Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Liver Disease Prediction Using Machine Learning: A Leakage Free Evaluation of SVM, Logistic Regression, Random Forest and XGBoost on the Indian Liver Patient Dataset

Nnaemeka Ugwumba

nnaemeka_ugwumba@uniport.edu.ng

University of Port Harcourt https://orcid.org/0009-0000-2493-9846

Clement Barisua Ndeekor

clement.ndeekor@uniport.edu.ng

University of Port Harcourt, Port Harcourt, Nigeria

Research Article

Keywords: liver disease prediction, machine learning, support vector machine, logistic regression, XGBoost, random forest, Indian Liver Patient Dataset, nested cross validation, data leakage, class imbalance, SHAP, algorithmic fairness, clinical decision support

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11662916/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Liver Disease Prediction Using Machine Learning: A Leakage Free Evaluation of SVM, Logistic Regression, Random Forest and XGBoost on the Indian Liver Patient Dataset

Nnaemeka Kingsley Ugwumba
nnaemeka_ugwumba@uniport.edu.ng
Department of Computer Science, Faculty of Computing
University of Port Harcourt, Port Harcourt, Nigeria

Clement Barisua Ndeekor
clement.ndeekor@uniport.edu.ng
Department of Computer Science, Faculty of Computing
University of Port Harcourt, Port Harcourt, Nigeria

Abstract

Liver disease prediction using machine learning has been studied widely on the Indian Liver Patient Dataset (ILPD), yet many reported accuracies come from a single train and test split, from resampling carried out before the data are divided, or from duplicated patient records that appear in both training and test data. This study presents a leakage free evaluation of five classifiers for liver disease detection from routine liver function tests: support vector machine (SVM), logistic regression, random forest, XGBoost and k nearest neighbours (kNN). After removing 13 duplicate records, 570 patients were analysed. Imputation, log transformation, scaling and minority oversampling were fitted inside each training fold, and performance was estimated by nested cross validation with three repeats of stratified fivefold outer validation and a fivefold inner grid search. Logistic regression reached the highest mean ROC AUC of 0.754 (95% CI 0.717 to 0.790) and the SVM reached 0.741 (0.683 to 0.800). No pairwise difference was significant under the corrected resampled t test with Holm adjustment. A preliminary single split experiment with the same SVM had given an AUC of 0.806, which shows how weak evaluation inflates published results. Handling class imbalance mattered more than the choice of classifier, since the SVM fell to an AUC of 0.690 without it. Every model detected disease less well in women than in men (SVM AUC 0.629 against 0.785). SHAP values and permutation importance ranked alanine aminotransferase, bilirubin, aspartate aminotransferase and alkaline phosphatase as the strongest predictors. On a second liver dataset for hepatitis C related disease, every model exceeded an AUC of 0.97, and the final SVM pipeline scores one patient in 1.6 ms. These findings give a realistic baseline for machine learning liver disease screening and show that sex stratified validation is needed before clinical use.

Keywords: liver disease prediction, machine learning, support vector machine, logistic regression, XGBoost, random forest, Indian Liver Patient Dataset, nested cross validation, data leakage, class imbalance, SHAP, algorithmic fairness, clinical decision support

1 Introduction

Liver disease causes about two million deaths each year, roughly four percent of all deaths worldwide, and about two thirds of these deaths occur in men [1, 15]. Much of this burden comes from conditions such as viral hepatitis, alcohol related liver disease and metabolic fatty liver disease, which progress for years with few symptoms. Early signs often appear first in routine blood chemistry. Bilirubin, the aminotransferases (ALT and AST), alkaline phosphatase and albumin are measured in most hospital laboratories at low cost, and clinical guidelines use them as the first step in evaluating suspected liver injury [25]. A model that flags likely liver disease from these values could help clinics with few specialists decide which patients to refer for imaging, serology or biopsy.

This possibility has made liver disease prediction a common task in applied machine learning. The Indian Liver Patient Dataset (ILPD), collected in Andhra Pradesh and released through the UCI Machine Learning Repository [34, 36], is the most used public benchmark for it. Published studies on this dataset report a wide range of results, from an accuracy of about 80 percent [17] to more than 90 percent [29], and accuracies near 98 percent were reported on an early version of the data [35]. Many other comparisons of classifiers have been published on the same records [2, 39]. Such a wide range on nearly the same records is a warning sign. Differences in evaluation design, rather than differences in algorithms, can explain much of it. Common practices include reporting a single random split, balancing the classes before splitting so that synthetic or copied samples reach the test data, tuning hyperparameters on the same data used to report accuracy, and ignoring the duplicate records in the file. Each of these practices is a form of data leakage, which makes models look better than they will be on new patients [23, 24].

Our own first experiments followed this pattern. An SVM tuned and tested on one 80/20 split, with duplicate records retained and oversampling applied to the whole training set before the tuning folds were formed, gave an accuracy of 0.735 and a ROC AUC of 0.806 on 117 test patients. We then rebuilt the experiment so that every data dependent step is fitted inside the training folds of a nested, repeated cross validation. Under this design the same SVM reached a mean AUC of 0.741. The present paper reports that rebuilt study.

The study asks four questions. First, how well do common classifiers detect liver disease on ILPD when evaluation is free of leakage? Second, are the differences between classifiers statistically reliable? Third, how much do class imbalance handling and the choice of decision threshold change clinical usefulness? Fourth, does performance differ between male and female patients, and which laboratory values drive the predictions?

The contributions are as follows.

1. A leakage free benchmark of SVM, logistic regression, random forest, XGBoost and kNN on ILPD after duplicate removal, using nested repeated cross validation and confidence intervals corrected for overlapping training sets.

2. A statistical comparison with three complementary tests and correction for multiple comparisons, which shows that no classifier is reliably better than logistic regression on this task.

3. An ablation of class imbalance strategies and an analysis of operating thresholds, calibration and sex

specific performance that clinical users need before deployment.

4. Model explanations with permutation importance and SHAP, a cross dataset check on a hepatitis C laboratory dataset, and a measurement of inference cost for the final pipeline. All code runs in a single notebook so that the results can be reproduced.

2 Literature Review

2.1 Machine learning on the Indian Liver Patient Dataset

Ramana, Babu and Venkateswarlu introduced the liver patient data from Andhra Pradesh and compared naive Bayes, C4.5, back propagation networks, kNN and SVM on it and on the BUPA liver disorders data from the United States [35]. On an early version of the Andhra Pradesh data with 751 records they reported accuracies between about 95 and 98 percent, with kNN performing best. A later paper compared the Indian and American patient groups and described the ILPD as released to the public, with 416 liver patients and 167 patients without liver disease [34, 36].

