the stability of automated feature selection across 20 nested cross-validated model fits, highlighting consistently selected original and engineered features. Figure 8 illustrates the confusion matrix of the engineered ensemble at the clinical decision threshold, showing the distribution of correctly and incorrectly classified liver-disease cases.

5.5. Screening errors, calibration, and explanation

At the screening-oriented threshold, the model recovered substantially more diseased cases than the balanced-threshold configuration but generated more false positives among non-disease patients. This trade-off is clinically meaningful: for a screening tool, false-negative reduction may be prioritized, but the observed specificity of 0.479 would lead to substantial downstream confirmatory testing. The Brier score of 0.169 indicates better probability calibration after engineering than the original-feature stack (0.172), but calibration does not compensate for limited discrimination caused by the small dataset. Figure 9 illustrates the global SHAP feature importance of the engineered LightGBM base learner, highlighting the most influential predictors of liver-disease classification.

Post hoc feature attribution with the LightGBM showed that the model considers the following features as important: Age, Direct Bilirubin, Albumin/Total-Protein Ratio, ALP, and ALT, followed by engineered enzyme and bilirubin representations. This is in addition to, not a substitute for, the nested feature-selection stability analysis. SHAP magnitudes depend on the model, and are not to be interpreted as causal effects.

6. Discussion

6.1. Main findings

This is one of a pair of observations made in this study. First, there is a tremendous variability in the liver-disease prediction accuracy reported in the literature, and the DOI-verified literature reports ILPD values from as low as the low 70's to above 98%. Second, the performance is much more modest if the same small ILPD data set is assessed using all the steps of the preprocessing, feature selection/tuning, out-of-fold stacking, calibration and training-only threshold choice in a fold-wise manner. The engineered ensemble proposed did not come close to a headline accuracy above 90% as it achieved ROC-AUC 0.761 and screening accuracy 0.695. That's not considered a failure since the absolute accuracy is lower, but the fact that it exists is proof that protocol design is affecting the apparent difficulty of this benchmark.

The Stage-3 SVM result is a good example of how to not only be misleading, but also dangerous. The model was able to learn the majority disease label well, and to achieve around 71.5% accuracy, while having sensitivity of 1.000 and a specificity of 0.004. The same majority collapse phenomenon has been observed when performing a leakage-aware ILPD analysis [19]. This illustrates the importance of reporting both sensitivity and specificity, as well as the other measures of accuracy such as the mean difference between the two, the MCC and the probability discrimination.

6.2. Why automated feature engineering helped

The engineered variables enhance the representation of clinically meaningful relationships, rather than add to dimensionality. Indirect bilirubin separates the unconjugated from the direct bilirubin; The direct-to-total bilirubin is a ratio and not a concentration; The De Ritis ratio is the relative balance between AST and ALT; Estimated globulin and albumin ratios are expressions of synthetic-protein relationships; and log transformations compress extreme biochemical values and retain the rank structure. The transforms were generated deterministically within the pipeline and the selector could have thrown out unhelpful transformations, so that the inclusion of these transformations did not cause the model to actually be using all of the engineered variables.

It is interesting to note that, once again, indirect bilirubin, bilirubin burden, enzyme sum and log transformed aminotransferases are selected. All of these variables were kept along with the original Age, ALT and AST in all nested-CV base-model fits. The agreement among the three methods (Extra Trees, LightGBM, Naive Bayes and logistic regression) implies that they represent signal that is available under varying inductive assumptions. But the frequency with which features are selected is an associative indicator and any derived biomarker would need to be clinically validated to be considered as a diagnostic indicator.

6.3. Reconciling the proposed results with 90–99% literature accuracy

The systematic comparison clearly reveals that “best accuracy” is not a protocol independent number. The combination of resampling, optimization, feature selection, synthetic data generation or the selection of a model for a particular subgroup is often used in high-performing studies [10, 12, 13, 15, 18]. All of these are legitimate if done within the training set, but will make the performance seem optimistic if any observations are made in the training set which are subsequently used as optimization decisions, synthetic examples, or feature rankings in the test set. On the other hand, repeated-CV and leakage-aware 2026 studies achieve accuracies of 0.69–0.73, and focus on specificity or robustness [19, 20] which are much closer to baseline behavior observed here.

The present design therefore does not claim to be superior with regards to raw accuracy. It is methodical, in the sense that it brings in an end-to-end pipeline that is auditable and in which the generation of the clinical features, the selection of the features, the search of hyper parameters, the stacking, the calibration and the optimizing of the thresholds are all restricted to the training data. No test folds were exploited and the engineered model resulted in an improved ROC-AUC and better clinical sensitivity/specificity trade-off. This is a more defensible basis to compare and validate externally in future.

6.4. Clinical interpretation

A liver-risk screening tool may have a higher threshold for sensitivity due to risk of a false negative from delaying further investigation. The outer-fold sensitivity of the sensitivity-constrained policy was 78.1%, which is within the reasonable range of how we expect a policy to perform when tested on untouched folds, and shows a normal generalization effect from the 80% training-constrained probability. The specificity was 47.9%, which means that the present model would give a lot of false positives. Thus, it is regarded as a prototype research system for making decisions in the process of research, not a clinical diagnostic system.

Even though the probability output is calibrated, it might still be useful for risk stratification as the probability can be used as a continuous risk score and therefore can be used in combination with a tiered referral threshold instead of a binary reference threshold. Validation of the decision curve analysis, cost sensitive thresholds, and validation in modern clinical cohorts with more comprehensive variables like platelet count, INR, glucose, comorbidities, imaging findings and disease etiology should be investigated in future studies.

