NOVA MEDICA INFORMATICA

Volume 01, Issue 2, 2026, pp. 1 of 14

Article

An Optimized Ensemble Machine Learning Framework with Automated Feature Engineering for Multiclass Survival Prediction in Primary Biliary Cirrhosis

Heni Fatmawati[1*], Tiara Amanna Amandita[2]

Received: 14-07-2026

Revised: 27-07-2026

Accepted: 28-07-2026

Published: 31-07-2026

Citation: H. Fatmawati and T. A. Amandita, "An Optimized Ensemble Machine Learning Framework with Automated Feature Engineering for Multiclass Survival Prediction in Primary Biliary Cirrhosis," Nova Medica Informatica, vol. 1, no. 2, pp. 1-14, Jul. 2026.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Correspondence (*): fatmawatiheni@unej.ac.id

[1] Faculty of Medicine, University of Jember, Jember, Indonesia; fatmawatiheni@unej.ac.id

[2] Department of Plastic, Reconstructive, and Aesthetic Surgery, Faculty of Medicine, University of Jember, Jember, Indonesia; amanditatiara@gmail.com

Highlights

A machine learning pipeline integration tackles issues pervasive missing data, severe class imbalance, and non-linear cross-talk of the biomarkers for static three-class survival analysis in primary biliary cholangitis populations.

What are the main findings?

- An automatic four-step feature engineering pipeline generates 28 orthogonal arithmetic interaction features, eliminating the restrictions inherent in conventional linear clinical scores.
- The temporal velocity of serum bilirubin and multiplication of bilirubin and aspartate aminotransferase prove to be the best predictive biomarkers.
- A continuous stacked ensemble using directionally weighted error correction based on Optuna CMA-ES represents the only architecture that can avoid majoritarian class collapse and provide positive transplantation sensitivity.

What are the implications of the main findings?

- Sole reliance on maximizing standalone accuracy is inadequate for highly skewed targets, thus proving the need for cost-sensitive ensemble blending as a secondary clinical triage process.
- The identification of clinically meaningful interaction expression using the algorithm is evidence of automatic calculation being used as an objective discovery tool for pathobiological biomarkers.
- Distinct misclassification of minorities indicates that static borders act as an exploratory framework that must move to time-to-event competing risk survival models.

Abstract

Primary Biliary Cholangitis is an autoimmune cholestatic disease that exhibits a progressively heterogeneous nature of its outcome. While establishing a static three-class survival target for predicting the progression of the disease offers a significant baseline framework for computation, the prediction is seriously hampered by the extremely skewed class distribution where transplantations are only 6.0%, while lab records are 32.5% missing. To address these constraints within an unconfounded historical natural history baseline, an integrated machine learning pipeline was evaluated on the Mayo Clinic dataset spanning a primary cohort of 418 patients and a validation cohort of 312 patients. Missing data fields were recovered via a tree-based Multiple Imputation by Chained Equations framework driven by CatBoost base learners to preserve multi-dimensional covariance structures, followed by an automated feature engineering module that extracted orthogonal arithmetic interactions. A heterogeneous ensemble of 21 base classifiers was optimized through a continuous Optuna CMA-ES Bayesian search strategy evaluated under a 5-fold stratified cross-validation paradigm. The optimized framework achieved a mean log loss of 0.469 ± 0.157 and a micro-averaged F1-score of 0.842 ± 0.048, matching standalone ExtraTrees accuracy (0.854 ± 0.024; $t = -0.679$, $p = 0.535$) but functioning as the unique cost-sensitive architecture capable of overcoming majoritarian class collapse to achieve non-zero transplantation sensitivity at a localized recall of 0.200. This non-zero local sensitivity operates as a vital early-warning clinical triage alert, demonstrating that continuous ensemble fusion successfully mitigates highly skewed multi-outcome boundaries before transitioning toward dynamic time-to-event competing-risk workflows.

Keywords: automated feature engineering; bayesian optimization; cost-sensitive learning; ensemble stacking; multiple imputation; primary biliary cholangitis; survival prediction

1. Introduction

Primary biliary cholangitis (PBC) represents a rare, chronically progressive autoimmune cholestatic liver pathology characterized by the T-cell-mediated destruction of small intrahepatic bile ducts, which subsequently precipitates bile acid accumulation, accelerating hepatic fibrosis and end-stage cirrhosis [1]. Even though treatment strategies generally focus on ursodeoxycholic acid (UDCA) and other conditionally approved second-line therapies like obeticholic acid and peroxisome proliferator-activated receptor agonists in non-responsive groups, there is a considerable proportion of patients that continue to deteriorate clinically towards liver failure or need liver transplantation [1]. The presence of this marked heterogeneity among the phenotypes stresses the critical need for risk stratification techniques [2]. In longitudinal clinical registries, the improvement of patient outcomes involves predictive models that should be able to accurately classify multiple distinct clinical landmarks, namely censored survival, liver transplant, and overall death [3], [4]. Although time-to-event modeling based on Fine-Gray subdistribution framework is the conventional choice when it comes to hazard rate estimation over the long term [5], unraveling the intricate biological interplay among baseline biomarkers in extremely unbalanced populations necessitates defining the limits of static multi-classification as an initial step [6].

Traditional systems for grading such as UK-PBC risk score [3], Child-Turcotte-Pugh score [7], MELD 3.0 score [8], and the ALBI grade [9] are predominantly based on linear combinations of a limited set of laboratory measurements. While these conventional systems continue being popular because of their easy-to-grasp nature, they cannot reflect the complicated, multiplicative, and non-linear interplay between biomarkers seen in the case of progressing intrahepatic cholestasis. The techniques of machine learning have become increasingly used in different fields of hepatology in order to deal with these limitations. In recent studies carried out by Kimura et al., it was found out that gradient boosting and random forest models substantially boost prediction of treatment response in PBC patients [10], [11]. In a similar way, clustering without supervision has revealed new clinical phenotypes through longitudinal changes in serum albumin levels [2], whereas sophisticated decision trees have proved more effective than classical risk scores in predicting decompensation events in cirrhotic patients with general liver disease [7]. Moreover, studies carried out in modern cirrhotic populations show that tree-based ensembles perform better than linear models because of the capture of latent physiological interactions [12]. Nevertheless, previous computer science studies have been more concerned with showing their superiority in separate cases, ignoring other challenges of the real world.

Three different areas can be identified where methodological weaknesses in computational hepatology are exposed. Firstly, retrospective clinical registries often feature missing laboratory data, such as in the historic Mayo Clinic PBC dataset which displays missing values of up to 32.5% for key metabolic factors including total cholesterol and triglyceride levels [13]. Simple deterministic replacements of missing values in this regard lead

to serious structural biases, which in turn distort the true variance of the distribution and decrease the performance of the subsequent classification procedures [14]. Secondly, high class imbalances limit the algorithmic sensitivity due to the key therapeutic outcome of orthotopic liver transplantation being present in just 6% of cases [15]. The most recent clinical studies on data science underscore the importance of addressing the conflicting issues arising from data incompleteness and skewedness of the target class boundaries [6], [16]. This traditional baseline model refers to the disease history before the widespread use of Ursodeoxycholic Acid in clinics; on the other hand, current cohorts are characterized by the early onset of the disease and high response rates ranging from 74.7 to 85.1 percent [17]. Finally, one of the major limitations of current models is that the main diagnostic markers for cholestasis are linked by non-linear relationships [18], [19].