Since then the dataset has served as a standard benchmark. Auxilia compared several machine learning techniques for predicting liver disease in Indian patients [2]. Singh, Bagga and Kaur combined feature selection methods with several classifiers in a software based prediction system [39]. Dritsas and Trigka evaluated a large set of models, including naive Bayes, logistic regression, SVM, decision trees, random forest, rotation forest, boosting, stacking and voting ensembles [17]. They oversampled the minority class with SMOTE to obtain 828 balanced instances and reported, under tenfold cross validation on this balanced set, an accuracy of 80.1 percent and an AUC of 0.884 for a soft voting ensemble. Md and colleagues applied iterative imputation, log transformation, upsampling to 832 records, several scaling methods and feature selection, and reported a test accuracy of 91.82 percent for an extra trees classifier [29]. Their description does not make clear whether upsampling and feature selection were fitted on the training data alone.

These studies show steady interest in the task, but they also show why results are hard to compare. They use different preprocessing steps, different splits and different metrics, mostly accuracy, which is misleading when about 71 percent of patients belong to one class. When resampling is applied before the data are divided, copies or interpolations of test patients enter the training data, and reported performance rises without any real gain in predictive ability.

2.2 Data leakage and model evaluation

Kaufman and colleagues defined leakage as the use of information in model training that would not be available when the model is used in practice, and showed how it arises in data mining competitions and applications [24]. Kapoor and Narayanan reviewed published machine learning based science across many fields and found leakage to be a leading cause of irreproducible results [23]. Their taxonomy includes preprocessing on the full dataset, duplicates across training and test data, and feature selection outside the validation loop, all of which are common in ILPD studies.

A second problem is model selection bias. When hyperparameters are tuned and performance is reported on the same folds, the estimate is optimistic. Varma and Simon showed that this bias can be large for small samples and recommended nested cross validation, where an inner loop selects the model and an outer loop measures it [42]. Cawley and Talbot reached the same conclusion and showed that the selection bias can be as large as the differences between learning algorithms [6].

Statistical comparison of classifiers needs care as well. Dietterich showed that the standard paired t test on resampled data has a high type I error and recommended McNemar’s test or a five by two cross validated test [16, 28]. Nadeau and Bengio corrected the variance of the resampled t test for the overlap between training sets [30]. For AUC, the DeLong test compares correlated ROC curves without distributional assumptions [13], and Sun and Xu gave a fast algorithm for it [41]. Demšar reviewed tests for comparing classifiers and stressed correction for multiple comparisons [14], for which the Holm procedure is a standard choice [22].

2.3 Class imbalance and clinical metrics

He and Garcia surveyed methods for learning from imbalanced data, including random oversampling, undersampling and cost sensitive learning [20]. SMOTE creates synthetic minority samples by interpolation between neighbours [7]. With imbalanced classes, accuracy and the ROC curve can hide poor detection of the minority class. Saito and Rehmsmeier argued that the precision recall plot is more informative in this setting [38], and Chicco and Jurman showed that the Matthews correlation coefficient (MCC) gives a more truthful single number than accuracy or F1 [9]. For clinical prediction models, calibration matters as much as discrimination, since predicted risks guide decisions [5], and the Brier score combines both aspects [4]. The TRIPOD statement sets out what a report of a prediction model should contain [11].

A broad systematic review by Christodoulou and colleagues covered 71 studies and found no performance benefit of machine learning over logistic regression for clinical prediction models, while noting weak methodology in many comparisons [10]. This finding is relevant to ILPD, where complex ensembles are often reported as best without a fair comparison to a regularised linear model.

2.4 Interpretability and fairness in medical prediction

Clinical users need to know why a model flags a patient. Permutation importance measures the loss in performance when a feature is shuffled [3, 19], and SHAP values attribute each prediction to the input features using Shapley values from game theory [26]. Rudin argued that for high stakes decisions an interpretable model should be preferred when it performs as well as a black box model [37].

Fairness is a related concern. Obermeyer and colleagues showed that a widely used health algorithm underestimated the needs of Black patients [31]. For liver disease, Straw and Wu analysed ILPD by sex and found that female patients had a higher false negative rate across all four classifiers they tested, with gaps of about 24 percentage points for logistic regression and 21 points for random forest [40]. Feature selection did not remove the disparity. Their work is one of the few that look beyond overall accuracy on this dataset.

2.5 Benchmark quality and research gap

Benchmark datasets can carry hidden problems. McDermott and Forsyth showed that most papers using the BUPA liver disorders data treat an unsuitable field as the diagnosis label [27]. Hoffmann and colleagues used a hepatitis C laboratory dataset to build diagnostic pathways with machine learning [21], and this dataset offers a second source of liver test data with confirmed diagnoses.

From this review we identify a gap. Few ILPD studies remove duplicate records, confine all preprocessing and resampling to training folds, use nested cross validation, report confidence intervals and formal tests, and examine calibration, thresholds and sex specific performance together. The present study addresses these points in one reproducible pipeline.

3 Materials and Methods

3.1 Data

The ILPD contains 583 records with ten input attributes and a binary label [34]. The inputs are age, sex, total bilirubin, direct bilirubin, alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), total proteins, albumin and the albumin to globulin (A/G) ratio. The label marks patients with liver disease (416 records) and without liver disease (167 records). Table 1 summarises the attributes. Four values of the A/G ratio are missing.

We found 13 exact duplicate rows and removed them, which left 570 patients: 406 with liver disease and 164 without, a prevalence of 71.2 percent. Of these, 430 were male and 140 female, and the prevalence of disease was 73 percent among men and 65 percent among women (Figure 1). Removing duplicates matters because identical records may otherwise fall into both training and test folds.

Table 1. Attributes of the ILPD (all 583 records before duplicate removal; values as recorded in the source data)

Attribute	Mean (SD)	Median	Range
Age (years)	44.75 (16.19)	45.0	4 to 90
Total bilirubin	3.30 (6.21)	1.0	0.4 to 75.0
Direct bilirubin	1.49 (2.81)	0.3	0.1 to 19.7
Alkaline phosphatase	290.58 (242.94)	208	63 to 2110
Alanine aminotransferase	80.71 (182.62)	35	10 to 2000
Aspartate aminotransferase	109.91 (288.92)	42	10 to 4929
Total proteins	6.48 (1.09)	6.6	2.7 to 9.6
Albumin	3.14 (0.80)	3.1	0.9 to 5.5
A/G ratio (4 missing)	0.95 (0.32)	0.93	0.3 to 2.8
Sex			441 male (75.6%), 142 female

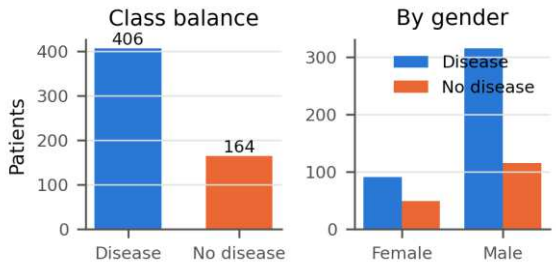


Figure 1. Class balance of the 570 unique ILPD patients overall and by sex.

