Research Square

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

Development of Bootstrap-Weighted Oblique Random Forest Algorithm for Multiclass Drug-Induced Liver Injury Prediction and Risk Assessment in Humans

Chung Yuan Yu
Arizona State University
Fahad Mostafa
fahad.mostafa@asu.edu

Arizona State University

Method Article

Keywords: DILI, bootstrap weighted oblique random forest, class imbalance, Mold2 descriptors, multiclass classification, interpretable machine learning

Posted Date: June 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9918997/v1

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

Additional Declarations: No competing interests reported.

Development of Bootstrap-Weighted Oblique Random Forest Algorithm for Multiclass Drug-Induced Liver Injury Prediction and Risk Assessment in Humans

Chung Yuan Yu , Fahad Mostafa *

School of Mathematical and Natural Sciences, Arizona State University
*Corresponding author's Email: fahad.mostafa@asu.edu/fahadraj.du@gmail.com

Abstract

Drug-induced liver injury (DILI) accounts for more safety-related withdrawals of approved drugs than any other adverse event. Early detection of DILI potential is a central objective of preclinical safety assessment, and computational models are now routinely applied alongside experimental assays. We propose the Bootstrap-Weighted Oblique Random Forest (BWORF), a tree ensemble that couples L_1-sparse oblique split learning, temperature-scaled weighted bootstrap sampling, and weighted leaf-level probability estimation into a single imbalance-aware design based on Mold2 descriptors. We designed three experiments to assess BWORF's performance by posing three critical questions: (1) does BWORF outperform conventional machine learning baselines on a three-class DILI risk stratification benchmark? (2) does the BWORF design produce mechanistic behavior under controlled conditions and generalize beyond the DILI domain? (3) could improved per-compound auditability be established? For question 1, we evaluated BWORF on a DILIrank-derived benchmark harmonized from LiverTox likelihood categories into three ordered risk tiers. BWORF achieved the highest balanced accuracy and macro F1 among six conventional baselines and a matched oblique-forest comparator, shifting predictive capacity toward intermediate and most-concern risk tiers. For question 2, BWORF recovered planted causal predictors earlier than axis-aligned and unweighted oblique comparators in simulation, led the forest-variant panel on a multiclass thyroid external benchmark, and led the precision-recall ranking on a population-scale BRFSS heart-attack held-out test, with the lead emerging at full sample size. For question 3, fold-local mutual-information screening and representative LIME case studies surfaced overlapping descriptor families, supporting per-compound auditability. Altogether, the developed BWORF model could serve as a class-balancing oblique-forest extension for imbalanced biomedical tabular prediction.

Keywords: DILI; bootstrap weighted oblique random forest; class imbalance; Mold2 descriptors; multiclass classification; interpretable machine learning.

1 Introduction

Drug-induced liver injury (DILI) accounts for more safety-related withdrawals of approved drugs than any other adverse event [1, 2] and remains a persistent challenge in drug development and post-market safety assessment because hepatotoxic risk is mechanistically heterogeneous, difficult to anticipate from preclinical studies, and clinically significant when missed [2–4]. Many hepatotoxic compounds are not flagged by conventional animal toxicology or *in vitro* hepatocyte assays and are identified only in clinical trials or after market approval [1, 5]. Because experimental hepatotoxicity screening is expensive, low-throughput, and of limited predictive power for idiosyncratic reactions, computational screening has become widely used alongside experimental hepatotoxicity testing in preclinical pipelines [5–7]. Curated benchmarks such as DILIrank, DILIst, and DILIrank 2.0 have made systematic structure-based modeling possible by providing harmonized hepatotoxicity annotations for large compound sets [1, 3, 8, 9], yet the prediction problem itself remains far from solved, and recent reviews continue to describe DILI screening as an area where current computational approaches are inadequate [2, 4, 5].

A substantial body of work has applied machine learning to DILI risk prediction from molecular descriptors and fingerprints. Quantitative structure-activity relationship (QSAR) models trained on FDA-approved drug labeling established the feasibility of descriptor-based DILI classification [10–13], and decision-forest variants on the DILIrank annotation extended this to three-class risk stratification [3, 6]. More recent efforts have introduced convolutional architectures on molecular fingerprints [14], substructure-space deep neural networks [15, 16], supervised chemical graph mining [17], transcriptional graph neural networks [18], QSAR modeling tied to adverse outcome pathways [19], and ensemble integrations of conventional and deep learners [20, 21], with DeepDILI [7], InterDILI [22], and MORF [23] advancing the integration of prediction quality with per-compound transparency. Across these studies, two algorithmic properties recur and constrain performance in the moderate-sample, class-imbalanced regime that characterizes curated DILI benchmarks. First, axis-aligned decision trees and their ensembles grow each internal node by thresholding a single feature at a time, which is efficient and interpretable but can require deeper trees to represent correlated combinations of descriptors common in Mold2 and similar structure-based encodings [24–28]. Second, uniform bootstrap aggregation reproduces the empirical class distribution in expectation and therefore leaves minority tiers thinly sampled during training [29–31]. Under class imbalance, these properties combine into a consistent failure mode: aggregate accuracy looks high because the majority No-concern class is easy to predict, while recall and F1 on the minority tiers that matter for safety triage remain weak, a pattern visible even in the strongest DILIrank-based studies [6, 7]. No existing DILI-specific method couples oblique split learning with weighted bootstrap sampling and weighted leaf-level probability estimation into a single imbalance-aware ensemble design, and none of the prior DILI-specific studies cited above addresses class imbalance at the ensemble-generation level.

Two further gaps in prior DILI modeling concern evaluation methodology and per-compound interpretability. On the evaluation side, accuracy and threshold-free discrimination dominate reported summaries in DILI work, yet these metrics can conceal minority-tier failure under class imbalance [31–33], and precision-recall framing has been argued to be more informative than receiver-operating framing in low-prevalence regimes [32, 33]. Protection against optimistic bias through strictly nested cross-validation, while standard in broader methodological work, is not uniformly adopted in DILI benchmarking [34, 35]. On the interpretability side, per-compound auditability is increasingly emphasized as a modeling objective instead of an optional add-on, with recent interpretable DILI models [22, 23] and reviews of explainable toxicity modeling [36–39] pushing toward evaluation at both the feature level and the case level. Surrogate explanation methods such as LIME [40] support case-level auditing of black-box classifiers but require disciplined selection of explanation cases to avoid cherry-picking, a concern rarely addressed explicitly in DILI-specific reporting.

To address these gaps, we developed the Bootstrap-Weighted Oblique Random Forest, the first tree ensemble to integrate oblique split learning [25–27], weighted bootstrap sampling [30, 31], and weighted leaf-level probability estimation into a single class-balanced ensemble design for DILI risk prediction. Each internal node depends on a learned combination of descriptors instead of a single feature; each bootstrap increases the expected training exposure of minority-tier compounds across trees; and each terminal node reports probabilities consistent with the weighted training objective. The three components share the same class-balanced weighting, so training exposure, probability estimation, and final prediction follow a common design principle rather than a sequence of bolted-on adjustments. This formulation advances a methodological progression from axis-aligned random forest [28] to oblique random forest [26, 27] to a bootstrap-weighted oblique variant and is compatible with recent oblique-forest work in adjacent biomedical modeling contexts [23, 41].

Our evaluation of BWORF comprises a comparison with six conventional machine learning baselines and a matched oblique-forest comparator on a three-class DILI risk stratification benchmark harmonized from LiverTox likelihood categories [3], a controlled simulation with known causal structure designed to isolate the mechanistic contributions of oblique splits and weighted bootstrap, and a two-regime external benchmark suite combining a multiclass in-regime test on the thyroid disease benchmark with a population-scale severely imbalanced binary test on the BRFSS 2022 heart-attack benchmark [42, 43]. We further demonstrated per-compound auditability through fold-local mutual-information screening and representative LIME case studies selected from pooled out-of-fold predictions of the finalized multiclass artifact [34, 40]. Taken together, BWORF offers a class-balancing alternative to standard tree ensembles for DILI risk prediction and demonstrates that ensemble-level class balancing, coupled with oblique split learning, can improve minority-tier recall without the tuning overhead of external resampling strategies.

2 Methods and Materials

2.1 Study design overview

This study evaluates BWORF under three complementary experimental components. The primary component is a three-class DILI risk stratification task that uses Mold2 molecular descriptors as predictors and LiverTox likelihood categories harmonized into ordered risk tiers as the supervised-learning target [3, 24].

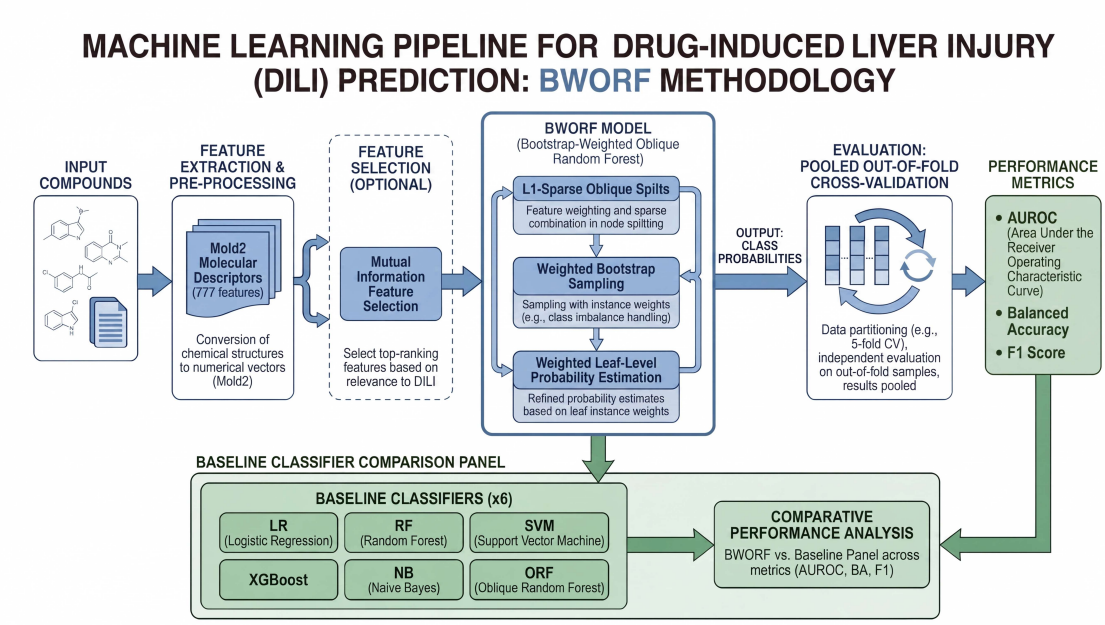


Figure 1: Pipeline of DILI prediction. Data preparation (DILIrank-derived descriptor table joined to harmonized LiverTox categories), fold-local preprocessing, BWORF training (oblique split learning with L_1-sparse projection, weighted bootstrap sampling, weighted leaf-level probability estimation), baseline panel, and pooled out-of-fold evaluation under stratified cross-validation.