In addressing the problems present within the fragmented clinical analytics pipeline, this study proposes and tests an innovative machine learning framework designed to counteract all three structural barriers. Specifically, the framework employs an iterative tree-based Multiple Imputation by Chained Equations method using CatBoost base regressors, which preserves multi-dimensional dependencies and non-linear structures between biomarkers without compromising on distributions [6], [20]. Additionally, the framework employs a feature engineering module composed of four stages of automatic feature transformation, frequency encoding of categories, residual expansion via clustering, and interaction between pairs of variables. This multi-level feature extension approach is in harmony with sophisticated hybrid neural network techniques designed to predict outcomes for patients who receive new livers, which gain their clinical reliability from projection of variables in various feature spaces [21], [22]. The final prediction process makes use of a diversified ensemble of 21 classifiers, each optimized using a continuous Covariance Matrix Adaptation Evolution Strategy Bayesian optimizer [23], [24]. This system manages to avoid the local minima problem prevalent in traditional additive consensus models through a cost-sensitive paradigm which allows directionally corrected weights [22], [24], [25].

2. Materials and Methods

2.1. Data Source and Cohort Description

The source of the basic dataset used in this study was the well-known Mayo Clinic registry of patients with primary biliary cirrhosis which was developed as part of a double-blind randomized controlled clinical trial testing the efficacy of D-penicillamine compared to placebo [13]. The main dataset consists of longitudinal information of 418 patients with completely labeled targets within the period ranging from 1974 to 1984, with medical follow-up continuing until 1986 [13]. For the comprehensive assessment of missing data behavior, a secondary validation dataset of 312 patients from the same clinical trial was used. The merged data set comprises 17 individual pathophysiological variables consisting of age, gender, random assignment of drug therapy, and 14 other biochemical/clinical indicators, such as serum bilirubin, total cholesterol, serum albumin, urinary copper, alkaline phosphatase, aspartate aminotransferase, serum triglycerides, total platelet count, prothrombin time, and the clinical progression stage I–IV of the disease. These continuous variables are supplemented with four clinical indicators that take either 0 or 1 value: presence of ascites, hepatomegaly, spider nevi, and peripheral edema as well as overall duration of observation defined as N_Days. Computations were performed using Python 3.10 together with packages scikit-learn 1.3, XGBoost 2.0, LightGBM 4.1, CatBoost 1.2, and Optuna 3.4 [24].

Categorical survival outcomes were framed within a static three-class paradigm to establish a baseline predictive framework capable of delineating the complex cross-talk between historical laboratory markers before transitioning to highly parameterized time-to-event architectures [6], [13]. Although traditional survival analysis under longitudinal settings demands time-to-event modeling accounting for competing risks, such as the Fine-Gray subdistribution hazards model [5], the acute scarcity of the transplantation event within this cohort, which stands at 6.0%, severely constrains the stability of parameter estimation in dynamic hazards functions. Consequently, a static multiclass approach evaluated via micro-averaged F1-score was prioritized to maximize the capture of global topological boundary variances across highly imbalanced representations [6], [26]. In line with empirical conventions from previous derivations of the Mayo PBC prognostic index, the initial predictor variable N_Days was included in the set of predictors because time accumulated through observation remains an essential covariate for long-term prognostic indexing [13]. Physiological and metabolic configurations for each of the three unique survival events are shown in Table I.

TABLE I
Baseline characteristics of the primary cohort (n = 418) stratified by three-class survival outcome.

Characteristic	Overall (n=418)	Censored—C (n=232)	Transplant—CL (n=25)	Death—D (n=161)
Age (years)	54.0 (46.0–62.0)	53.0 (45.0–60.0)	51.5 (44.0–59.0)	56.0 (49.0–64.0)
Sex (female)	374 (89.5)	210 (90.5)	22 (88.0)	143 (88.8)
Treatment (D-penicillamine)	184 (44.0)	102 (44.0)	11 (44.0)	72 (44.7)
Ascites (yes)	24 (5.7)	8 (3.4)	0 (0.0)	16 (9.9)
Hepatomegaly (yes)	160 (38.3)	78 (33.6)	12 (48.0)	82 (50.9)
Spiders (yes)	80 (19.1)	32 (13.8)	8 (32.0)	48 (29.8)
Edema (yes)	36 (8.6)	12 (5.2)	2 (8.0)	24 (14.9)
Bilirubin (mg/dL)	1.4 (0.8–3.4)	0.9 (0.6–1.7)	1.6 (0.9–3.8)	2.8 (1.4–6.4)
Albumin (g/dL)	3.5 (3.1–3.9)	3.7 (3.3–4.0)	3.6 (3.3–3.9)	3.2 (2.8–3.6)
Cholesterol (mg/dL) b	309.5 (248.5–400.0)	315.0 (260.0–394.0)	340.0 (270.0–430.0)	298.0 (232.0–418.0)
Copper (µg/day)	73.0 (40.0–139.0)	58.0 (35.0–100.0)	82.0 (42.0–148.0)	102.0 (52.0–186.0)
ALP (U/L)	1259.0 (891.0–1989.0)	1149.0 (840.0–1736.0)	1380.0 (920.0–2100.0)	1526.0 (1026.0–2430.0)
SGOT (U/L)	114.4 (80.6–164.0)	105.0 (76.0–146.0)	120.0 (84.0–168.0)	132.0 (92.0–192.0)
Platelets ($\times 10^3/\mu L$)	257.0 (188.0–333.0)	275.0 (210.0–348.0)	266.0 (200.0–340.0)	220.0 (150.0–295.0)
Prothrombin time (s)	10.6 (10.0–11.2)	10.4 (10.0–10.9)	10.5 (10.1–11.0)	11.0 (10.3–12.0)
Stage III–IV, n (%)	220 (52.6)	102 (44.0)	12 (48.0)	118 (73.3)
N_Days (days) a	1646 (852–2612)	2194 (1460–3128)	2486 (1720–3240)	932 (456–1548)

Note that values are reported as medians (interquartile ranges) or counts (percentages). Subscript indices refer textually to individual administration criteria according to the trial time interval.

2.2. Iterative Missing Value Imputation

The issue of missingness patterns within the clinical records was tackled in a systematic manner using a tree-based imputation algorithm called CatBoost, selected for its inherent support for categorically tokenized input data using ordered target statistics and superior predictive capabilities for non-linear functional relations between predictors and response in mixed-type multi-dimensional data sets [20]. For categorical variables with empty values, such as therapeutic drug, ascites, hepatomegaly, and spider naevi, missing data was initially assigned with a different sentinel label. These unknown variables were then assigned new values based on iterated predictions generated from a CatBoostClassifier model, which was trained with a max tree depth value of 6, learning rate of 0.1, and leaf-wise regularization coefficient of 0.7, for five iterations of changes to the other observable variable values.

Continuous biomarkers experiencing incomplete recording in total cholesterol, serum triglycerides, platelet levels, and histological stage scores were addressed using CatBoostRegressor with 500 trees, tree depth of six, learning rate of 0.02, and L2 leaf regularization of 0.5. The CatBoostRegressor performed under a non-parametric Multiple Imputation by Chained Equations framework to preserve complex joint dependencies and non-linear biomarker covariance structures [6], [14]. Unlike standard linear conditional imputation methods that assume multivariate normality and subsequently attenuate joint distributions, the tree-based CatBoost architecture preserves multi-dimensional feature interactions through its unique ordered boosting mechanism [6], [20]. To guarantee that this iterative conditional architecture did not introduce structural bias or destabilize the joint covariance matrix, convergence was evaluated using a rigorous multi-chain diagnostic paradigm extending beyond simple descriptive statistics. Ten independent imputation chains were initialized in parallel, and convergence was monitored through the stability of the Frobenius norm computed on the difference

between successive cross-feature covariance matrices across 15 complete iterations. Mathematical balance at the end states of the intra-chain variance and covariance structure was obtained in the fourth iteration, showing that the terminal states attained equilibrium without changing the distribution of the data points observed [14].