6.5. Strengths and limitations

Key strengths include the DOI-restricted evidence based, explicit protocol comparison, preservation of the raw dataset, duplicate sensitivity analysis, fold-wise preprocessing, nested feature selection and optimization, out-of-fold stacking, probability calibration, training-only threshold selection, automatic clinical feature engineering, multi-metric evaluation, and two-level interpretability. Using an original-feature and an engineered-feature model with the same outer folds allows for an ablation-style evaluation of the contribution due to the engineered-feature.

There are a number of constraints to consider. ILPD is small, historical and geographically limited and does not specify etiology or severity of disease. In the legacy distribution there are four A/G-ratio values missing, and there are exact duplicates in the distribution. The systematic review was limited to English-language studies with a DOI and thus intentionally does not include other non-DOI studies that may be relevant. No registered systematic review should be submitted based on search counts from subscription bibliographic databases as this is a recalibration of the PRISMA flow. Two points of note: there was no external hospital cohort, and generalization to anything beyond ILPD is unknown. However, the engineered ratios are clinically interpretable, but still deterministic transformations of the same small set of

predictors. Finally, SHAP was used for interpretation of a base learner, and it does not imply a causal effect on biomarker.

Conclusion

A DOI-verified systematic synthesis and a leakage-aware ILPD benchmark were combined to examine why liver-disease prediction studies report widely different accuracy. The literature contains headline accuracies from approximately 71% to 99.6%, but protocol heterogeneity including resampling, optimization, feature selection, synthetic generation, and validation design precludes direct ranking by accuracy alone. In our nested experimental framework, automated clinical feature engineering expanded 10 original biomarkers to a 24-feature pool and automated mutual-information selection identified stable original and derived predictors without exposing outer test folds to feature-ranking decisions. The engineered ensemble improved ROC-AUC from 0.744 to 0.761 and, at the sensitivity-constrained operating point, improved accuracy from 0.684 to 0.695, specificity from 0.443 to 0.479, balanced accuracy from 0.612 to 0.630, and MCC from 0.225 to 0.258 while maintaining 0.781 sensitivity. These gains are modest but reproducible under stricter validation. The main conclusion is therefore not that ILPD can be made to exceed a predetermined accuracy threshold, but that robust feature engineering, leakage prevention, calibration, and transparent class-specific evaluation provide a more credible basis for medical-AI research. External multicenter validation is required before clinical use.

Declarations

Not Applicable

Funding Declaration

No funding was received for this research

Author Contribution

Sukhpreet Singh conceived the study, conducted the systematic review, performed the data analysis and machine-learning experiments, developed the proposed methodology, prepared the figures and tables, and wrote the manuscript. He reviewed and approved the final version of the manuscript.

References

1. Asrani SK, Devarbhavi H, Eaton J, Kamath PS (2019) Burden of liver diseases in the world. *J Hepatol* 70(1):151–171. 10.1016/j.jhep.2018.09.014
2. Ahn JC, Connell A, Simonetto DA, Hughes C, Shah VH (2021) Application of artificial intelligence for the diagnosis and treatment of liver diseases. *Hepatology* 73(6):2546–2563. 10.1002/hep.31603
3. Khan RA, Luo Y, Wu F-X (2022) Machine learning based liver disease diagnosis: A systematic review. *Neurocomputing* 468:492–509. 10.1016/j.neucom.2021.08.138
4. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71. 10.1136/bmj.n71
5. Fathi M, Nemati M, Mohammadi SM, Abbasi-Kesbi R (2020) A machine learning approach based on SVM for classification of liver diseases. *Biomedical Engineering: Appl Basis Commun* 32(3):2050018. 10.4015/S1016237220500180