A controlled simulation study with configurable linkage disequilibrium and planted interaction effects complements the real-data benchmark by providing ground truth against which causal-variant recovery can be measured. An external validation suite tests whether the BWORF profile generalizes outside the DILI domain and outside the Mold2 descriptor representation, and is organized into two complementary regimes: a multiclass in-regime test on the thyroid disease benchmark, which is the closest analog to the primary three-class DILI task, and a population-scale binary test on the BRFSS 2022 heart-attack benchmark under severe class imbalance, which probes whether BWORF's class-balancing components retain measurable leverage at modern data scale [42, 43]. Three smaller binary external datasets (breast cancer recurrence, heart failure clinical records, and Pima Indians diabetes) are reported in Additional File 1 (S4) for continuity with prior external benchmarking practice.

All real-data experiments use pooled out-of-fold (OOF) evaluation under repeated strat-

ified cross-validation with fold-local preprocessing to prevent leakage, and baseline comparators are fitted with hyperparameters fixed in advance (Section 2.4) so that performance differences reflect model family rather than tuning budget [34]. Figure 1 summarizes the full evaluation workflow from data preparation through modeling and pooled OOF evaluation.

2.2 Data and multiclass label construction

To curate a three-class DILI risk stratification benchmark for BWORF development and evaluation, we employed the DILIrank 1.0 descriptor table, a curated small-molecule drug collection with FDA label-derived DILI annotations, as the source of predictor compounds [1, 3]. The primary multiclass benchmark contains $N = 574$ small-molecule compounds drawn from that DILIrank 1.0-derived descriptor table and represented by 777 Mold2 molecular descriptors [24]. Mold2 is an open-access software that calculates 777 chemical-physical-based one- and two-dimensional descriptors from chemical structure [24]. The supervised-learning target is constructed separately from curated LiverTox likelihood annotations joined to the descriptor table via PubChem compound identifier (CID). Raw LiverTox annotations expressed as letter-grade categories (A–E, with optional bracket modifiers such as [HD] for hepatotoxicity observed in high-dose settings) were first harmonized to their base category and then mapped deterministically to three ordered risk tiers: Most DILI concern (class 2: A, B), Likely DILI concern (class 1: C, D), and No DILI concern (class 0: E, E*). The special category E* was preserved during parsing and grouped with E in the final low-concern tier. One compound was excluded because a valid likelihood label was not available after harmonization, leaving $N = 574$. The resulting dataset is imbalanced, with 333 compounds (58.0%) in the No-concern tier, 101 (17.6%) in the Likely-concern tier, and 140 (24.4%) in the Most-concern tier (Table 1).

Table 1: Primary multiclass DILI benchmark summary. Predictor features are the 777 Mold2 molecular descriptor columns (D001–D777). Labels were harmonized from LiverTox likelihood categories via the mapping A/B→2, C/D→1, E/E*→0.

Class (meaning)	Count	Percent	LiverTox source
0 (No DILI concern)	333	58.0%	E, E*
1 (Likely DILI concern)	101	17.6%	C, D
2 (Most DILI concern)	140	24.4%	A, B
Total	**574**	**100.0%**	—

2.3 Bootstrap-Weighted Oblique Random Forest

BWORF extends the standard random forest [28, 29] along two axes: (i) axis-aligned splits are replaced with learned linear-combination splits following the oblique decision tree tradition [25–27], and (ii) uniform bootstrap sampling is replaced with a weighted bootstrap policy paired with weighted leaf-level probability estimation to improve minority-tier treatment under class imbalance [30, 31]. Both extensions are integrated into the

ensemble-generation process rather than applied as external corrections. Figure 2 shows the overall BWORF forest architecture, with weighted bootstrap sampling, oblique split induction at each internal node, and weighted leaf-level probability estimation integrated into a single training pipeline.

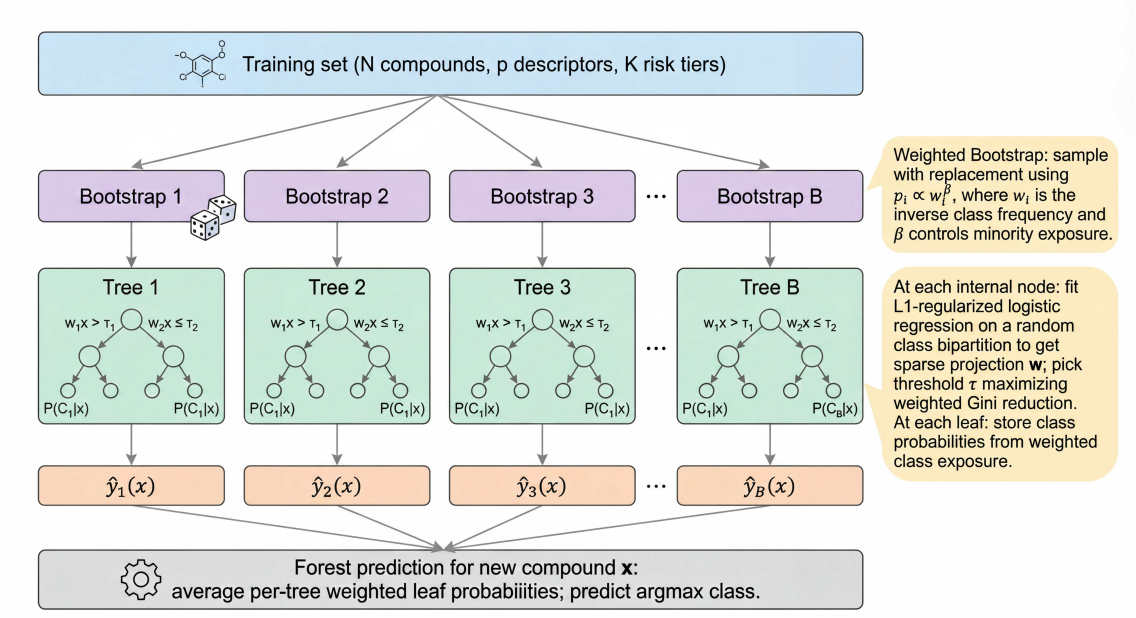


Figure 2: BWORF forest architecture. A training fold is resampled into B class-balanced bootstraps using temperature-scaled weights, and each bootstrap grows one oblique decision tree. At every internal node, an L_1-regularized logistic regression is fitted to a randomly sampled class bipartition to obtain a sparse projection $\mathbf{w}$; a threshold τ is then chosen along $\mathbf{w}^\top \mathbf{x}$ to maximize weighted Gini reduction on the original multiclass labels. Terminal nodes store weighted class probabilities rather than raw counts. The forest prediction averages per-tree weighted leaf probabilities and returns the argmax class, with ties broken by raw counts.

Let N denotes the total sample count for a dataset and n denotes a within-fold sample count; K is the number of classes and p is the number of predictors. For the BWORF ensemble, B is the number of trees, $d_{\max}$ the maximum tree depth, $n_{\min}$ the minimum leaf size, n_{tries} the number of class-bipartition split candidates evaluated per internal node, λ the L_1 regularization strength of the oblique projection, and T the bootstrap temperature scalar governing how strongly class-balanced weights concentrate sampling mass. Per-sample quantities use subscript i: $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector, $y_i \in \{0, \ldots, K-1\}$ the class label, w_i the class-balanced base weight, and p_i the bootstrap sampling probability. The oblique split is parameterized by a sparse projection $\mathbf{w} \in \mathbb{R}^p$ and a scalar threshold τ, giving the score $z_i = \mathbf{w}^\top \mathbf{x}_i$, where $\top$ denotes vector transpose. To distinguish the bootstrap temperature from the transpose operator, T is rendered as an upright exponent T when applied to weights (w_i^{T}).

Algorithm 1 BWORF training

Require: Training data (X, y) with n samples and K classes; number of trees B; max depth $d_{\max}$; minimum leaf size $n_{\min}$; number of split tries per node n_{tries}; L_1 strength λ; temperature T

Ensure: Trained forest $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_B\}$

Compute class-balanced base weights $w_i \leftarrow n/(K \cdot n_{\text{class}(i)})$
Form sampling probabilities $p_i \leftarrow w_i^{\mathrm{T}} / \sum_{j=1}^{n} w_j^{\mathrm{T}}$
for $b = 1, \ldots, B$ **do**
    Draw bootstrap indices $\mathcal{I}_b \sim \text{Multinomial}(n, p)$
    Grow tree $\mathcal{T}_b$ on $(X[\mathcal{I}_b], y[\mathcal{I}_b])$ using GrowNode$(\cdot)$
end for
return $\mathcal{F}$

function GrowNode(X_{node}, y_{node}, depth)
    if depth $\geq d_{\max}$ **or** $|y_{\text{node}}| < 2n_{\min}$ **or** y_{node} is pure **then**
        return leaf with $P(k) \propto \sum_{i:\, y_i = k} w_i$
    end if
    $(\mathbf{w}^*, \tau^*, G^*) \leftarrow (\varnothing, \varnothing, +\infty)$
    for $r = 1, \ldots, n_{\text{tries}}$ **do**
        Sample bipartition $\{C_0, C_1\}$ of classes present at node
        Form binary target $\tilde{y}_i \leftarrow \mathbb{I}(y_i \in C_1)$
        Fit L_1-regularized logistic regression with strength $\lambda \rightarrow$ sparse $\mathbf{w}^{(r)}$
        Compute projections $z_i \leftarrow (\mathbf{w}^{(r)})^\top \mathbf{x}_i$
        for each candidate threshold τ over sorted unique z_i **do**
            Compute weighted Gini G_τ on the original multiclass labels
            if $G_\tau < G^*$ and both children satisfy $n_{\min}$ **then**
                $(\mathbf{w}^*, \tau^*, G^*) \leftarrow (\mathbf{w}^{(r)}, \tau, G_\tau)$
            end if
        end for
    end for
    if $\mathbf{w}^* = \varnothing$ **then**
        return leaf with $P(k) \propto \sum_{i:\, y_i = k} w_i$
    end if
    Split: $L \leftarrow \{i : (\mathbf{w}^*)^\top \mathbf{x}_i \leq \tau^*\}$, $R \leftarrow$ complement
    return internal node with split $(\mathbf{w}^*, \tau^*)$ and children GrowNode(X_L, y_L, depth+1), GrowNode(X_R, y_R, depth+1)
end function

Oblique split learning with L_1-sparse projection. At each internal node, BWORF evaluates candidate oblique splits of the form

$$z = \mathbf{w}^\top \mathbf{x}, \qquad \text{send left if } z \leq \tau, \text{ right otherwise,} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^p$ is a sample's feature vector, $\mathbf{w} \in \mathbb{R}^p$ is a learned projection vector, and τ is a scalar threshold. In Equation (1), if $\mathbf{w}$ had only one nonzero entry the split would reduce to an ordinary single-feature threshold; in the oblique case, multiple entries of $\mathbf{w}$ may be nonzero so that the split depends on a weighted combination of predictors.