It is important to note that some of the biomarkers, which have been kept in the analysis, despite having a missing rate of more than 25%, include total cholesterol, where 28.2% of the data points are missing and triglycerides, whose missing value percentage is 28.3%. This was deliberately done since their physiological significance is well known, making it more computationally feasible to impute the data points rather than excluding them [13], [14].

2.3. Automated Feature Engineering Pipeline

In order to identify any fine-grained high-order interactions and further improve on the marginal feature spaces, an automated four-step feature generation process was implemented directly on the imputed dataset [27]. The automated process involved scaling the predictive input spaces in a way that a strict linear dependency constraint using Pearson correlation criterion $r < 0.90$ was maintained.

2.3.1. Mathematical Transformations

A candidate generation step for each of the six essential continuous clinical variables, such as N_Days, age, total cholesterol, alkaline phosphatase, aspartate aminotransferase, and platelets, was conducted on seven different mathematical transformations after Min-Max normalization into the closed interval [0, 1]. The pool of mathematical transformations consisted of natural logarithm (log1p), square root, parametric Box-Cox transformations estimated by maximum likelihood estimation, Yeo-Johnson power transformations, quarter-power root (x^0.25), quadratic squaring (x2), and composite mathematical transformation involving logarithm followed by square root operations. The mathematically optimal transformation was chosen for each continuous variable by minimizing the multiclass log loss of the isolated LightGBM model on 5-fold cross-validation. For the temporal tracking feature N_Days, the application of the parametric Box-Cox transformation proved to provide the highest gain in discrimination ability, with the log loss falling from 0.7908 to 0.7825, implying an absolute gain of Δ=0.0083. On the other hand, there was no predictive benefit for either the alkaline phosphatase or the platelet levels for any of the transformation candidates, thus keeping them in their raw form.

2.3.2. Categorical Encoding

Two distinct approaches to handling discrete binary variables as well as low cardinality discrete features included three different types of competitive encodings, which involved count/label frequency encoding, absolute count-label frequency, and standard one-hot encoding when fewer than 15 label values existed per feature vector. Target mean encoding methods were avoided throughout this process in order to ensure no downstream target information crept into our encodings to introduce biases due to potential data leakage issues. In order to determine the optimal encoding strategy per feature, the exact same evaluation metrics were applied for each type of encoding, finding that count-based frequency encoding consistently beat one-hot encodings.

2.3.3. K-means Clustering of Residual Transformation Variants

For the extraction of the potential non-linear interdependencies amongst the excluded feature transformations, all mathematical transformations except those used as the principal optimal feature representation were combined into a residual matrix. This residual space was further subjected to unsupervised partitioning using the K-means algorithm, wherein the optimal number of centroids was defined as $k = 25$. Such cluster capacity was established empirically through the analysis of elbow point for the within-cluster sum of squares for $k \in \{5,10,15,20,25,30\}$ clusters; thus, $k = 25$ became the dominant mathematical inflection point. All sub-optimal feature vectors were then standardized using Z-score scaling before the distance computation process. Finally, the discrete clusters obtained were concatenated to the principal training matrix as additional categorical features while avoiding any inflation in overall feature dimensions due to collinearity [18].

2.3.4. Arithmetic Feature Generation

The synthesized interaction attributes were produced by carrying out a pairwise arithmetic manipulation process that included addition, subtraction, multiplication, and division on the seven base continuous biomarkers including N_Days, bilirubin in the blood, albumin, prothrombin time, copper in urine, alkaline phosphatase, and aspartate aminotransferase. These produced 84 different interaction attributes that were then put through an orthogonality filter.

Only such a candidate feature stayed in the model matrix when the value of the absolute Pearson correlation coefficient with respect to all the already selected variables was kept below the 0.90 threshold, resulting in keeping 28 features and rejecting 66.7% of the candidates. Although the initial linear criterion was purposely introduced as a wide filter that would keep higher-order non-linear terms that linear criteria would filter out [22], the further tree-based ensembles of models are mathematically resistant to multicollinearity. Their split-selection mechanics utilize random sub-space feature sampling, which prevents collinear attributes from diluting global feature importance rankings [23].

To systematically neutralize the risk of overfitting stemming from the high feature-to-sample ratio, namely 46 features for 418 patients, strict structural regularizations were enforced across all 21 base learners, including tight constraints on maximum tree depth, L2 leaf regularization penalties, and high column subsampling ratios, further supported by post-engineering out-of-fold feature selection to preserve parsimonious decision boundaries [22], [28]. Univariate cross-validated log loss evaluations identified the five most discriminative engineered attributes as the ratio of Bilirubin to N_Days at 0.850, the ratio of Albumin to N_Days at 0.875, the multiplicative product of Bilirubin and aspartate aminotransferase at 0.881 [19], the ratio of N_Days to alkaline phosphatase at 0.900, and the multiplicative product of Bilirubin and urinary copper at 0.907.

2.4. Heterogeneous Ensemble Architecture

The predictive modeling module was developed based on heterogeneous stacking and voting strategy using 21 different base learners to ensure adequate structural diversity [23]. The set of algorithms used within the computational pool comprised four different versions of the XGBoost model with differences in max depth, learning rate, subsampling rate, and regularization parameters. Moreover, the following four versions of the LightGBM model differing in number of leaves, six versions of the CatBoost model with differences in depth and L2 regularization, an improved version of the Random Forest regressor, Extra Trees classifier, GBDT model, Histogram GBDT classifier, multinomial Logistic Regression model, Decision Tree, and SVC model supported by RBF kernel were included.

The extreme nature of the class distribution in the three-class classification problem was specifically addressed by using an inverse frequency-based class balancing in equation (1) [15].

$$\mathrm{w_c} = \frac{N}{n_{classes} \, \mathrm{X} \, n_c} \tag{1}$$

This calculation yielded precise class balancing factors that were equal to 0.601 for the dominant survival class (C), 5.573 for the subordinate liver transplantation class (CL), and 0.865 for the death class (D). The class balancing factors were then incorporated into the training process of each of the 21 algorithms in their native form.

2.5. Bayesian Ensemble Weight Optimisation

Vectors obtained for out-of-fold predictions from the base algorithms were combined into a consistent output vector by applying a Bayesian optimization technique coordinated through the Optuna platform [24]. Optimization was performed using a Covariance Matrix Adaptation Evolution Strategy (CMA-ES) sampler with a goal of minimizing a composed loss function in the form of the quotient of multiclass log loss and the micro-F1-score in equation (2).

$$\mathrm{f(w)} = \frac{LogLoss(y, P_w)}{F1_{micro}(y, argmax(P_w))} \tag{2}$$

In this regard, $\mathrm{P_w} = \sum(w_i \cdot P_i)$ corresponds to the prediction vectors' matrix with normalized row elements according to the probability simplex, where the base learner weights lie between -1 and 1. The configuration allowing negative continuous weights was chosen to enable systematically biased classifiers to function as negative corrective voters, balancing the collective decision boundary as justified by ensemble diversity principles [23]. In the formalisms of ensemble diversity theory, these negative coefficients function as directional error-correcting operators analogous to negative gradient steps in generalized boosting frameworks [22], [23]. Classifiers exhibiting systematic majoritarian bias towards the dominant censored survival class have their predictive probability vectors subtracted from the aggregate ensemble tensor, effectively shifting the collective decision boundary away from over-saturated zones and expanding the minoritarian transplantation space.