6. Singh J, Bagga S, Kaur R (2020) Software-based prediction of liver disease with feature selection and classification techniques. *Procedia Comput Sci* 167:1970–1980. 10.1016/j.procs.2020.03.226
7. Mostafa F, Hasan E, Williamson M, Khan H (2021) Statistical machine learning approaches to liver disease prediction. *Livers* 1(4):294–312. 10.3390/livers1040023
8. Amin R, Yasmin R, Ruhi S, Rahman MH, Reza MS (2023) Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms. *Inf Med Unlocked* 36:101155. 10.1016/j.imu.2022.101155
9. Dritsas E, Trigka M (2023) Supervised machine learning models for liver disease risk prediction. *Computers* 12(1):19. 10.3390/computers12010019
10. Alauthman M et al (2023) Tabular data generation to improve classification of liver disease diagnosis. *Appl Sci* 13(4):2678. 10.3390/app13042678
11. Rehman AU, Butt WH, Ali TM et al (2024) A Machine Learning-Based Framework for Accurate and Early Diagnosis of Liver Diseases: A Comprehensive Study on Feature Selection, Data Imbalance, and Algorithmic Performance. *Int J Intell Syst* 2024:6111312. 10.1155/2024/6111312
12. Ganie SM, Pramanik PKD (2024) A comparative analysis of boosting algorithms for chronic liver disease prediction. *Healthc Analytics* 5:100313. 10.1016/j.health.2024.100313
13. Islam R, Sultana A, Tuhin MN (2024) A comparative analysis of machine learning algorithms with a tree-structured Parzen estimator for liver disease prediction. *Healthc Analytics* 6:100358. 10.1016/j.health.2024.100358
14. Ganji A, Usha D, Rajakumar PS (2025) Scalable Computing: Pract Experience 26(3):1104–1115. 10.12694/scpe.v26i3.4254. Enhanced early diagnosis of liver diseases using feature selection and machine learning techniques on the Indian Liver Patient Dataset
15. Rani R, Jaiswal G, Nancy, Lipika, Bhushan S, Ullah F, Singh P, Diwakar M (2025) Enhancing liver disease diagnosis with hybrid SMOTE-ENN balanced machine learning models—an empirical analysis of Indian patient liver disease datasets. *Front Med* 12:1502749. 10.3389/fmed.2025.1502749
16. El Atifi W, El Rhazouani O, Khan FM, Sekkat H (2025) Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare. *PLoS ONE* 20(8):e0330899. 10.1371/journal.pone.0330899
17. Mohapatra RK, Jolly L, Dakua SP (2026) From raw to refined: the influence of preprocessing on ML performance for liver disease detection. *Discover Artif Intell* 6:27. 10.1007/s44163-025-00659-1
18. Saraf E, Yang M, Sakthivel R et al (2026) Hybrid intelligent systems for liver disease prediction: a demographic-aware machine learning framework. *Front Med* 12:1728061. 10.3389/fmed.2025.1728061
19. Eray O (2026) Liver disease prediction using hybrid feature selection: a comparative analysis of machine learning models. *Appl Sci* 16(13):6726. 10.3390/app16136726
20. Leite GA, Azevedo BF, Ferreira SR, Pacheco MF, Fernandes FP, Pereira AI (2026) Liver disease diagnosis: fuzzy c-means and scoring ranking approach. *SN Comput Sci* 7:470. 10.1007/s42979-026-05026-0
21. Mithu K, Begum S, Uddin MN, Huda MN (2026) Data Model 7(1):1–12. 10.25163/data.7110695. Interpretable machine learning data modeling for liver disease risk profiling: insights from the Indian Liver Patient Dataset
22. Ramana BV, Venkateswarlu N ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. 10.24432/C5D02C
23. Jia W, Sun M, Lian J, Hou S (2022) Feature dimensionality reduction: a review. *Complex Intell Syst* 8(3):2663–2693. 10.1007/s40747-021-00637-x
24. Mienye ID, Sun Y (2022) A survey of ensemble learning: concepts, algorithms, applications, and prospects. *IEEE Access* 10:99129–99149. 10.1109/ACCESS.2022.3207287

25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 16:321–357. 10.1613/jair.953

26. Lundberg SM, Erion G, Chen H et al (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2:56–67. 10.1038/s42256-019-0138-9

27. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42. 10.1007/s10994-006-6226-1

