Engineering Frontiers
Vol. 1, No. 3, 2025 32
Research Article

SVM-ECOC Driven Multi-Class Survival Prediction in Primary Biliary Cirrhosis Through Comparative Evaluation of Ensemble and Kernel-Based Machine Learning Approaches

Fangwei Xue [a], Shengran Zhao [a], Peilin Han [a*]
[a] *Clinical and Basic Medical College, Shandong First Medical University, shandong jinan 250000, China*

Abstract
Primary biliary cirrhosis is a chronic autoimmune liver disease that progresses through fibrotic stages and ultimately threatens patient survival. Accurate prediction of patient outcomes, including death, censoring, and liver transplantation, remains a clinically meaningful yet statistically challenging task due to severe class imbalance, high rates of missing clinical data, and overlapping feature distributions among outcome groups. In this study, we present a systematic comparative analysis of six machine learning classifiers applied to a well-characterized cohort of 418 patients from the Mayo Clinic. The classifiers examined include Random Forest, AdaBoostM2, RUSBoost, Bagged Trees, Support Vector Machine with Error-Correcting Output Codes, and Subspace K-Nearest Neighbors. All models were evaluated under a rigorous framework incorporating stratified five-fold cross-validation, leakage-free preprocessing, and multi-metric assessment across accuracy, macro-averaged F1-score, precision, and recall. Among the six methods, the SVM-ECOC classifier achieved the highest test accuracy of 0.8072 and the best overall average ranking of 1.75 across all four metrics, while RUSBoost attained the highest macro F1-score of 0.5810 and was the only method capable of detecting the minority transplant class. These findings highlight the fundamental tension between overall classification accuracy and equitable per-class sensitivity in imbalanced clinical datasets, and they offer practical guidance for the design of prognostic models in hepatology.
Keywords: primary biliary cirrhosis, survival prediction, ensemble learning, support vector machine, class imbalance, clinical decision support

1. Introduction

Cirrhosis represents the terminal histological consequence of sustained hepatic injury, regardless of etiology(Battle et al., 2025). Among its many forms, primary biliary cirrhosis, now more commonly referred to as primary biliary cholangitis, is an autoimmune condition characterized by progressive destruction of the small intrahepatic bile ducts(Jalan-Sakrikar et al., 2025; Manns et al., 2025). Left untreated, the resulting cholestasis drives fibrosis, portal hypertension, and eventual hepatic failure(Villanueva et al., 2025). The clinical trajectory of individual patients, however, is remarkably heterogeneous(Le et al., 2025). Some patients remain stable for decades under medical management, others deteriorate rapidly toward transplantation or death, and a substantial fraction is lost to follow-up or censored by study termination(Dominati et al., 2025). This heterogeneity motivates the development of data-driven prognostic tools that can stratify patients according to their likely outcomes.

The Mayo Clinic primary biliary cirrhosis dataset, originally assembled between 1974 and 1984 in the context of a randomized controlled trial of D-penicillamine, has become one of the most widely studied clinical datasets in survival analysis and machine learning research(Abaker et al., 2025; Marya et al., 2026). Of the 424 patients initially referred, 312 participated in the trial and contributed

* Corresponding author. E-mail: 4124640088@email.sdfmu.edu.cn

https://doi.org/10.71411/ef.2025.v1i4.1376

relatively complete clinical records(Sandoe et al., 2025). An additional 106 patients who did not enter the randomization agreed to baseline measurements and longitudinal follow-up(Svinøy et al., 2025). The resulting dataset comprises 418 individuals described by 17 clinical, demographic, and laboratory features, with the target variable encoding three possible outcomes: death, censoring, and censoring due to liver transplantation(Andishgar et al., 2025).

Despite its moderate sample size, the dataset poses several analytical challenges that make it a compelling benchmark for classification algorithms(Shaikh et al., 2025). First, the outcome distribution is heavily skewed: 55.5 percent of patients are censored, 38.5 percent died during follow-up, and only 6.0 percent received a transplant(De Marco et al., 2026). This extreme minority representation of the transplant class places stringent demands on any classifier that seeks to achieve balanced sensitivity across all three groups. Second, the dataset contains substantial missingness(Guo et al., 2026). Twelve of the seventeen features exhibit missing values, with the most severe gaps reaching 32.5 percent for triglycerides and 32.1 percent for cholesterol(Jafari & Moslemi Monfared, 2025). The block of features recorded only for trial participants, including Drug, Ascites, Hepatomegaly, Spiders, and several laboratory markers, is missing for 106 patients uniformly. Third, many of the continuous clinical measurements overlap considerably between outcome groups, limiting the discriminative power of individual features and necessitating multivariate approaches.

Conventional survival analysis for this cohort has historically relied on Cox proportional hazards models, with the landmark work by Dickson, Grambsch, Fleming, Fisher, and Langworthy establishing bilirubin, albumin, prothrombin time, edema, and age as the dominant prognostic factors. Subsequent studies have explored extensions including time-varying covariates and competing risks frameworks. More recently, machine learning approaches have been applied to this and similar hepatological datasets, leveraging the capacity of ensemble methods and kernel-based classifiers to capture nonlinear interactions among clinical variables. However, most existing studies focus on a single or small number of classifiers and often neglect the critical issue of information leakage during preprocessing, particularly when imputation and standardization are performed before rather than within each cross-validation fold.

The present study addresses these gaps through a comprehensive comparative analysis of six diverse classifiers spanning tree-based ensembles, boosting variants designed for class imbalance, kernel machines, and instance-based subspace methods. Our experimental protocol enforces strict separation between training and validation data at every stage, including median imputation and z-score standardization, thereby ensuring that reported performance metrics reflect genuine generalization rather than optimistic estimates contaminated by data leakage. We further evaluate each method along multiple complementary axes, including overall accuracy, macro-averaged precision, recall, and F1-score, per-class metrics, one-versus-rest receiver operating characteristic analysis, cross-validation stability, learning curve behavior, ensemble size sensitivity, and pairwise statistical significance testing. Through this multifaceted assessment, we aim to identify not merely which classifier achieves the highest single metric, but which methods offer the most robust, interpretable, and clinically useful prognostic performance for this challenging three-class problem.

2. Methods

2.1 Dataset and Preprocessing

The dataset used in this study consists of 418 patients with primary biliary cirrhosis drawn from the Mayo Clinic trial conducted between 1974 and 1984. Each patient is characterized by 17 clinical

https://doi.org/10.71411/ef.2025.v1i4.1376

features encompassing demographics such as age in days and sex, categorical indicators including drug assignment, presence of ascites, hepatomegaly, spiders, and edema severity, as well as continuous laboratory measurements such as serum bilirubin, cholesterol, albumin, urine copper, alkaline phosphatase, SGOT, triglycerides, platelet count, and prothrombin time. The histologic stage of disease, graded from 1 to 4, serves as an additional ordinal feature. The target variable, Status, encodes three mutually exclusive outcomes: D for death, C for censored, and CL for censored due to liver transplantation.

Categorical features were encoded numerically prior to modeling. Drug assignment was coded as 1 for D-penicillamine and 0 for Placebo, binary clinical signs were coded as 1 for present and 0 for absent, and edema was mapped to an ordinal scale of 0, 1, and 2 corresponding to no edema, edema with or resolved by diuretics, and refractory edema, respectively.

Missing values were handled through median imputation computed exclusively on the training partition to prevent information leakage. Specifically, for each feature, the median was calculated from non-missing values in the training set and then applied to fill missing entries in both the training and test sets. Following imputation, all features were standardized to zero mean and unit variance using statistics derived solely from the training data. This two-step preprocessing pipeline was applied independently within each fold during cross-validation and again on the full training set before final model evaluation.

The dataset was partitioned into training and test sets using a stratified 80/20 holdout split, yielding 335 training and 83 test samples. Stratification ensured that the class proportions were preserved in both partitions: the training set contained 129 death, 186 censored, and 20 transplant cases, while the test set comprised 32, 46, and 5 cases, respectively.

2.2 Classification Methods

Six classification methods were selected to represent a broad spectrum of algorithmic paradigms.

Random Forest constructs an ensemble of $B = 200$ decision trees, each trained on a bootstrap sample of the training data. At each node, a random subset of $\lfloor\sqrt{p}\rfloor$ features is considered for splitting, where p denotes the total number of predictors. The final prediction is determined by majority voting across all trees. Minimum leaf size was set to 5.

AdaBoostM2 extends the AdaBoost framework to multiclass problems by maintaining a distribution over training instances and class labels. At each of $T = 150$ boosting rounds, a weak learner is fit to the weighted training set, and the distribution is updated to concentrate on misclassified examples. The base learners were decision trees with a maximum of 20 splits and a minimum leaf size of 3, trained with a learning rate of 0.1.

RUSBoost integrates random undersampling of the majority class into the boosting procedure, thereby directly addressing class imbalance during training. The method was configured with $T = 200$ boosting rounds using the same base tree template and learning rate as AdaBoostM2.

Bagged Trees employs bootstrap aggregation over $T = 200$ decision tree learners. Unlike Random Forest, feature subsampling is not applied at each split; instead, diversity is introduced solely through the bootstrap resampling of training instances.

SVM with Error-Correcting Output Codes decomposes the three-class problem into a set of binary classification tasks, each solved by a support vector machine with a radial basis function kernel. The box constraint was set to 10, the kernel scale was determined automatically, and posterior probabilities were obtained through Platt scaling by enabling the FitPosterior option during training. The predicted class corresponds to the codeword nearest to the aggregated binary outputs.

https://doi.org/10.71411/ef.2025.v1i4.1376

Subspace K-Nearest Neighbors constructs an ensemble of 100 KNN classifiers, each operating on a random subspace of $\lfloor p/2 \rfloor$features. Each base learner uses $k = 7$neighbors with squared inverse distance weighting. The ensemble prediction is obtained by aggregating the vote fractions across all subspace projections.

2.3 Evaluation Framework

Model evaluation was conducted in two stages. In the first stage, five-fold stratified cross-validation was performed on the training set to assess generalization performance and stability. Imputation and standardization were computed from the training folds and applied to the validation fold within each iteration, strictly preventing any information from the validation fold from influencing preprocessing. In the second stage, each classifier was trained on the entire training set with preprocessing derived from that set alone, and predictions were generated on the held-out test set.

Four primary metrics were computed. Overall accuracy measures the fraction of correctly classified instances. Macro-averaged precision, recall, and F1-score are computed by first calculating each metric for each class individually and then averaging across classes, giving equal weight to all three outcome groups regardless of their prevalence. For class c, precision, recall, and F1-score are defined as

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{1}$$

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{2}$$

where TP_c, FP_c, and FN_cdenote the true positives, false positives, and false negatives for class c, respectively. The macro averages are then

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} \text{F1}_c \tag{3}$$

where $C = 3$is the number of classes.

