# LOSS-GUIDED STOCHASTIC EXPLORATION FOR LEARNING A SINGLE DECISION TREE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Decision trees provide predictions through compact decision paths, but greedy construction often commits to locally optimal splits, creating a performance gap between single trees and ensembles. We introduce a loss-guided stochastic node exploration method for learning stronger single-tree structures. The method augments original features with data-derived directions, generates randomized and criterion-guided split candidates, and evaluates their downstream utility through staged search. This allows later splits to inform earlier decisions while using only one tree at inference time. Across diverse tabular classification datasets, the resulting trees outperform conventional and recent single-tree baselines, achieving state-of-the-art performance among the evaluated methods. Their mean Macro-F1 was approximately 1–2 percentage points lower than that of the tuned tree ensembles. These results suggest that improving feature representation and tree-construction search can narrow the performance gap between single trees and ensembles while retaining an explicit decision structure.

## 1 INTRODUCTION

Decision trees (DTs) are nonparametric models for classification and regression that recursively partition the feature space through a sequence of tests and assign a prediction to each terminal region. Their explicit rules can capture nonlinear relationships and feature interactions while allowing users to inspect how an individual prediction is obtained (Breiman et al., 1984). These advantages make compact DTs attractive when interpretability and efficient inference are important. However, learning a globally optimal tree is a combinatorial optimization problem that is NP-hard in general (Hyafil & Rivest, 1976; Bertsimas & Dunn, 2017). Practical methods such as classification and regression trees (CART; Breiman et al. 1984) therefore rely on greedy split selection. Although computationally efficient, this strategy relies on immediate improvements and may overlook splits whose benefits appear only in later steps.

Despite their interpretability, individual DTs often provide lower predictive accuracy and less stable predictions than tree ensembles. They are therefore widely used as base learners in random forests (RFs; Breiman 2001) and gradient-boosted trees (Friedman, 2001), where aggregating many trees improves generalization. This gain comes at a cost: an ensemble prediction is distributed across many trees and can no longer be represented by a single decision path. A central question is therefore whether a stronger construction procedure can recover part of the ensemble advantage while retaining one explicit tree. Existing work approaches this question through optimal or lookahead search (Babbar et al., 2025; Chaouki et al., 2025) and through differentiable optimization of tree parameters (Marton et al., 2024; Mao et al., 2025). These methods show the value of moving beyond greedy induction, but leave room for a search procedure that directly evaluates the downstream consequences of early split decisions.

Motivated by this gap, we introduce a loss-guided stochastic exploration procedure for constructing a single decision tree. Instead of selecting an early split only by its immediate impurity reduction, the procedure evaluates that split through complete downstream subtrees sampled from that node. Randomized proposals broaden the structural search, while criterion-guided proposals provide locally informed candidates. The search then fixes promising prefixes in stages, resamples their remaining subtrees, and refines a beam of complete-tree candidates. Multiple trees are explored only during training; the final model is one selected tree with a single explicit prediction path. The search



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

can also consider splits based on learned linear combinations of features, in addition to standard single-feature splits.

The proposed objective is designed to balance predictive performance and structural simplicity. It combines a term for predictive accuracy with a class-balanced term, ensuring that performance is not dominated by majority classes. A regularization penalty discourages unnecessary tree complexity and favors compact, interpretable structures.

Our contributions are: (i) an explicit construction procedure combining downstream root assessment, staged prefix search, and complete-tree refinement; (ii) a composite objective combining predictive performance, class-balanced probability error, and tree complexity (iii) an empirical evaluation assessing the predictive performance of our approach relative to tree ensembles and recent single-tree baselines.

## 2 Related work

**Greedy trees and ensembles.** CART constructs trees through local impurity reduction and controls their size through pruning (Breiman et al., 1984). RFs combine trees trained with randomized observations and feature subsets to reduce prediction variability (Breiman, 2001). Gradient boosting instead adds trees sequentially to improve a predictive objective (Friedman, 2001); XGBoost (XGB), LightGBM (LGB), and CatBoost (CAT) develop this approach through regularization, efficient split search, and mechanisms for handling categorical data (Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018). These approaches improve prediction through aggregation, whereas the optimization of an individual tree concerns the quality of a single partition of the input space.

**Lookahead and optimized single trees.** Optimal classification trees formulate tree construction as constrained optimization (Bertsimas & Dunn, 2017). Separable trees with dynamic programming (STreeD; van der Linden et al. 2023) identifies conditions for optimizing objectives and constraints through separate subtree problems. Branches uses AND/OR graph search with an AO*-type algorithm and optimality guarantees (Chaouki et al., 2025). SParse Lookahead for Interpretable Trees (SPLIT; Babbar et al. 2025) reduces search cost by combining root-level lookahead with greedy construction near the leaves. Dynamic programming decision trees (DPDT; Kohler et al. 2025) uses a Markov decision process formulation and heuristic split generation to restrict the search while optimizing a regularized training loss. These methods differ in search cost, admissible splits, and optimality guarantees.

**Differentiable single trees.** A separate line of work optimizes tree parameters using gradient information. GradTree learns hard, axis-aligned trees through gradient-based optimization (Marton et al., 2024). DTSemNet provides a neural representation of oblique trees that can be trained with vanilla gradient descent (Panda et al., 2024), while the ReLU+Argmin-based differentiable decision tree (Mao et al. 2025) introduces another reformulation for differentiable training. These approaches differ in their split representations and treatment of discrete routing, but share the goal of learning a single predictive tree without relying solely on greedy impurity reduction.

**Predictive objectives and complexity.** Tree construction also depends on how predictive quality and structural size are measured. Pruning and sparse-tree optimization balance prediction error against tree complexity (Breiman et al., 1984; Bertsimas & Dunn, 2017; Chaouki et al., 2025). Probability-based criteria capture information that hard-label predictions discard: the Brier score measures squared error in predicted class probabilities (Brier, 1950). STreeD further shows how the structure of an objective determines whether it can be optimized through independent subtree problems (van der Linden et al., 2023). Objective design and search strategy are therefore closely connected.

Within this literature, our approach uses sampled complete-tree continuations to assess early splits and guide staged structural search. It evaluates discrete candidate trees directly, without a differentiable relaxation or a certificate of global optimality.