Figures

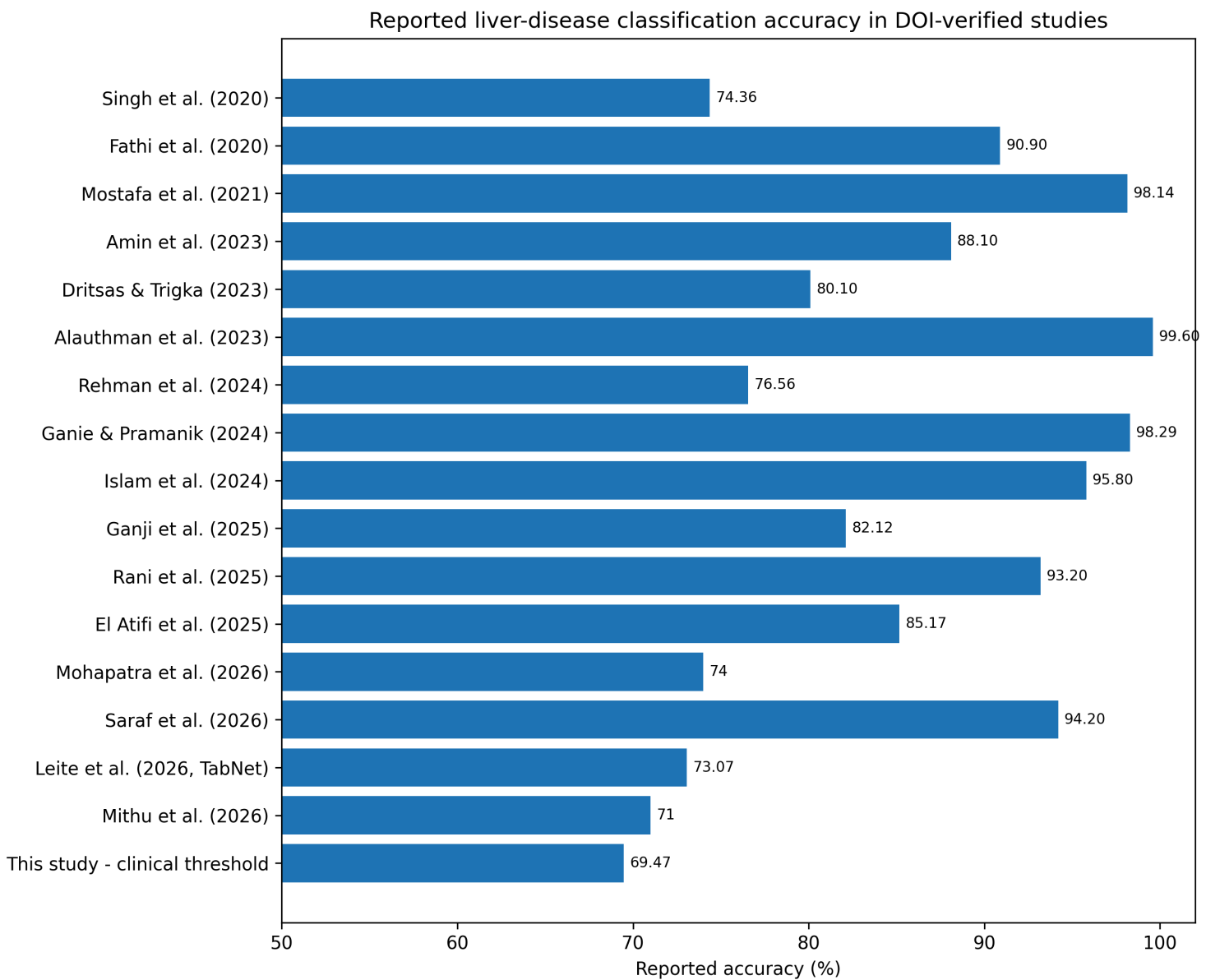


Figure 1

Headline accuracy reported by selected DOI-verified studies and the present study