The enzyme and bilirubin values are strongly right skewed, with maxima tens of times larger than the medians. Figure 2 shows the Spearman correlations. Total and direct bilirubin are almost perfectly correlated, ALT and AST correlate at 0.8, and total proteins, albumin and the A/G ratio form a second correlated group. The two groups are only weakly related, which suggests that the data hold two broad signals: hepatocellular injury and cholestasis on one side, and synthetic function on the other.

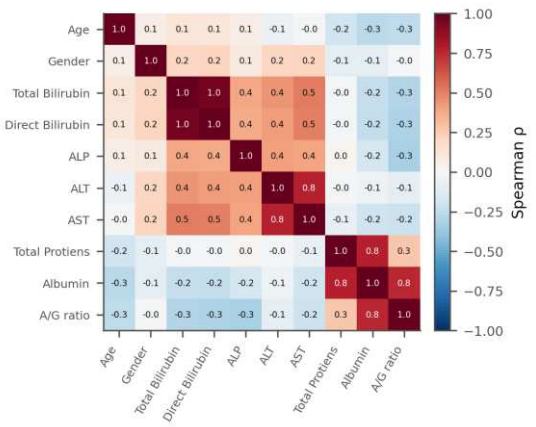


Figure 2. Spearman correlation between ILPD attributes.

For the cross dataset check we used the hepatitis C dataset of Hoffmann and colleagues [21], obtained from Kaggle. It contains 615 records with age, sex and ten laboratory values (albumin, ALP, ALT, AST, bilirubin, cholinesterase, cholesterol, creatinine, gamma glutamyl transferase and protein). We removed seven records labelled as suspect blood donors and defined the positive class as hepatitis, fibrosis or cirrhosis. This left 608 records with 75 positives. No duplicates were found.

3.2 Preprocessing inside the training folds

All data dependent steps were placed in one pipeline that is fitted on training data only and then applied unchanged to the matching test data. The pipeline has four steps. Missing values are replaced with the median of the training fold. Every feature is transformed with $\log(1 + x)$ to reduce skew, after clipping negative values at zero. Features are standardised to zero mean and unit variance using training fold statistics. Finally, the minority class in the training fold is randomly oversampled with replacement until both classes are the same size. Sex is coded as one for male and zero for female.

Because the pipeline is refitted for every inner and outer training fold, no test patient influences imputation values, scaling constants or resampled records. This removes the forms of leakage described in Section 2.2.

3.3 Classifiers and hyperparameter search

We compared five classifiers implemented in scikit learn [32] and the XGBoost library [8]. The SVM uses a radial basis function kernel [12], with probability estimates obtained by Platt scaling [33]. Logistic regression uses L2 regularisation. The random forest has 300 trees [3]. XGBoost uses 200 boosted trees with row and column subsampling of 0.8. The kNN classifier uses Euclidean distance on the standardised features. Table 2 lists the search grids. Every grid was searched exhaustively in the inner loop with ROC AUC as the selection criterion.

Table 2. Hyperparameter grids searched in the inner loop

Model	Grid
SVM	$C \in \{0.1, 1, 10, 100\}$, $\gamma \in \{\text{scale}, 0.01, 0.05, 0.1\}$
Logistic regression	$C \in \{0.01, 0.1, 1, 10\}$
Random forest	max depth $\in \{\text{none}, 4, 8\}$, min leaf $\in \{1, 5\}$, max features $\in \{\sqrt{p}, 0.5\}$
XGBoost	max depth $\in \{2, 3, 4\}$, learning rate $\in \{0.03, 0.1\}$, min child weight $\in \{1, 5\}$
kNN	$k \in \{5, 11, 21, 31\}$, weights $\in \{\text{uniform}, \text{distance}\}$

3.4 Nested repeated cross validation

Performance was estimated with nested cross validation [6, 42]. The outer loop is stratified fivefold cross validation repeated three times with different shuffles, which gives 15 outer test folds of about 114 patients each. Inside each outer training set, a

stratified fivefold inner loop selects hyperparameters. The selected configuration is refitted on the whole outer training set and scored once on the outer test fold. Hyperparameters were therefore never chosen with the help of the data on which they were scored. All random processes used a fixed seed of 42.

3.5 Metrics and statistical analysis

On each outer test fold we computed ROC AUC, the area under the precision recall curve (PR AUC), balanced accuracy, accuracy, sensitivity, specificity, precision, F1, MCC and the Brier score, with a default threshold of 0.5 for the threshold based metrics. We report the mean over the 15 folds with a 95 percent confidence interval based on the variance correction of Nadeau and Bengio [30], which accounts for the overlap between training sets in repeated cross validation. Intervals are truncated at 1.0 where the correction exceeds the range of a metric.

The SVM was taken as the reference model, since it was the starting point of the study. It was compared with each other model in three ways: the corrected resampled t test on the 15 paired fold AUCs [30]; the DeLong test on pooled out of fold predictions from the first repeat [13, 41]; and the exact McNemar test on the pooled predicted labels at a threshold of 0.5 [16, 28]. Each family of p values was adjusted with the Holm procedure [22].

3.6 Imbalance ablation, thresholds, calibration and subgroups

To measure the effect of imbalance handling, the SVM was evaluated under the same nested design with four strategies: no correction, class weights inversely proportional to class frequency, random oversampling and SMOTE [7]. For the SVM we examined operating points between thresholds of 0.27 and 0.70 on the pooled out of fold predictions, including the threshold that maximises the Youden index [43] and the lowest threshold that keeps sensitivity at or above 0.90. Calibration was assessed with reliability curves over eight quantile bins and with the Brier score [5]. For each model and each sex we computed ROC AUC, sensitivity and specificity on the pooled out of fold predictions, with 95 percent bootstrap intervals from 1,000 resamples [18].

3.7 Interpretation, external data and deployment

Permutation importance was computed on each held out fold of the first repeat as the mean drop in ROC AUC over ten shuffles of each feature [19]. This measures importance on data the model did not see. SHAP values [26] were computed for a final SVM tuned on all 570 patients, using the kernel explainer with a background of 20 k means centroids, applied to 150 randomly selected patients.

On the hepatitis C data we ran the same five pipelines with two repeats of fivefold nested cross validation. Finally, the tuned SVM pipeline was saved and timed on four CPU cores: single patient latency over 300 calls after a warm up, and the time to score a batch of 1,000 patients. The experiments ran in Python 3.13.16 with scikit learn 1.6.1, NumPy 2.1.3 and pandas 2.3.3 on a Kaggle notebook, and the full run took 10.6 minutes.