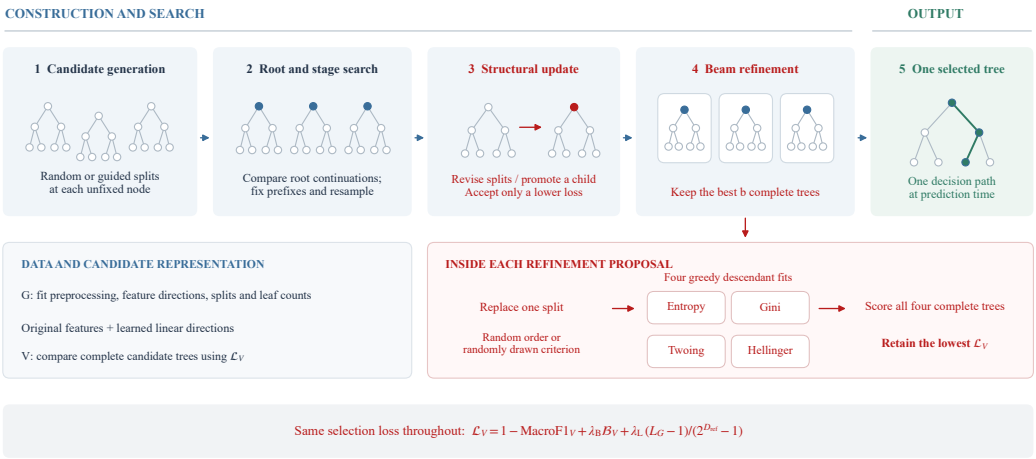


Figure 1: Overview of tree construction and refinement. Complete candidate trees are evaluated through root continuations and progressively fixed prefixes. Structural updates and beam refinement use the same whole-tree loss $\mathcal{L}_V$. Each refinement proposal compares four subtrees grown with entropy, Gini, twoing, or Hellinger and retains the lowest-loss completion. One tree is returned for prediction. Numbered panels indicate operations.

# 3 Loss-guided stochastic exploration

## 3.1 Overview

Figure 1 summarizes the proposed method. The procedure constructs complete candidate trees using random and criterion-guided splits, then assesses promising roots by sampling their descendant subtrees. At each later stage, it fixes a selected tree prefix and resamples the subtrees below it; each stage evaluates complete trees rather than adding one depth level. The retained tree then undergoes structural updates and beam-based refinement using the same whole-tree loss. Each split replacement compares four criterion-specific subtrees and retains the completion with the lowest validation loss. Tree-construction data determine split candidates and leaf counts, while separate structure-validation data guide tree selection. Algorithm 1 summarizes the procedure.

## 3.2 Tree Construction Setup

Let $G$ be the tree-construction set and $V$ a disjoint structure-validation set. We fit preprocessing, feature coefficients, and leaf counts on $G$, and use $V$ to compare candidate trees. Hyperparameters are chosen using a separate holdout within the outer training partition. The outer test set is used only for evaluation.

**Feature dictionary.** Let $\widetilde{x} \in \mathbb{R}^d$ denote the preprocessed original input. The tree can use both original coordinates and learned linear combinations of them. We write the available features as

$$z(x) = [\widetilde{x};\, \phi_1(\widetilde{x}); \ldots; \phi_J(\widetilde{x})]\,, \qquad \phi_j(\widetilde{x}) = w_j^\top \widetilde{x} + b_j. \tag{1}$$

The initial learned features include principal component analysis (PCA) and ridge-regression directions. Centering and scaling are absorbed into $w_j$ and $b_j$. Additional directions can be proposed during the loss-consistent structural update described below, so the number $J$ of learned features can increase. Their coefficients are fitted on $G$, while $V$ helps determine whether the resulting tree edits are retained. Thus validation labels guide feature selection but are not regression targets for fitting the coefficients. Appendix C describes how these additional directions are formed.

Each internal node tests $z_j(x) \le t$. An original coordinate gives a single-feature split; a learned coordinate gives a split based on a linear combination of features. Both produce one decision path in the final tree. A dense linear combination can be harder to interpret than a single-feature test, even when the tree is small.

**Algorithm 1** Loss-guided stochastic exploration of a single decision tree

**Require:** construction set $G$, structure-validation set $V$, depth limit $D$, stages $S$, stage budget $B$, guided-split probability $q$, beam width $b$, refinement budget $R$

```
Fit preprocessing and initial feature directions on G
Sample and score B trees using q; set T* to the lowest-L_V tree
Sample elite-root continuations and update T*
Set the prefix to the root minimizing U(r)
for s = 2, ..., S do
    Sample and score B prefix-conditioned trees using q; update T*
    Extend the prefix by elite consensus if s < S and extension is permitted
end for
Apply L_V-based structural edits; initialize the beam and incumbent
for at most R split replacements across the beam do
    Alternate between random ordering and ranking by a uniformly sampled criterion
    Grow descendants with each of the four criteria; select by L_V
    Retain the best b distinct trees and update the incumbent
end for
Optionally recalibrate the incumbent's leaf counts on the outer training set
return the incumbent tree
```

For a leaf $\ell$, let $n_{\ell,k}$ be its number of construction examples from class $k$, and let $n_\ell = \sum_k n_{\ell,k}$. Tree depth counts edges from the root, so a root-only tree has depth zero. With $K$ task classes and smoothing parameter $\alpha > 0$, the probabilities used to score a tree are

$$p_k(x;T) = \frac{n_{\ell(x),k} + \alpha}{n_{\ell(x)} + K\alpha}, \qquad \hat{y}(x;T) = \arg\max_k p_k(x;T). \tag{2}$$

An empty leaf receives a uniform distribution. After hyperparameter selection and the final structure search, leaf counts can be recomputed on the full outer training set while keeping every split fixed.

### 3.3 A WHOLE-TREE OBJECTIVE

Let $C_V$ contain the classes represented in $V$, and let $n_c$ count its examples from class $c$. Define the class-balanced Brier score

$$\mathcal{B}_V(T) = \frac{1}{|C_V|} \sum_{c \in C_V} \frac{1}{n_c} \sum_{i \in V : y_i = c} \sum_{k=1}^{K} \big(p_k(x_i;T) - \mathbf{1}[y_i = k]\big)^2. \tag{3}$$

Equation 3 uses the sum of squared class-probability errors, without dividing it by two. Class balancing prevents majority-class frequency from directly dominating this term, although it can increase variance when a class has few structure-validation examples. Let $L_G(T)$ denote the number of leaves occupied by tree-construction observations, and let $D_{\mathrm{ref}}$ be a fixed reference depth at least as large as the allowed candidate depths. Our objective is

$$\mathcal{L}_V(T) = 1 - \mathrm{MacroF1}_V(T) + \lambda_{\mathrm{B}} \mathcal{B}_V(T) + \lambda_{\mathrm{L}} \frac{L_G(T) - 1}{2^{D_{\mathrm{ref}}} - 1}. \tag{4}$$

Macro-F1 uses the fixed task class set, with zero assigned to undefined classwise F1 values. The normalization assigns zero complexity to a one-leaf tree and a common incremental cost to an additional occupied leaf. Normalizing by each candidate's own depth would instead make the same leaf count cheaper at a larger depth. Keeping the reference fixed avoids that dependence. The coefficients control distinct tradeoffs and are selected using the hyperparameter-validation holdout, not test performance.