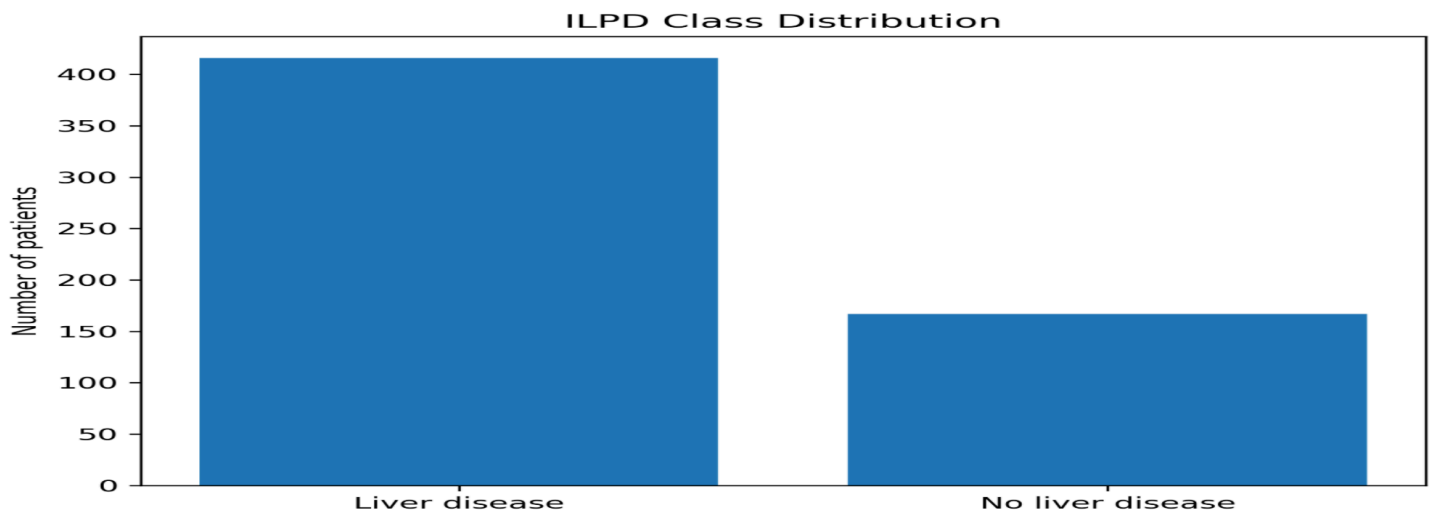


Figure 2

Class distribution of the 583-record ILPD cohort

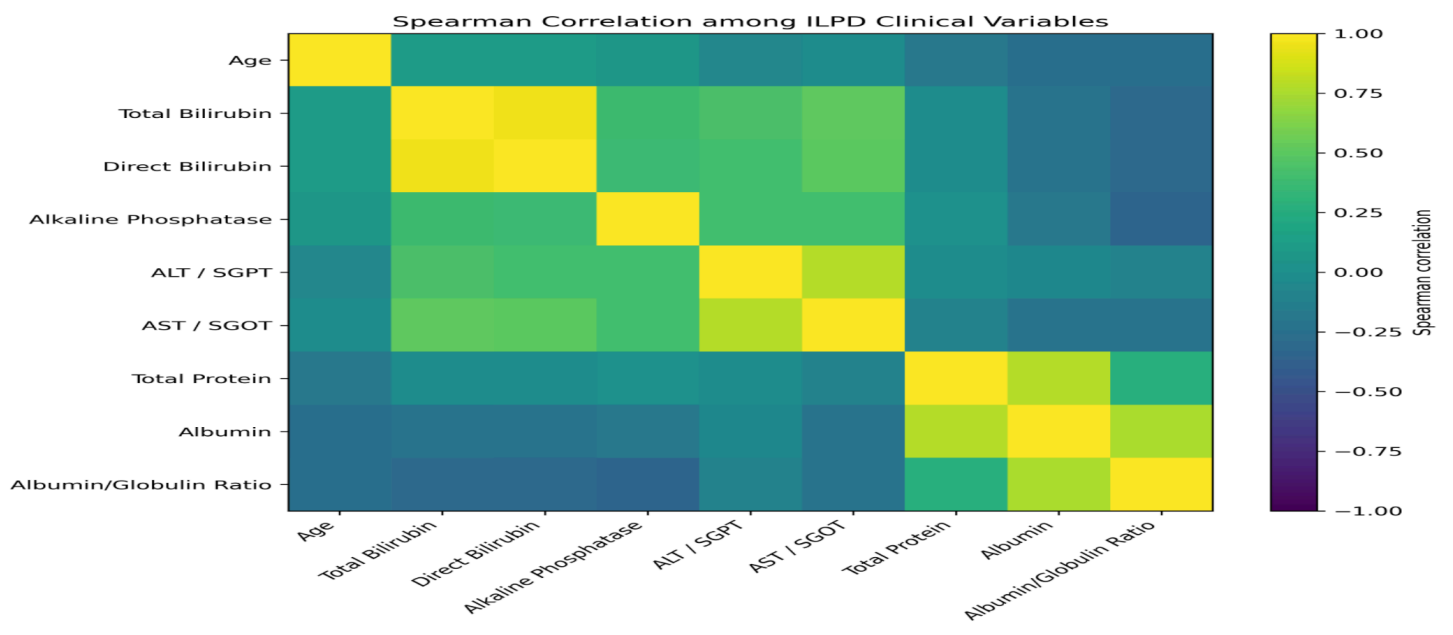


Figure 3

Spearman correlation matrix for numeric ILPD clinical variables

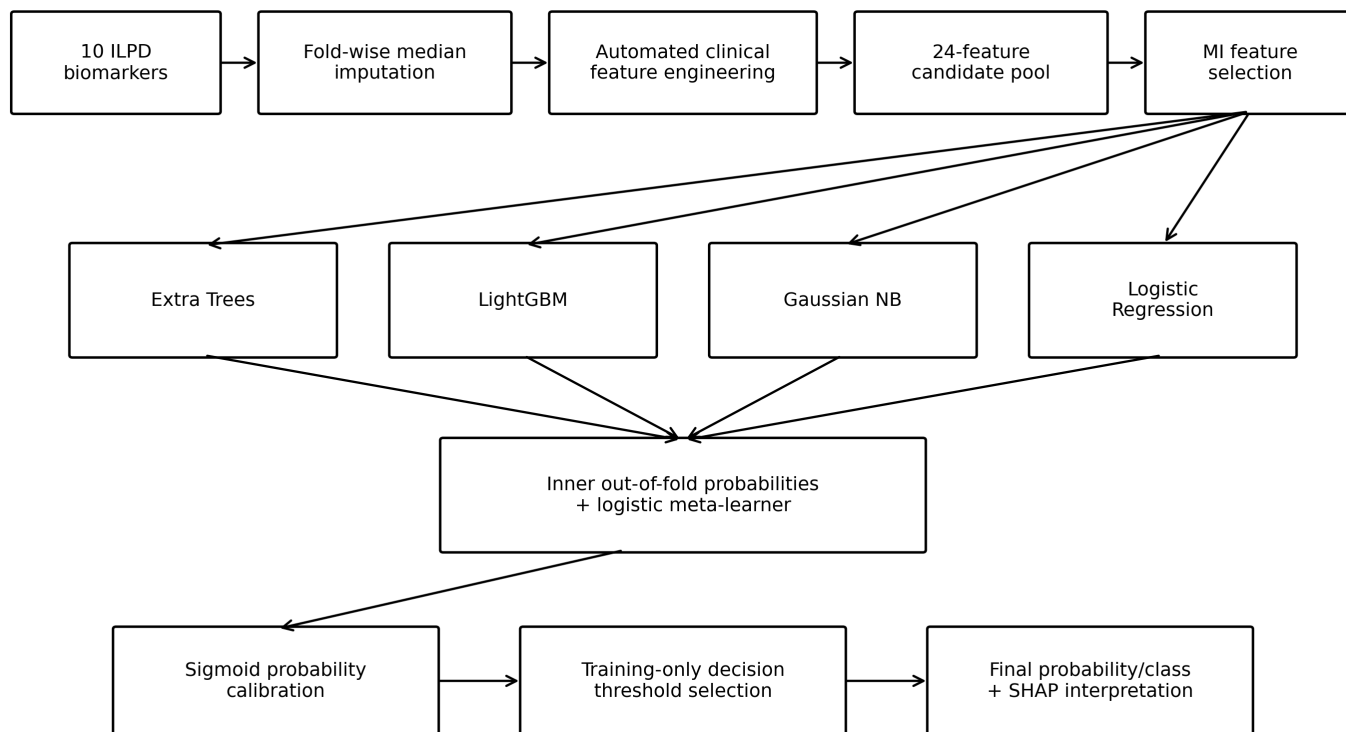