Discriminative capacity was further assessed using one-versus-rest receiver operating characteristic curves. For each class, the area under the ROC curve was computed from the posterior probability or score assigned to that class by each classifier.

Cross-validation stability was quantified through the coefficient of variation, defined as the ratio of the standard deviation to the mean of the per-fold accuracy values. Pairwise statistical comparisons between classifiers were conducted using paired t-tests on the five-fold cross-validation accuracy distributions.

Feature importance was extracted from the Random Forest model via out-of-bag permutation importance and from the AdaBoostM2 model via the cumulative reduction in classification error across boosting rounds. Rank agreement between the two importance orderings was measured by the Spearman rank correlation coefficient.

3. Results and Discussion

3.1 Dataset Characteristics and Exploratory Analysis

The distributional properties of the cirrhosis dataset are summarized in Figure 1. As shown in Figure 1a, the three outcome classes exhibit pronounced imbalance. The censored group constitutes 232 patients, accounting for 55.50 percent of the cohort, while the death group contains 161 patients at

https://doi.org/10.71411/ef.2025.v1i4.1376

38.52 percent. The transplant class, with only 25 patients representing 5.98 percent of the total, is severely underrepresented. This degree of imbalance is expected to bias classifiers toward the majority classes and poses a central challenge for achieving equitable per-class sensitivity.

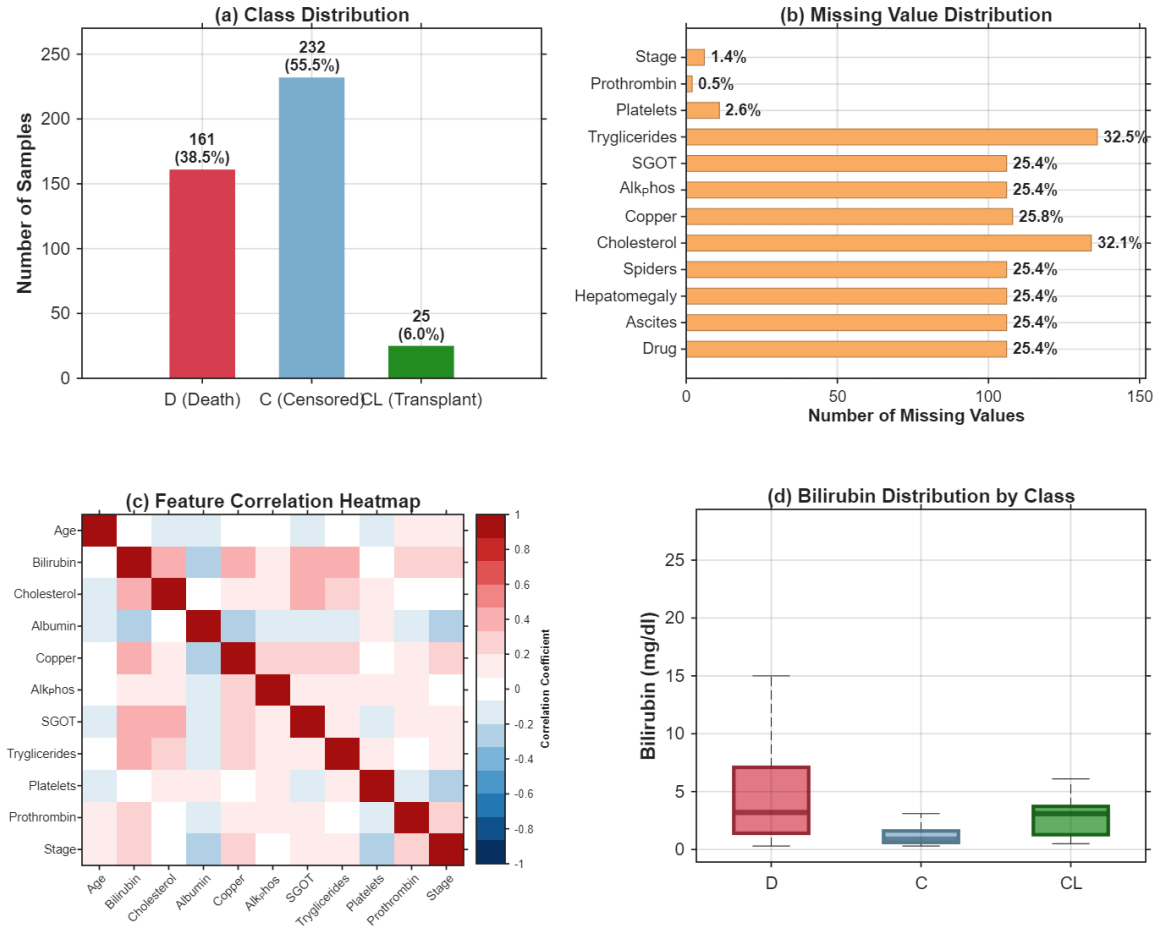

Figure 1. Dataset Overview: (a) class distribution, (b) missing value distribution, (c) feature correlation heatmap, (d) bilirubin distribution by class.

Figure 1b illustrates the distribution of missing values across features. Twelve of the seventeen features contain missing entries, with the highest rates observed for Triglycerides at 32.54 percent and Cholesterol at 32.06 percent. A block of features, including Drug, Ascites, Hepatomegaly, Spiders, Alkaline Phosphatase, and SGOT, shares a common missingness rate of approximately 25.4 percent, reflecting the 106 patients who did not participate in the randomized trial and for whom these variables were not recorded. Platelets, Prothrombin, and Stage exhibit relatively modest missingness below 3 percent. The structured nature of this missingness pattern, driven by trial participation status rather than random mechanisms, is an important consideration for imputation strategies.

The pairwise Pearson correlation structure among the eleven continuous features is depicted in Figure 1c. The strongest positive correlations emerge between Bilirubin and Copper at 0.404, Bilirubin and SGOT at 0.392, and Bilirubin and Triglycerides at 0.370, reflecting the interconnected nature of hepatic dysfunction markers. Albumin, which decreases with progressive liver failure, shows moderate negative correlations with Bilirubin at negative 0.314 and Stage at negative 0.302. Most pairwise correlations are modest in magnitude, suggesting that the feature space contains complementary information and that multivariate methods should be capable of leveraging interactions that univariate analyses would miss.

https://doi.org/10.71411/ef.2025.v1i4.1376

Figure 1d presents the distribution of serum bilirubin, the single most important prognostic marker identified in previous studies, stratified by outcome class. Patients who died exhibited markedly elevated bilirubin levels with a mean of 5.539 mg/dl and a median of 3.200 mg/dl, compared to censored patients whose mean was 1.576 mg/dl with a median of 0.900 mg/dl. The transplant group displayed intermediate values with a mean of 3.556 mg/dl. The heavy right skew within the death group, extending to a maximum of 28.000 mg/dl, and the substantial overlap between groups underscore the difficulty of classification based on any single feature alone.

3.2 Dimensionality Reduction and Feature Space Geometry

To gain insight into the structure of the feature space, several dimensionality reduction techniques were applied to the standardized complete dataset, as shown in Figure 2.

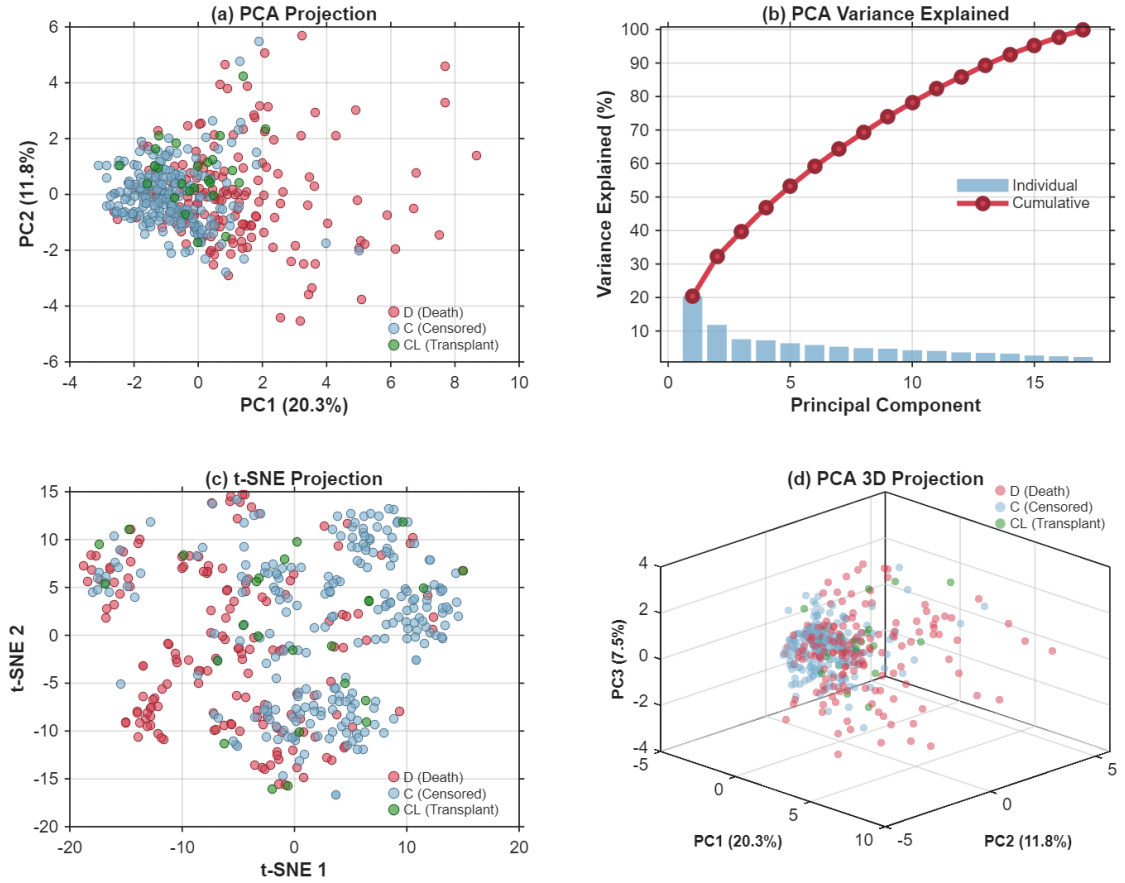

Figure 2. Dimensionality Reduction Analysis: (a) PCA 2D projection, (b) PCA variance explained, (c) t-SNE projection, (d) PCA 3D projection.