4 Results

4.1 Discrimination and overall performance

Table 3 gives the nested cross validation results on ILPD. All five classifiers fall in a narrow band of ROC AUC between 0.723 and 0.754. Logistic regression has the highest mean ROC AUC (0.754), PR AUC (0.894), balanced accuracy (0.701), specificity (0.803) and MCC (0.366). The SVM follows with a ROC AUC of 0.741 and the widest interval of the five. Random forest and XGBoost give slightly higher sensitivity and F1 than the SVM but lower specificity, and kNN is lowest on most measures. Every PR AUC is well above the prevalence baseline of 0.71, so all models carry real signal, but none separates the classes well.

Table 3. Nested cross validation results on ILPD (570 patients). Each cell shows the mean over 15 outer folds with the 95% corrected confidence interval below it. Threshold based metrics use a threshold of 0.5. Lower Brier scores are better.

Model	ROC AUC	PR AUC	Bal. acc.	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	Brier
SVM	0.741 (0.683, 0.800)	0.885 (0.854, 0.916)	0.690 (0.644, 0.736)	0.655 (0.610, 0.700)	0.608 (0.541, 0.674)	0.772 (0.678, 0.866)	0.870 (0.825, 0.916)	0.714 (0.667, 0.761)	0.346 (0.262, 0.429)	0.207 (0.187, 0.227)
Logistic regression	**0.754** (0.717, 0.790)	**0.894** (0.878, 0.910)	**0.701** (0.667, 0.735)	0.657 (0.613, 0.702)	0.599 (0.521, 0.676)	**0.803** (0.720, 0.886)	**0.885** (0.848, 0.922)	0.712 (0.659, 0.764)	**0.366** (0.304, 0.428)	0.203 (0.191, 0.216)
Random forest	0.727 (0.684, 0.769)	0.883 (0.863, 0.902)	0.668 (0.600, 0.735)	0.656 (0.618, 0.693)	0.640 (0.545, 0.734)	0.695 (0.493, 0.898)	0.848 (0.777, 0.918)	0.724 (0.678, 0.769)	0.308 (0.188, 0.428)	**0.201** (0.189, 0.213)
XGBoost	0.726 (0.681, 0.770)	0.876 (0.856, 0.896)	0.665 (0.607, 0.724)	**0.658** (0.615, 0.701)	**0.648** (0.588, 0.708)	0.683 (0.553, 0.813)	0.838 (0.784, 0.892)	**0.729** (0.689, 0.768)	0.303 (0.196, 0.409)	0.204 (0.189, 0.218)
kNN	0.723 (0.679, 0.767)	0.876 (0.850, 0.903)	0.671 (0.633, 0.709)	0.623 (0.577, 0.669)	0.557 (0.482, 0.633)	0.784 (0.703, 0.866)	0.867 (0.828, 0.905)	0.676 (0.622, 0.730)	0.312 (0.243, 0.382)	0.225 (0.203, 0.246)

Figure 3 shows the mean ROC and precision recall curves over the outer folds. The curves overlap across most of their range, and the shaded band around the SVM curve covers the other models. The ROC curves rise steeply at low false positive rates, so the models identify a subset of clearly abnormal patients with few false alarms, and then flatten.

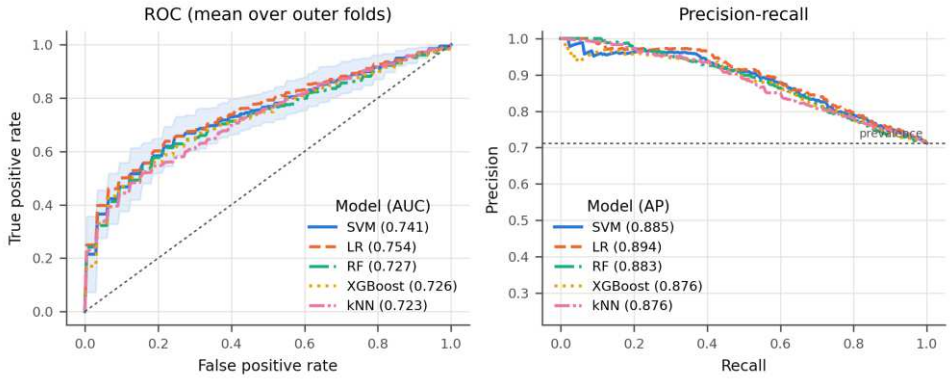


Figure 3. Mean ROC curves (left) and precision recall curves (right) over the 15 outer folds. The shaded band shows one standard deviation around the SVM curve. The dotted line on the right marks the prevalence of 0.71.

Figure 4 plots the main metrics with their corrected confidence intervals. The intervals overlap for every metric and every pair of models. The largest contrast is in the balance between sensitivity and specificity. The tree ensembles detect more patients with disease at a threshold of 0.5 but raise more false alarms, while logistic regression, the SVM and kNN are more conservative.

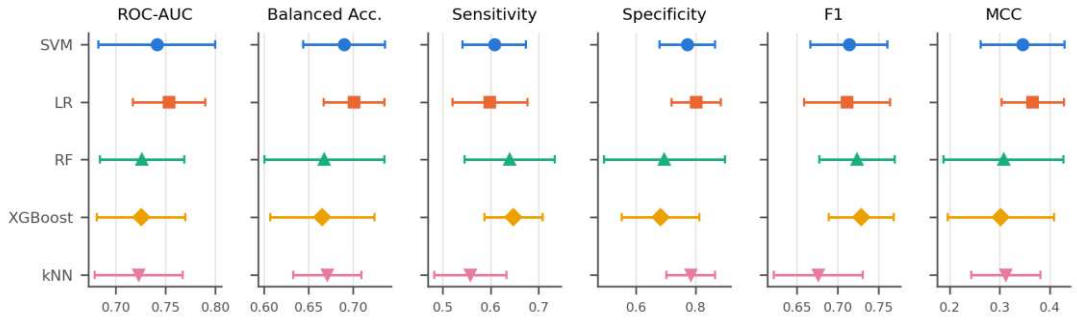


Figure 4. Mean performance with 95% corrected confidence intervals over 15 outer folds.

The hyperparameters chosen by the inner loop point the same way. The SVM chose $\gamma = 0.01$, a smooth decision boundary, in 11 of 15 outer folds, and logistic regression most often chose the strongest regularisation ($C = 0.01$, 7 of 15). Random forest chose a maximum depth of 4 in 11 folds, XGBoost chose a learning rate of 0.03 in 14 folds and a depth of 2 in 13 folds, and kNN chose $k = 31$ in 10 folds. In every family the search preferred simple models, which suggests that the data hold limited structure beyond what a smooth or nearly linear boundary can capture.