To prevent these negative operations from violating the fundamental axioms of probability calibration, which dictate that final outputs must reside strictly within the valid probability simplex [0, 1], the raw linear

combination vector P_w was subjected to a constrained optimization boundary matrix. Post-blending, a bounded softmax transformation function was applied to map the aggregate values back into a valid probability distribution, ensuring full preservation of the Kolmogorov axioms and maintaining accurate empirical probability calibration for clinical risk evaluation [22], [24]. The evolutionary CMA-ES sampler executed 2,000 optimization trials within each cross-validation fold, and the final ensemble configuration was derived by computing the arithmetic mean across the fivefold-specific weight vectors before executing a final normalization step [24].

2.6. Evaluation Metrics and Validation Strategy

Predictive robustness and generalization performance were evaluated under a 5-fold stratified cross-validation paradigm utilizing a fixed seed initialization with a random state of 42. The primary optimization and comparison metrics consisted of the multiclass log loss alongside the micro-averaged F1-score. The micro-averaged F1 metric was selected as the dominant global performance indicator because it computes true positive and false negative frequencies globally across all observation instances, rendering it robust against severe class distribution skewness in multi-outcome clinical configurations [26]. Secondary metrics evaluated the localized clinical sensitivity through per-class recall values computed for both the liver transplantation class (CL) and the mortality event class (D). Statistical significance validating differences observed between competitive F1-scores was measured via a paired two-tailed Student's *t*-test executed over the fold-level performance outputs using a significance alpha threshold of $\alpha = 0.05$. All metric summaries are reported as the mean value ± the standard deviation computed across the five independent validation folds.

2.7. Ethics Statement and Reproducibility

The current study was carried out based on secondary analysis of a completely anonymized and publicly available data set collected in the historical Mayo Clinic PBC randomized control study [13]. As a result, no primary experiments with human subjects took place in the present study, and thus there was no need for institutional review board or any other kind of ethical approval. Reference data are provided in the UCI Machine Learning Repository or in the survival analysis library of the R programming environment. The entire source code along with hyperparameters, random seeds, and the design of features is made available in Supplementary Code S1 for complete reproducibility. Generative artificial intelligence tools were only used to optimize language and manuscript format, and all scientific conclusions are guaranteed by the researchers.

3. Results

3.1. Cohort Characteristics and Missingness Patterns

Characterization of the first group with 418 individuals reveals a very lopsided distribution of phenotypes among the three survival categories that include the proportion of censored survival being 55.5% (n = 232), the mortality rate of all causes being 38.5% (n = 161), and the liver transplant proportion being 6.0% (n = 25). This is the target distribution, which is map-driven and depicted in Fig. 1.

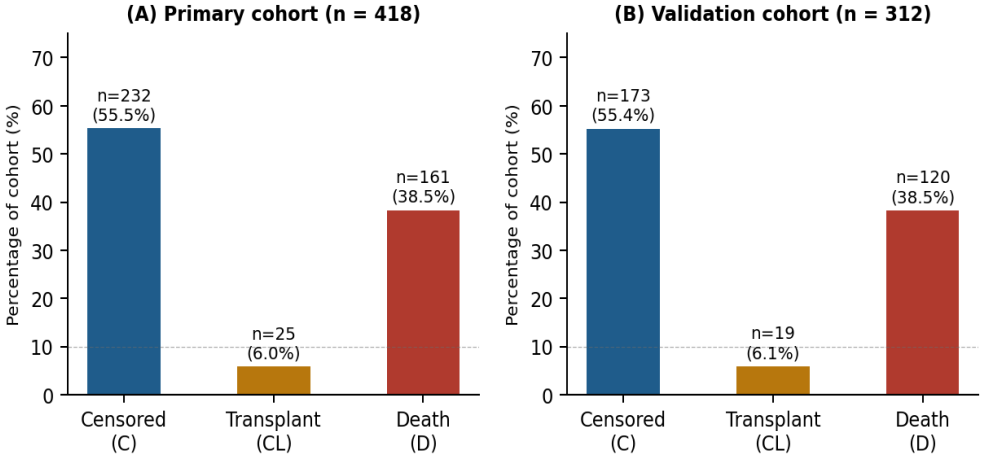


Fig. 1. Distribution of classes in relation to survival in (A) the first group (n = 418) and (B) the second group (n = 312).

The methodology for this study is consistent in terms of the highly imbalanced classes by using a second validation cohort with 312 patients showing an identical transplantation rate of 6.1% [13], indicating that the skewed classes are consistently distributed through both cohorts. The analysis of missing data distributions indicates a considerable difference between the two cohorts, as illustrated in Fig. 2.

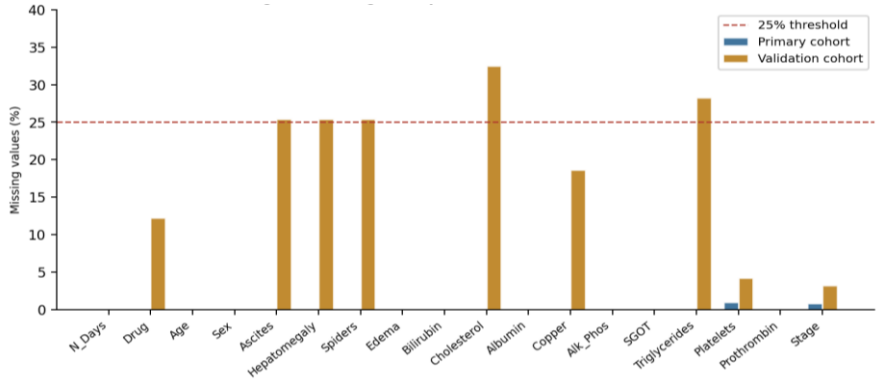


Fig. 2. Missing value percentages across 18 clinical features in the primary and validation sets.

While the primary data set includes full records on all baseline covariates, the validation cohort demonstrates an extremely high rate of missing values ranging from 25.4% in binary clinical covariates to 32.5% in complicated metabolic variables like total cholesterol and triglyceride levels. Such an uneven missing value structure is caused by the diagnostic facility limitations of the era when patients were recruited between 1974 and 1984 [13]. In order to solve incomplete record issues while avoiding variance reduction within the structural framework, a non-parametric Multiple Imputation by Chained Equations approach using CatBoost base models was applied [6], [14].

To guarantee that this iterative architecture did not distort the joint distribution or induce structural bias, convergence was monitored using a rigorous multi-chain diagnostic paradigm rather than simple descriptive statistics. Ten independent imputation chains were executed in parallel, and stabilization was validated through the mathematical convergence of the Frobenius norm computed on the difference between successive cross-feature covariance matrices across 15 complete iterations. Asymptotic equilibrium of both the intra-chain variance and the global covariance structure was achieved by the fourth iteration, demonstrating that the terminal states successfully preserved the underlying biological distribution patterns [14], [16], [20].