The projection $\mathbf{w}$ is learned through an L_1-regularized linear classifier, which promotes sparsity so that a relatively small subset of predictors drives each split [44]. For multiclass nodes, BWORF draws n_{tries} random bipartitions of the class set; for each bipartition, a sparse linear separator is fitted to the temporary binary target, candidate thresholds are enumerated over sorted unique projected scores $z_i = \mathbf{w}^\top \mathbf{x}_i$, and the $(\mathbf{w}, \tau)$ pair maximizing impurity reduction (subject to minimum-leaf constraints) is retained. Algorithm 1 below summarizes the full training procedure end-to-end; each sample i carries a class-balanced base weight w_i and is drawn into per-tree bootstraps with probability $p_i \propto w_i^{\mathrm{T}}$, where T is the bootstrap temperature. Algorithm 1 summarizes the full training procedure end-to-end, including bootstrap sampling, tree growth, and the node-level oblique split routine; the remainder of this subsection describes each component in detail. In the below, three important steps has been discussed to support above algorithm for DILI prediction.

- Each tree in the ensemble is trained on a bootstrap resample drawn with replacement from the fold training data [29]. In BWORF, each training sample i receives a class-balanced base weight w_i computed from inverse class frequency within the fold, and the sampling probability is

$$p_i = \frac{w_i^{\mathrm{T}}}{\sum_{j=1}^{n} w_j^{\mathrm{T}}}, \qquad T > 0, \tag{2}$$

  where the temperature parameter T in Equation (2) controls how strongly class-balanced weights concentrate sampling mass on higher-weight observations [30, 31]. Values $T > 1$ amplify weight differences and increase minority-class exposure; $0 < T < 1$ flatten differences toward uniform sampling. This procedure changes only the expected training exposure of different classes across bootstrap samples; it does not alter the observed labels or the prevalence against which the model is ultimately evaluated.

- To maintain consistency between weighted training and probabilistic prediction, terminal nodes store class probabilities computed from weighted class exposure rather than from raw counts. Forest-level probability estimates are obtained by averaging these weighted leaf probabilities across trees, and hard class prediction follows the same weighted probability representation, with ties resolved using raw leaf counts to avoid arbitrary dependence on class index ordering.

- BWORF can optionally apply fold-safe univariate mutual information (MI) screening to reduce dimensionality before oblique split learning [45, 46]. For each cross-validation fold, MI scores are computed on the training partition only, a top-k subset of features is retained, and the same feature indices are applied to the held-out fold. When MI screening is disabled, the full feature matrix is passed forward unchanged. We report results for both the MI-screened variant (BWORF + MI) and the pass-through variant so that the weighted-bootstrap contribution can be assessed separately from univariate feature screening.

2.4 Baseline model panel and hyperparameter configuration

Baseline comparators comprise logistic regression (LR) [47], random forest (RF) [28], XGBoost [48], support vector machine with RBF kernel (SVM) [47], Gaussian naïve Bayes, and an oblique random forest (ORF) variant that shares BWORF's oblique split learner with weighted bootstrap disabled. The ORF comparator isolates the contribution of weighted bootstrap sampling from oblique split learning itself: the gap between ORF and BWORF is attributable to the weighted-bootstrap and weighted-leaf components, whereas the gap between RF and ORF is attributable to replacing axis-aligned splits with L_1-sparse oblique splits.

Hyperparameters were fixed in advance for reproducibility. All unspecified parameters follow library defaults (scikit-learn 1.3 for LR, RF, SVM, NB; XGBoost 1.7 for gradient-boosted trees). Table 2 reports the explicitly set hyperparameters for each model in the main multiclass benchmark panel. The two BWORF variants reported differ only in whether fold-local MI screening is applied: BWORF uses the full 777-feature matrix, whereas BWORF with MI retains the top 50 features by training-fold MI score before oblique split learning.

Table 2: Explicitly set hyperparameters for the primary multiclass DILI benchmark and the BRFSS 2022 external benchmark.

Model	Explicitly set hyperparameters
LR	`solver=lbfgs`; `C=1.0`; `max_iter=4000`; standardized inputs
RF	`n_estimators=100`; `max_depth=6`; `min_samples_split=10`; `min_samples_leaf=1`; `max_features=sqrt`
SVM (RBF)	`kernel=rbf`; `C=1.0`; `gamma=scale`; `probability=True`; standardized inputs
XGBoost	`n_estimators=100`; `max_depth=6`; `learning_rate=0.1`; `subsample=1.0`; `colsample_bytree=1.0`; `objective=multi:softprob`; `eval_metric=mlogloss`; `num_class=3`
Naïve Bayes	`GaussianNB` with default priors
ORF	`n_estimators=100`; `max_depth=6`; `min_samples_split=10`; `min_samples_leaf=1`; `l1_strength=0.2`; `n_tries=10`; `weighted_bootstrap=False`
BWORF	`n_estimators=100`; `max_depth=6`; `min_samples_split=10`; `min_samples_leaf=1`; `l1_strength=0.2`; `n_tries=10`; `weighted_bootstrap=True`; `bootstrap_temperature=1.0`; full feature pass-through
BWORF + MI	Same as BWORF with `mi_top_k=50` fold-local MI screening
BRFSS 2022 external benchmark panel (NB and SVM omitted at this scale)	
LR	`solver=lbfgs`; `class_weight=balanced`; standardized inputs
RF	`n_estimators=100`; `class_weight=balanced`
XGBoost	`n_estimators=100`; `scale_pos_weight` set to inverse positive prevalence
ORF	`n_estimators=100`; `max_depth=5`; `min_samples_split=50`; `min_samples_leaf=5`; `l1_strength=1.0`; `n_tries=2`; `weighted_bootstrap=False`
BWORF	Same as ORF with `weighted_bootstrap=True`; `bootstrap_temperature=1.0`

Note: All unspecified parameters follow library defaults in Table 2. The ORF comparator shares the oblique split learner and tree hyperparameters with BWORF but disables the weighted bootstrap; the two BWORF variants differ only in whether fold-local MI screening is enabled.

To evaluate BWORF on the BRFSS 2022 external benchmark introduced in Section 2.1,

we configured a five-model panel reflecting the dataset's larger scale and lower feature dimensionality, with hyperparameters retuned from the primary DILI configuration as shown in Table 2. To make full-N BWORF fits feasible at this scale, we developed a byte-equivalent parallel implementation that preserves bit-exact identity with the serial reference and ran the five-seed ten-fold protocol as a 50-task array on the ASU Sol HPC cluster [49].

2.5 Cross-validation protocol and evaluation

All experiments on the primary benchmark use stratified 10-fold cross-validation repeated across five fixed seeds, with predictions pooled across held-out folds for OOF evaluation [34]. Fold-local preprocessing (standardization and, where enabled, MI screening) is fitted only on training-fold data and then applied to the held-out fold. Baseline models are reported as mean $\pm$ standard deviation (SD) across seeds using seed-level pooled OOF predictions. For BWORF, we report a representative best-accuracy coherent operating point alongside the seed-averaged baselines; the corresponding seed-averaged BWORF distribution is provided in the Supplementary Information so that both reporting conventions are available for inspection.

Ranking-based discrimination is summarized using macro-averaged one-vs-rest area under the receiver operating characteristic curve (AUROC) [50] and area under the precision-recall curve (AUPRC) [32, 33]. Operational summaries (accuracy, macro F1, balanced accuracy, Matthews correlation coefficient (MCC), log loss) are computed under the default decision rule (argmax for multiclass) and interpreted descriptively. For the multiclass setting, balanced accuracy is particularly informative because the three risk tiers are unequally prevalent and overall accuracy can conceal minority-tier failure.

For the BRFSS 2022 external benchmark, a two-stage evaluation protocol is used because the dataset's size makes a held-out confirmatory test practical and consistent with reporting standards for clinical prediction models [51, 52]. The cleaned dataset is partitioned once into a modeling set and a held-out test set under stratified sampling using a fixed preparation seed drawn independently of the cross-validation seeds, so that the held-out test set is uncovered only at the end of the development process [34, 52]. Stratified 10-fold cross-validation repeated across five fixed seeds is then run on the modeling set and reported analogously to the primary DILI benchmark. After cross-validation, each model is refit on the entire modeling set, once per cross-validation seed, and scored on the held-out test set; the held-out scores are reported as the primary external evaluation. This protocol mirrors the temporal-split refit-and-evaluate strategy used by Li et al. [7] for DeepDILI, with the BRFSS modeling and held-out partitions serving the same role as their pre-1997 training and post-1997 test sets.

2.6 Local interpretability of DILI prediction

To support case-level inspection of individual BWORF predictions, we generated local explanations using Local Interpretable Model-agnostic Explanations (LIME) [36, 40]. LIME approximates the behavior of the trained classifier in the neighborhood of a specific prediction by fitting a sparse interpretable surrogate model and reporting the per-feature contributions to the local decision. For each explanation, the locked multiclass BWORF + MI artifact (seed 43) was used as the black-box predictor, and LIME was applied to individual compounds from the pooled OOF predictions. We selected a representative set of nine case studies spanning three confidence regimes, namely high-confidence correct, borderline correct, and notable error, to illustrate how local support patterns vary across prediction quality and risk-tier assignment [37, 38]. These explanations are diagnostic summaries of model behavior, not causal attributions of hepatotoxicity, and should be interpreted accordingly [36].

3 Results

The following six subsections report BWORF's performance on the primary multiclass DILI benchmark (section 3.1), mechanistic validation in a controlled simulation (section 3.2), generalization to external biomedical tabular benchmarks (section 3.3), and three interpretability analyses covering prediction confidence, feature-level importance, and case-level local explanations (sections 3.4–3.6).

3.1 Multiclass DILI risk stratification

On the primary multiclass DILI risk stratification benchmark, under stratified 10-fold cross-validation with pooled OOF evaluation, BWORF achieved the highest balanced accuracy (0.433) and a macro F1 (0.422) tied with the strongest baseline (XGBoost) after rounding, while shifting the confusion-matrix profile toward more even treatment of the intermediate and most-concern risk tiers (Table 3). The selected BWORF operating point (BWORF + MI, seed 43) achieved an accuracy of 0.476 (95% Wilson CI [0.435, 0.516]). Among the aggregated baselines, XGBoost achieved the highest mean accuracy (0.601 ± 0.013), macro AUROC (0.649 ± 0.015), and macro AUPRC (0.458 ± 0.015), while RF retained the lowest mean log loss (0.921 ± 0.006). BWORF allocates more predictive capacity to the Likely-concern and Most-concern tiers, consistent with the weighted bootstrap design redistributing training exposure toward under-represented classes [30, 31]. Figures 3 and 4 visualize the class-wise one-vs-rest discrimination profiles and the cross-seed distribution of pooled OOF performance, respectively.