4.2 Statistical comparison

Table 4 reports the tests against the SVM. The corrected resampled t test found no difference in fold AUC for any pair (all unadjusted p values above 0.46). The DeLong test on the pooled predictions of the first repeat found the SVM ahead of XGBoost (adjusted $p = 0.010$) and close to the threshold of significance against random forest (adjusted $p = 0.054$). The McNemar test found no difference in error patterns after adjustment. The DeLong test treats one set of predictions as fixed and ignores the variation that comes from retraining on different folds, so it tends to be more liberal than the corrected t test. Taking the three tests together, we find no reliable difference among the classifiers, and we found none in favour of the SVM over logistic regression.

Table 4. SVM against each other model. Δ is SVM minus the other model. t: corrected resampled t test on 15 fold AUCs. DeLong and McNemar: pooled out of fold predictions of the first repeat. All p values are Holm adjusted; b/c counts patients classified correctly by only the SVM or only the other model.

Comparison	ΔAUC (folds)	p (t)	ΔAUC (pooled)	p (DeLong)	b/c	p (McNemar)
SVM vs LR	−0.012	1.000	0.014	0.175	19/12	0.562
SVM vs RF	0.015	1.000	0.027	0.054	26/22	0.665
SVM vs XG Boost	0.016	1.000	0.038	0.010	45/31	0.406
SVM vs kNN	0.018	1.000	0.014	0.372	49/28	0.088

4.3 Effect of class imbalance handling

Table 5 shows the SVM under four imbalance strategies. Without correction, the SVM labels almost every patient as diseased: sensitivity is 0.965 but specificity is only 0.091, balanced accuracy is close to chance at 0.528, and the ROC AUC drops to 0.690. Class weights restore the ranking ability (AUC 0.754) but leave the default threshold badly placed, with a specificity of 0.238. Random oversampling and SMOTE give almost the same results, with balanced accuracy near 0.69 and MCC near 0.35. The gain from adding any imbalance correction (about 0.05 to 0.06 in AUC) is larger than the spread between the five classifiers in Table 3 (0.031).

Table 5. SVM with four class imbalance strategies under the same nested design (mean with 95% corrected CI; upper limits truncated at 1.0)

Strategy	ROC AUC	Bal. acc.	Sensitivity	Specificity	MCC
None	0.690 (0.643, 0.737)	0.528 (0.498, 0.558)	0.965 (0.924, 1.000)	0.091 (0.008, 0.174)	0.124 (0.021, 0.227)
Class weight	0.754 (0.716, 0.791)	0.577 (0.540, 0.614)	0.916 (0.866, 0.967)	0.238 (0.150, 0.326)	0.213 (0.126, 0.300)
Oversampling	0.741 (0.683, 0.800)	0.690 (0.644, 0.736)	0.608 (0.541, 0.674)	0.772 (0.678, 0.866)	0.346 (0.262, 0.429)
SMOTE	0.753 (0.717, 0.789)	0.695 (0.659, 0.731)	0.613 (0.554, 0.673)	0.776 (0.693, 0.860)	0.354 (0.288, 0.420)

4.4 Operating thresholds and calibration

At the default threshold the SVM misses many patients with disease. The pooled confusion matrix in Figure 5 shows 156 false negatives among 406 patients with liver disease, a sensitivity of 0.616, against 34 false positives among 164 patients without disease. The threshold that maximises the Youden index is 0.531, almost the same as the default, and gives a similar matrix.

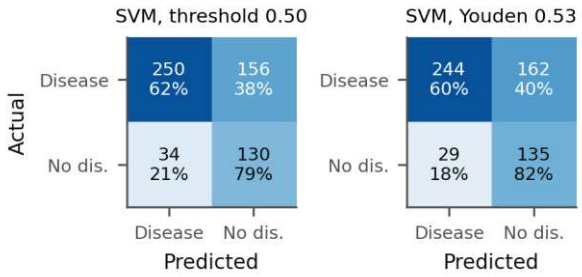


Figure 5. Pooled out of fold confusion matrices of the SVM for the first repeat at the default threshold and at the Youden threshold.

Table 6 and Figure 6 show the trade off across thresholds. To reach a sensitivity of 0.90, which would suit a screening setting, the threshold must fall to 0.273. Specificity then drops to 0.232, so about three in four patients without disease would be flagged. At a threshold of 0.40, sensitivity is 0.702 and specificity 0.640, a more balanced compromise. These thresholds were chosen on the same predictions on which they are reported, so the values in Table 6 are slightly optimistic and would need confirmation on new data.

Table 6. SVM operating points on pooled out of fold predictions (570 patients, first repeat)

Threshold	TP	FP	FN	TN	Sens.	Spec.	PPV	NPV

Threshold	TP	FP	FN	TN	Sens.	Spec.	PPV	NPV
0.273 (sens. ≥ 0.90)	367	126	39	38	0.904	0.232	0.744	0.494
0.300	354	111	52	53	0.872	0.323	0.761	0.505
0.400	285	59	121	105	0.702	0.640	0.828	0.465
0.500 (default)	250	34	156	130	0.616	0.793	0.880	0.455
0.531 (Youden)	244	29	162	135	0.601	0.823	0.894	0.455
0.600	217	21	189	143	0.534	0.872	0.912	0.431
0.700	180	14	226	150	0.443	0.915	0.928	0.399

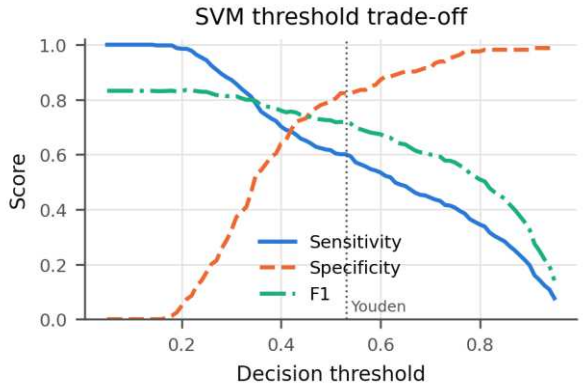


Figure 6. Sensitivity, specificity and F1 of the SVM as the decision threshold changes. The dotted line marks the Youden threshold.

Figure 7 shows that all models are poorly calibrated in the same direction. The observed rate of disease lies above the predicted probability in almost every bin. For example, patients given a predicted probability near 0.3 have an observed disease rate near 0.5. This is the expected result of training on data balanced to a 50 percent prevalence while the true prevalence is 71 percent. The ranking of patients is unaffected, but the raw probabilities should not be read as risks without recalibration to the target population. The Brier scores lie between 0.20 and 0.23 for all models.