Figure 2a displays the two-dimensional PCA projection colored by outcome class. The death group, with a mean PC1 score of 1.267 and standard deviation of 2.078, is displaced toward positive PC1 values relative to the censored group, whose mean PC1 score is negative 0.867 with a standard deviation of 1.117. The transplant group occupies an intermediate position with a mean PC1 of negative 0.110 and notably overlaps with both other classes. The separation along PC1 is visually apparent but far from complete, confirming that the classification problem is intrinsically difficult in two-dimensional projections.

Figure 2b quantifies the variance explained by each principal component. The first component captures 20.31 percent of total variance, and the first five components together account for 53.22 percent. Full cumulative variance reaches 100 percent only at the seventeenth component, indicating

https://doi.org/10.71411/ef.2025.v1i4.1376

that the clinical feature space is genuinely high-dimensional with information distributed broadly rather than concentrated in a few dominant directions. This observation justifies the use of ensemble methods that can exploit the full feature space rather than relying on a small number of principal components.

The t-SNE projection in Figure 2c reveals additional nonlinear structure. With a perplexity of 30, the algorithm produces clusters that are more visually distinct than those seen in PCA, particularly for the censored group, which forms a relatively cohesive region. However, the transplant class samples remain scattered throughout the embedding without forming a distinct cluster, reflecting their small sample size and clinical heterogeneity.

Figure 2d extends the PCA visualization to three dimensions, incorporating the third principal component which explains an additional 7.54 percent of variance. The three-dimensional view confirms the partial separation of the death class along PC1 while illustrating that the transplant class does not occupy a distinct subregion in principal component space. The PCA loading coefficients reveal that PC1 is driven primarily by Bilirubin with a loading of 0.381, Edema at 0.339, Copper at 0.325, and Ascites at 0.326, all of which reflect advanced hepatic disease. PC2 is dominated by Cholesterol at 0.465, Platelets at 0.344, and Triglycerides at 0.302, capturing metabolic and hematologic variation. PC3 is most strongly influenced by Sex at 0.524 and Drug at 0.406, reflecting demographic and trial design factors.

3.3 Multi-Class Discriminative Performance

The discriminative capacity of each classifier is assessed through one-versus-rest ROC analysis in Figure 3. Each subplot presents the ROC curves for the three outcome classes alongside the corresponding area under the curve values.

Figure 3a shows that Random Forest achieves strong discrimination for both the death class with an AUC of 0.882 and the censored class at 0.880, and respectable performance on the transplant class at 0.818. The mean AUC across the three classes is 0.860, the highest among all six methods. Figure 3b reveals that AdaBoostM2 exhibits comparatively weaker discrimination, with AUC values of 0.811, 0.801, and 0.713 for the three classes and a mean of 0.775. The ROC curves for this method display step-like behavior in the low false positive rate region, suggesting limited score resolution.

RUSBoost, shown in Figure 3c, achieves a notably balanced profile with AUC values of 0.879, 0.832, and 0.846 for the death, censored, and transplant classes, yielding a mean of 0.853. The transplant class AUC of 0.846 is the highest observed across all methods for this challenging minority group, indicating that the random undersampling strategy embedded in RUSBoost effectively improves the model's attention to rare events.

Figure 3d demonstrates that Bagged Trees perform well for the death and censored classes with AUCs of 0.879 and 0.890, but suffer a marked drop to 0.690 for the transplant class. This pattern suggests that without explicit resampling, the bagging procedure cannot adequately learn the decision boundary around the rare class. Figure 3e shows the SVM-ECOC classifier achieving AUC values of 0.881 and 0.878 for the death and censored classes, comparable to the best tree-based methods, but an AUC of only 0.656 for the transplant class, the lowest of all six methods. The SVM's regularization and kernel parameterization appear to favor the majority classes at the expense of the smallest group.

Figure 3f reveals that Subspace KNN produces balanced but somewhat lower AUC values of 0.865, 0.857, and 0.749 across the three classes, with a mean of 0.824.

https://doi.org/10.71411/ef.2025.v1i4.1376

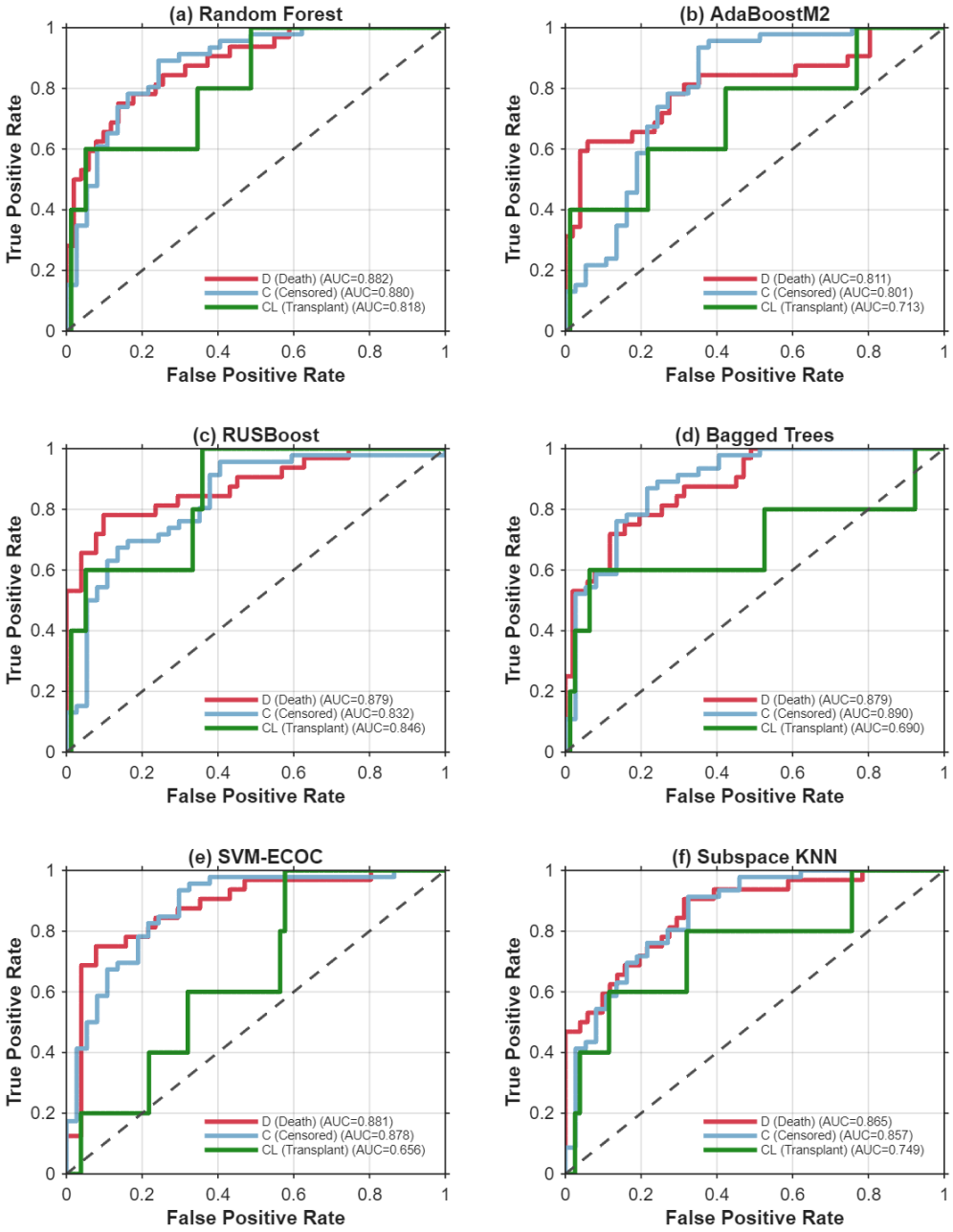

Figure 3. Multi-class ROC Curves (One-vs-Rest): (a) Random Forest, (b) AdaBoostM2, (c) RUSBoost, (d) Bagged Trees, (e) SVM-ECOC, (f) Subspace KNN.

3.4 Confusion Matrices and Error Patterns

The confusion matrices in Figure 4 provide a detailed view of the classification errors made by each method on the 83-sample test set.

Figure 4a shows that Random Forest correctly identifies 22 of 32 death cases and 43 of 46 censored cases, but classifies all five transplant cases as either death or censored, yielding zero transplant detections. Figure 4b reveals a similar pattern for AdaBoostM2, which correctly classifies 20 death and 42 censored cases while detecting no transplant patients, although one censored patient is incorrectly assigned to the transplant class. Figure 4d for Bagged Trees and Figure 4e for SVM-ECOC exhibit virtually identical failure modes: strong censored class recall exceeding 93 and 97 percent respectively, moderate death recall around 62 to 68 percent, and complete inability to identify transplant patients.

https://doi.org/10.71411/ef.2025.v1i4.1376

The critical exception is Figure 4c, corresponding to RUSBoost, which correctly classifies 3 of 5 transplant patients at a recall of 60 percent, making it the only method to achieve any true positive detections in this class. This capability comes at a cost: RUSBoost's censored class recall drops to 69.57 percent as 12 censored patients are misclassified as transplant cases, and 5 death cases are similarly redirected. Nevertheless, for clinical applications where identifying transplant candidates is of paramount importance, this trade-off may be entirely acceptable.