Table 3: Internal validation performance on the primary multiclass DILI benchmark under stratified 10-fold cross-validation repeated across five fixed seeds. Baseline rows are mean ± standard deviation (SD) across seeds using seed-level pooled OOF predictions. The BWORF row reports the coherent best-accuracy operating point (BWORF + MI, seed 43), with a 95% Wilson confidence interval for the accuracy entry. The corresponding seed-averaged BWORF distribution is provided in the Supplementary Information. Bal. acc.: balanced accuracy; MCC: Matthews correlation coefficient; AUROC/AUPRC: macro one-vs-rest.

Model	Accuracy	Bal. acc.	Macro F1	MCC	Log loss	AUROC	AUPRC
LR	0.482 ± 0.010	0.396 ± 0.004	0.396 ± 0.005	0.106 ± 0.009	2.531 ± 0.033	0.572 ± 0.002	0.392 ± 0.005
RF	0.587 ± 0.007	0.372 ± 0.005	0.330 ± 0.005	0.130 ± 0.019	0.921 ± 0.006	0.637 ± 0.007	0.438 ± 0.010
SVM (RBF)	0.582 ± 0.004	0.342 ± 0.003	0.269 ± 0.004	0.058 ± 0.028	0.956 ± 0.004	0.557 ± 0.008	0.380 ± 0.007
XGBoost	0.601 ± 0.013	0.430 ± 0.020	0.422 ± 0.026	0.219 ± 0.032	1.112 ± 0.031	0.649 ± 0.015	0.458 ± 0.015
Naïve Bayes	0.277 ± 0.002	0.334 ± 0.002	0.230 ± 0.003	0.008 ± 0.003	8.711 ± 0.033	0.498 ± 0.002	0.336 ± 0.001
ORF	0.528 ± 0.013	0.347 ± 0.011	0.328 ± 0.013	0.039 ± 0.027	1.153 ± 0.030	0.574 ± 0.004	0.396 ± 0.007
BWORF	**0.476** [0.435, 0.516]	**0.433**	**0.422**	0.164	1.017	0.595	0.415

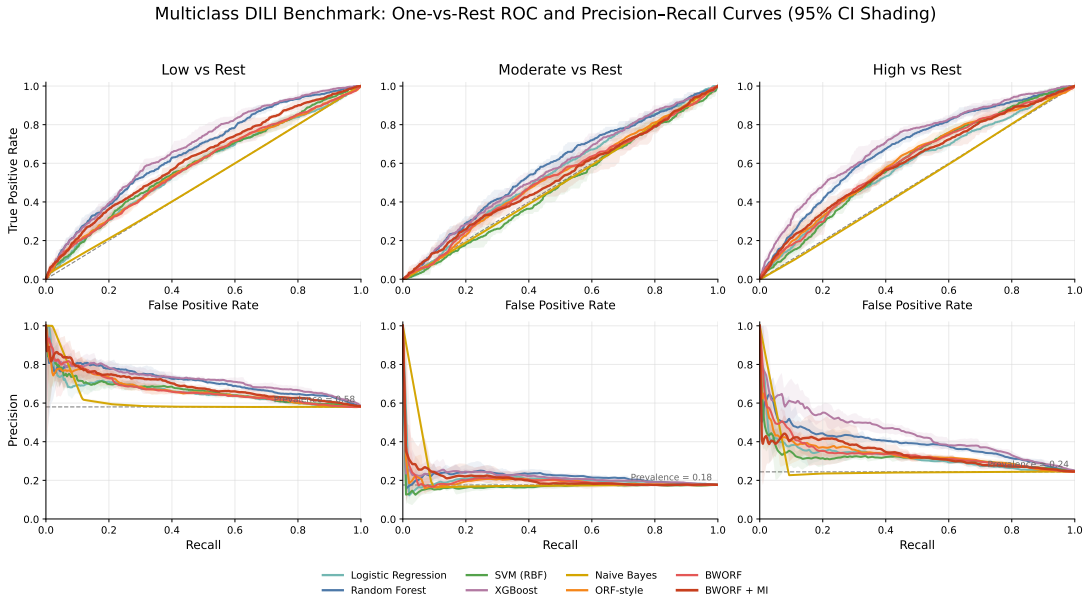


Figure 3: One-vs-rest ROC and precision–recall curves for the primary multiclass DILI benchmark. Curves were computed from pooled out-of-fold predictions under stratified 10-fold cross-validation repeated across five fixed seeds. Shaded bands indicate seed-level 95% interval envelopes. Scalar macro AUROC and AUPRC summaries are reported in Table 3.

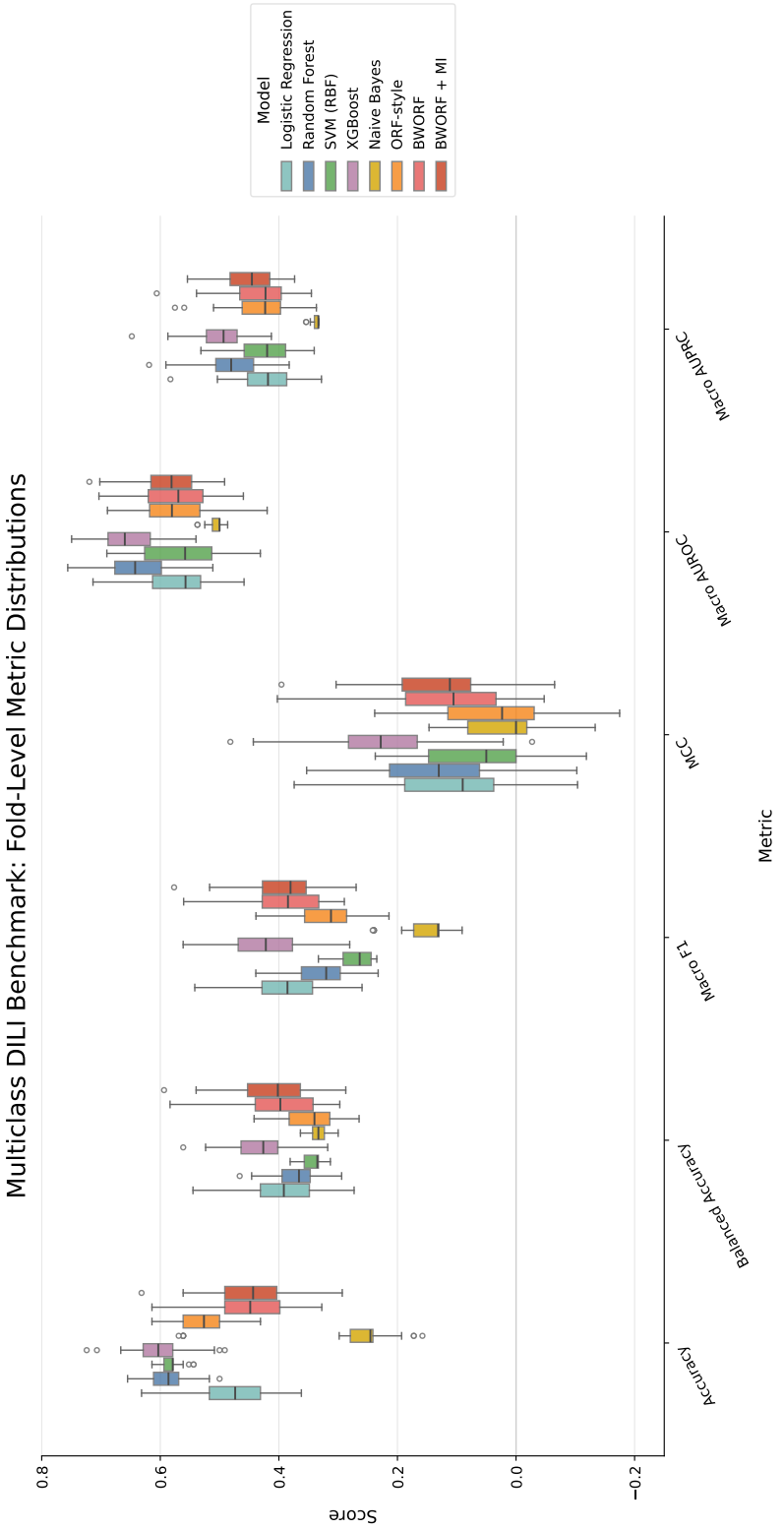


Figure 4: Seed-level pooled out-of-fold performance distributions across all evaluated models. Metrics shown are Accuracy, Balanced Accuracy, Macro F1, MCC, Macro AUROC, Macro AUPRC. Boxes indicate the interquartile range with median line; whiskers extend to 1.5× the interquartile range. BWORF with and without mutual-information screening are displayed separately.

3.2 Simulation study

To isolate the mechanistic contributions of oblique split learning and weighted bootstrap sampling from confounds present in the real-data benchmark, we evaluated BWORF and its two forest baselines on a simulation grid with known causal structure. The simulation scenario grid varied sample size $N \in \{500, 2500\}$, linkage-disequilibrium correlation $\rho \in \{0.2, 0.8\}$, number of causal interaction pairs $m \in \{2, 5, 10\}$, heritability $h^2 \in [0.025, 0.200]$, and the linear-to-interaction signal ratio, with five replicates per scenario and five fixed seeds per method. Three methods were compared: the axis-aligned RF baseline, the unweighted ORF comparator, and the proposed BWORF with weighted bootstrap enabled.

Table 4 reports predictive performance across the grid. RF and ORF remained close in grid-averaged AUROC (0.527 and 0.524) and grid-averaged Brier score (0.251 for both), consistent with oblique split learning alone providing little discrimination advantage when prevalence and signal structure are held fixed. The BWORF row reports the accuracy-selected coherent operating point across the grid, which occurred under a moderate-sample, high-signal regime ($N = 500$, $m = 2$, $h^2 = 0.15$, $\rho = 0.8$, signal ratio $= \infty$(pure-linear signal ratio)) and achieved AUROC 0.717 and Brier score 0.221. In the interaction-heavy setting (signal ratio = 0.5) at $h^2 = 0.15$, per-scenario AUROC differences between BWORF and ORF ranged from -0.016 to 0.046 with a mean difference of 0.014, indicating that the BWORF advantage is concentrated in regimes where discriminative signal is carried by combinations of correlated predictors.

Table 4: Predictive performance in the simulation study. RF and ORF are reported as grid-averaged summaries over the full scenario grid; BWORF is reported at its accuracy-selected coherent operating point.

Model	AUROC	Brier score
RF	0.527	0.251
ORF	0.524	0.251
BWORF	0.717	0.221

Because the simulator records the true causal predictors, permutation-importance rankings can be compared directly against ground truth. On a grid-averaged basis, RF and ORF recovered causal variants relatively late (mean rank 433.1 and 443.6, respectively over 1000 candidates), with recall@20 values of 0.130 and 0.100. Under the same accuracy-selected BWORF configuration, causal recovery was substantially stronger: mean rank of the causal variants was 45.8 and recall@20 reached 0.500 (Table 5). The near-tenfold improvement in causal-variant ranking accompanies the AUROC gain reported above, suggesting that weighted bootstrap sampling not only rescales the decision threshold but also improves the exposure of individual trees to the informative minority-class observations that drive oblique split learning toward true causal structure.