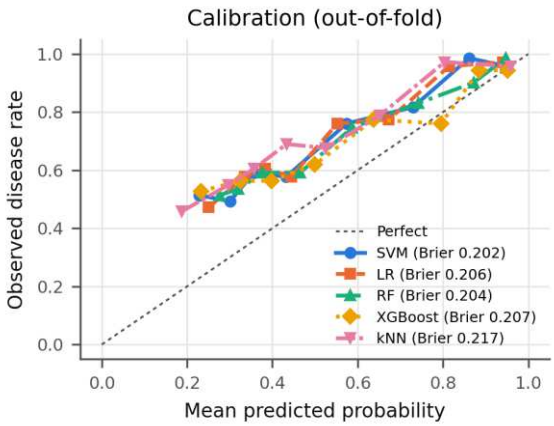


Figure 7. Calibration of out of fold predicted probabilities (first repeat, eight quantile bins).

4.5 Performance by sex

Table 7 and Figure 8 compare male and female patients. Every model discriminates less well for women. For the SVM, the ROC AUC is 0.785 for men and 0.629 for women, and the bootstrap intervals barely touch. Sensitivity for women is 0.418 against 0.673 for men, so the SVM misses more than half of the female patients with liver disease at the default threshold, while its specificity for women is higher (0.898). The same pattern holds for all five classifiers, with female ROC AUC between 0.586 and 0.630 and female sensitivity between 0.352 and 0.462. The female subgroup is small (140 patients, 91 with disease), so its intervals are wide, but the direction of the gap is consistent across models.

Table 7. Performance by sex on pooled out of fold predictions (first repeat) with 95% bootstrap intervals. Men: n = 430, prevalence 0.73. Women: n = 140, prevalence 0.65.

Model	Sex	ROC AUC	Sensitivity	Specificity
SVM	Male	0.785 (0.737, 0.831)	0.673 (0.621, 0.723)	0.748 (0.664, 0.825)
	Female	0.629 (0.539, 0.717)	0.418 (0.316, 0.525)	0.898 (0.809, 0.967)
LR	Male	0.764 (0.710, 0.810)	0.644 (0.595, 0.697)	0.757 (0.677, 0.833)
	Female	0.630 (0.543, 0.721)	0.451 (0.351, 0.554)	0.857 (0.762, 0.951)
RF	Male	0.758 (0.710, 0.809)	0.686 (0.635, 0.736)	0.713 (0.626, 0.795)
	Female	0.586 (0.492, 0.690)	0.418 (0.313, 0.529)	0.816 (0.705, 0.920)
XGBoost	Male	0.746 (0.696, 0.795)	0.698 (0.643, 0.749)	0.643 (0.557, 0.727)
	Female	0.591 (0.500, 0.682)	0.462 (0.362, 0.563)	0.612 (0.468, 0.761)
kNN	Male	0.768 (0.723, 0.816)	0.625 (0.575, 0.678)	0.783 (0.704, 0.854)
	Female	0.611 (0.512, 0.702)	0.352 (0.253, 0.451)	0.816 (0.698, 0.923)

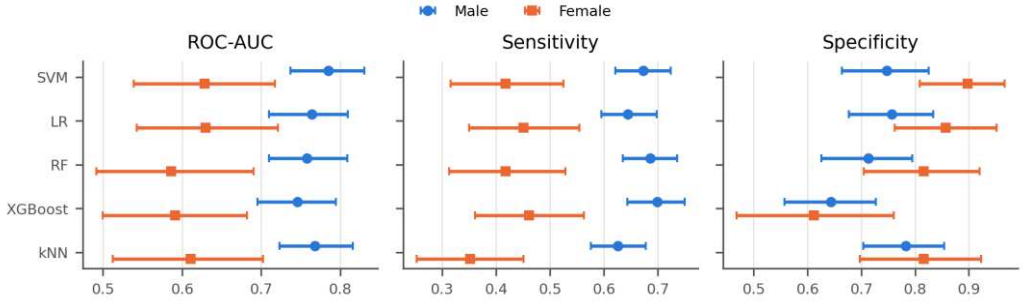


Figure 8. ROC AUC, sensitivity and specificity by sex with 95% bootstrap intervals.

4.6 Feature importance and explanations

Figure 9 gives permutation importance measured on held out folds. For the SVM, shuffling ALT causes the largest drop in AUC (0.027), followed by direct bilirubin, AST, age and ALP. Total bilirubin adds little once direct bilirubin is present, which fits their near perfect correlation. Total proteins, the A/G ratio and sex have importance close to zero. The other models agree on the leading group, although logistic regression gives more weight to direct bilirubin (0.057) and albumin (0.039). Sex has importance close to zero for every model (at most 0.009), so the lower performance for women is not caused by the models relying on the sex variable itself.

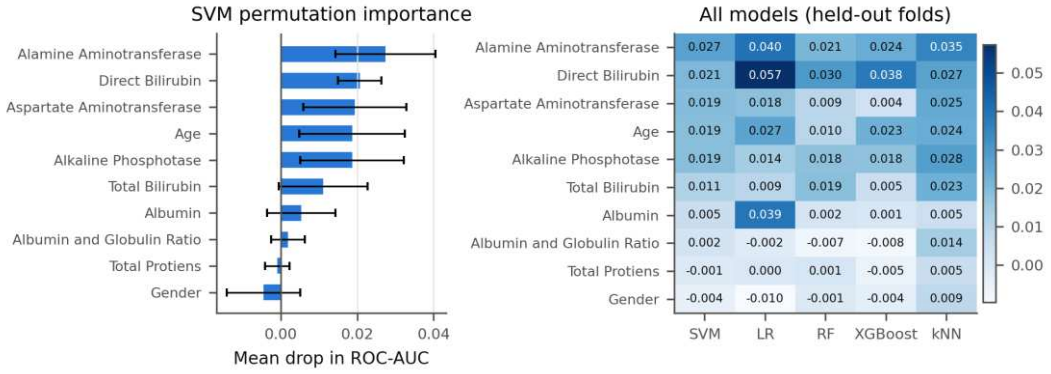


Figure 9. Permutation importance (mean drop in ROC AUC) on held out folds. Left: SVM with one standard deviation across folds. Right: all models.

The SHAP summary for the final SVM in Figure 10 shows the direction of each effect. High values of ALT, direct and total bilirubin, ALP and AST push predictions towards disease, and low values push them away. Older age raises the predicted probability. Low albumin and a low A/G ratio raise it as well, which matches reduced hepatic synthetic function. These directions agree with how clinicians read liver function tests [25], which supports the face validity of the model.