Figure 4

Proposed leakage-aware architecture with automated clinical feature engineering

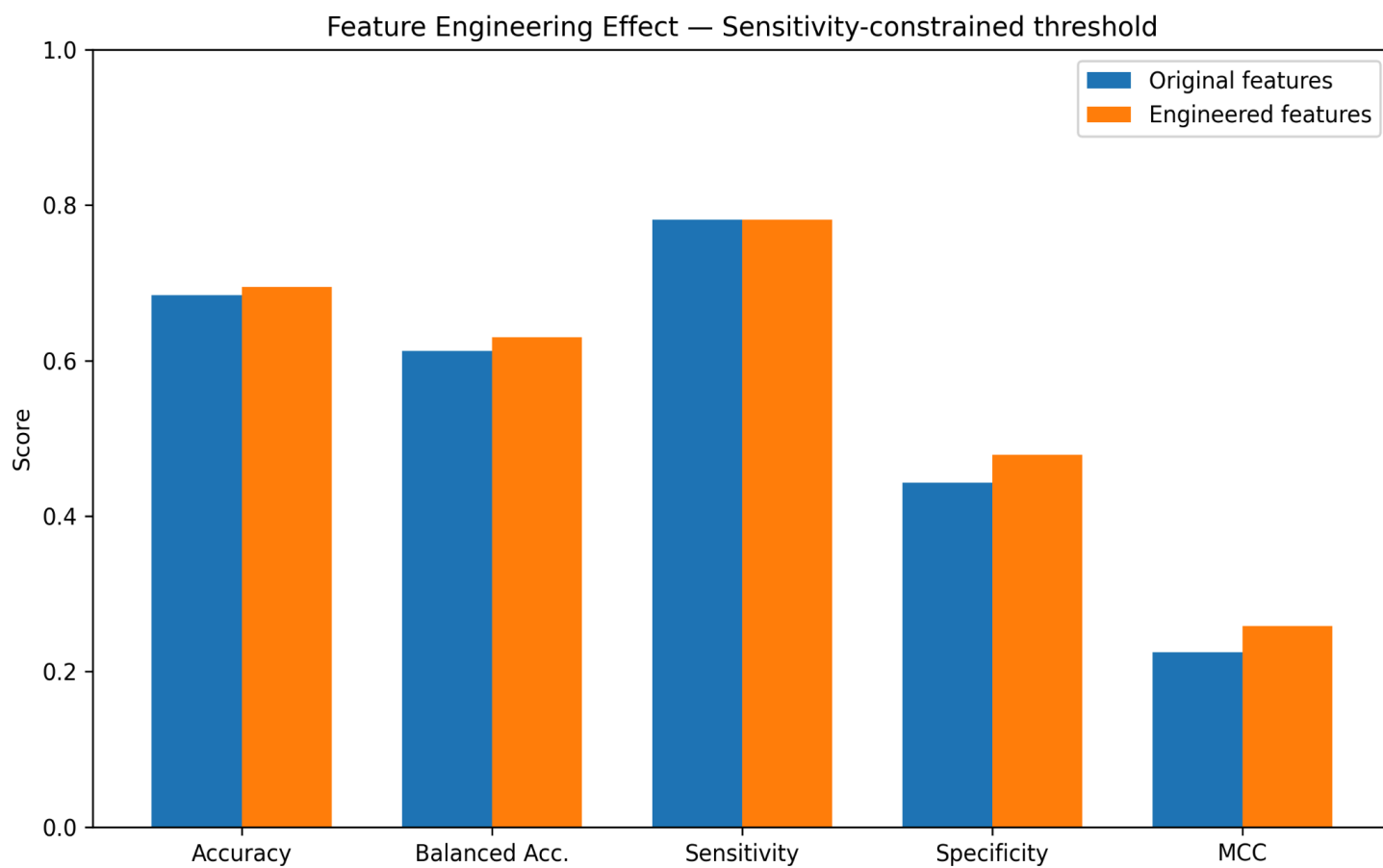


Figure 5

Effect of automated feature engineering under the sensitivity-constrained threshold policy