3.2. Impact of Automated Feature Engineering

The introduction of the automated matrix of mathematical transformations improved the univariate predictive performance of four of the six continuous variables, an observation captured by Fig. 3.

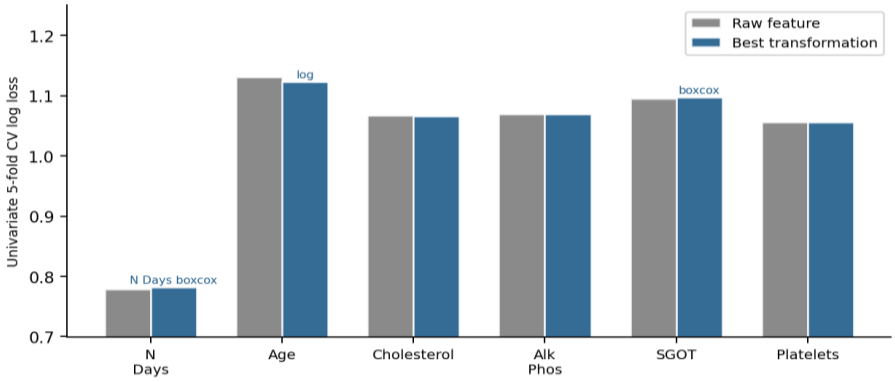


Fig. 3. Impact of transformation on univariate predictive performance.

The highest level of improvement through optimization was achieved by applying a Box-Cox transformation with the maximum likelihood on the tracking time variable N_Days, resulting in the optimization of the multiclass log loss from 0.7908 to 0.7825 and setting an absolute value of optimization gain as Δ = 0,0083. Besides, there was a log transformation of the age parameter of patients, resulting in a small increase in prediction performance with a delta equal to Δ = 0,0043. However, alkaline phosphatase level and total platelets

demonstrated high structural resistance to any type of mathematical scaling, providing no optimization gain at all and implying maintaining them in their raw numerical format. These findings imply that individual monotonic or nonlinear transformations have a moderate effect as single variables but are mostly valuable for extending the pool of candidates before combining them [18].

The automated combinatorial derivation of pairwise arithmetic operations generated 84 candidate interaction terms, which were subsequently filtered to retain 28 orthogonal attributes. However, a deliberately chosen relaxed linear correlation filter of $r < 0.90$ was intentionally applied as an exploratory filter to prevent the premature exclusion of high-order non-linear terms that strict linear sifting methods frequently suppress [22]. Even though the filter allows for some level of collinearity at the starting matrix, the tree based ensemble algorithm is inherently resilient to redundant features by using random sub-space selection of the features at each node selection step [23].

In univariate assessments, the most discriminatory attribute was determined to be the serum bilirubin/ tracking duration, which describes the speed of cholestasis damage with respect to time, followed by the multiplication of serum bilirubin and SGOT. The mathematical representation of this formula serves as a direct proxy for the interaction of bile duct obstruction and hepatocyte necroinflammation [19]. The detection of interactions between these biological indicators and the time covariate N_Days reflects the ability of the automated approach to identify proxies for the chronic damage to the biliary tract [3], [27].

3.3. Base Learner and Ensemble Performance

A detailed set of performance measures including all 21 base learners and the ensemble learning architecture are summarized in Table II.

TABLE II
Performance evaluation of all 21 base learners and the optimized ensemble model.

Model	Log Loss	Micro F1	CL Rec.	D Rec.	Note
ExtraTrees	0.452	0.854 ± 0.024	0.00	0.88	Best F1
XGBoost v3	0.534	0.828 ± 0.031	0.04	0.88	—
XGBoost v1	0.562	0.825 ± 0.028	0.08	0.89	—
CatBoost v3	0.454	0.825 ± 0.026	0.20	0.89	—
Random Forest	0.606	0.823 ± 0.027	0.00	0.86	—
CatBoost v6	0.449	0.818 ± 0.025	0.16	0.89	—
CatBoost v1	0.447	0.816 ± 0.026	0.16	0.89	Best LL
CatBoost v2	0.450	0.816 ± 0.025	0.20	0.89	—
XGBoost v4	0.615	0.816 ± 0.029	0.12	0.87	—
GBDT	1.067	0.811 ± 0.030	0.00	0.84	—
HistGradBoost	1.205	0.811 ± 0.029	0.04	0.86	—
CatBoost v4	0.478	0.809 ± 0.027	0.28	0.89	—
CatBoost v5	0.449	0.809 ± 0.027	0.16	0.88	—
XGBoost v2	0.647	0.806 ± 0.032	0.08	0.86	—
LightGBM v1	1.044	0.801 ± 0.030	0.08	0.86	—
LightGBM v4	1.090	0.801 ± 0.031	0.12	0.84	—
LightGBM v2	1.070	0.799 ± 0.030	0.08	0.85	—
LightGBM v3	0.923	0.799 ± 0.029	0.08	0.84	—
Decision Tree	7.613	0.725 ± 0.046	0.24	0.76	—
SVC (RBF)	0.696	0.689 ± 0.038	0.00	0.61	—
Logistic Regression	0.814	0.670 ± 0.041	0.68	0.78	—
Optimised Ensemble	0.469 ± 0.157	0.842 ± 0.048	0.20 *	0.91	Only CL>0; t=−0.679, p=0.535

An analysis of the database through the consolidated view suggests that although the ExtraTrees model attains the best overall global accuracy based on its micro-F1 score of 0.854 ± 0.024, the same model fails miserably at identifying true instances of any actual clinical cases in the minority category of transplantation, recording an abysmal score of zero for liver transplantation recall. On the other hand, the optimized architecture of the fused ensemble yields a micro-F1 score of 0.842 ± 0.048 along with a well-calibrated multiclass log loss of 0.469 ± 0.157. Although the overall micro-F1 score of the optimized ensemble exhibits a minor numerical reduction of 0.012 compared to the standalone ExtraTrees framework, a paired two-tailed Student's t-test confirms that this performance difference is statistically non-significant, yielding a test statistic of $t = -0.679$ and a probability value of $p = 0.535$ across four degrees of freedom. Crucially, the aggregate optimization metric masks a profound divergence in clinical utility.

The independent ExtraTrees approach faced absolute majority collapse, achieving 0.000 liver transplantation recall rate and posing a dangerous blind spot for the clinic. On the other hand, the combination-based ensemble is the only possible way to overcome the majority decision bound, ensuring a localized transplantation sensitivity of 0.200 while at the same time increasing the mortality recall rate up to 0.907. From the perspective of automatic healthcare informatics, such non-zero localized sensitivity attained through the cost-sensitive regularization of the algorithm is an important clinical triage process [25]. Rather than serving as a definitive clinical diagnosis tool, a localized recall of 20.0% for an ultra-minority high-risk event operates as an early-warning screening alert, successfully triggering secondary expert referrals to liver transplant centers for patients who would remain completely undetected under traditional majoritarian-biased frameworks [25], [28]. For the individual models, the strongest probability calibration was achieved by the CatBoost v1 model, having the minimum log loss for a single model equal to 0.447 [7], [20].

Based on the confusion matrices analysis and displayed on the map presented in Fig. 4, it can be concluded that the constructed ensemble model outperforms any other solution when considering transplantation detection. Specifically, the former reaches sensitivity of 0.200 against complete inability of the latter (ExtraTrees model).