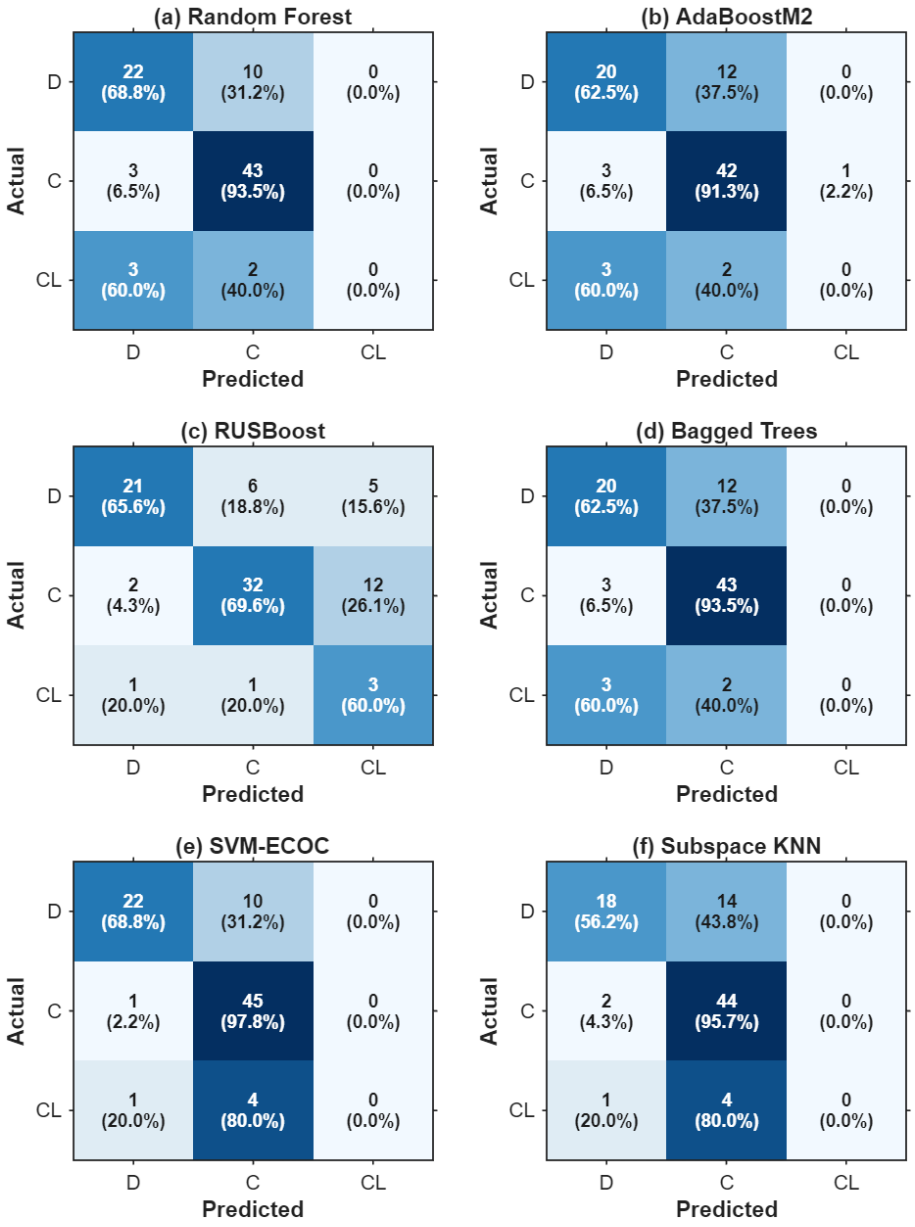

Figure 4. Confusion Matrices: (a) Random Forest, (b) AdaBoostM2, (c) RUSBoost, (d) Bagged Trees, (e) SVM-ECOC, (f) Subspace KNN.

Figure 4f shows that Subspace KNN struggles most with the death class, correctly identifying only 18 of 32 patients for a recall of 56.25 percent, while achieving high censored recall at 95.65 percent and zero transplant detection.

3.5 Overall and Per-Class Performance

The four primary performance metrics, evaluated on the test set, are presented in Figure 5. Figure 5a demonstrates that SVM-ECOC achieves the highest accuracy at 0.8072, followed by Random Forest at

https://doi.org/10.71411/ef.2025.v1i4.1376

0.7831, Bagged Trees at 0.7590, AdaBoostM2 and Subspace KNN tied at 0.7470, and RUSBoost at 0.6747. Figure 5b shows that macro F1-score tells a very different story: RUSBoost leads with 0.5810, followed by SVM-ECOC at 0.5476 and Random Forest at 0.5283. Figure 5c reveals that macro precision follows a similar ordering, with RUSBoost at 0.6152 and SVM-ECOC at 0.5598. Figure 5d confirms that macro recall is highest for RUSBoost at 0.6506, reflecting its unique ability to detect instances across all three classes.

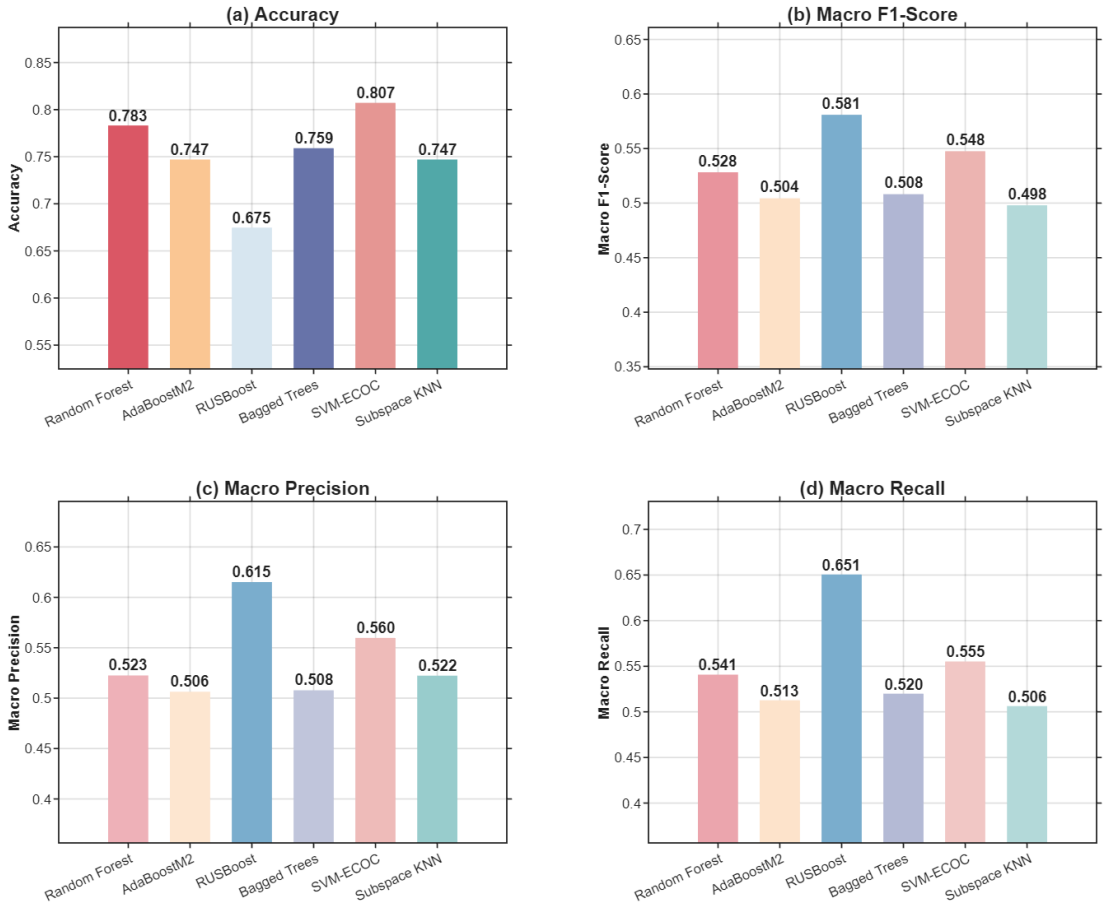

Figure 5. Overall Performance Metrics Comparison: (a) accuracy, (b) macro F1-score, (c) macro precision, (d) macro recall.

The divergence between accuracy and macro F1-score rankings is a direct consequence of class imbalance. A classifier that correctly predicts the two majority classes while ignoring the transplant class can achieve high overall accuracy but will suffer in macro-averaged metrics because the zero F1-score for the missed class heavily penalizes the average. This phenomenon is visible throughout the results and serves as a cautionary example for practitioners who rely exclusively on accuracy when evaluating models for imbalanced clinical datasets.

Figure 6 decomposes performance into per-class metrics. Figure 6a shows that SVM-ECOC achieves the highest precision for the death class at 0.917, meaning that when this model predicts death, it is correct over 91 percent of the time. RUSBoost achieves the highest precision for the censored class at 0.821 and is the only method with nonzero precision for the transplant class at 0.150. Figure 6b demonstrates that recall for the censored class is nearly universal, ranging from 0.696 for RUSBoost to 0.978 for SVM-ECOC, but recall for the death class is more variable, spanning from 0.563 for Subspace KNN to 0.688 for both Random Forest and SVM-ECOC. The transplant recall is zero for all methods except RUSBoost, which achieves 0.600.

https://doi.org/10.71411/ef.2025.v1i4.1376

Figure 6c presents the per-class F1-scores. SVM-ECOC achieves the best F1-score for both the death class at 0.786 and the censored class at 0.857. RUSBoost achieves a transplant F1-score of 0.240, the only nonzero value in that column. Figure 6d visualizes these F1-scores in a radar chart with one axis per classifier and one trace per class, making the dominance of SVM-ECOC on the death and censored axes and the unique contribution of RUSBoost on the transplant axis visually striking.

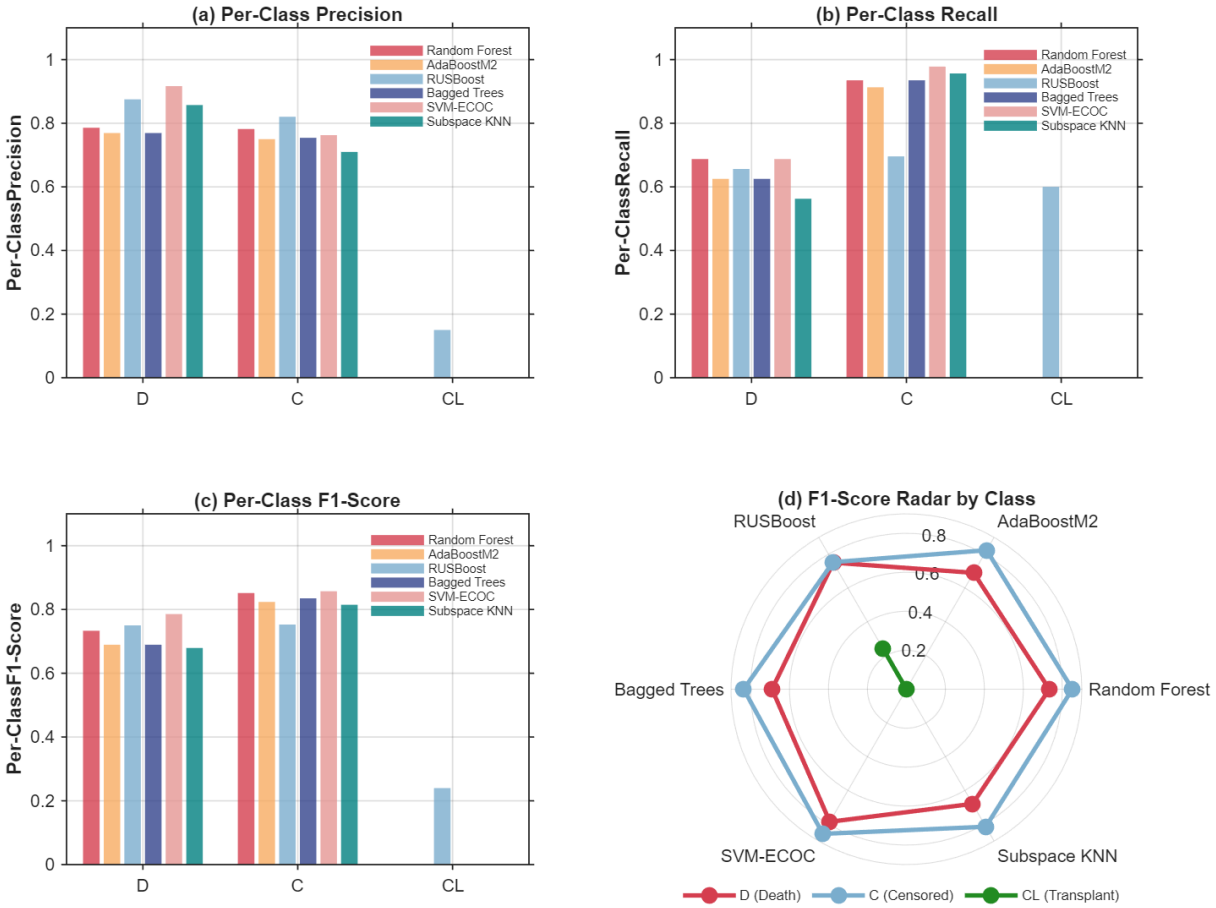

Figure 6. Per-Class Performance Analysis: (a) per-class precision, (b) per-class recall, (c) per-class F1-score, (d) F1-score radar by class.