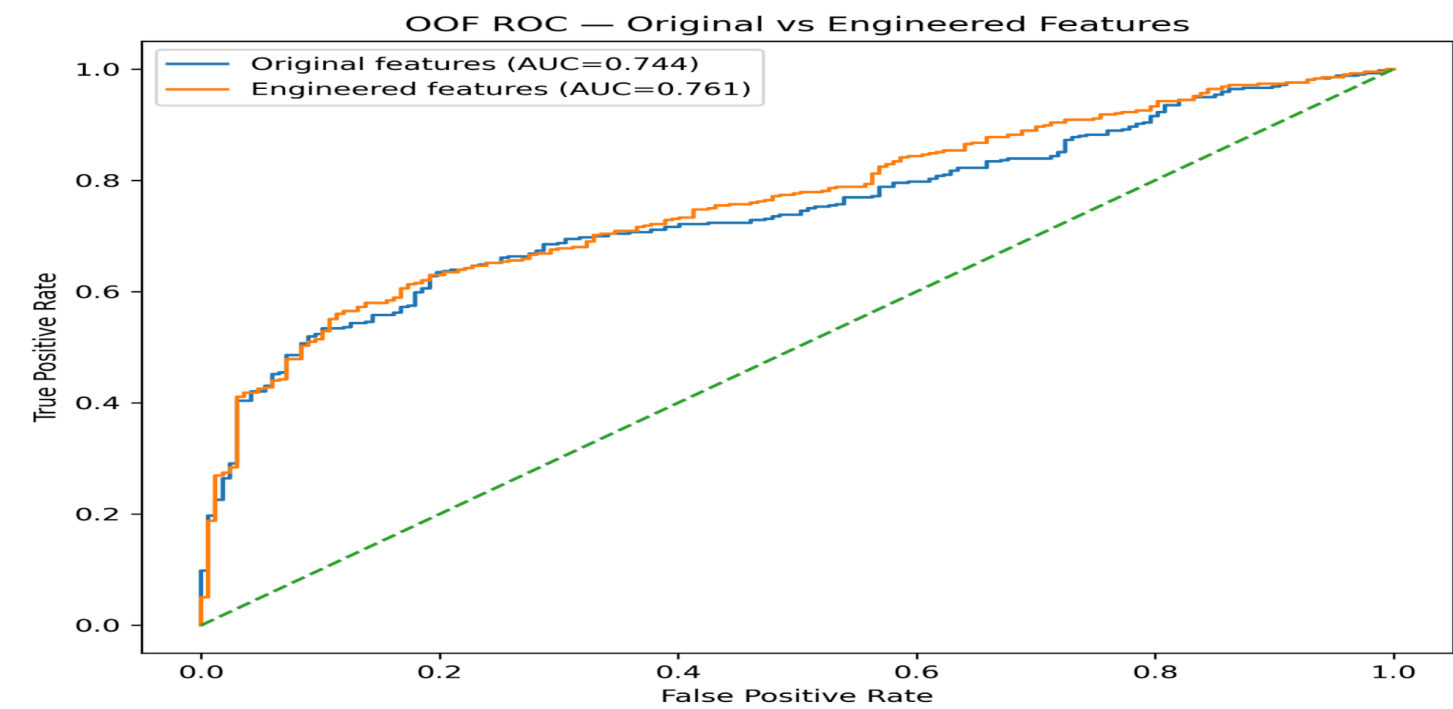


Figure 6

Out-of-fold ROC comparison between the original and engineered feature sets

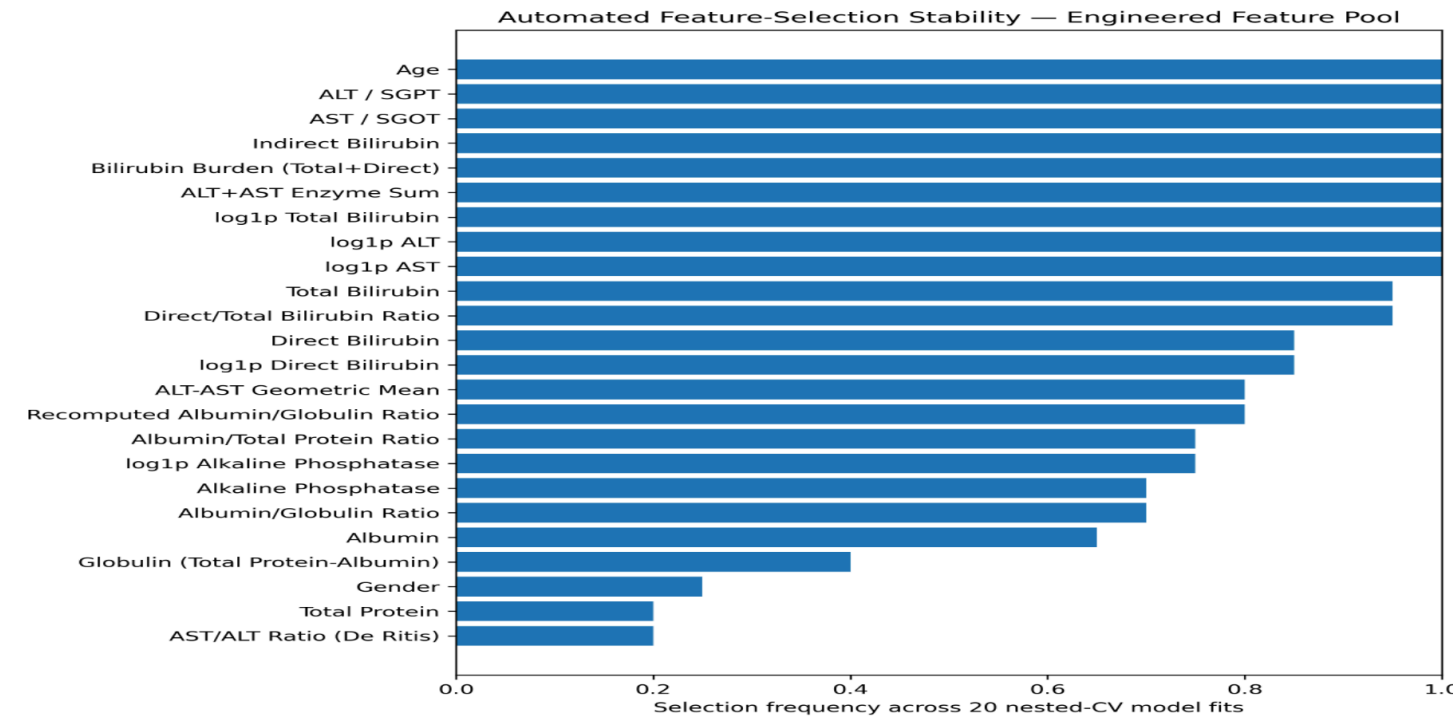


Figure 7

Selection frequency across the 24-feature engineered candidate pool

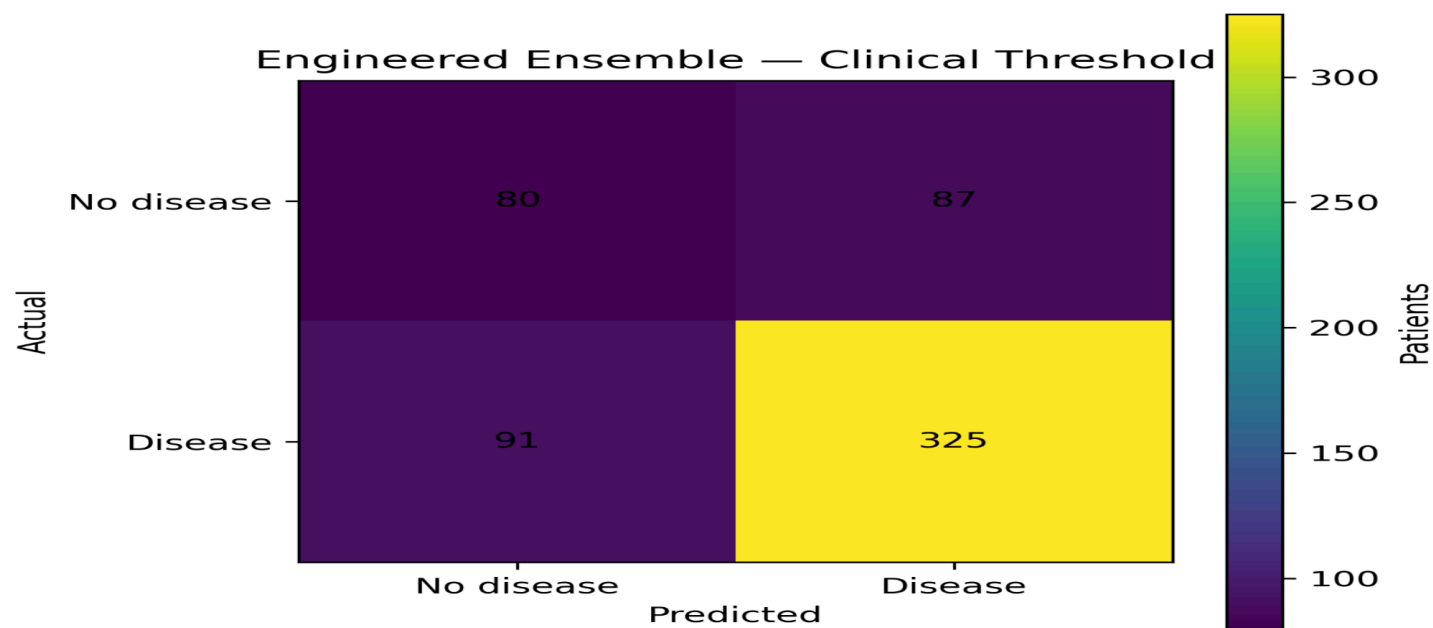