The objective is discontinuous because routing and hard predictions can change abruptly. We therefore evaluate candidate structures directly. The Brier term supplies probability-sensitive information without requiring a differentiable relaxation, while the leaf term measures structural size rather than the complexity of an oblique predicate. The search is approximate and does not guarantee a globally optimal tree.

### 3.4 CANDIDATE TREE GENERATION AND ROOT ASSESSMENT

Candidate generation and root assessment are performed as two consecutive steps. The method first generates a pool of complete trees subject to the depth and purity stopping rules. It then uses the strongest trees in this pool to identify and evaluate promising root predicates.

**$\epsilon$-greedy-style split proposals.** When generating the initial and staged candidate pools, the method applies the following proposal rule independently at every node $v$ whose split has not already been fixed:

$$S_v = \begin{cases} S_v^{\text{guided}}, & \text{with probability } q, \\ S_v^{\text{random}}, & \text{with probability } 1-q. \end{cases} \tag{5}$$

Here, $q$ is the probability of using a criterion-guided proposal, and the corresponding exploration probability is $\epsilon = 1 - q$. The rule therefore applies to both the root and its descendants. During staged search, nodes belonging to the fixed prefix are retained, while the rule is applied to the remaining unfixed nodes.

A random proposal selects a feature uniformly and samples a threshold uniformly between its smallest and largest values at the node. Proposals that do not separate the observations are resampled. A guided proposal draws uniformly from entropy (Quinlan, 1986), Gini impurity and twoing (Breiman et al., 1984), and Hellinger distance (Cieslak et al., 2012), and then searches for a split according to the selected criterion. Entropy and Gini favor pure child nodes, twoing emphasizes separation between the class distributions of the two children, and Hellinger is less sensitive to class imbalance. If no valid guided split is found, the method falls back to a random proposal. The resulting complete trees are evaluated using $\mathcal{L}_V$, allowing each proposed split to be assessed together with its downstream subtree. The final refinement step uses a separate proposal rule, described in the next subsection.

After generating the initial pool, the method extracts promising roots from an elite subset of complete trees. For each root predicate $r$, it fixes $r$ and samples a set of descendant completions $\mathcal{C}(r)$. Let $Q_\tau(r)$ denote the empirical $\tau$-percentile of their losses, and define

$$A_\tau(r) = \{T \in \mathcal{C}(r) : \mathcal{L}_V(T) \le Q_\tau(r)\}.$$

The root utility is

$$U(r) = \frac{1}{|A_\tau(r)|} \sum_{T \in A_\tau(r)} \mathcal{L}_V(T). \tag{6}$$

This score averages the losses in the strongest part of the sampled continuation distribution, including ties at the percentile threshold. It therefore measures whether a root repeatedly leads to strong completed trees, rather than judging it from either its immediate split quality or a single best completion. Root proposals using the same feature are combined before continuation sampling, as detailed in Appendix C.

### 3.5 STAGED PREFIX SEARCH AND COMPLETE-TREE REFINEMENT

After root selection, the search fixes an upper-tree prefix and samples new subtrees below it. In each region, it retains the most frequent split feature among selected low-loss trees and averages its thresholds. The prefix grows up to a prescribed limit, after which later stages resample beneath it. Each stage generates a pool of complete trees rather than adding a depth level. Although the best loss within each pool may fluctuate, the search retains the lowest-loss tree across all stages and root evaluations. After the staged search, the retained tree is updated through split replacement, child promotion that can remove branches, and optional learned feature directions. Each edit is evaluated using the same $\mathcal{L}_V$ as staged search and accepted only if it lowers the objective.

Final refinement maintains a beam of the best $b$ complete trees. At an eligible node, candidate feature–threshold pairs are visited by alternating a random ordering with a ranking drawn uniformly from entropy, Gini, twoing, and Hellinger. Each proposal replaces one split. The method then builds four descendant completions, one with each criterion, within the remaining depth limit and retains the completed tree with the smallest $\mathcal{L}_V$. Duplicate tree states are removed, and the best $b$ candidates remain in the beam. Because the current best tree remains eligible, the lowest loss found during refinement cannot increase. Only the final selected tree is used for prediction. Appendix D analyzes the computational cost.

# 4 EXPERIMENTS

## 4.1 EVALUATION PROTOCOL

We evaluated the method on 22 tabular classification datasets using ten frozen stratified 80/20 train–test splits per dataset. Hyperparameter optimization (HPO) used one class-preserving 70/30 holdout within each outer training set, yielding 56% inner fit, 24% inner validation, and 20% sealed test data. For the proposed tree, the inner-fit portion was further divided into 42% tree-construction and 14% structure-validation data. After HPO, we reconstructed the selected configuration using 60% tree-construction and 20% structure-validation data, recalibrated leaf probabilities on the full training set, and refitted the ensemble baselines on the same rows. The test set was accessed only after the fit lock was written.

All methods received 64 trials per outer split. We used five search stages ($S = 5$) in all experiments. We used the tree-structured Parzen estimator (TPE; Bergstra et al. 2011) implemented in Optuna (Akiba et al., 2019). Model selection maximized validation Macro-F1, broke ties by accuracy, and then preferred the earlier trial. The proposed-tree search spaces are detailed in Appendix B. For the ensemble tree comparison, we tuned RF, XGB, LGB, and CAT with the same split and trial protocol. For the single-tree comparison, we tuned classification and regression trees (CART; Breiman et al. 1984), an extremely randomized tree (ExtraTree; Geurts et al. 2006), dynamic programming decision trees (DPDT; Kohler et al. 2025), and separable trees with dynamic programming (STreeD; van der Linden et al. 2023) using the same outer splits, inner holdout, and 64-trial budget. Appendix B gives the selected-configuration frequencies and baseline spaces.

We reported Macro-F1, accuracy, multiclass log loss, and average rank, averaging results first across ten repetitions per dataset and then across datasets. Statistical comparisons used dataset-level means with the Friedman test and two-sided Wilcoxon signed-rank tests with Holm correction.

## 4.2 PERFORMANCE ACROSS THE FIVE SEARCH STAGES

We ran the source search once per split and saved the complete tree at each of five stages. We replayed Stages 1–4 through the same downstream construction and refinement used for Stage 5 while keeping preprocessing, the outer split, the HPO-selected configuration, the model seed, and the test set fixed.

Table 1 reports the resulting 220 matched evaluations. No single stage dominates every criterion. Stage 3 has the highest mean Macro-F1 and accuracy, Stage 2 has the lowest mean log loss, and Stage 4 has the best average log-loss rank. Counting ties, the best Macro-F1 occurs at Stage 1 on six datasets, Stage 2 on six, Stage 3 on five, Stage 4 on five, and Stage 5 on one. The corresponding accuracy counts are 7, 8, 7, 5, and 3.