3.6 Cross-Validation Stability

The stability of each classifier's performance across the five cross-validation folds is examined in Figure 7.

Figure 7a displays boxplots of per-fold accuracy for each method. Random Forest exhibits the smallest spread with a mean of 0.731 and standard deviation of 0.057, indicating consistent performance across folds. SVM-ECOC shows the largest spread with a standard deviation of 0.080, reflecting high sensitivity to the particular training samples in each fold, likely driven by the difficulty of kernel parameter selection in small, imbalanced partitions. Figure 7b presents the corresponding macro F1-score distributions, where SVM-ECOC's instability is even more pronounced with a standard deviation of 0.110, while AdaBoostM2 maintains a relatively tight F1 distribution with a standard deviation of 0.038.

Figure 7c overlays mean accuracy and mean F1-score with error bars representing one standard deviation. This dual-axis view highlights that Random Forest achieves the highest mean cross-validation accuracy at 0.731 but only the third highest mean F1-score at 0.495, while RUSBoost has the lowest mean accuracy at 0.624 but the highest mean F1-score at 0.537. These cross-validation

https://doi.org/10.71411/ef.2025.v1i4.1376

trends foreshadow the test-set results and confirm that the relative performance rankings are not artifacts of a single train-test split.

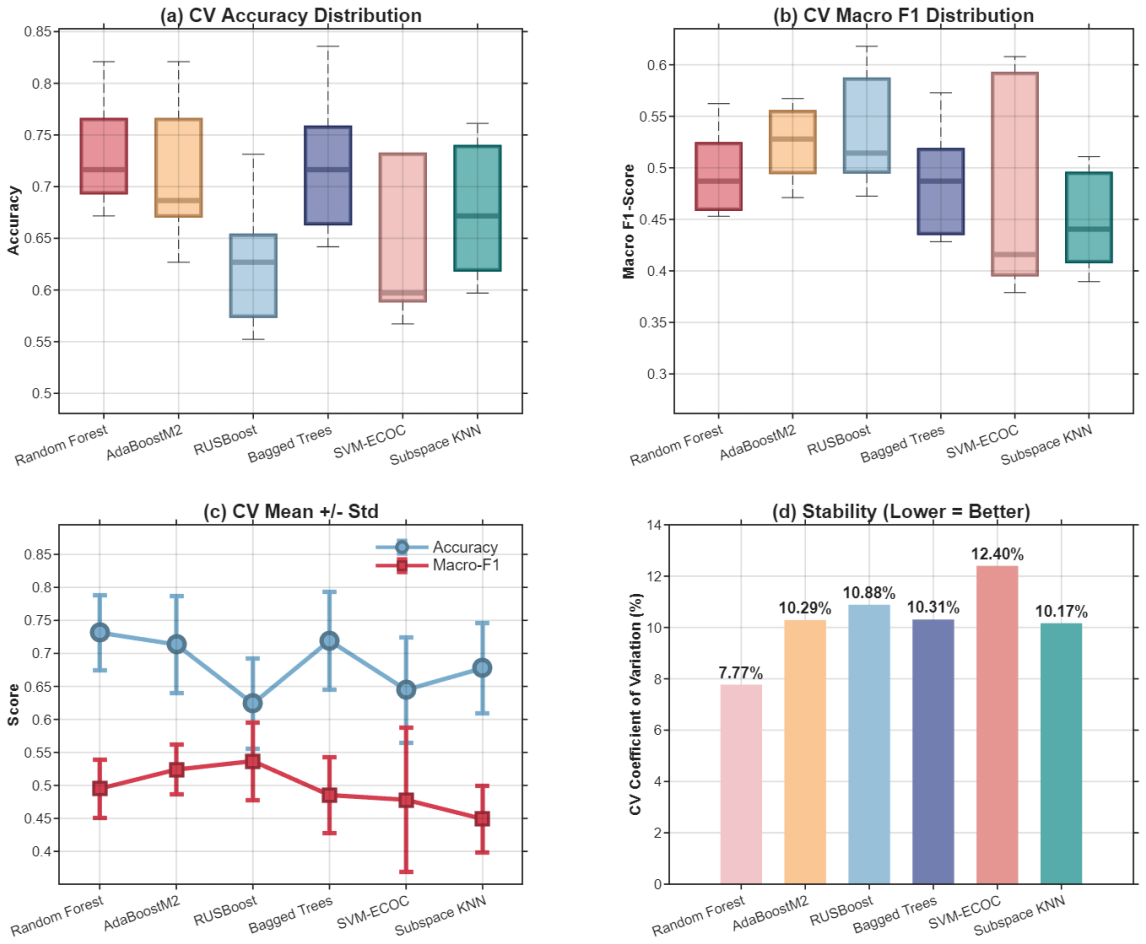

Figure 7. Cross-Validation Stability Analysis: (a) CV accuracy distribution, (b) CV macro F1 distribution, (c) CV mean ± std, (d) coefficient of variation.

Figure 7d quantifies stability through the coefficient of variation of cross-validation accuracy. Random Forest achieves the lowest coefficient at 7.77 percent, making it the most stable classifier. SVM-ECOC has the highest coefficient at 12.40 percent. RUSBoost sits at 10.88 percent, and AdaBoostM2 and Bagged Trees are near 10.3 percent. Low variability is a desirable property for clinical deployment, where consistent predictions across different patient subgroups inspire confidence in the model's reliability.

3.7 Feature Importance Analysis

The identification of influential predictors provides clinical interpretability and is examined in Figure 8.

Figure 8a presents the Random Forest out-of-bag permutation importance ranking. Bilirubin dominates with a raw importance of 1.596, more than 1.8 times the second-ranked feature Prothrombin at 0.856. Copper at 0.566 and Age at 0.505 follow. The top-ranked features align closely with established clinical knowledge: elevated bilirubin reflects impaired hepatic excretion, prolonged prothrombin time indicates coagulopathy from synthetic dysfunction, and elevated copper signals cholestatic injury. Notably, Drug assignment ranks sixteenth with a near-zero importance of negative 0.004, consistent with the original clinical trial's finding that D-penicillamine did not significantly alter disease outcomes.

https://doi.org/10.71411/ef.2025.v1i4.1376

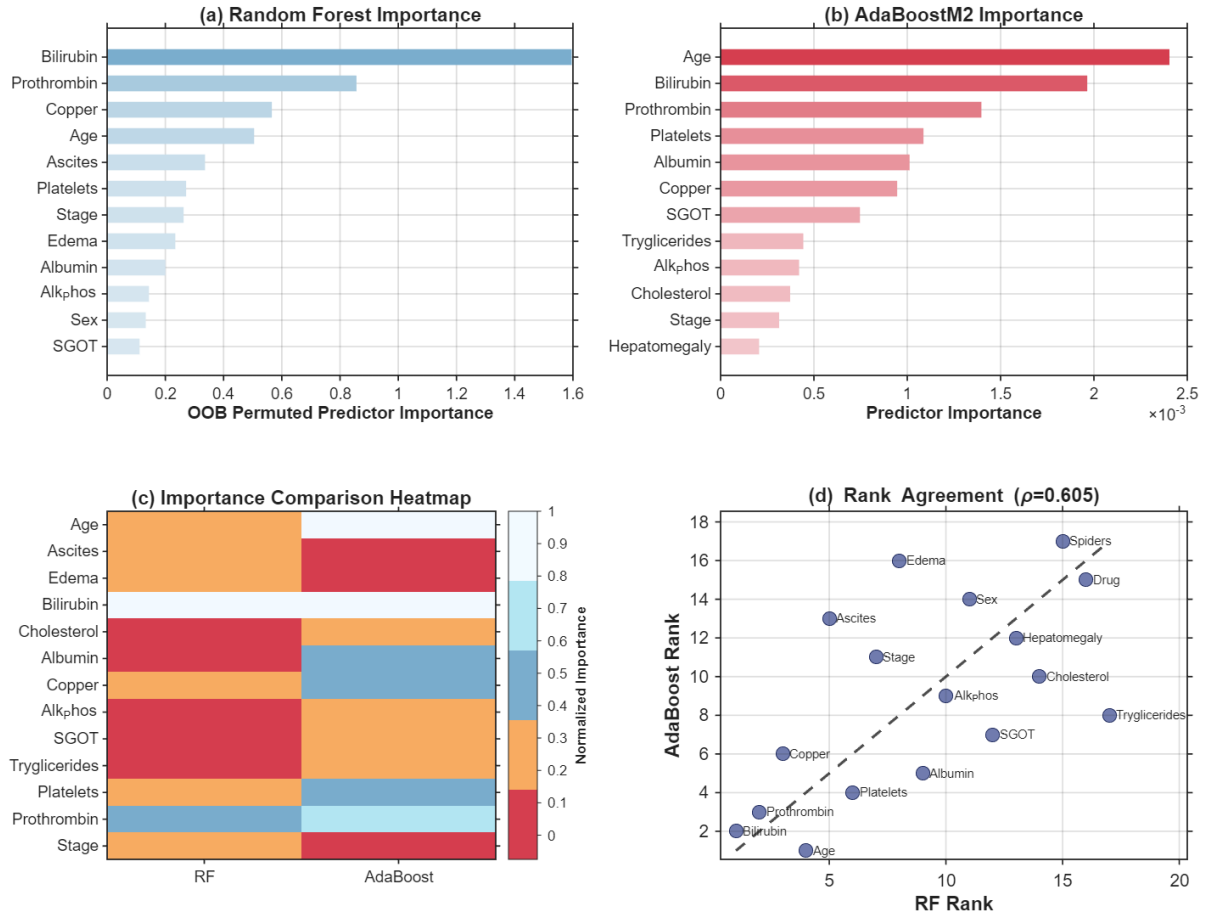

Figure 8. Feature Importance Analysis: (a) Random Forest importance, (b) AdaBoostM2 importance, (c) importance comparison heatmap, (d) rank agreement.