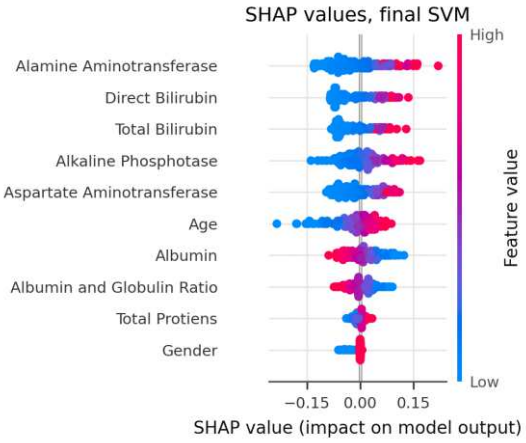


Figure 10. SHAP values for the final SVM on 150 patients. Each point is one patient; colour shows the feature value.

4.7 Cross dataset check and inference cost

Table 8 compares ROC AUC on ILPD and on the hepatitis C data. On the hepatitis C data, every model reaches an AUC between 0.978 and 0.990, with XGBoost highest. The same pipelines that reach only about 0.74 on ILPD perform very well on a dataset where the positive class has confirmed hepatitis, fibrosis or cirrhosis and the negative class consists of blood donors. This contrast suggests that the modest ILPD results come mainly from the dataset and its labels and not from a weakness of the pipeline.

Table 8. ROC AUC by dataset (mean with 95% corrected CI; upper limits truncated at 1.0). ILPD: 15 outer folds. Hepatitis C: 10 outer folds.

Model	ILPD (n = 570)	Hepatitis C (n = 608)
SVM	0.741 (0.683, 0.800)	0.984 (0.958, 1.000)
Logistic regression	0.754 (0.717, 0.790)	0.983 (0.952, 1.000)
Random forest	0.727 (0.684, 0.769)	0.988 (0.962, 1.000)
XGBoost	0.726 (0.681, 0.770)	0.990 (0.967, 1.000)
kNN	0.723 (0.679, 0.767)	0.978 (0.933, 1.000)

The final SVM pipeline, tuned on all 570 patients, selected $C = 1$ and $\gamma = 0.01$ with an inner cross validation AUC of 0.752. Table 9 shows its cost. Scoring one patient, including imputation, transformation and scaling, takes a median of 1.59 ms, and a batch of 1,000 patients takes 26.3 ms, about 38,000 patients per second. The saved model occupies 57.1 KB. A pipeline of this size can run on a basic clinic computer or a low cost server without special hardware.

Table 9. Inference cost of the final SVM pipeline (Kaggle notebook, 4 CPU cores)

Measure	Value
Single patient latency, median	1.59 ms
Single patient latency, 95th percentile	1.80 ms
Batch of 1,000 patients	26.3 ms
Throughput	37,976 patients/s
Serialised model size	57.1 KB
Support vectors	548

5 Discussion

5.1 Evaluation design changes the conclusion

The most important result of this study is the gap between the preliminary single split estimate (AUC 0.806, accuracy 0.735) and the nested estimate (AUC 0.741). Both use the same algorithm and the same data. The difference comes from retained duplicates, oversampled records shared across the tuning folds, and the chance variation of a single split of 117 patients. The corrected interval of the SVM spans 0.683 to 0.800, so a single split can land almost anywhere in that range. Studies that report one split, or that balance the data before dividing it [17, 29], cannot be compared directly with leakage free estimates, and their figures should be read as upper bounds. This agrees with the wider findings of Kapoor and Narayanan [23] and with the selection bias described by Varma and Simon [42] and Cawley and Talbot [6].

5.2 Simple models are enough on this task

Logistic regression matched or beat the SVM, random forest, XGBoost and kNN on every ranking metric, and no classifier was reliably better than another. The inner search also favoured smooth and shallow models. This matches the systematic review of Christodoulou and colleagues, who found no benefit of machine learning over logistic regression in clinical prediction [10]. For ILPD, a regularised logistic regression is therefore a strong default. It is also easier to inspect and to explain to clinicians, which supports the case made by Rudin for interpretable models in high stakes settings [37]. We kept the SVM as the reference because the study began with it, but our results do not support choosing it over logistic regression for clinical use.

5.3 Imbalance handling and thresholds are clinical decisions

Imbalance handling changed the SVM AUC by up to 0.064, which is twice the spread between classifiers. Without correction the SVM became almost useless at the default threshold. With class weights the ranking recovered but the default threshold remained poorly placed. These results show that the threshold must be chosen as part of the design, using the cost of a missed diagnosis against the cost of a false referral. For a screening tool, sensitivity is usually the priority, but on ILPD a sensitivity of 0.90 costs a specificity of only 0.23. In practice the model would then act mainly as a rule out tool for patients with clearly normal results, not as a diagnostic test. The poor calibration in Figure 7 adds a further caution: predicted probabilities from a model trained on balanced data should be recalibrated to local prevalence before they are shown as risks [5].

5.4 Lower performance for women

All five models performed worse for women, with female sensitivity between 0.35 and 0.46. This confirms, under a leakage free design and with confidence intervals, the sex disparity reported by Straw and Wu [40]. Since the sex variable itself had no importance, the gap more likely reflects differences in how disease presents in laboratory values, the smaller number of female patients in training, and possibly different referral patterns in the source hospitals. A model deployed without checking this would miss more liver disease in women. Reporting performance by sex, and setting thresholds or training data so that sensitivity is acceptable in both groups, should be part of any validation of such tools [11, 31].

5.5 Why ILPD is hard

The near perfect results on the hepatitis C data show that the pipeline can learn a strong signal when one exists. ILPD differs in several ways. Its negative class consists of hospital patients who were judged not to have liver disease, not healthy donors, so many of them may have abnormal values for other reasons. It lacks markers such as gamma glutamyl transferase and cholinesterase that are present in the hepatitis C data. Its labels come from a single regional source with no published diagnostic criteria. Together these factors likely set a ceiling on what any model can reach from these ten attributes, and they suggest that further gains will come from better data rather than from more complex algorithms. The hepatitis C task is also easier by design, since it compares established disease with healthy donors, so its results show that the pipeline works and do not by themselves show that it generalises.

5.6 Limitations

This study has several limitations. All ILPD data come from one region of India, and we had no independent test set from another hospital for the same task. The hepatitis C data test a related but different problem. The operating points in Table 6 were selected on the same out of fold predictions they describe. The DeLong and McNemar tests and the subgroup analysis used the first repeat only. The female subgroup is small, so its intervals are wide. SVM probabilities rely on Platt scaling within each training set, which adds noise for small samples. Finally, we did not study clinical outcomes or how clinicians would act on the predictions, which would be needed before any deployment.