Figure 2 clearly shows these differences across datasets. For example, Car Evaluation reaches its best Macro-F1 at Stage 3 and then loses performance at Stage 4, while Hill Valley peaks at Stage 2. Balance Scale peaks earlier at Stage 2. Because running the search to completion does not produce monotonic improvements across all datasets but rather yields dataset-dependent trade-offs, the optimal stopping stage should be determined via validation performance for each dataset, rather than uniformly adopting the final stage.

| Metric | Stage 1 | Stage 2 | Stage 3 | Stage 4 | Stage 5 |
|---|---|---|---|---|---|
| Macro-F1 (%) | 78.35 | 78.45 | **78.78** | 78.27 | 78.57 |
| Accuracy (%) | 84.83 | 84.91 | **84.99** | 84.76 | 84.73 |
| Log loss ↓ | 0.425 | **0.420** | 0.424 | 0.426 | 0.424 |
| F1 average rank ↓ | 3.27 | 2.89 | **2.52** | 3.02 | 3.30 |
| Accuracy average rank ↓ | 3.23 | 2.75 | **2.57** | 3.09 | 3.36 |
| Log-loss average rank ↓ | 3.41 | 2.77 | 3.00 | **2.75** | 3.07 |

Table 1: Dataset-balanced performance of saved stage checkpoints over 22 datasets and ten repetitions.

| Cohort | Metric | Stage 3 | Stage 5 | RF | XGB | LGB | CAT | Ensemble mean |
|---|---|---|---|---|---|---|---|---|
| 22 datasets | Macro-F1 (%) | **78.78** | 78.57 | 77.87 | 78.34 | **78.71** | **78.85** | 78.44 |
| 22 datasets | Accuracy (%) | 84.99 | 84.73 | **85.25** | 85.18 | **85.43** | **85.64** | 85.38 |
| 22 datasets | Log loss ↓ | 0.424 | 0.424 | 0.393 | **0.373** | **0.380** | **0.371** | 0.379 |
| 22 datasets | F1 rank ↓ | 4.18 | 4.45 | **3.41** | 3.45 | **3.05** | **2.45** | 3.09 |
| Without Hill Valley | Macro-F1 (%) | 78.26 | 78.04 | 78.94 | **79.29** | **79.77** | **80.05** | 79.51 |
| Without Hill Valley | Accuracy (%) | 84.77 | 84.50 | **86.67** | 86.45 | **86.80** | **87.15** | 86.77 |
| Without Hill Valley | Log loss ↓ | 0.429 | 0.429 | 0.378 | **0.352** | **0.355** | **0.347** | 0.358 |
| Without Hill Valley | F1 rank ↓ | 4.33 | 4.57 | **3.33** | 3.48 | **3.00** | **2.29** | 3.03 |

Table 2: Dataset-balanced comparison of Stage 3, final Stage 5, and four tuned ensembles. Red, blue, and green indicate the top three individual models. Since Stage 3 exceeds the best ensemble on Hill Valley by 31.25 Macro-F1 points, results excluding this influential dataset are also reported.

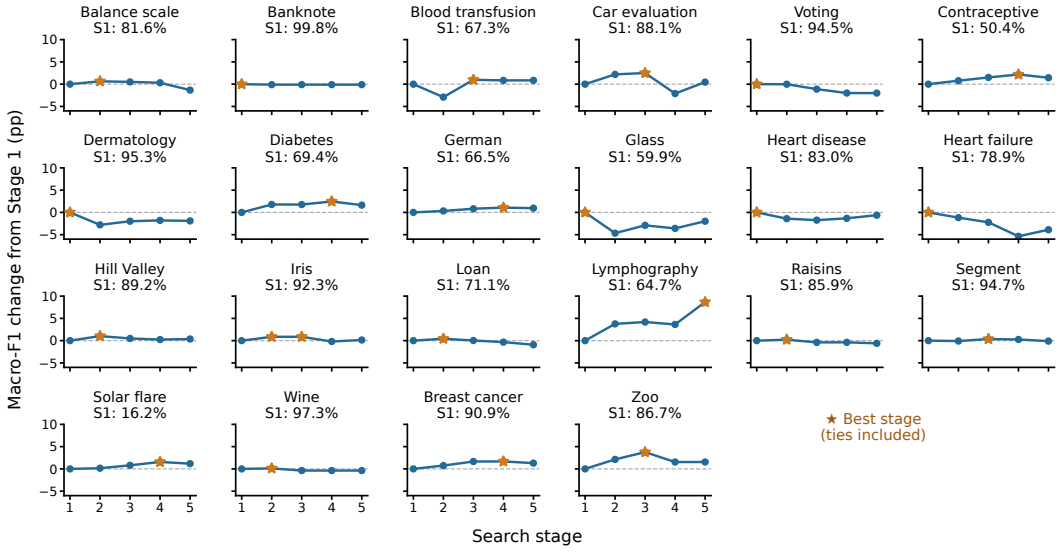


Figure 2: Dataset-level changes in Macro-F1 across search stages, relative to Stage 1 (percentage points). All panels use the same scale; titles report the Stage 1 score. Stars mark the best stage, including ties.

### 4.3 Proposed tree versus tuned ensembles

Table 2 presents the comparison of the final tree obtained after completing Stage 5 and the Stage 3 tree, which achieved the highest average Macro-F1 among the five search stages, with four tuned tree ensembles and their mean. With Hill Valley included, Stage 3 and Stage 5 were 0.07 and 0.28 percentage points below the best individual ensemble, respectively. The mean Macro-F1 of the four ensembles was 78.44%; Stage 3 exceeded this mean by 0.34 points, and Stage 5 exceeded it by 0.13 points. Accuracy and log loss favored the ensembles. Hill Valley had a substantial influence on the aggregate comparison: Stage 3 achieved 89.74% Macro-F1, compared with 58.49% for the best ensemble, a difference of 31.25 percentage points. We therefore also reported the comparison without Hill Valley. In this sensitivity analysis, Stage 3 and Stage 5 trailed the mean ensemble Macro-F1 of 79.51% by 1.25 and 1.47 points, respectively.

The Friedman tests indicated significant differences among the six methods. However, after Holm correction, the pairwise Macro-F1 differences between the proposed trees and the ensembles were not significant, either with or without Hill Valley. Accuracy and log loss generally favored the ensembles, with stronger differences after excluding Hill Valley. Complete test statistics, adjusted $p$-values, and win/tie/loss counts are reported in Appendix E.

| Metric | Stage 3 | Stage 5 | CART | ExtraTree | DPDT | STreeD |
|---|---|---|---|---|---|---|
| Macro-F1 (%) | **77.42** | **77.21** | 73.88 | 72.49 | **74.88** | 74.06 |
| Accuracy (%) | **84.26** | **84.00** | 82.91 | 82.02 | 82.69 | **82.99** |
| F1 rank ↓ | **2.50** | **2.95** | 3.90 | 4.40 | **3.55** | 3.70 |
| Accuracy rank ↓ | **2.75** | **3.15** | **3.38** | 4.15 | 4.00 | 3.58 |

Table 3: Single-tree comparison result.

| Metric | Stage 3 | Stage 5 | SPLIT |
|---|---|---|---|
| Macro-F1 (%) | **80.71** | **80.40** | **80.31** |
| Accuracy (%) | **83.67** | **83.40** | **84.19** |

Table 4: Matched comparison with SPLIT on ten supported binary datasets.