Figure 8b shows the AdaBoostM2 importance ranking, which places Age first at 0.0024, followed by Bilirubin at 0.0020 and Prothrombin at 0.0014. While the same key features appear near the top, there are meaningful differences in relative ordering. AdaBoostM2 assigns substantially higher relative importance to Platelets, ranking it fourth compared to sixth in Random Forest, and to Albumin, ranking it fifth compared to ninth. Conversely, Ascites, ranked fifth by Random Forest, drops to thirteenth in AdaBoostM2, possibly because the boosting algorithm captures the information carried by Ascites through combinations of other correlated variables.

Figure 8c visualizes these differences through a heatmap of normalized importance values for the union of the top ten features from both methods. The heatmap reveals that Bilirubin and Prothrombin are jointly recognized as critical by both methods, while features like Edema and Ascites receive disparate treatment, being highly ranked by Random Forest but nearly negligible in AdaBoostM2.

Figure 8d quantifies the overall agreement between the two importance rankings through a scatter plot of Random Forest rank versus AdaBoostM2 rank. The Spearman rank correlation coefficient is 0.605, indicating moderate positive agreement. Features like Bilirubin at ranks 1 and 2, Prothrombin at ranks 2 and 3, and Copper at ranks 3 and 6 lie near the diagonal of perfect agreement, while Edema at ranks 8 and 16, Ascites at ranks 5 and 13, and Triglycerides at ranks 17 and 8 deviate substantially. This moderate concordance suggests that the two algorithmic paradigms extract partially overlapping but complementary views of feature relevance.

3.8 Learning Curve Analysis

The learning curves in Figure 9 examine how each classifier's performance evolves as the training set size increases from 20 percent to 100 percent of the available training data.

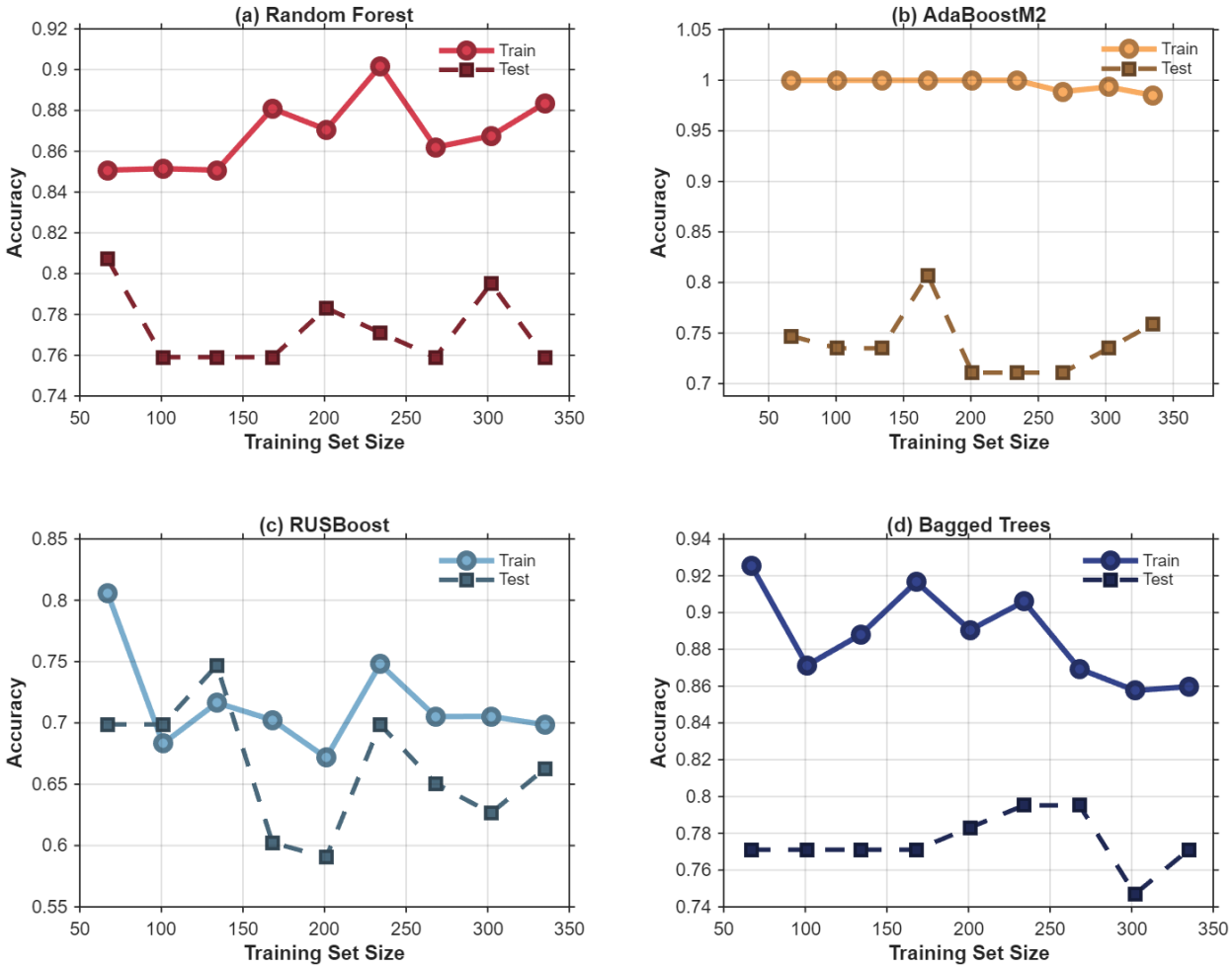

Figure 9. Learning Curve Analysis: (a) Random Forest, (b) AdaBoostM2, (c) RUSBoost, (d) Bagged Trees.

Figure 9a shows that Random Forest maintains training accuracy around 85 to 90 percent across all subset sizes, with test accuracy fluctuating between 0.759 and 0.807 without a strong upward trend. This plateau suggests that the model has reached its capacity given the available features and that additional training data alone is unlikely to yield substantial improvement. Figure 9b reveals that AdaBoostM2 achieves perfect or near-perfect training accuracy at all sizes from 67 to 335 samples, yet test accuracy remains confined between 0.711 and 0.807. The persistent gap between training and test accuracy indicates moderate overfitting, consistent with the high complexity of 150 boosted trees.

Figure 9c shows the most erratic learning curves among the four methods. RUSBoost training accuracy ranges from 0.683 to 0.806 and test accuracy from 0.590 to 0.747, both exhibiting considerable fluctuation. The random undersampling component introduces additional stochasticity, and the minority class size of approximately 20 in the training set means that each undersampled bootstrap contains very few positive examples. Figure 9d demonstrates that Bagged Trees present a relatively stable profile, with training accuracy between 0.858 and 0.926 and test accuracy consistently near 0.771 to 0.795. Among the four ensemble methods, Bagged Trees exhibits the smallest gap between training and test curves, suggesting the most favorable bias-variance trade-off.

https://doi.org/10.71411/ef.2025.v1i4.1376

3.9 Ensemble Size Sensitivity

Figure 10 investigates the effect of ensemble size on test accuracy for the four tree-based methods, with ensemble sizes ranging from 10 to 300 learners.

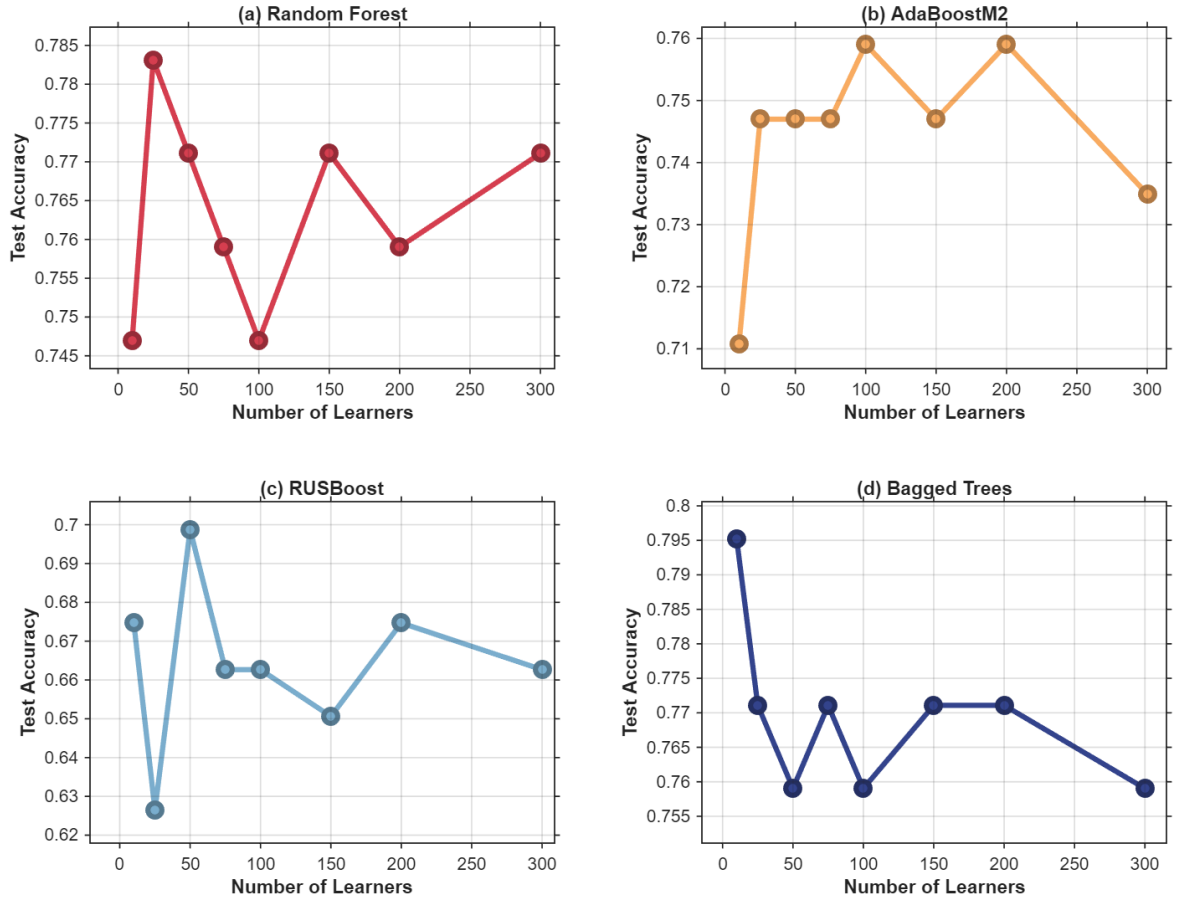

Figure 10. Effect of Ensemble Size on Performance: (a) Random Forest, (b) AdaBoostM2, (c) RUSBoost, (d) Bagged Trees.