Figure 8

Out-of-fold confusion matrix for the engineered ensemble under the sensitivity-constrained threshold.

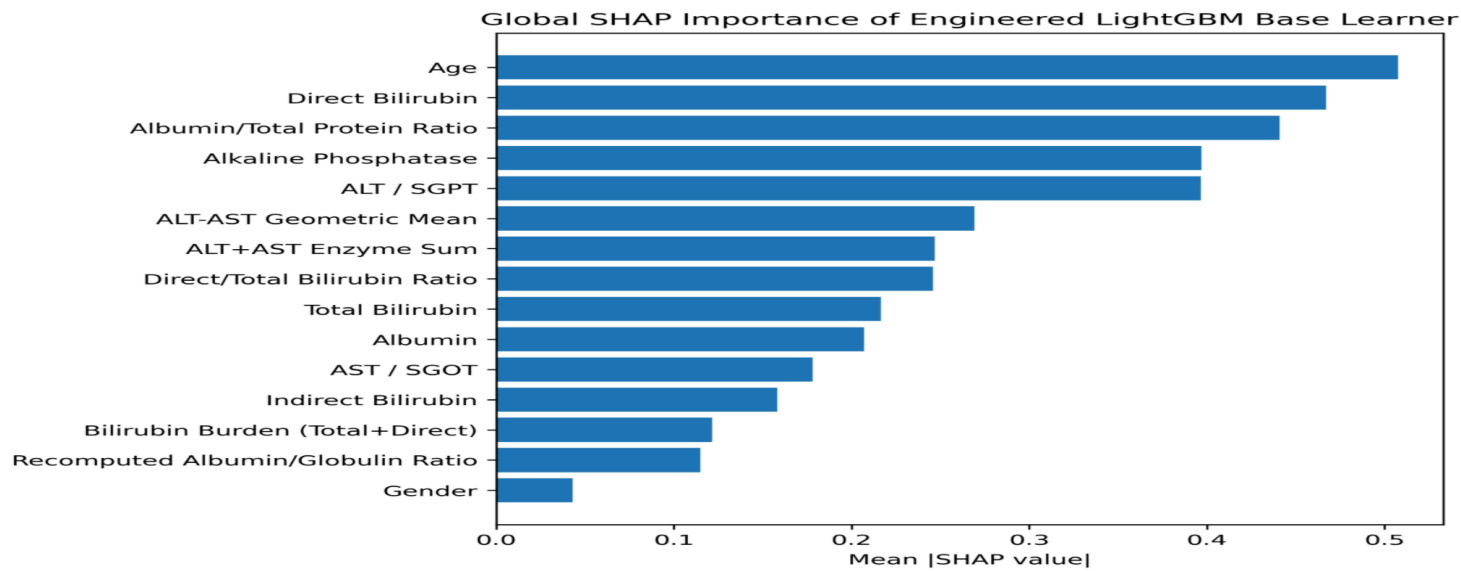


Figure 9

Post-hoc mean absolute SHAP importance for a tuned LightGBM base learner fitted to the engineered feature pool.