### 4.4 Proposed tree versus single trees

We compared the proposed trees with CART, ExtraTree, DPDT, and STreeD using the same outer splits and HPO budget. As shown in Table 3, Stage 3 achieved the highest Macro-F1 and accuracy across the 20 datasets, exceeding the strongest baselines by 2.54 and 1.27 percentage points, respectively. Stage 5 followed closely, consistent with the dataset-dependent stage behavior in Section 4.2.

Because SPLIT supports only binary classification, Table 4 compares the methods on ten supported binary datasets. Stage 3 exceeded SPLIT by 0.40 points in Macro-F1, whereas SPLIT led Stage 3 by 0.52 points in accuracy. These small, statistically non-significant differences indicate a metric-dependent trade-off. Overall, the proposed search outperformed the conventional single-tree baselines and remained competitive with SPLIT.

### 4.5 Loss ablation

The ablation fixed the selected depth, stage multiplier, guided-split probability, refinement budget, beam width, preprocessing, data split, and random seed. It removed only the Brier term, only the leaf penalty, or both terms while holding the remaining settings constant. Table 5 reports the resulting Macro-F1, accuracy, and log loss.

The full objective achieved the highest Macro-F1 and accuracy. Removing both the Brier term and the leaf penalty slightly improved standard log loss but reduced Macro-F1 by 0.39 percentage points. These results suggest that the auxiliary terms primarily guided the discrete search toward trees with stronger hard-label predictions and lower structural complexity, rather than uniformly improving every probability-based metric.

### 4.6 Stability under new tree-search seeds

We fixed each dataset/run's HPO-selected configuration and changed only the tree-search seed. Across 220 reruns, mean Macro-F1 changed from 78.57% to 78.44% and accuracy from 84.73% to 84.86%. Mean absolute per-run changes were 4.13 and 2.77 percentage points, respectively, indicating variation across runs despite stable aggregate performance.

Original and rerun predictions matched on 90.81% of held-out examples (correlation 0.849; Cohen's $\kappa = 0.791$). Across all rows, agreement was 92.89% and correlation was 0.882. Structural similarity was lower: feature Jaccard, weighted feature Jaccard, and path-feature match were 0.276, 0.246, and 0.083, respectively. Feature-based structures matched exactly in 1.82% of paired trees; rounded trees matched in 0.45%. Comparisons among the ten original seeds showed similar prediction agreement (90.87%; correlation 0.847), despite low feature Jaccard (0.219) and path match (0.041). Predictions were therefore more stable than tree structures.

### 4.7 Effect of the shared feature representation on ensembles

Table 6 isolates the effect of the input representation. We trained RF, XGB, LGB, and CAT with the augmented feature dictionary used by our method. With Hill Valley included, the augmented input

| Objective | Macro-F1 (%) | Accuracy (%) | Log loss |
|---|---|---|---|
| Full loss | **78.57** | **84.73** | **0.424** |
| No Brier | 77.68 | 84.59 | 0.425 |
| No leaf penalty | 77.99 | 84.51 | 0.430 |
| Base: $1 - \mathrm{MacroF1}$ | **78.18** | **84.65** | **0.423** |

Table 5: Stage-5 loss ablation over all 22 datasets. Each value is dataset-balanced over ten repetitions.

| Model | Macro-F1 | ΔF1 | Accuracy | ΔAcc. | Log loss ↓ | ΔLL |
|---|---|---|---|---|---|---|
| *Hill Valley included (22 datasets)* | | | | | | |
| RF | 80.16 | +2.29 | 86.39 | +1.14 | 0.481 | +0.088 |
| XGB | 80.14 | +1.80 | 86.28 | +1.10 | 0.441 | +0.068 |
| LGB | 80.62 | +1.91 | 86.54 | +1.11 | 0.480 | +0.100 |
| CAT | 81.11 | +2.26 | 87.10 | +1.46 | 0.341 | −0.030 |
| **Ensemble mean** | **80.51** | **+2.07** | **86.58** | **+1.20** | **0.436** | **+0.057** |
| *Hill Valley excluded (21 datasets)* | | | | | | |
| RF | 79.72 | +0.78 | 86.25 | −0.42 | 0.488 | +0.110 |
| XGB | 79.43 | +0.14 | 85.87 | −0.58 | 0.456 | +0.104 |
| LGB | 79.93 | +0.16 | 86.13 | −0.67 | 0.496 | +0.141 |
| CAT | 80.41 | +0.36 | 86.69 | −0.46 | 0.351 | +0.004 |
| **Ensemble mean** | **79.87** | **+0.36** | **86.23** | −0.54 | **0.448** | **+0.090** |

Table 6: Ensemble performance with the augmented input used by our method. Δ denotes augmented minus original input. Values are dataset-balanced means over ten repetitions per dataset.

increased the ensemble mean Macro-F1 by 2.47 percentage points and accuracy by 1.75 points. Hill Valley had a large influence: after excluding it, the mean Macro-F1 gain was 0.70 points and mean accuracy changed by $-0.05$ points. The augmented features therefore improved classification on average, but most of the aggregate gain came from Hill Valley, and log loss did not improve consistently. This control indicates that the feature representation itself contributed to the predictive results and that the full improvement should not be attributed to tree search alone.

## 5 Limitations and Future Work

Our method requires more training computation than greedy decision-tree induction because it evaluates many candidate structures. Its stochastic search can also produce different tree structures across runs, although their predictions are generally similar. In addition, transformed feature directions may reduce the simplicity of individual decision rules.

Future work will focus on reducing the search cost, improving structural stability, and constraining feature transformations. Extending the method to regression, larger datasets, and deeper trees is another important direction.

## 6 Conclusion

In this paper, we introduce a loss-guided stochastic search method that constructs a single decision tree by evaluating the downstream effects of candidate splits. Across 22 datasets, the final trees obtained after five search stages achieved a mean Macro-F1 of 78.57%. Their pairwise Macro-F1 differences from the four ensemble baselines were not statistically significant after Holm correction. The best search stage varied across datasets, and additional stages did not consistently improve performance. Moreover, repeated searches with the same hyperparameters often produced different tree structures while maintaining similar predictive performance. These results support assessing splits through their downstream consequences and show that the proposed search can produce diverse tree structures with comparable predictive quality.

## Reproducibility Statement

We provide the experimental protocol, data splits, preprocessing steps, hyperparameter search spaces, and implementation details in the appendix. Per-split results and selected configurations were recorded for every experiment, and all reported tables were generated from these records. The accompanying code includes scripts for reproducing the main results and stage-wise analyses.