Table 5: Causal-signal recovery in the simulation study based on permutation-importance rankings. Lower mean rank indicates earlier recovery of planted causal variants; recall@20 reports the fraction of causal variants appearing in the top 20 ranked features.

Model	Mean rank of causal variants	Recall for top 20 features
RF	433.1	0.130
ORF	443.6	0.100
BWORF	45.8	0.500

3.3 External validation

To test whether the BWORF profile generalizes outside the DILI domain and outside the Mold2 descriptor representation, we evaluated the method on two complementary external regimes: a multiclass in-regime test on the thyroid disease benchmark, which is the closest analog to the primary three-class DILI task, and a population-scale binary test on the BRFSS 2022 heart-attack benchmark, which probes whether BWORF's class-balancing components retain measurable leverage at modern data scale [42, 53]. Three smaller binary external datasets (breast cancer recurrence, heart failure clinical records, Pima Indians diabetes) were also evaluated for continuity with prior practice and are reported in Additional File 1.

3.3.1 Multiclass in-regime transfer: thyroid disease

On the multiclass thyroid task, BWORF achieved the highest macro F1 in the comparison panel (0.975), exceeding the unweighted ORF comparator (0.963) and the axis-aligned RF baseline (0.940) (Table 6). The accuracy-selected BWORF operating point reached 0.981 (95% Wilson confidence interval (CI) [0.953, 0.993]). All three forest variants achieved macro AUROC values above 0.995, indicating that ranking-based discrimination is near the ceiling on this dataset and the meaningful comparison is on macro F1, where class-balancing leverage is most visible. This profile mirrors the minority-tier-forward signature observed on the primary DILI benchmark and is consistent with the class-balancing design of the method [30, 31].

Table 6: External validation on the multiclass thyroid disease benchmark. RF and ORF entries are seed-averaged pooled out-of-fold summaries across five fixed seeds; the BWORF entry reports the accuracy-selected coherent operating point. F1 denotes macro F1.

Model	AUROC	Macro F1
RF	0.998	0.940
ORF	0.995	0.963
BWORF	0.995	0.975

3.3.2 Population-scale severe imbalance: BRFSS 2022 heart attack

On the BRFSS 2022 held-out test set (N = 49,198, refit on the modeling set of N = 196,788 per Section 2.5), BWORF achieved the highest area under the precision-recall curve (AUPRC, 0.421), the highest area under the receiver operating characteristic curve (AUROC, 0.893), and the highest balanced accuracy (0.810) among the five evaluated models (Table 7). Logistic regression and XGBoost achieved the leading values on Matthews correlation coefficient (MCC) and positive-class F1 respectively, with BWORF within a small margin on each. Under 5.46% positive prevalence, AUPRC is the most discriminating ranking-based metric, and the BWORF AUPRC margin of 2.7% over the next-best model is the largest in the panel [32, 33].

Table 7: Held-out test set performance on the BRFSS 2022 heart-attack benchmark. Models were refit on the modeling set per cross-validation seed and scored on the held-out test set; entries are seed-averaged across five fixed seeds. Bal. acc.: balanced accuracy; MCC: Matthews correlation coefficient; $F1_+$: F1 on the positive class.

Model	AUROC	AUPRC	Bal. acc.	MCC	$F1_+$	Log loss
LR	0.892	0.410	0.806	0.348	0.334	0.423
RF	0.888	0.400	0.801	0.328	0.311	0.469
XGBoost	0.886	0.403	0.802	0.346	0.335	0.381
ORF	0.885	0.387	0.798	0.313	0.292	0.482
BWORF	**0.893**	**0.421**	**0.810**	0.337	0.316	0.410

A scale-dependent comparison clarifies where the BWORF advantage arises. At a 4,000-row subsample of the same cleaned dataset, BWORF cross-validation AUROC is 0.852, in the middle of the panel and behind RF (0.878), ORF (0.873), and LR (0.872). At full modeling-set scale (N = 196,788), BWORF rises to 0.889 (tied with LR). The AUROC gain for BWORF between the two scales is +0.041, the largest in the panel together with XGBoost (+0.041), well above the gains observed for LR (+0.017), RF (+0.005), and ORF (+0.008).

3.4 Prediction confidence analysis

Applying the confidence-binning approach introduced by Hong et al. [6], we assessed whether BWORF predictions with higher model confidence are more reliable by binning the 574 pooled OOF predictions from the BWORF + MI (seed 43) operating point into probability bins defined by the maximum predicted class probability (Figure 5). Overall accuracy increased monotonically with confidence, rising from 0.295 in the lowest-confidence bin ([0.3, 0.4)) to 0.461, 0.633, and 0.735 in successively higher bins, and reaching 1.000 in the sparsely populated [0.7, 0.8) bin (n = 4), demonstrating that the model's probabilistic output provides a meaningful signal for prediction triage. No-concern recall showed a corresponding increase from 0.152 to 0.821 across the same range, consistent with the majority class becoming easier to identify at higher confidence levels.

The majority of predictions (334 of 574, 58.2%) fell in the [0.4, 0.5) bin, reflecting the inherent difficulty of multiclass DILI risk stratification on a moderate-sized benchmark with

three imbalanced tiers. Among the 128 predictions (22.3%) with confidence ≥ 0.5, overall accuracy ranged from 0.633 to 1.000, indicating that a confidence-based triage threshold could identify a reliable prediction subset for downstream use. Likely-concern recall was highest in the lowest-confidence bin (0.500) and in the [0.6, 0.7) bin (0.667), suggesting that intermediate-tier compounds are identified across a broader confidence range rather than concentrating at high confidence. These results support the use of prediction confidence as an operational parameter for interpreting BWORF risk assignments in practice [6].

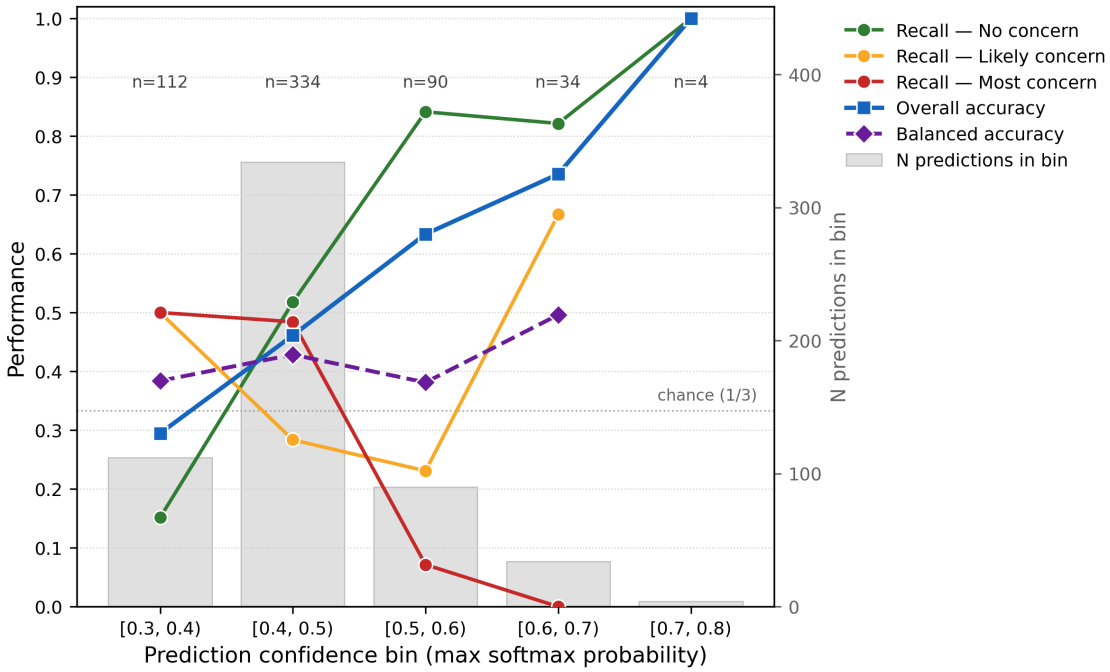


Figure 5: Prediction confidence analysis for the BWORF + MI (seed 43) operating point. Predictions from pooled out-of-fold evaluation ($N = 574$) were binned by maximum predicted class probability (confidence). Lines show per-class recall rates and overall accuracy; gray bars indicate the number of predictions in each bin. Overall accuracy increases monotonically with confidence, supporting the use of prediction confidence for triage of model outputs.

3.5 Feature importance and chemical interpretation

The descriptors most frequently retained by fold-local MI screening across the 50 held-out folds provide context about which molecular properties the BWORF + MI pipeline treats as most informative (Figure 6). Among the top-ranked Mold2 descriptors, several encode molecular size and shape properties, including mean atomic van der Waals volume and average molecular weight, while others capture electronic and physicochemical characteristics such as atomic polarizability and Sanderson electronegativity [24]. This pattern is consistent with prior DILI prediction studies using the same descriptor family: Hong et al. [6] identified molecular size, shape, and physicochemical properties as the most informative Mold2 features for DILI risk discrimination in their 3-class Decision Forest model, and Li et

al. [7] reported overlapping descriptors among the top contributors to DeepDILI. Molecular size and shape are established determinants of drug transport and metabolism [6], while polarizability and electronegativity relate to off-target binding capacity relevant to hepatotoxic potential [6]. The convergence of informative descriptors across independent models and datasets supports the view that BWORF is learning chemically meaningful structure rather than fitting noise.