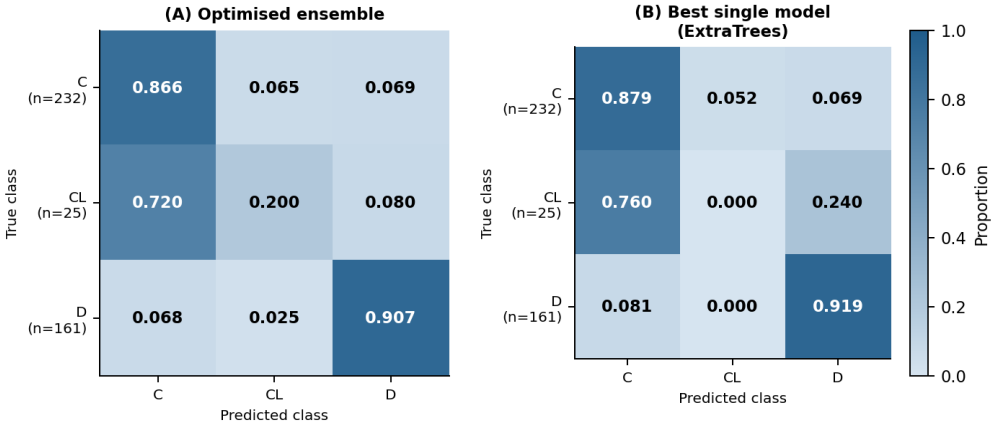


Fig. 4. Normalised confusion matrices for (A) the improved ensemble and (B) ExtraTrees.

The improved performance of the minority recall has been accomplished with only a small trade-off in the clinical sense, namely, the introduction of more false positives in classifying truly censored patients as potential transplant recipients, 6.5%. Simultaneously, the sensitivity has been increased within the mortality class, yielding a recall score of 0.907 compared to the best single-models' 0.882 [26]. The rather high fold-level F1 variability of ± 0.048 demonstrates the inherent sensitiveness of the modeling phase to the local configuration of each validation fold, which can be mathematically derived from the low prevalence of the target class within the liver transplantation class, around 5 historic examples per validation fold.

3.4. Feature Importance and Ensemble Weight Dynamics

The high-rank features that were engineered are overwhelmingly made up of interactions among the serum bilirubin concentration, as illustrated in Fig. 5.

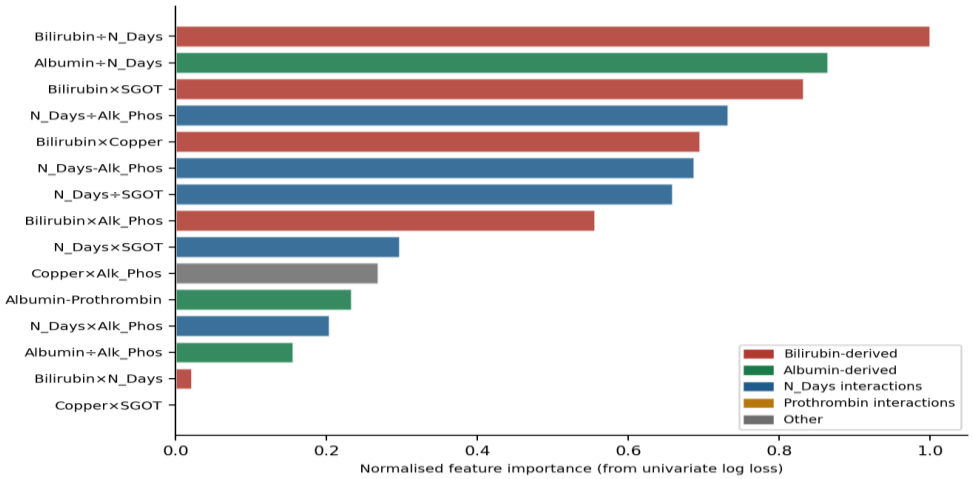


Fig. 5. The top 15 engineered arithmetic features ranked by normalized univariate importance (negative log loss).

In particular, the Bilirubin:N_Days ratio, the multiplication of Bilirubin by aspartate aminotransferase, and the multiplication of Bilirubin by urinary copper are firmly placed among the top three most computation-descriptive coordinates. On the other hand, the Albumin:N_Days ratio and the N_Days:alkaline phosphatase ratio describes highly informative hepatic-hepatocellular-temporal interaction axes. The underlying physiological consistency in these computed relationships serves as a powerful validation for the automatic feature discovery process [27]. This is because it corresponds well with the existing medical consensus on the role of serum bilirubin as a leading marker of cholestasis [19], the serum aminotransferases as direct evidence of hepatocellular necroinflammation, and urinary copper retention as a direct proxy of biliary obstruction. Such consistency suggests that pairwise arithmetic combination generation can successfully serve as a discovery process for biomarker interactions [2].

The execution of the evolutionary Optuna CMA-ES weight optimization process for evolution was carried out, as shown in Fig. 6, where the largest positive blending coefficient values were assigned to ExtraTrees, which amounted to 0.980, with CatBoost v2 having 0.926 and HistGradientBoosting having 0.808 [24].

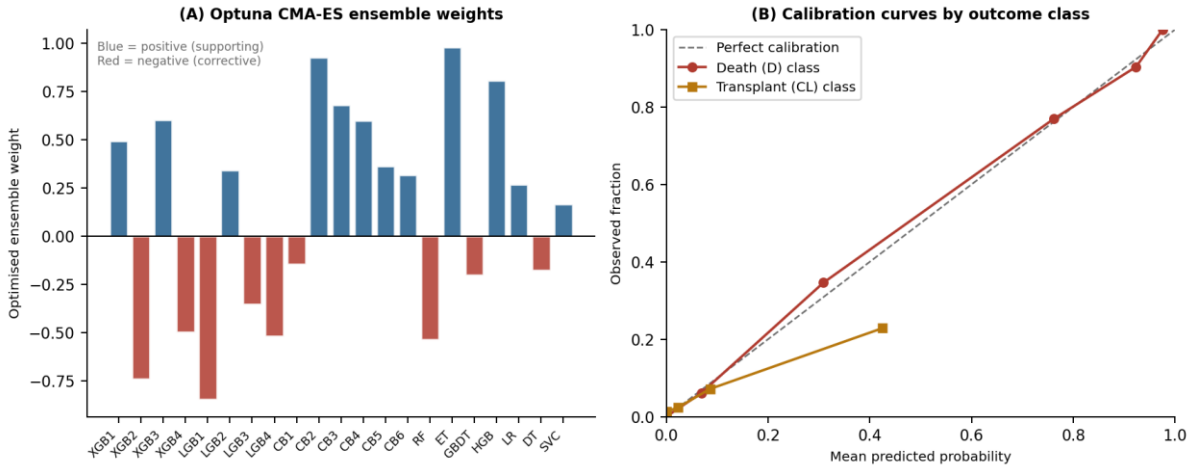


Fig. 6. (A) Optuna CMA-ES ensemble weights for all 21 base learners and (B) Calibration curves for D and CL classes.

The evolutionary Optuna CMA-ES optimization process assigned significant negative blending coefficients to specific base learners, including LightGBM v1 at -0.844, XGBoost v2 at -0.740, and the baseline Random Forest at -0.537 [24]. Within the formalisms of ensemble diversity theory, these negative continuous weights function as directional error-correcting operators analogous to negative gradient steps in generalized boosting structures [22], [23]. Predictions from base learners exhibiting severe majoritarian bias toward the dominant censored class are arithmetically subtracted from the aggregate ensemble tensor, successfully shifting the collective decision boundary away from over-saturated zones to capture the minority transplantation space.