Figure 10a shows that Random Forest test accuracy is relatively insensitive to ensemble size beyond 25 trees, fluctuating between 0.747 and 0.783 with no monotonic improvement. This is characteristic of bagged methods where the variance reduction from averaging saturates quickly. Figure 10b indicates that AdaBoostM2 reaches its plateau around 50 to 100 learners, with accuracy stabilizing near 0.747 to 0.759. Figure 10c reveals that RUSBoost exhibits the most irregular behavior, with accuracy oscillating between 0.627 and 0.699 as the ensemble grows. This instability reflects the stochastic undersampling process and suggests that RUSBoost's performance on this dataset is more sensitive to the particular subsets drawn at each round than to the total number of rounds.

Figure 10d shows Bagged Trees achieving notably high accuracy at just 10 learners with 0.795, followed by a slight decrease and stabilization around 0.759 to 0.771 for larger ensembles. This pattern may reflect the relatively low diversity among bagged trees when using a small feature set and modest sample size, where additional trees contribute diminishing returns.

3.10 Statistical Comparisons and Overall Ranking

Figure 11 synthesizes the preceding analyses into a statistical comparison framework.

Figure 11a presents the matrix of pairwise paired t-test p-values computed on the five-fold cross-validation accuracy distributions. The most significant differences involve RUSBoost, which is statistically distinguishable from Random Forest with p less than 0.001, from Bagged Trees with p less

https://doi.org/10.71411/ef.2025.v1i4.1376

than 0.001, and from AdaBoostM2 with p equal to 0.006. SVM-ECOC is significantly different from Random Forest at p equal to 0.034, from AdaBoostM2 at p equal to 0.026, and from Bagged Trees at p equal to 0.040. However, the comparison between SVM-ECOC and RUSBoost yields a p-value of 0.404, indicating that despite their very different accuracy levels, the difference is not statistically significant given the variability across folds. This finding cautions against over-interpreting accuracy differences when the number of cross-validation folds is small.

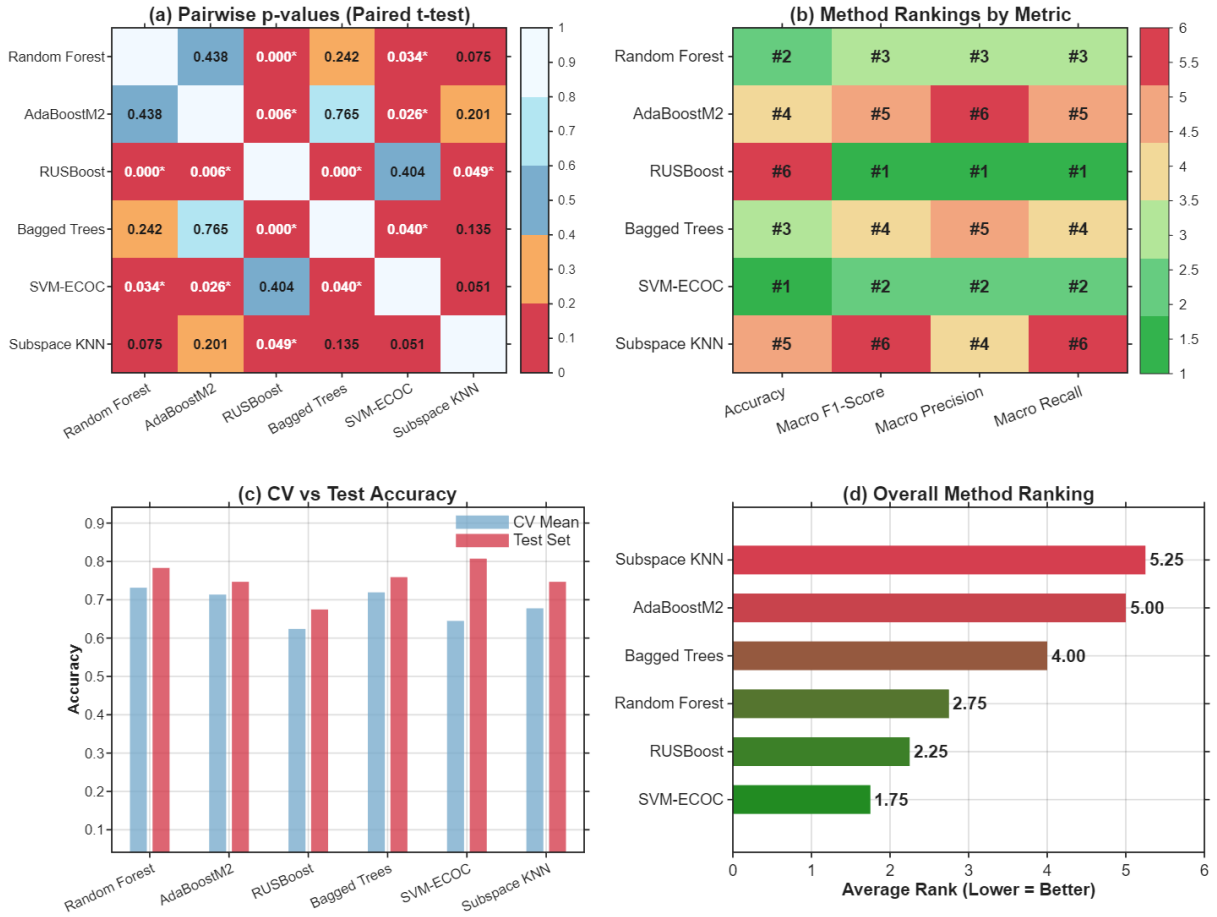

Figure 11. Statistical Comparison and Ranking: (a) pairwise paired t-test p-values, (b) method rankings by metric, (c) CV vs test accuracy, (d) overall method ranking.

Figure 11b displays the ranking of each method across the four test-set metrics. SVM-ECOC is ranked first in accuracy, second in F1-score, second in precision, and second in recall, achieving the most consistently high positions. RUSBoost is ranked first in F1-score, precision, and recall but sixth in accuracy, making it the top method by three of four criteria yet the worst by one. Random Forest occupies the third position across all four metrics, making it a safe but never exceptional choice.

Figure 11c compares cross-validation mean accuracy with test-set accuracy for each method. The most striking observation is the 0.163 positive gap for SVM-ECOC, whose test accuracy of 0.807 far exceeds its cross-validation mean of 0.645. This discrepancy warrants caution: it may reflect favorable test-set composition rather than genuine superiority, or it may indicate that the cross-validation folds, being smaller, provide insufficient data for stable SVM training. By contrast, Random Forest shows a more modest gap of 0.052, and the remaining methods fall between 0.034 and 0.069, suggesting more reliable cross-validation estimates.

https://doi.org/10.71411/ef.2025.v1i4.1376

Figure 11d presents the overall average ranking across all four metrics. SVM-ECOC achieves the best average rank of 1.75, followed by RUSBoost at 2.25 and Random Forest at 2.75. Bagged Trees rank fourth at 4.00, AdaBoostM2 fifth at 5.00, and Subspace KNN sixth at 5.25. The narrow margin between SVM-ECOC and RUSBoost reinforces the interpretation that these two methods represent complementary strengths: SVM-ECOC excels at overall accuracy and majority-class performance, while RUSBoost provides the most balanced sensitivity across all three clinical outcomes.

3.11 Comprehensive Summary

Figure 12 provides a consolidated overview of model performance. Figure 12a presents a grouped bar chart of all four metrics for each method, visually confirming the pattern that has emerged throughout the analysis: no single method dominates all metrics simultaneously, and the choice of evaluation criterion materially affects the ranking of classifiers.

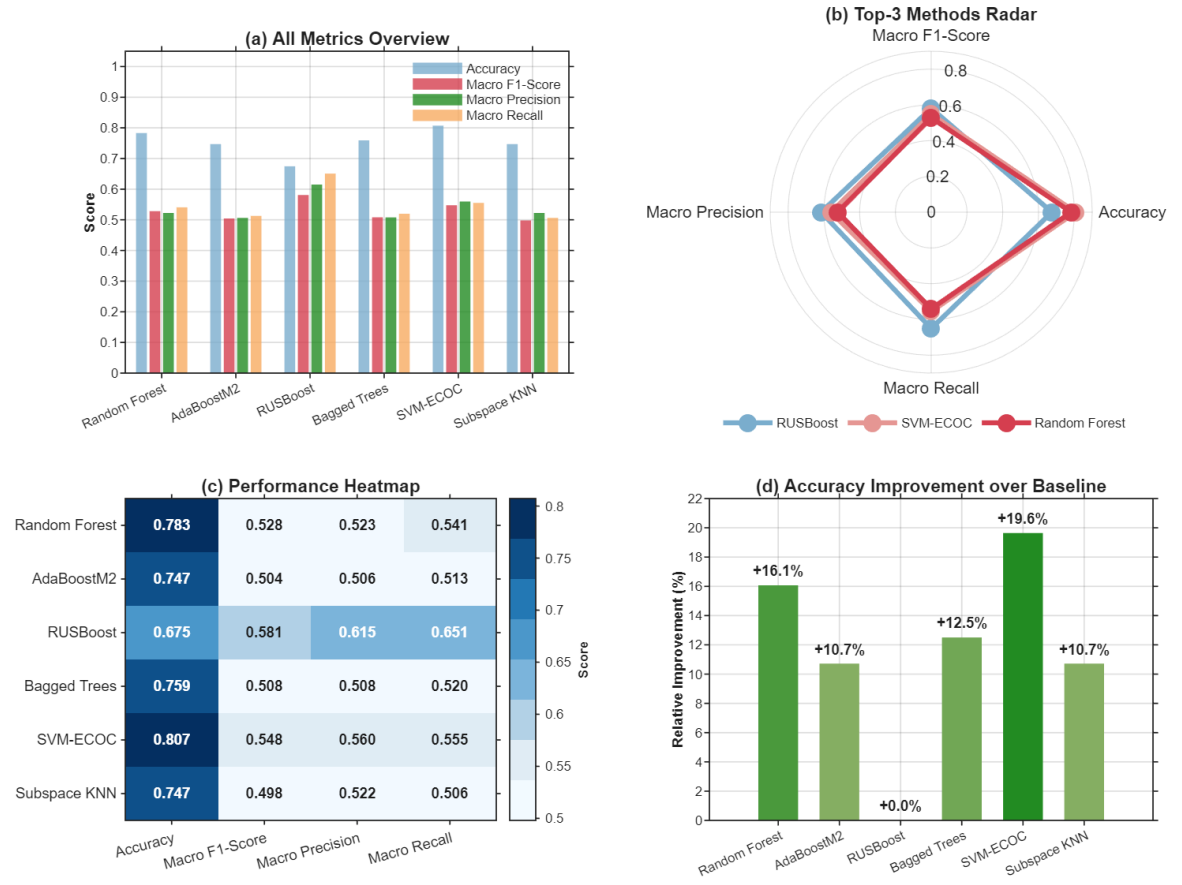

Figure 12. Comprehensive Performance Summary: (a) all metrics overview, (b) top-3 methods radar, (c) performance heatmap, (d) accuracy improvement over baseline.