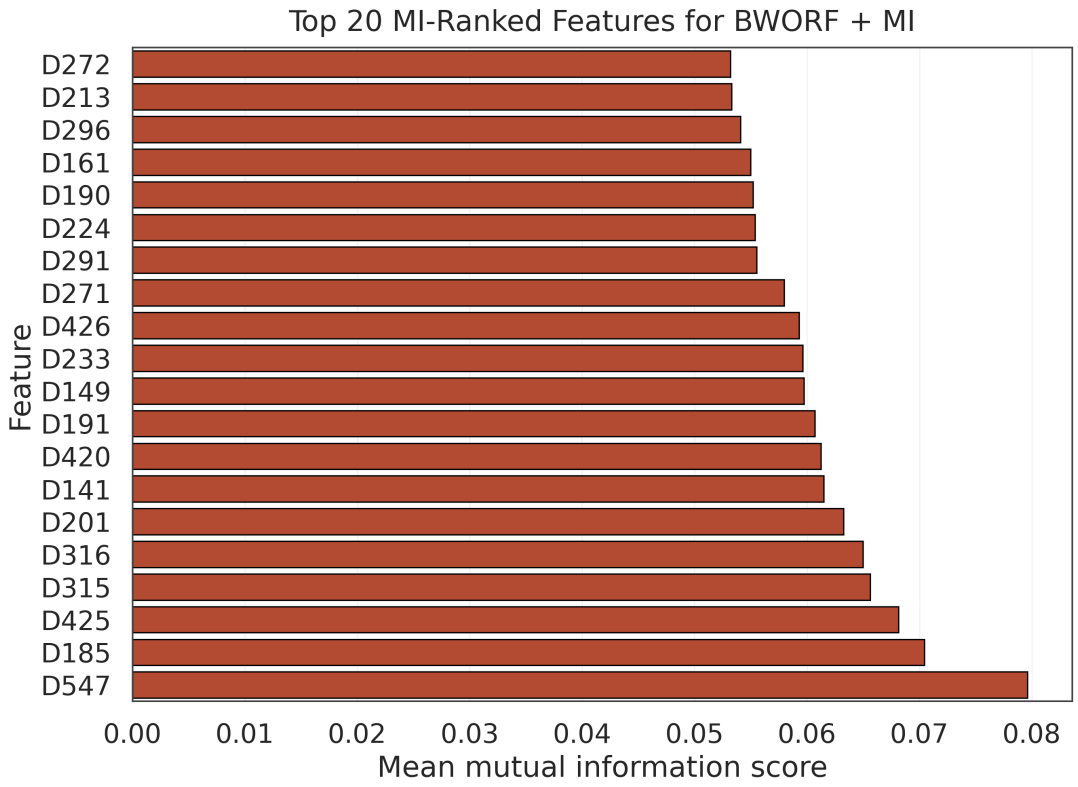


Figure 6: Top-20 Mold2 descriptors by fold-local mutual information screening frequency in the BWORF + MI pipeline. Bars show how often each descriptor was retained across the 50 held-out folds of stratified 10-fold cross-validation repeated across five seeds.

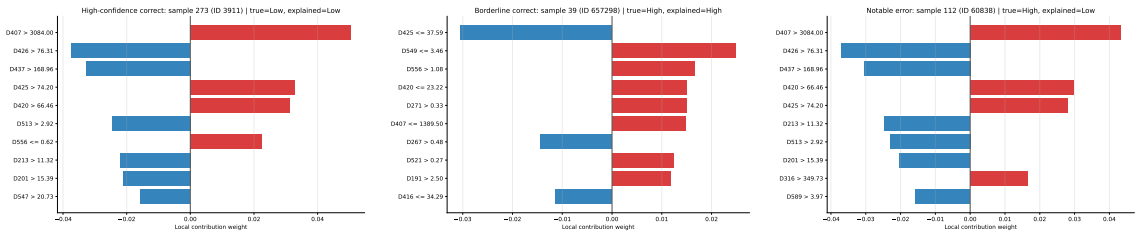
3.6 Local interpretability

To examine whether BWORF predictions can be meaningfully inspected at the individual-compound level, we generated LIME explanations for a representative set of nine compounds from the pooled OOF predictions of the BWORF + MI (seed 43) operating point, spanning three confidence regimes: high-confidence correct, borderline correct, and notable error (Figure 7). High-confidence correct predictions display concentrated local support, with a small number of descriptors contributing strongly in the direction of the assigned tier and minimal opposing contributions. This pattern is consistent with a decisive class assignment in which the local decision surface is dominated by a few informative features. Borderline correct predictions show narrower decision margins and more mixed-sign feature contributions, reflecting operation near the boundary between adjacent risk tiers where the

classifier's local support is split across competing cues. Notable errors are especially diagnostic: they show competing local contributions that pull the prediction toward a different tier, reinforcing the broader finding that BWORF's class-balancing behavior produces local decisions that can reflect tension between neighboring risk tiers rather than a single dominant signal [37, 40].

These explanations demonstrate that the finalized BWORF artifact is sufficiently inspectable to distinguish decisive predictions, boundary cases, and diagnostically informative failures at the individual-compound level, supporting the auditability requirements emphasized in recent DILI and toxicity-prediction literature [22, 36, 39]. Additional LIME case studies are provided in the Supplementary Information.

A natural follow-on question is how the descriptors surfaced by these local explanations relate to the global feature importance reported earlier. The descriptors that LIME surfaces at the case level provide a useful comparison against the global MI ranking reported in Section 3.5. Across the nine-case LIME set, 28 distinct descriptors appeared at least once in the top ten local contributions, and 14 of these also appeared in the MI top-20 used to construct the BWORF + MI pipeline, giving a 70% agreement rate between global relevance and case-level salience. Descriptor D425, which is ranked first by MI, appeared in the top ten local contributions of every one of the nine case studies, providing the strongest point of convergence between the two views. Other descriptors that recurred frequently in the LIME panels, including D407, D420, D426, D547, and D271, are also within the MI top-20. Fourteen LIME-surfaced descriptors, including D213, D437, and D556, were not in the MI top-20 but carried visible local weight in individual cases; these are features whose contribution depends on the neighborhood of a specific compound rather than on marginal association with the risk label. The two views therefore act as complementary rather than redundant summaries: MI identifies descriptors that are broadly informative for the risk-stratification task, while LIME identifies descriptors that matter for particular compounds, and the overlap on the highest-MI descriptors supports the interpretation that the screened feature set captures the descriptors the model most consistently relies on.



(a) High-confidence correct (b) Borderline correct (c) Notable error

Figure 7: Representative LIME explanations for three compounds from the BWORF + MI (seed 43) pooled out-of-fold predictions. (a) A high-confidence correct prediction showing concentrated local support from a small number of descriptors. (b) A borderline correct prediction with narrower decision margin and mixed-sign contributions. (c) A notable error showing competing local contributions pulling toward an adjacent risk tier.

4 Discussion

On the primary multiclass DILI benchmark, BWORF achieved the highest balanced accuracy (0.433) and a macro F1 (0.422) tied with the strongest baseline after rounding, and was within the same performance band as the leading tree-ensemble baselines on overall accuracy, macro AUROC, and macro AUPRC [3, 4]. The substantive pattern behind these summaries is a redistribution of predictive capacity: higher per-class recall on the intermediate (Likely concern) and highest (Most concern) risk tiers than any baseline in the comparison panel, with predictive emphasis shifted away from the already-well-predicted majority No-concern tier. This trade-off is the intended behavior of the weighted-bootstrap and weighted-leaf components, which together increase minority-class exposure during training and carry that exposure through to the probabilistic output, rather than concentrating learning on the majority class that dominates most curated DILI label distributions [29–31]. In a preclinical screening setting where identifying compounds with elevated hepatotoxic risk is the operative evaluation criterion, a method that prioritizes minority-tier recall while remaining competitive on aggregate metrics aligns with the practical requirements of DILI risk assessment [1, 2, 5].

The controlled simulation study supports this class-balancing interpretation by providing a regime in which causal structure is specified by design and class prevalence can be varied independently of signal strength. Under moderate-sample, interaction-heavy conditions, BWORF's grid-averaged discrimination advantage over the axis-aligned RF baseline and the unweighted ORF comparator was clearest; the two forest baselines remained close in grid-averaged AUROC, consistent with oblique split learning alone providing little discrimination benefit when prevalence and signal structure are held fixed [26–28]. More striking is the causal-recovery profile: under the accuracy-selected operating point, BWORF recovered planted causal predictors with a mean rank approximately an order of magnitude earlier than either RF or ORF (Table 5), and its recall@20 of causal variants was multiple-fold higher. The near-tenfold improvement in causal-variant ranking accompanies the AUROC gain reported above, suggesting that weighted bootstrap sampling does not merely rescale the decision threshold but improves the exposure of individual trees to the informative minority-class observations that drive oblique split learning toward true causal structure [25, 30, 31]. Evidence from the external validation suite supports the class-balancing interpretation along two complementary axes. On the multiclass thyroid benchmark, the closest in-regime analog to the primary DILI task, BWORF led the forest-variant panel on macro F1 while all three forest variants performed near the ceiling on macro AUROC (Table 6). The proximity to the AUROC ceiling implies that ranking-based discrimination saturates on this dataset, leaving macro F1 as the metric where class-balancing leverage is most visible; the BWORF advantage there reproduces the minority-tier-forward signature observed on the primary DILI benchmark [32, 33]. On the three smaller binary external datasets where the multiclass imbalance structure is absent, BWORF produced operating points competitive with a matched RF baseline, consistent with its class-balancing leverage

being most visible in the imbalanced multiclass regime for which it was designed [30, 31].

The BRFSS 2022 heart-attack benchmark probes a different generalization axis: severe binary imbalance at population scale. On the held-out test set, BWORF achieved the highest AUPRC, AUROC, and balanced accuracy of the five evaluated models (Table 7). The AUPRC ranking is the most informative under 5.46% positive prevalence, where AUROC saturation and threshold-derived metrics carry less information than ranking on the positives [32, 33]. The scale-dependent comparison between subsample and full-N cross-validation is methodologically informative: BWORF's AUROC gain across the two scales is the largest in the panel together with XGBoost, while the other forest comparators gain at most one quarter as much. This pattern is consistent with weighted bootstrap and oblique split learning needing sufficient minority-class density per leaf to express their advantage; at small N the minority class is too sparse for these components to fully contribute, and at large N they can [26, 27, 30, 31, 53]. Taken together, the simulation, the multiclass thyroid result, the population-scale BRFSS result, and the scale-dependent comparison indicate that BWORF's profile is a function of task regime and data scale rather than an artifact of the DILI dataset in isolation.

Beyond raw discrimination, BWORF's probabilistic output carries a reliability signal with practical implications for screening pipelines. Overall accuracy rose monotonically across the well-populated confidence bins, from 0.295 in the lowest-confidence bin to 0.461, 0.633, and 0.735 in successively higher bins, with further improvement at the highest confidence level on a smaller subset of predictions (Figure 5). The 128 predictions at confidence at least 0.5, corresponding to 22.3% of the benchmark, exhibited accuracy of 0.633 or above, indicating that a confidence-based operating threshold can isolate a subset of BWORF outputs substantially more reliable than aggregate accuracy suggests [6]. This pattern mirrors the confidence-binning analysis reported by Hong et al. for Decision Forest models on a related 3-class DILI benchmark [6] and is consistent with broader work on probabilistic calibration for classifier deployment under class imbalance [31, 32]. In practice, a screening pipeline built on BWORF could prioritize high-confidence predictions for immediate downstream evaluation while routing low-confidence compounds to additional experimental testing [2, 4, 5].

Turning from model behavior to the feature-level picture, the descriptors most frequently retained by fold-local MI screening converge with prior findings in the DILI prediction literature. Hong et al. [6] identified molecular size, shape, and physicochemical properties as the most informative Mold2 descriptors for DILI risk discrimination in their 3-class Decision Forest model, and Li et al. [7] reported overlapping descriptor families among the top contributors to DeepDILI's model-level representation. The appearance of the same descriptor families in BWORF's screened feature set (Figure 6), despite differences in split geometry, bootstrap policy, and label construction, is consistent with the interpretation that the method is capturing chemically meaningful structure associated with hepatotoxic risk rather than fitting dataset-specific noise [3, 22–24].