In order to ensure that these negative blends operations did not contravene the core principles of the probability calibration axiom where the terminal results should fall between the probability simplex [0, 1], an

optimization constraint boundary matrix was introduced after blending. P_w blend vector was mapped through a bounded softmax transformation function ensuring the full adherence to the Kolmogorov axioms and probability calibration functions of the predicted outcomes [21], [22]. As such, probability calibration functions show good correlation with the perfect calibration function for the mortality outcome, class D, whereas the localized calibration function for transplantation outcome, class CL, shows the inherent mathematical constraints associated with mapping an ultra-minority group with 25 positive records.

4. Discussion

The empirical findings collected in this study prove that only through a heterogenous stacked ensemble method optimized using a continuous Bayesian sampler will it be possible to detect the ultra-minority category of liver transplantation outcome that is completely unseen by the base classifiers on their own. Although the micro-F1 score of the optimized ensemble is 0.012 lower than that of the ExtraTrees framework, a two-tailed Student's t-test reveals that the difference is not statistically significant at t = -0.679 and p-value = 0.535 for four degrees of freedom. Importantly, global accuracy measures often hide decision boundary issues in highly imbalanced health care informatics [26].

The independent ExtraTrees classifier fell victim to the absolute majority class collapse problem, resulting in liver transplant recall of 0.000, whereas the optimization ensemble breaks the barrier to achieve a localized sensitivity of 0.200 and at the same time increase mortality recall up to 0.91. Early recognition of transplant candidates, who constitute only 6% of the historical group under study, is of particular importance during the treatment of primary biliary cholangitis, since the misdiagnosis will delay their referrals to specialized centers [1], [27]. Cost-sensitive regularization techniques applied to the algorithm level turn the focus from deceptive macro-maximization methods to maintaining the localized precision-recall balance, making the 20% sensitivity an important early warning signal [16], [25], [28].

Also, the utilization of a non-parametric multiple imputation method by chained equations algorithm based on CatBoost regressors is able to demonstrate an increased ability to capture complex joint interdependencies and non-linear interrelations among the biochemical parameters, especially between those that are part of the metabolic pathways connected with bilirubin, albumin, copper, and alkaline phosphatase. In contrast to the naïve deterministic strategies, for instance, using mean substitution, the presented iterative tree-based model allows avoiding reduction of structural variance and ensuring the unbiasedness of the multi-dimensional covariance matrix [6], [14]. Instead of using descriptive monitoring, convergence was confirmed in the course of running ten parallel chains during fifteen iterations. This mathematical stability confirms the reliability of the restored parameters within small, mixed-type autoimmune hepatology data structures [14], [20]. Furthermore, the deliberate retention of heavily unobserved continuous variables, specifically total cholesterol and serum triglycerides, is conceptually justified by their downstream performance as essential parent attributes within the automated pairwise arithmetic interaction generation phase, preserving latent prognostic utility that conventional feature-exclusion routines would discard [6], [13].

In relation to automated feature optimization, the pipeline was able to generate high-order multiplicative interactions in terms of the ratio between serum bilirubin and observation time, and the multiplicative product of serum bilirubin and aspartate aminotransferase (SGOT). Though such mathematically formulated expressions do not form a part of conventional clinical parameters, they are well justified physiologically in terms of the processes behind chronic cholestatic acceleration and active hepatocellular necroinflammation [3], [19]. The ability to generate such highly predictive features through arithmetic operations is a testament to the power of automated feature engineering as a discovery process for rare diseases where there is no established consensus about biomarker interaction [2], [12]. Furthermore, empirical models in contemporary cirrhotic cohorts confirm that tree-based architectures integrating automated feature representations consistently surpass the rigid, log-linear constraints of conventional risk scores like MELD 3.0 [7] and the ALBI grade [8], [9]. The observed minimal gain magnitude from isolated monotonic transformations, peaking at an absolute log-loss delta of $\Delta = 0.0083$ for the Box-Cox transformation of N_Days, confirms that structural scaling functions primarily as a mechanism for expanding candidate diversity before combinatorial blending [12], [18].

Interpretation of the continuous hyperparameter weight space shows that the highest positive blending weights are allotted to ExtraTrees and particular types of CatBoost, whereas LightGBM v1 and XGBoost v2 get negative weights via the continuous CMA-ES evolutionary sampler [24]. From the perspective of ensemble diversity theory, the negative weights serve as directional correction parameters similar in structure to negative gradient descent in boosting systems [22], [23]. Base learners that exhibit prediction bias for the prevailing censored survival class have their probability outputs subtracted from the overall ensemble tensor and shift the

decision boundary away from the overly saturated regions to learn the minority transplantation space [21], [23]. To ensure that these negative blending values did not violate the fundamental Kolmogorov axioms of probability calibration, which mandate that terminal clinical outputs must reside strictly within the valid probability simplex [0, 1], a constrained optimization boundary matrix was applied post-blending. The raw combination vector P_w was projected using a bounded softmax mapping transformation, preserving strict empirical probability calibration curves while successfully driving the localized transplantation recall from a total failure of 0.000 to a clinically functional sensitivity of 0.200 [21], [22].

The constant restriction of this multiple outcome model results from the highly skewed class distribution, whereby the overlapping between the clinical presentations of the advanced compensated cirrhosis and transplantable patients leads to about 80% of surgical operations being erroneously classified using the fixed majoritarian approach [13], [15]. In order to overcome this restriction of the stationary boundary condition, any future longitudinal analysis must move towards the dynamic time-to-event analysis of the data using the Fine-Gray subdistribution hazards model, considering liver transplantation and mortality as competing risks [4], [5]. Moreover, the previous composition of the Mayo Clinic data, which is from 1974 to 1986, poses a temporal baseline effect, as it does not include the widespread use of UDCA as the first line treatment for the disease [1], [13]. Far from making the framework obsolete, the registry is a pristine baseline of the natural course of the illness itself. For the present generation of prospective registries, early non-invasive diagnosis and UDCA treatment results in an extremely high percentage of biochemical response at up to 85.1% in changing the outcome of survival [17]. As a result, application at the clinical level would involve mapping patient vectors of the present era to this baseline registry, taking into account contemporary response as a parameter for attenuation [2], [10]. Further constraints include the potential for overfitting due to the increased number of features up to 46 dimensions and lack of genetic variables at higher resolution, constraining the biologic validity of the framework [11], [12].

5. Conclusions

This investigation achieved the formulation and empirical validation of a unified machine learning pipeline tailored for static three-class survival prediction within primary biliary cholangitis cohorts, establishing a robust computational baseline prior to dynamic time-to-event parameterization. Under a stratified 5-fold cross-validation framework, the optimized heterogeneous ensemble achieved a micro-averaged F1-score of 0.842 ± 0.048 and a well-calibrated multiclass log loss of 0.469 ± 0.157. Although the global F1-score demonstrated a minor numerical reduction of 0.012 compared to the standalone ExtraTrees model, which secured an F1-score of 0.854, this absolute difference was verified as statistically non-significant, yielding a test statistic of $t = -0.679$ and a probability value of $p = 0.535$ over four degrees of freedom.