6 Conclusion

This paper evaluated five machine learning classifiers for liver disease prediction on the Indian Liver Patient Dataset with a design that removes duplicate records, fits all preprocessing and resampling inside training folds, and uses nested repeated cross validation with corrected confidence intervals and formal tests. Under this design, all classifiers reached a ROC AUC between 0.72 and 0.75. Logistic regression performed best on most metrics, and no classifier was reliably better than the others. A single split estimate for the same SVM was about 0.065 higher, which shows how much weak evaluation inflates reported results. Imbalance handling and threshold choice affected clinical usefulness more than the choice of algorithm, predicted probabilities needed recalibration, and every model detected disease less well in women. The predictions relied on ALT, bilirubin, AST, ALP, age and albumin in directions that agree with clinical knowledge, and the final pipeline is small and fast enough for routine use.

Future work should validate these models on liver test data from other regions, including Nigerian hospitals, add markers such as gamma glutamyl transferase, study threshold setting that keeps sensitivity acceptable for both sexes, and test whether such a tool changes referral decisions in practice. We recommend that future ILPD studies report leakage free nested estimates with confidence intervals and results by sex, so that their findings can be compared and trusted.

Data and Code Availability

The ILPD is available from the UCI Machine Learning Repository [34] and Kaggle, and the hepatitis C data are available on Kaggle. The complete experiment, including all figures and tables in this paper, is produced by a single Python script that runs in one Kaggle notebook cell with a fixed random seed.

References

[1] Sumeet K. Asrani, Harshad Devarbhavi, John Eaton, and Patrick S. Kamath. 2019. Burden of liver diseases in the world. *Journal of Hepatology* 70, 1 (2019), 151–171. doi:10.1016/j.jhep.2018.09.014

[2] L. Alice Auxilia. 2018. Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease. In *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018)*. IEEE, 45–50. doi:10.1109/ICOEI.2018.8553682

[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. doi:10.1023/A:1010933404324

[4] Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78, 1 (1950), 1–3.

[5] Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. 2019. Calibration: the Achilles heel of predictive analytics. *BMC Medicine* 17 (2019), 230. doi:10.1186/s12916-019-1466-7

[6] Gavin C. Cawley and Nicola L. C. Talbot. 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* 11 (2010), 2079–2107.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. doi:10.1613/jair.953

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*. ACM, 785–794. doi:10.1145/2939672.2939785

[9] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21 (2020), 6. doi:10.1186/s12864-019-6413-7

[10] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* 110 (2019), 12–22. doi:10.1016/j.jclinepi.2019.02.004

[11] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G. M. Moons. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 350 (2015), g7594. doi:10.1136/bmj.g7594

[12] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297. doi:10.1007/BF00994018

[13] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 3 (1988), 837–845. doi:10.2307/2531595

[14] . 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research* 7 (2006), 1–30.

[15] Harshad Devarbhavi, Sumeet K. Asrani, Juan Pablo Arab, Yvonne Ayerki Nartey, Elisa Pose, and Patrick S. Kamath. 2023. Global burden of liver disease: 2023 update. *Journal of Hepatology* 79, 2 (2023), 516–537. doi:10.1016/j.jhep.2023.03.017

[16] Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* 10, 7 (1998), 1895–1923. doi:10.1162/089976698300017197

[17] Elias Dritsas and Maria Trigka. 2023. Supervised Machine Learning Models for Liver Disease Risk Prediction. *Computers* 12, 1 (2023), 19. doi:10.3390/computers12010019

[18] Bradley Efron and Robert J. Tibshirani. 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

[19] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* 20, 177 (2019), 1–81.

[20] Haibo He and Edwardo A. Garcia. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21, 9 (2009), 1263–1284. doi:10.1109/TKDE.2008.239

[21] Georg Hoffmann, Andreas Bietenbeck, Ralf Lichtinghagen, and Frank Klawonn. 2018. Using machine learning techniques to generate laboratory diagnostic pathways: a case study. *Journal of Laboratory and Precision Medicine* 3, 6 (2018). doi:10.21037/jlpm.2018.06.01

[22] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[23] Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4, 9 (2023), 100804. doi:10.1016/j.patter.2023.100804

[24] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Transactions on Knowledge Discovery from Data* 6, 4 (2012), Article 15. doi:10.1145/2382577.2382579

[25] Paul Y. Kwo, Stanley M. Cohen, and Joseph K. Lim. 2017. ACG Clinical Guideline: Evaluation of Abnormal Liver Chemistries. *American Journal of Gastroenterology* 112, 1 (2017), 18–35. doi:10.1038/ajg.2016.517

[26] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

[27] James McDermott and Richard S. Forsyth. 2016. Diagnosing a disorder in a classification benchmark. *Pattern Recognition Letters* 73 (2016), 41–43. doi:10.1016/j.patrec.2016.01.004

[28] Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (1947), 153–157. doi:10.1007/BF02295996

[29] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi. 2023. Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease. *Biomedicines* 11, 2 (2023), 581. doi:10.3390/biomedicines11020581

[30] Claude Nadeau and Yoshua Bengio. 2003. Inference for the Generalization Error. *Machine Learning* 52, 3 (2003), 239–281. doi:10.1023/A:1024068626366

[31] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. doi:10.1126/science.aax2342

[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[33] John C. Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*. MIT Press, 61–74.

[34] Bendi Ramana and N. Venkateswarlu. 2022. ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. doi:10.24432/C5D02C

[35] Bendi Venkata Ramana, M. Surendra Prasad Babu, and N. B. Venkateswarlu. 2011. A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *International Journal of Database Management Systems* 3, 2 (2011), 101–114. doi:10.5121/ijdms.2011.3207

[36] Bendi Venkata Ramana, M. S. Prasad Babu, and N. B. Venkateswarlu. 2012. A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis. *International Journal of Computer Science Issues* (May 2012).

[37] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. doi:10.1038/s42256-019-0048-x

[38] Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* 10, 3 (2015), e0118432. doi:10.1371/journal.pone.0118432

[39] J. Singh, S. Bagga, and R. Kaur. 2020. Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques. *Procedia Computer Science* 167 (2020), 1970–1980. doi:10.1016/j.procs.2020.03.226

[40] Isabel Straw and Honghan Wu. 2022. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics* 29, 1 (2022), e100457. doi:10.1136/bmjhci-2021-100457

[41] Xu Sun and Weichao Xu. 2014. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters* 21, 11 (2014), 1389–1393. doi:10.1109/LSP.2014.2337313

[42] Sudhir Varma and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7 (2006), 91. doi:10.1186/1471-2105-7-91

[43] W. J. Youden. 1950. Index for rating diagnostic tests. *Cancer* 3, 1 (1950), 32–35.