## Statement on the use of AI

AI tools were utilized solely for language editing, text polishing, and code debugging. All conceptualization, experimental design, analytical decisions, and empirical validations were conducted exclusively by the authors.

## References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019. doi: 10.1145/3292500.3330701.

Varun Babbar, Hayden McTavish, Cynthia Rudin, and Margo Seltzer. Near-optimal decision trees in a SPLIT second. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2114–2175, 2025. URL https://proceedings.mlr.press/v267/babbar25a.html.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106:1039–1082, 2017. doi: 10.1007/s10994-017-5633-9.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

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

Ayman Chaouki, Jesse Read, and Albert Bifet. Branches: Efficiently seeking optimal sparse decision trees via AO*. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7430–7484, 2025. URL https://proceedings.mlr.press/v267/chaouki25a.html.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

David A. Cieslak, T. Ryan Hoens, Nitesh V. Chawla, and W. Philip Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1): 136–158, 2012. doi: 10.1007/s10618-011-0222-1.

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

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Laurent Hyafil and Ronald L. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, 1976. doi: 10.1016/0020-0190(76)90095-8.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hector Kohler, Riad Akrour, and Philippe Preux. Breiman meets bellman: Non-greedy decision trees with MDPs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2025. doi: 10.1145/3711896.3736868.

Qiangqiang Mao, Jiayang Ren, Yixiu Wang, Chenxuanyin Zou, Jingjing Zheng, and Yankai Cao. Differentiable decision tree via "ReLU+Argmin" reformulation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sascha Marton, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. GradTree: Learning axis-aligned decision trees with gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14323–14331, 2024. doi: 10.1609/aaai.v38i13.29345.

Subrat Panda, Blaise Genest, Arvind Easwaran, and Ponnuthurai N. Suganthan. Vanilla gradient descent for oblique decision trees. In *ECAI 2024*, pp. 1140–1147, 2024. doi: 10.3233/FAIA240607.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986. doi: 10.1007/BF00116251.

Jacobus G. M. van der Linden, Mathijs M. de Weerdt, and Emir Demirović. Necessary and sufficient conditions for optimal decision trees using dynamic programming. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9173–9212, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1d5fce9627e15c84db572a66e029b1fc-Paper-Conference.pdf.

# A DATASETS AND COMPLETED COHORTS

The primary stage, ensemble, loss-ablation, pruning, and seed-stability cohorts contain 22 datasets with ten completed outer runs each: Balance Scale, Banknote Authentication, Blood Transfusion, Car Evaluation, Congressional Voting Records, Contraceptive Method Choice, Dermatology, Diabetes, German Credit, Glass, Heart Disease, Heart Failure, Hill Valley, Iris, Loan Approval, Lymphography, Raisins, Segment, Solar Flare, Wine, Wisconsin Breast Cancer, and Zoo. Every primary comparison therefore contains 220 matched dataset–run pairs. SPLIT is evaluated only on the ten binary datasets.

# B DETAILED EXPERIMENTAL SETTINGS

## B.1 OUTER SPLIT, HPO, AND FINAL REFIT

Outer splits use stratified 80/20 train/test partitions with random states 1–10. Within each outer-training partition, a single class-preserving 70/30 holdout is fixed using the outer-run seed. The proposed tree, RF, XGB, LGB, CAT, and each single-tree baseline receive 64 tree-structured Parzen estimator (TPE) trials with four startup trials and no trial pruning. The proposed-tree search jointly tunes maximum depth, stage multiplier, guided-split probability $q$, refinement budget, beam width, $\lambda_{\mathrm{B}}$, and $\lambda_{\mathrm{L}}$. Validation Macro-F1 selects the configuration, with validation accuracy and then trial order used as tie breakers. The selected configuration is refitted without reading the test set; a stored method lock records the split, configuration, source hash, and evaluation rule before test evaluation.

For the proposed tree, the HPO inner fit is divided into 42% tree-construction and 14% structure-validation portions of the full data; preprocessing and learned directions are fitted on grow observations. The final refit uses 60% tree-construction and 20% structure-validation data, after which leaf counts are recalibrated on outer-train80. Ensemble and single-tree baselines refit on outer-train80.

## B.2 PROPOSED-TREE SEARCH SPACE AND SELECTED CONFIGURATIONS

| Parameter | Search space | Selections over 220 runs |
|---|---|---|
| Maximum depth | $\{3, 4, 5, 6\}$ | 42, 43, 58, 77 |
| Stage multiplier | $\{1, 2\}$ | 94, 126 |
| Guided probability | $\{0, .01, .025, .05\}$ | 46, 50, 60, 64 |
| Refinement budget | $\{32, 64, 128, 256\}$ | 57, 46, 56, 61 |
| Beam width | $\{1, 2, 4\}$ | 72, 72, 76 |
| $\lambda_{\mathrm{B}}$ | continuous $[0, .5]$ | mean .245, SD .150 |
| $\lambda_{\mathrm{L}}$ | continuous $[0, .08]$ | mean .041, SD .023 |

Table 7: HPO search space and empirical selections. Counts follow the order of the listed discrete values. The coefficient ranges are continuous uniform distributions.

There are five complete-tree stage pools. For $n$ tree-construction plus structure-validation rows and $p$ input features, the base stage budget is

$$B_0(n, p) = \min\Big\{2048, \max\Big[64, 64\left\lceil \sqrt{n/100}\log_2(p+1)\right\rceil\Big]\Big\}.$$

A selected stage multiplier $m \in \{1, 2\}$ gives $B_s = \min(8192, mB_0)$ at every stage; it changes the number of complete candidates evaluated per stage, not the number of stages. At each unconstrained node, the guided branch in Equation 5 is used with probability $q \in \{0, .01, .025, .05\}$ and random exploration with probability $1 - q$. Guided proposals draw independently and uniformly from entropy, Gini, twoing, and Hellinger. Root assessment considers at most 30 distinct elite features and up to 64 continuation trees per feature, subject to plateau stopping. Leaf probabilities use additive smoothing $\alpha = 0.5$. The construction objective is Equation 4, with the common complexity denominator $2^6 - 1 = 63$.

### B.3 Ensemble search spaces

RF searches 100–600 trees, unrestricted or bounded depth, minimum leaf sizes 1–5, and several feature-sampling fractions. XGB searches 100–600 trees, depth 2–10, learning rate 0.02–0.3, row and feature fractions 0.7–1.0, $L_2$ regularization 0.01–10, and minimum child weight 1–5. LGB searches 100–600 trees, unrestricted or bounded depth, 7–63 leaves, learning rate 0.02–0.2, minimum child samples 5–50, feature fractions 0.7–1.0, and $L_2$ regularization 0–10. CAT searches 200–1000 iterations, depth 3–10, learning rate 0.02–0.2, leaf regularization 1–20, and random strength 0–2.