Local case-level analysis through LIME adds a complementary view to the global

feature-importance picture. The nine representative case studies distinguish decisive predictions (concentrated local support from a small number of descriptors) from boundary cases (narrower decision margins, mixed-sign contributions) and from diagnostically informative failures (competing contributions pulling toward adjacent risk tiers) [37, 38, 40]. This three-regime pattern indicates that the finalized BWORF artifact is sufficiently inspectable to support per-compound auditing consistent with recent interpretable DILI work [22, 23, 39]. More concretely, 14 of the 20 descriptors retained by MI screening also appeared in the top ten local contributions of at least one of the nine case studies, and one descriptor (D425, ranked first by MI) recurred in every case. The convergence between a univariate training-time screen and post-hoc case-level attribution is a useful diagnostic, because it indicates that global feature selection and local decision support are not drawing on disjoint evidence, which is a known failure mode for pipelines that combine screening with downstream interpretability tooling [36, 45, 46].

Taken together, these findings place BWORF within an active line of tree-ensemble work on DILI risk prediction. Hong et al. [6] established that a 3-class Decision Forest on Mold2 descriptors provides higher-resolution risk discrimination than a binary formulation on DILIrank-derived data and argued for prediction confidence as an operational interpretation parameter, a position the present analysis reinforces [3]. Li et al. [7] showed that a deep meta-classifier built on model-level representations can improve discrimination beyond individual conventional classifiers, and more recent interpretable approaches including InterDILI [22] and MORF [23] have advanced the integration of prediction quality with per-compound transparency. BWORF enters this trajectory from a different direction: rather than adding representational depth, it targets the imbalance failure mode directly at the ensemble level by coupling L_1-sparse oblique split learning [25–27] with weighted bootstrap sampling and weighted leaf-level probability estimation [29–31], a design compatible with recent oblique-forest work in adjacent biomedical modeling contexts [41]. The net contribution is a forest whose design priorities are aligned with the risk tiers that safety screening is meant to catch: minority-tier recall is promoted as a first-class evaluation criterion at the ensemble level, rather than recovered post-hoc through threshold adjustment or external reweighting [2, 4, 5, 8].

5 Conclusion

DILI remains a challenging prediction task due to its multifactorial nature, involving complex interactions among chemical structure, biological pathways, and patient-specific variability. This study introduced the BWORF, a tree ensemble that couples L_1-sparse oblique split learning with temperature-scaled weighted bootstrap sampling and weighted leaf-level probability estimation, and evaluated it on a three-class DILI risk stratification task against conventional and oblique tree-ensemble baselines [27, 28, 30]. On the primary benchmark, BWORF achieved the highest balanced accuracy (0.433) and a macro F1 (0.422) tied with the strongest baseline, with higher per-class recall on the intermediate and Most-

concern risk tiers than any baseline and comparable performance on overall accuracy and threshold-free discrimination [3]. Supporting evidence from a controlled simulation study, a multiclass thyroid external benchmark, and a population-scale binary BRFSS heart-attack benchmark indicates that this class-balancing profile is a function of task regime and data scale rather than an artifact of the DILI dataset in isolation [31]. Several limitations should be noted. All DILI analyzes used Mold2 descriptors alone, which prevents separating algorithmic contributions from representational ones [7, 24]. The multiclass endpoint was constructed by collapsing LiverTox likelihood categories into three ordered tiers, which impose hard boundaries on labels derived from heterogeneous evidence, and compounds near category boundaries may be intrinsically difficult to classify [2, 4]. The dataset is moderate in size ($N = 574$) and imbalanced; repeated stratified cross-validation reduces evaluation bias but does not eliminate sensitivity to fold composition in small-to-moderate datasets [34]. Node-level split search currently reduces multiclass separation to temporary binary groupings during oblique split induction, which may underuse the ordered structure of the three-tier DILI endpoint. The highest-priority extension is external validation on independent chemical DILI datasets with compatible label definitions; DILIrank 2.0 expands the public benchmark base [3, 8], and recent DILI modeling work increasingly combines structural information with additional biological or assay-derived signals [5, 21]. Representation enrichment, comparing Mold2 with fragment-based, fingerprint, or multimodal descriptors, would help clarify whether the current performance ceiling is primarily algorithmic or representational [7]. Ordinal-aware multiclass split strategies that respect the ordered risk-tier structure represent a natural methodological extension of the current BWORF formulation. Causal inference applied to LIME attributions could further distinguish descriptors that drive BWORF decisions from those that reflect confounding structure in the Mold2 descriptor space [36, 40].

Acknowledgments

Authors are thankful to ASU supercomputing center-SOL for simulation study.

Code, Data and supplementary material availability:

Cleaned data will be available upon request, raw data available at FDA website. The code and supplementary materials for this work are available in the public GitHub repository, which is described as the Bootstrap-Weighted Oblique Random Forest Algorithm repository and is distributed under the MIT License.

Declaration of Interests

The authors have no conflict of interest to report.

Ethics Approval

There is no ethical approval needed due to the use of simulated and publicly available data.

Funding Statement

The authors do not have funding to report.

Clinical Trial Registration

The authors did not use clinical trial data directly. The authors used publicly available data with proper references in the text.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

References

[1] Shraddha Thakkar, Ting Li, Zhichao Liu, Leihong Wu, Ruth Roberts, and Weida Tong. Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity. *Drug Discovery Today*, 25(1):201–208, January 2020. ISSN 1878-5832. doi: 10.1016/j.drudis.2019.09.022.

[2] Jack Uetrecht. DILI prediction in drug development: present and future. *Expert Opinion on Drug Metabolism & Toxicology*, 21(6):665–676, June 2025. ISSN 1742-5255. doi: 10.1080/17425255.2025.2495955. URL https://doi.org/10.1080/17425255.2025.2495955. _eprint: https://doi.org/10.1080/17425255.2025.2495955.

[3] Minjun Chen, Ayako Suzuki, Shraddha Thakkar, Ke Yu, Chuchu Hu, and Weida Tong. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discovery Today*, 21(4):648–653, April 2016. ISSN 1359-6446. doi: 10.1016/j.drudis.2016.02.015. URL https://www.sciencedirect.com/science/article/pii/S1359644616300411.

[4] Fahad Mostafa and Minjun Chen. Computational models for predicting liver toxicity in the deep learning era. *Frontiers in Toxicology*, 5, January 2024. ISSN 2673-3080. doi: 10.3389/ftox.2023.1340860. URL https://www.frontiersin.org/journals/toxicology/articles/10.3389/ftox.2023.1340860/full.

[5] Yves M. K. Kantagba, Seydou G. Barro, Serge L. W. Nikiema, and Pascal Staccini. Artificial intelligence for drug-induced liver injury prediction: A systematic review. *Studies in Health Technology and Informatics*, 328:183–187, 2025. doi: 10.3233/SHTI250698.

[6] Huixiao Hong, Shraddha Thakkar, Minjun Chen, and Weida Tong. Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-approved Drugs. *Scientific Reports*, 7(1):17311, December 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-17701-7. URL https://www.nature.com/articles/s41598-017-17701-7.

[7] Ting Li, Weida Tong, Ruth Roberts, Zhichao Liu, and Shraddha Thakkar. DeepDILI: Deep Learning-Powered Drug-Induced Liver Injury Prediction Using Model-Level Representation. *Chemical Research in Toxicology*, 34(2):550–565, February 2021. ISSN 0893-228X. doi: 10.1021/acs.chemrestox.0c00374. URL https://doi.org/10.1021/acs.chemrestox.0c00374.

[8] AyoOluwa O. Olubamiwa, Yanyan Qu, Skylar Connor, Weida Tong, Dongying Li, and Minjun Chen. Dilirank 2.0: An updated and expanded database for drug-induced liver injury risk based on fda labeling and a literature review. *Drug Discovery Today*, 30 (11):104485, 2025. doi: 10.1016/j.drudis.2025.104485.

[9] National Center for Toxicological Research. Drug Induced Liver Injury Rank (DILIrank 2.0) Dataset. *FDA*, September 2025. URL https://www.fda.gov/science-research/liver-toxicity-knowledge-base-ltkb/drug-induced-liver-injury-rank-dilirank-20-dataset.

[10] Minjun Chen, Huixiao Hong, Hong Fang, Reagan Kelly, Guangxu Zhou, Jürgen Borlak, and Weida Tong. Quantitative structure-activity relationship models for predicting drug-induced liver injury based on FDA-approved drug labeling annotation and using a large collection of drugs. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, 136(1):242–249, November 2013. ISSN 1096-0929. doi: 10.1093/toxsci/kft189.

[11] Shao-Huan Huang, Chun-Wei Tung, Ferenc Fülöp, and Jia-Hsin Li. Developing a QSAR model for hepatotoxicity screening of the active compounds in traditional Chinese medicines. *Food and Chemical Toxicology*, 78:71–77, 2015. doi: 10.1016/j.fct.2015.01.020.

[12] Dennis Mulliner, Friedlieb Schmidt, Manuela Stolber, Ingo Gröne, Juan Carlos Gómez-Manzano, Detlef Lang, Tim Geppert, Daria Novikova, Michael Reutlinger, Alfonso T. Garcia-Sosa, Gisbert Schneider, and Lars Friedrich. Computational models for human and animal hepatotoxicity with a global application scope. *Chemical Research in Toxicology*, 29(5):757–767, 2016. doi: 10.1021/acs.chemrestox.5b00465.

[13] Hui Zhang, Lili Ding, Yi Zou, Shao-Qing Hu, Hai-Guo Huang, Wei-Bo Kong, and Ji Zhang. Predicting drug-induced liver injury in human with Naïve Bayes classifier approach. *Journal of Computer-Aided Molecular Design*, 30(10):889–898, 2016. doi: 10.1007/s10822-016-9972-6.

[14] Thanh-Hoang Nguyen-Vo, Loc Nguyen, Nguyet Do, Phuc H. Le, Thien-Ngan Nguyen, Binh P. Nguyen, and Ly Le. Predicting Drug-Induced Liver Injury Using Convolutional Neural Network and Molecular Fingerprint-Embedded Features. *ACS Omega*, 5(39): 25432–25439, October 2020. doi: 10.1021/acsomega.0c03866. URL https://doi.org/10.1021/acsomega.0c03866.

[15] Youjun Xu, Ziwei Dai, Fangjin Chen, Shuaishi Gao, Jianfeng Pei, and Luhua Lai. Deep learning for drug-induced liver injury. *Journal of Chemical Information and Modeling*, 55(10):2085–2093, 2015. doi: 10.1021/acs.jcim.5b00238.

[16] Myung-Gyun Kang and Nam Sook Kang. Predictive Model for Drug-Induced Liver Injury Using Deep Neural Networks Based on Substructure Space. *Molecules*, 26 (24):7548, January 2021. ISSN 1420-3049. doi: 10.3390/molecules26247548. URL https://www.mdpi.com/1420-3049/26/24/7548.