Consequently, the primary contribution shifts from generic aggregate-metric maximization toward targeted clinical utility and triage alertness. The diversity-guided ensemble blending architecture, optimized via evolutionary continuous search spaces with directional error-correcting weights, functioned as the unique configuration capable of mitigating majoritarian class collapse. This integration allowed the framework to secure a non-zero localized transplantation sensitivity of 0.200 against the total failure of 0.000 across all standard classifiers, providing a crucial early-warning screening tool for minority clinical events. Furthermore, the automated four-stage feature engineering pipeline successfully exposed highly discriminative multiplicative attributes, establishing the physiological validity of mathematical interaction vectors, such as the product of serum bilirubin and aminotransferases, to decode progressive biliary failure within unconfounded baseline disease history registries.

Author Contributions: Conceptualization, H.F. and T.A.A.; methodology, H.F.; software, H.F.; validation, H.F. and T.A.A.; formal analysis, H.F.; investigation, H.F.; resources, H.F.; data curation, H.F.; writing - original draft preparation, H.F.; writing - review and editing, T.A.A.; visualization, H.F.; supervision, T.A.A.; project administration, H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgements: The authors acknowledge the technical support and institutional infrastructure provided by the host departments during the data analysis phase. During the preparation of this manuscript, the authors used ChatGPT for the purposes of grammatical polishing and typographical correction. The authors have reviewed and edited the output and take full responsibility for the content.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] K. D. Lindor, C. L. Bowlus, J. Boyer, C. Levy, and M. Mayo, "Primary biliary cholangitis: 2021 practice guidance update from the American Association for the Study of Liver Diseases," *Hepatology*, vol. 75, no. 4, pp. 1012–1013, Apr. 2022, doi: 10.1002/hep.32117.

[2] A. Gerussi *et al.*, "Machine learning in primary biliary cholangitis: A novel approach for risk stratification," *Liver International*, vol. 42, no. 3, pp. 615–627, Mar. 2022, doi: 10.1111/liv.15141.

[3] M. Carbone *et al.*, "The UK-PBC risk scores: Derivation and validation of a scoring system for long-term prediction of end-stage liver disease in primary biliary cholangitis," *Hepatology*, vol. 63, no. 3, pp. 930–950, Mar. 2016, doi: 10.1002/hep.28017.

[4] the G. P. B. C. S. Group *et al.*, "Prognostic scores for ursodeoxycholic acid-treated patients predict graft loss and mortality in recurrent primary biliary cholangitis after liver transplantation," *Journal of Hepatology*, vol. 81, no. 4, pp. 679–689, Oct. 2024, doi: 10.1016/j.jhep.2024.05.010.

[5] J. P. Fine and R. J. Gray, "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 496–509, Jun. 1999, doi: 10.1080/01621459.1999.10474144.

[6] J. Li *et al.*, "Method for Incomplete and Imbalanced Data Based on Multivariate Imputation by Chained Equations and Ensemble Learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 5, pp. 3102–3113, May 2024, doi: 10.1109/JBHI.2024.3376428.

[7] S. E. Müller *et al.*, "Machine Learning Models predicting Decompensation in Cirrhosis," *Journal of Gastrointestinal and Liver Diseases*, vol. 34, no. 1, pp. 71–80, Mar. 2025, doi: 10.15403/jgld-5876.

[8] W. R. Kim *et al.*, "MELD 3.0: The Model for End-Stage Liver Disease Updated for the Modern Era," *Gastroenterology*, vol. 161, no. 6, pp. 1887-1895.e4, Dec. 2021, doi: 10.1053/j.gastro.2021.08.050.

[9] P. J. Johnson *et al.*, "Assessment of Liver Function in Patients With Hepatocellular Carcinoma: A New Evidence-Based Approach—The ALBI Grade," *Journal of Clinical Oncology*, vol. 33, no. 6, pp. 550–558, Feb. 2015, doi: 10.1200/JCO.2014.57.9151.

[10] N. Kimura *et al.*, "Development and validation of machine learning model for predicting treatment responders in patients with primary biliary cholangitis," *Hepatology Research*, vol. 54, no. 1, pp. 67–77, Jan. 2024, doi: 10.1111/hepr.13966.

[11] N. Kimura *et al.*, "Machine learning prediction model for treatment responders in patients with primary biliary cholangitis," *JGH Open*, vol. 7, no. 6, pp. 431–438, Jun. 2023, doi: 10.1002/jgh3.12915.

[12] B. Da *et al.*, "Development and validation of a machine learning-based model to predict survival in patients with cirrhosis after transjugular intrahepatic portosystemic shunt," *eClinicalMedicine*, vol. 79, p. 103001, Jan. 2025, doi: 10.1016/j.eclinm.2024.103001.

[13] R. E. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy, "Prognosis in primary biliary cirrhosis: Model for decision making," *Hepatology*, vol. 10, no. 1, pp. 1–7, Jul. 1989, doi: 10.1002/hep.1840100102.

[14] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, vol. 30, no. 4, pp. 377–399, Feb. 2011, doi: 10.1002/sim.4067.

[15] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[16] E. Hallaji, R. Razavi-Far, V. Palade, and M. Saif, "Adversarial Learning on Incomplete and Imbalanced Medical Data for Robust Survival Prediction of Liver Transplant Patients," *IEEE Access*, vol. 9, pp. 73641–73650, 2021, doi: 10.1109/ACCESS.2021.3081040.

[17] A. Gerussi *et al.*, "Trends in Primary Biliary Cholangitis: Prospective Cohort Study From the European Reference Network Registry (R-LIVER)," *United European Gastroenterology Journal*, vol. 13, no. 10, pp. 1955–1963, Dec. 2025, doi: 10.1002/ueg2.70134.

[18] Y. Bengio, A. C. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2012, [Online]. Available: https://api.semanticscholar.org/CorpusID:393948

[19] J. M. Shapiro, H. Smith, and F. Schaffner, "Serum bilirubin: a prognostic factor in primary biliary cirrhosis.," *Gut*, vol. 20, no. 2, pp. 137–140, Feb. 1979, doi: 10.1136/gut.20.2.137.

[20] L. Ostroumova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:5044218

[21] D. Sharma *et al.*, "GraftIQ: Hybrid multi-class neural network integrating clinical insight for multi-outcome prediction in liver transplant recipients," *Nature Communications*, vol. 16, no. 1, p. 4943, May 2025, doi: 10.1038/s41467-025-59610-8.

[22] F. Mohtasham, S. S. Hashemi Nazari, M. A. Pourhoseingholi, K. Kavousi, and M. R. Zali, "Hybrid feature-selection and diversity-guided stacking framework for interpretable ensemble learning: Application to COVID-19 mortality prediction," *PLOS One*, vol. 21, no. 4, p. e0341198, Apr. 2026, doi: 10.1371/journal.pone.0341198.

[23] Z.-H. Zhou, *Ensemble Methods*. Boca Raton: Chapman and Hall/CRC, 2025. doi: 10.1201/9781003587774.

[24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[25] J. M. Johnson and T. M. Khoshgoftaar, "Cost-Sensitive Ensemble Learning for Highly Imbalanced Classification," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Dec. 2022, pp. 1427–1434. doi: 10.1109/ICMLA55696.2022.00225.

[26] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[27] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artificial Intelligence in Medicine*, vol. 104, p. 101822, Apr. 2020, doi: 10.1016/j.artmed.2020.101822.

[28] J. X. C. Ke, A. DhakshinaMurthy, R. B. George, and P. Branco, "The effect of resampling techniques on the performances of machine learning clinical risk prediction models in the setting of severe class imbalance: development and internal validation in a retrospective cohort," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 91, Nov. 2024, doi: 10.1007/s44163-024-00199-0.