### B.4 Single-tree search spaces

CART and ExtraTree search depths $\{3, 4, 5, 6, 8, 10, 12\}$, minimum leaf sizes 1–12, and Gini or entropy criteria; CART also tunes cost-complexity pruning in $[0, .03]$, while ExtraTree tunes the feature fraction. DPDT uses the same depth range, node layouts $\{8\text{–}3, 16\text{–}4, 32\text{–}8\}$, minimum leaf sizes 1–8, and Gini or entropy. STreeD searches depth 3–6, 3/5/10 thresholds, minimum leaf sizes 1/2/5, and log-uniform cost complexity $10^{-4}$–0.05. SPLIT searches depth 3–6, lookahead 2/3, and log-uniform regularization $10^{-4}$–0.03. SPLIT is binary-only in this experiment.

## C Search implementation details

**Root and prefix selection.** The initial pool is ranked by whole-tree loss. Root proposals are collected from the best 30 trees. If a root sends at most one construction observation to one child, the split of the other child is considered instead, when available. Proposals sharing a feature are combined by averaging their thresholds. Each resulting root is assessed with up to 64 sampled continuations, and its score is the mean loss at or below the fifth percentile. The root with the lowest score supplies the first fixed split. Subsequent prefix steps use the best 30 trees in the current pool, select the most frequent available feature in each region, and average the associated thresholds. The implementation includes child substitution when a split leaves a nearly empty region. Prefix extension stops at the smaller of four levels and the depth limit; remaining stages resample below the fixed prefix.

**Loss-consistent structural update.** The structural update minimizes the same $\mathcal{L}_V$ as candidate search and final refinement. It does not use a separate weighted combination of training and validation Macro-F1. At a visited node, each construction example is passed through both existing child subtrees. The resulting predictions define class-weighted error costs for routing the example left or right. These costs generate threshold proposals; they do not replace $\mathcal{L}_V$ when choosing a completed tree.

The update considers splits on existing features, new learned directions, and replacing a node by either of its children. A new direction is fitted by ridge regression on $G$: the target identifies the child with lower error cost, and the example weight is the absolute difference between the two costs. Its coefficients are fitted without validation labels. After changing a split or promoting a child, leaf counts are recomputed on $G$ and the completed tree is scored on $V$. An edit is accepted only if it reduces $\mathcal{L}_V$ by more than the numerical tolerance $10^{-12}$. Promoting a child can remove branches. The former separate validation-F1 pruning pass and binary threshold sweep are not performed by this update; binary routing-cost calculations use the fixed probability threshold $0.5$. Candidate selection uses the smoothed probabilities and argmax predictions in the whole-tree objective.

**Four-criterion refinement.** For each eligible node and each available feature, the procedure considers at most seven thresholds between consecutive distinct construction values. The candidate list has four rankings, one each for entropy, Gini, twoing, and Hellinger, plus a seeded random permutation. It alternates a criterion-ranked proposal with a random-order proposal. For each ranked turn, a criterion is drawn uniformly and the next unused split from its ranking is proposed. Already proposed feature–threshold pairs are skipped. This rule is separate from the probability $q$ used in source-tree generation; the realized fractions can differ from an exact half-and-half split as lists are exhausted.

Let $n_v$ be the number of construction examples reaching node $v$, $e_v$ their error count under its current subtree, and $h_v$ the number of proposals already drawn there. The adaptive rule chooses an available node maximizing

$$\frac{e_v + \sqrt{n_v}}{\sqrt{|G|}\sqrt{1+h_v}}.$$

For each split replacement, four descendant completions are grown, one with each criterion held fixed throughout that completion. Both child subtrees are rebuilt greedily within the remaining depth limit, stopping at pure nodes, singleton nodes, or regions without a valid split. These fits use the same restricted threshold candidates described above. Each completion is inserted into the full tree and evaluated by $\mathcal{L}_V$; the completion with the smallest loss is selected, with ties broken by total terminal-leaf count and then a deterministic tree fingerprint.

The refinement budget $R$ counts split replacements, not individual complete-tree evaluations. Each replacement evaluates four completed trees, including identical completions if different criteria produce the same structure. The initial tree is scored once separately. Each beam state attempts at most four replacements per round, subject to the remaining budget. After the four-way comparison, an already-seen winning tree is discarded. The next beam contains the best $b$ distinct trees from the current beam and the new candidates. The lowest-loss tree found so far is retained, so its loss cannot increase during refinement. The leaf-count tie breaker counts all terminal leaves, whereas the objective counts only leaves occupied by construction examples.

## D COMPUTATIONAL COST

Training consists of feature preparation, staged search, loss-based structural updates, final refinement, and leaf recalibration:

$$C_{\text{train}} = C_{\text{prep}} + C_{\text{source}} + C_{\text{update}} + C_{\text{refine}} + C_{\text{leaf}}.$$

Let $n_G = |G|$, $n_V = |V|$, $D$ be the maximum depth, $K$ the number of classes, and $L$ the number of terminal leaves. We count tree operations separately from feature fitting and transformation. With feature values available, evaluating $\mathcal{L}_V$ routes both $G$ and $V$, obtains construction-class counts, and computes validation predictions and the loss. Its routing and class-probability arithmetic cost is

$$C_{\mathcal{L}} = O((n_G + n_V)D + n_V K + LK)\,. \tag{7}$$

The same evaluation applies to binary and multiclass trees. Split edits, branch promotion, and pruning use this objective. Their cost is included in $C_{\text{update}}$, together with candidate construction and any additional feature-direction fits.

**Staged candidate search.** Let $p$ be the number of available features during source search and $q = 1 - \epsilon$ the probability of a criterion-guided split. A random proposal scans one selected feature, whereas a guided proposal searches features and thresholds. With bounded expected retries for invalid random splits and fixed $K$, a sorting-based approximation for constructing and scoring one complete candidate is

$$C_{\text{cand}} = O(n_G D + qpn_G D \log n_G) + C_{\mathcal{L}}. \tag{8}$$

This accounts for the different costs of random and guided proposals. Cached calculations and restricted threshold sets can reduce the work.

Let $S$ be the number of search stages, $B_s(a)$ the stage budget at multiplier $a$, $R_0$ the number of consolidated root proposals, and $H$ the maximum number of continuations per root. Then

$$M_{\text{source}}(a) \le \sum_{s=1}^{S} B_s(a) + HR_0, \qquad B_s(a) = \min(aB_0, B_{\max})\,. \tag{9}$$

The resulting source-search cost is approximately

$$C_{\text{source}} = O(M_{\text{source}}(a)[n_G D + (1-\epsilon)pn_G D \log n_G + C_{\mathcal{L}}])\,. \tag{10}$$

Increasing $a$ raises the stage-generation cost approximately linearly until the cap is reached; increasing $q$ raises the share of guided splits. These parameters do not scale feature preparation, structural updates, or refinement in the same way.