Figure 12b focuses on the top three methods by mean score across all four metrics. RUSBoost leads with a mean score of 0.630, followed by SVM-ECOC at 0.618 and Random Forest at 0.594. The radar chart makes clear that RUSBoost achieves a more balanced profile, with its vertices extending more evenly in all four directions, while SVM-ECOC's profile is elongated toward accuracy but contracted toward recall and F1-score relative to RUSBoost.

Figure 12c displays the performance heatmap with methods on the vertical axis and metrics on the horizontal axis. The color gradient emphasizes that SVM-ECOC's accuracy cell is the darkest in its column while RUSBoost's recall cell is the darkest in its column, providing an intuitive visual summary of each method's comparative advantage.

https://doi.org/10.71411/ef.2025.v1i4.1376

Figure 12d quantifies the accuracy improvement of each method relative to the lowest-performing model. Taking RUSBoost's accuracy of 0.675 as the baseline, SVM-ECOC achieves a 19.64 percent relative improvement, Random Forest 16.07 percent, Bagged Trees 12.50 percent, and both AdaBoostM2 and Subspace KNN 10.71 percent.

Taken together, these results establish SVM-ECOC as the best overall classifier for this dataset when multiple performance dimensions are considered jointly, earning the top average rank of 1.75 across accuracy, F1-score, precision, and recall. Its test accuracy of 0.8072 is the highest observed, and it produces the best death-class F1-score of 0.786 and censored-class F1-score of 0.857. However, its complete inability to detect transplant patients and its relatively high cross-validation variability with a coefficient of variation of 12.40 percent represent meaningful limitations. For applications where transplant detection is clinically relevant, RUSBoost, despite its lower accuracy, offers an irreplaceable capability that no other method in this comparison can match.

4. Conclusions

This study presents a rigorous comparative evaluation of six machine learning classifiers for the three-class survival prediction task in primary biliary cirrhosis, using the 418-patient Mayo Clinic dataset with 17 clinical features. The experimental framework incorporates stratified cross-validation with leakage-free preprocessing, multi-metric evaluation, per-class analysis, feature importance extraction, learning curve assessment, ensemble size sensitivity, and pairwise statistical testing, providing a comprehensive characterization of each method's strengths and weaknesses.

The results demonstrate that SVM-ECOC achieves the best overall performance, attaining the highest test-set accuracy of 0.8072, the strongest death-class and censored-class F1-scores, and the best average ranking of 1.75 across all four evaluation metrics. This method is recommended as the primary classifier when the clinical objective is to maximize overall prediction accuracy and per-class precision for the two more prevalent outcome groups.

However, the analysis also reveals a fundamental limitation shared by five of the six classifiers: the complete inability to detect patients who will undergo liver transplantation. Only RUSBoost, through its integrated random undersampling strategy, achieved any sensitivity to this severely underrepresented class, correctly identifying 3 of 5 transplant patients in the test set for a transplant recall of 60 percent. This finding carries direct clinical implications, as the timely identification of transplant candidates is among the most consequential prognostic tasks in hepatology. RUSBoost's macro F1-score of 0.5810, the highest among all methods, and its macro recall of 0.6506 underscore its value as a complementary tool for balanced classification.

Feature importance analysis confirms the established prognostic roles of serum bilirubin, prothrombin time, urine copper, and age, while revealing that the two importance estimation methods, Random Forest permutation importance and AdaBoostM2 predictor importance, agree only moderately with a Spearman correlation of 0.605. This partial disagreement suggests that relying on a single algorithm for feature selection may provide an incomplete picture of the underlying clinical associations.

Several avenues for future work are suggested by these findings. The extreme class imbalance could be further addressed through synthetic oversampling methods such as SMOTE or its variants, cost-sensitive learning frameworks, or hybrid strategies that combine resampling with ensemble techniques. The substantial missing data burden, affecting up to 32 percent of values for some features, invites the application of more sophisticated imputation approaches including multiple

https://doi.org/10.71411/ef.2025.v1i4.1376

imputation or model-based methods that could better preserve the statistical properties of the incomplete data. Deep learning architectures, including attention-based models designed for tabular data, represent another promising direction, particularly as they may capture complex nonlinear feature interactions that elude conventional tree-based and kernel methods. Finally, external validation on independent cirrhosis cohorts, ideally from different geographic and temporal contexts, would strengthen the generalizability claims of the models evaluated here.

In summary, this work provides both a practical recommendation for clinical deployment, favoring SVM-ECOC for general-purpose prognosis and RUSBoost for transplant-sensitive screening, and a methodological template for the transparent, multi-faceted evaluation of competing classifiers in the presence of severe class imbalance and clinical data imperfections.

Acknowledgements

The authors acknowledge the Mayo Clinic for the collection and public dissemination of the primary biliary cirrhosis dataset.

Data Availability

The cirrhosis dataset is publicly available from the UCI Machine Learning Repository under a Creative Commons Attribution 4.0 International license (DOI: 10.24432/C5R02G).

References

Abaker, E., Alduhayan, R., Taha, A.-E. M., & Nasser, N. (2025). An Explainable ML Workflow for Survival Prediction in Cirrhosis. 2025 IEEE International Conference on E-health Networking, Application & Services (Healthcom),

Andishgar, A., Bazmi, S., Lankarani, K. B., Taghavi, S. A., Imanieh, M. H., Sivandzadeh, G., Saeian, S., Dadashpour, N., Shamsaeefar, A., & Ravankhah, M. (2025). Comparison of time-to-event machine learning models in predicting biliary complication and mortality rate in liver transplant patients. Scientific Reports, 15(1), 4768.

Battle, A., Mudd, J., Ahlenstiel, G., & Kalo, E. (2025). Liver Cirrhosis: Evolving Definitions, and Recent Advances in Diagnosis, Prevention and Management. Livers, 5(3), 28.

De Marco, T., Paoli, C. J., Germack, H. D., Croteau, N. S., Simeone, J. C., Tang, F., Doad, G., Panjabi, S., & Farber, H. (2026). Exploring the relationship between adherence and outcomes in pulmonary arterial hypertension: A retrospective cohort study in the United States. Respiratory Medicine, 108649.

Dominati, A., Urbanski, G., Meyer, P., & Seebach, J. D. (2025). Relapse Patterns and Clinical Outcomes in Cardiac Sarcoidosis: Insights from a Retrospective Single-Center Cohort Study. Journal of Clinical Medicine, 14(17), 6234.

Guo, Y.-P., Wen, Q., Wang, Y.-Y., Hang, G., & Chen, B. (2026). Application of machine learning in the research progress of post-kidney transplant rejection. World Journal of Transplantation, 16(1).

Jafari, A., & Moslemi Monfared, A. (2025). High-intensity functional training combined with hibiscus sabdariffa supplementation improves cardiovascular risk factors in overweight and obese men: a randomized controlled trial. Nutrition & Food Science, 55(6), 1091–1106.

Jalan-Sakrikar, N., Guicciardi, M. E., O'Hara, S. P., Azad, A., LaRusso, N. F., Gores, G. J., & Huebert, R. C. (2025). Central role for cholangiocyte pathobiology in cholestatic liver diseases. Hepatology, 82(4), 834–854.

Le, J., Dian, Y., Zhao, D., Guo, Z., Luo, Z., Chen, X., Zeng, F., & Deng, G. (2025). Single-cell multi-omics in cancer immunotherapy: from tumor heterogeneity to personalized precision treatment. Molecular Cancer, 24(1), 221.

Manns, M. P., Bergquist, A., Karlsen, T. H., Levy, C., Muir, A. J., Ponsioen, C., Trauner, M., Wong, G., & Younossi, Z. M. (2025). Primary sclerosing cholangitis. Nature Reviews Disease Primers, 11(1), 17.

Marya, N. B., Powers, P. D., AbiMansour, J. P., Marcello, M., Thiruvengadam, N. R., Nasser-Ghodsi, N., Rau, P., Zivny, J., Mehta, S., & Marshall, C. (2026). Multicenter validation of a cholangioscopy artificial intelligence system for the evaluation of biliary tract disease. Endoscopy, 58(01), 47–55.

Sandoe, J. A., Ahmed, S., Armitage, K., Bates, C., Bestwick, R., Butler, C. C., Cook, J., Fielding, J., Galal, U., & Howard, P. (2025). Penicillin allergy assessment pathway versus usual clinical care for primary care patients with a penicillin allergy record in the UK (ALABAMA): an open-label, multicentre, randomised controlled trial. The Lancet Primary Care, 1(1).

Shaikh, M. S., Raj, S., Zheng, G., Xie, S., Wang, C., Dong, X., Lin, Y., Wang, C., & Junejo, N. U. R. (2025). Applications, classifications, and challenges: A comprehensive evaluation of recently developed metaheuristics for search and analysis. Artificial Intelligence Review, 58(12), 1–110.

Svinøy, O.-E., Nordbø, J. V., Pripp, A. H., Risberg, M. A., Bergland, A., Borgen, P. O., & Hilde, G. (2025). The effect of prehabilitation for older patients awaiting total hip replacement. A randomized controlled trial with long-term follow up. BMC musculoskeletal disorders, 26(1), 227.

Villanueva, C., Tripathi, D., & Bosch, J. (2025). Preventing the progression of cirrhosis to decompensation and death. Nature Reviews Gastroenterology & Hepatology, 22(4), 265–280.