[17] Sangsoo Lim, Youngkuk Kim, Jeonghyeon Gu, Sunho Lee, Wonseok Shin, and Sun Kim. Supervised chemical graph mining improves drug-induced liver injury prediction. *iScience*, 26(1), January 2023. ISSN 2589-0042. doi: 10.1016/j.isci.2022.105677. URL https://www.cell.com/iscience/abstract/S2589-0042(22)01949-6.

[18] Hehuan Ma, Weizhi An, Yuhong Wang, Hongmao Sun, Ruili Huang, and Junzhou Huang. Deep Graph Learning with Property Augmentation for Predicting Drug-Induced Liver Injury. *Chemical Research in Toxicology*, 34(2):495–506, February 2021. ISSN 1520-5010. doi: 10.1021/acs.chemrestox.0c00322.

[19] Domenico Gadaleta, Serena Manganelli, Alessandra Roncaglioni, Cosimo Toma, Emilio Benfenati, and Enrico Mombelli. QSAR Modeling of ToxCast Assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. *Journal of Chemical Information and Modeling*, 58(8):1501–1517, August 2018. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00297. URL https://doi.org/10.1021/acs.jcim.8b00297.

[20] Muhammad Zafar Irshad Khan, Jia-Nan Ren, Cheng Cao, Hong-Yu-Xiang Ye, Hao Wang, Ya-Min Guo, Jin-Rong Yang, and Jian-Zhong Chen. Comprehensive hepatotoxicity prediction: ensemble model integrating machine learning and deep learning. *Frontiers in Pharmacology*, 15, August 2024. ISSN 1663-9812. doi: 10.3389/fphar.2024.1441587. URL https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2024.1441587/full.

[21] Srijit Seal, Dominic Williams, Layla Hosseini-Gerami, Manas Mahale, Anne E. Carpenter, Ola Spjuth, and Andreas Bender. Improved Detection of Drug-Induced Liver Injury by Integrating Predicted In Vivo and In Vitro Data. *Chemical Research in Toxicology*, 37(8):1290–1305, August 2024. ISSN 0893-228X. doi: 10.1021/acs.chemrestox.4c00015. URL https://doi.org/10.1021/acs.chemrestox.4c00015.

[22] Soyeon Lee and Sunyong Yoo. InterDILI: interpretable prediction of drug-induced liver injury through permutation feature importance and attention mechanism. *Journal of Cheminformatics*, 16(1):1, January 2024. ISSN 1758-2946. doi: 10.1186/s13321-023-00796-8. URL https://doi.org/10.1186/s13321-023-00796-8.

[23] Binsheng Sui, Qingzhuo He, Bowei Yan, Kunhong Liu, Yong Xu, Song He, and Xiaochen Bo. MORF: Multi-view oblique random forest for hepatotoxicity prediction.

iScience, 28(2), February 2025. ISSN 2589-0042. doi: 10.1016/j.isci.2025.111755. URL https://www.cell.com/iscience/abstract/S2589-0042(25)00014-8.

[24] Huixiao Hong, Qian Xie, Weigong Ge, Feng Qian, Hong Fang, Leming Shi, Zhenqiang Su, Roger Perkins, and Weida Tong. Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *Journal of Chemical Information and Modeling*, 48(7):1337–1344, July 2008. ISSN 1549-9596. doi: 10.1021/ci800038f. URL https://doi.org/10.1021/ci800038f.

[25] S. K. Murthy, S. Kasif, and S. Salzberg. A System for Induction of Oblique Decision Trees. *Journal of Artificial Intelligence Research*, 2:1–32, August 1994. ISSN 1076-9757. doi: 10.1613/jair.63. URL https://www.jair.org/index.php/jair/article/view/10121.

[26] Tyler M. Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse L. Patsolic, Benjamin Falk, Carey E. Priebe, Jason Yim, Randal Burns, Mauro Maggioni, and Joshua T. Vogelstein. Sparse Projection Oblique Randomer Forests. *Journal of Machine Learning Research*, 21(104):1–39, 2020. ISSN 1533-7928. URL http://jmlr.org/papers/v21/18-664.html.

[27] Rakesh Katuwal, P.N. Suganthan, and Le Zhang. Heterogeneous oblique random forest. *Pattern Recognition*, 99:107078, March 2020. ISSN 00313203. doi: 10.1016/j.patcog.2019.107078. URL https://linkinghub.elsevier.com/retrieve/pii/S0031320319303796.

[28] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[29] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996. ISSN 1573-0565. doi: 10.1007/BF00058655. URL https://doi.org/10.1007/BF00058655.

[30] Chao Chen, Andy Liaw, and Leo Breiman. Using Random Forest to Learn Imbalanced Data, 2004. URL https://digicoll.lib.berkeley.edu/record/85556. Num Pages: 12.

[31] Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009. ISSN 1558-2191. doi: 10.1109/TKDE.2008.239. URL https://ieeexplore.ieee.org/document/5128907.

[32] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*,

ICML '06, pages 233–240, New York, NY, USA, June 2006. Association for Computing Machinery. ISBN 978-1-59593-383-6. doi: 10.1145/1143844.1143874. URL https://dl.acm.org/doi/10.1145/1143844.1143874.

[33] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432, March 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0118432. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432.

[34] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, February 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-91. URL https://doi.org/10.1186/1471-2105-7-91.

[35] Bradley Efron and Robert Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997. doi: 10.1080/01621459.1997.10474007.

[36] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://www.nature.com/articles/s42256-019-0048-x.

[37] Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif Cifci, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S. Alkhawaldeh, Sadiq Hussain, Bilal Alatas, Afshin Shoeibi, Hossein Moosaei, Milan Hladík, Saeid Nahavandi, and Panos M. Pardalos. A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, 118:109370, August 2024. ISSN 0045-7906. doi: 10.1016/j.compeleceng.2024.109370. URL https://www.sciencedirect.com/science/article/pii/S0045790624002982.

[38] Maria Sahakyan, Zeyar Aung, and Talal Rahwan. Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access*, 9:135392–135422, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3116481. URL https://ieeexplore.ieee.org/document/9551946.

[39] Donghyeon Kim and Jinhee Choi. Ai-based toxicity prediction models using toxcast data: Current status and future directions for explainable models. *Toxicology*, 517: 154230, 2025. doi: 10.1016/j.tox.2025.154230.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

[41] Byron C. Jaeger, Sawyer Welden, Kristin Lenoir, Jaime L. Speiser, Matthew W. Segar, Ambarish Pandey, and Nicholas M. Pajewski. Accelerated and Interpretable Oblique Random Survival Forests. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 33(1):192–207, 2024. ISSN 1061-8600. doi: 10.1080/10618600.2023.2231048. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11343578/.

[42] Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System (BRFSS) 2022 Survey Data. U.S. Department of Health and Human Services, Atlanta, Georgia, 2023. URL https://www.cdc.gov/brfss/annual_data/annual_2022.html. Annual Survey Data, 2022; public domain.

[43] Kamil Pytlak. Personal Key Indicators of Heart Disease (2022 Update). Kaggle dataset, 2023. URL https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. Cleaned redistribution of CDC BRFSS 2022 survey data; file `heart_2022_no_nans.csv`, SHA-256 `f5eddf47d85170f2f7bc3ba523c275d705ddf22518f95d50aa33ddcfd786a94f`.

[44] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[45] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, July 1994. ISSN 1941-0093. doi: 10.1109/72.298224. URL https://ieeexplore.ieee.org/document/298224.

[46] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005. ISSN 1939-3539. doi: 10.1109/TPAMI.2005.159. URL https://ieeexplore.ieee.org/document/1453511.

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

[48] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[49] Douglas M. Jennewein, Johnathan Lee, Chris Kurtz, Will Dizon, Ian Shaeffer, Alan Chapman, Alejandro Chiquete, Josh Burks, Amber Carlson, Natalie Mason, Arhat Kobwala, Thirugnanam Jagadeesan, Praful Barghav, Torey Battelle, Rebecca Belshe, Debra McCaffrey, Marisa Brazil, Chaitanya Inumella, Kirby Kuznia, Jade Buzinski,

Sean Dudley, Dhruvil Shah, Gil Speyer, and Jason Yalim. The Sol Supercomputer at Arizona State University. In *Practice and Experience in Advanced Research Computing*, PEARC '23, pages 296–301, New York, NY, USA, Jul 2023. Association for Computing Machinery. ISBN 9781450399852. doi: 10.1145/3569951.3597573.

[50] David J. Hand and Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2): 171–186, November 2001. ISSN 1573-0565. doi: 10.1023/A:1010920819831. URL https://doi.org/10.1023/A:1010920819831.

[51] Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G.M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMC Medicine*, 13:1, 2015. doi: 10.1186/s12916-014-0241-z.

[52] Luciana Ferrer, Odette Scharenborg, and Tom Bäckström. Good practices for evaluation of machine learning systems, 2024. URL https://arxiv.org/abs/2412.03700.

[53] Simon Bin Akter, Rakibul Hasan, Sumya Akter, Md. Mahadi Hasan, and Tanmoy Sarkar. Improving heart disease probability prediction sensitivity with a grow network model. medRxiv preprint, 2024. URL https://www.medrxiv.org/content/10.1101/2024.02.28.24303495v1. GrowNet-based imbalance handling on BRFSS 2022 heart disease data.

Additional files

Additional file 1 — Supplementary Methods, Results, and Figures (PDF). Contains: (S1) full BWORF training and prediction pseudocode; (S2) seed-averaged BWORF distribution on the primary multiclass benchmark alongside the best-accuracy coherent operating point shown in Table 3; (S3) extended external benchmark results, comprising (S3a) full per-seed BRFSS 2022 cross-validation and held-out test metrics for all five evaluated models including precision, recall, log loss, and Brier score; (S3b) the BRFSS bootstrap-temperature sensitivity sweep over $T \in \{0.5, 1.0, 1.5, 2.0, 3.0\}$ at the 4,000-sample subsample scale; (S3c) full per-dataset results for the three smaller binary external benchmarks (breast cancer recurrence, heart failure clinical records, Pima Indians diabetes), including AUPRC and Wilson confidence intervals for the BWORF operating points; and (S3d) byte-equivalence validation results for the parallel BWORF implementation used at full BRFSS scale; (S4) full simulation scenario grid specification, data-generating process, and extended results including per-scenario AUROC differences between BWORF and ORF; (S5) complete hyperparameter configurations for baseline models; (S6) representative LIME case studies for high-confidence correct, borderline correct, and notable error predictions; (S7) fold-local preprocessing specification and reproducibility controls.

Additional file 2 — Source code and analysis pipeline. Python implementation of BWORF, evaluation scripts, and analysis notebooks used to generate all tables and figures reported in this manuscript and in Additional file 1. Distributed as a tagged GitHub release at https://github.com/arthuryu123/BWORF/releases/tag/v1.0. The release archive includes the BWORF implementation, all benchmark runner scripts, the BRFSS data preparation pipeline, the parallel-implementation byte-equivalence validation script, simulation scripts, and a README describing how to reproduce each table and figure.