**Four-criterion refinement.** The refinement budget $R$ counts split-replacement proposals across the entire beam. For each replacement, descendants are grown separately using entropy, Gini, twoing, and Hellinger. All four completed trees are scored with $\mathcal{L}_V$. If $R_{\text{eff}} \leq R$ replacements are evaluated, the number of complete-tree loss evaluations in refinement is

$$M_{\text{refine}} = 1 + 4R_{\text{eff}} \leq 1 + 4R,$$

including the starting tree. Let $C_{\text{grow}}(r, c)$ be the cost of rebuilding both child subtrees for replacement $r$ with criterion $c$. We write

$$C_{\text{refine}} = C_{\text{enum}} + C_{\text{manage}} + C_{\mathcal{L}} + \sum_{r=1}^{R_{\text{eff}}} \sum_{c=1}^{4} \left[ C_{\text{grow}}(r, c) + C_{\mathcal{L}} \right]. \tag{11}$$

Here $C_{\text{enum}}$ includes threshold enumeration and the four criterion rankings; $C_{\text{manage}}$ includes copying, duplicate checks, and beam selection. The descendant builders search up to seven thresholds per available feature at each eligible node. Beam width $b$ changes the states retained and the enumeration and storage costs, but does not multiply the shared evaluation budget by $b$. The source probability $q$ does not reduce the four descendant fits per replacement.

Under the settings in Appendix B, $S = 5$, $B_0 \leq 2048$, $a \in \{1, 2\}$, $B_{\max} = 8192$, $H = 64$, $R_0 \leq 30$, and $R \leq 256$. Hence the configured stage budget is at most $4096$, giving at most $22{,}400$ source candidates and $1{,}024$ refinement variants, plus the refinement starting-tree evaluation. Structural updates, pruning, and repeated evaluations have separate costs; these counts describe the source and refinement budgets.

**Comparison with ensembles and prediction cost.** For an ensemble of $N$ trees using $n$ fitting rows, depth at most $D$, and $m$ features per split, a sorting-based approximation is $O(NmnD \log n)$ for fixed $K$. Feature subsampling reduces $m$, while histogram-based boosting uses bin-based split search and requires different accounting. Mostly random source trees can be cheaper to construct than greedy base trees, but our method explores many candidates and adds full-tree validation and refinement. These operation counts do not imply a wall-clock speed ratio.

After feature transformation, the selected tree requires at most $D$ routing tests per example, compared with at most $ND$ for $N$ trees of comparable depth. Computing $J$ dense learned directions from $d$ preprocessed inputs adds $O(dJ)$, giving $O(dJ + D + K)$ for feature transformation, routing, and a class-probability output. The implementation forms the feature dictionary before prediction; it does not assume on-demand evaluation of only the directions along the selected path.

## E STATISTICAL TESTS

The following tables report every prespecified pairwise comparison used in the main text. W/T/L counts are based on dataset means. Holm correction is applied within each metric, cohort, and comparison family as stated in each caption.

| Stage | Comparator | Adv. | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | RF | +0.92 | 6/0/16 | 0.1207 | 0.6327 |
| Stage3 | XGB | +0.44 | 6/0/16 | 0.1870 | 0.7048 |
| Stage3 | LGB | +0.07 | 7/0/15 | 0.1762 | 0.7048 |
| Stage3 | CAT | -0.07 | 6/0/16 | 0.0684 | 0.4791 |
| Stage5 | RF | +0.70 | 8/0/14 | 0.1762 | 0.7048 |
| Stage5 | XGB | +0.23 | 7/0/15 | 0.1982 | 0.7048 |
| Stage5 | LGB | -0.14 | 7/0/15 | 0.1055 | 0.6327 |
| Stage5 | CAT | -0.28 | 5/0/17 | 0.0462 | 0.3693 |

Table 8: 22 datasets: paired Macro-F1 comparisons. Advantage is Stage minus baseline for hard metrics and baseline minus Stage for log loss, so positive values favor the proposed tree. Friedman $p = 0.00464$. Holm correction covers the eight Stage-versus-ensemble comparisons for this metric and cohort.

| Stage | Comparator | Adv. | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | RF | -0.26 | 4/1/17 | 0.0093 | 0.0371 |
| Stage3 | XGB | -0.18 | 5/0/17 | 0.0172 | 0.0516 |
| Stage3 | LGB | -0.43 | 2/3/17 | 0.0035 | 0.0207 |
| Stage3 | CAT | -0.65 | 2/0/20 | 0.0012 | 0.0086 |
| Stage5 | RF | -0.52 | 5/2/15 | 0.0178 | 0.0516 |
| Stage5 | XGB | -0.45 | 7/0/15 | 0.0194 | 0.0516 |
| Stage5 | LGB | -0.69 | 5/1/16 | 0.0043 | 0.0214 |
| Stage5 | CAT | -0.91 | 3/0/19 | 0.0006 | 0.0047 |

Table 9: 22 datasets: paired accuracy comparisons. Advantage is Stage minus baseline for hard metrics and baseline minus Stage for log loss, so positive values favor the proposed tree. Friedman $p = 2.15e-06$. Holm correction covers the eight Stage-versus-ensemble comparisons for this metric and cohort.

| Stage | Comparator | Adv. | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | RF | -0.030 | 6/0/16 | 0.0542 | 0.1085 |
| Stage3 | XGB | -0.050 | 4/0/18 | 0.0066 | 0.0317 |
| Stage3 | LGB | -0.044 | 4/0/18 | 0.0032 | 0.0228 |
| Stage3 | CAT | -0.053 | 5/0/17 | 0.0053 | 0.0317 |
| Stage5 | RF | -0.031 | 6/0/16 | 0.0542 | 0.1085 |
| Stage5 | XGB | -0.051 | 4/0/18 | 0.0074 | 0.0317 |
| Stage5 | LGB | -0.044 | 3/0/19 | 0.0029 | 0.0228 |
| Stage5 | CAT | -0.053 | 5/0/17 | 0.0053 | 0.0317 |

Table 10: 22 datasets: paired log loss comparisons. Advantage is Stage minus baseline for hard metrics and baseline minus Stage for log loss, so positive values favor the proposed tree. Friedman $p = 6.03e-05$. Holm correction covers the eight Stage-versus-ensemble comparisons for this metric and cohort.

| Stage | Comparator | Adv. | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | RF | -0.68 | 5/0/16 | 0.0384 | 0.1922 |
| Stage3 | XGB | -1.03 | 5/0/16 | 0.0701 | 0.2586 |
| Stage3 | LGB | -1.51 | 6/0/15 | 0.0646 | 0.2586 |
| Stage3 | CAT | -1.79 | 5/0/16 | 0.0175 | 0.1228 |
| Stage5 | RF | -0.90 | 7/0/14 | 0.0646 | 0.2586 |
| Stage5 | XGB | -1.25 | 6/0/15 | 0.0760 | 0.2586 |
| Stage5 | LGB | -1.72 | 6/0/15 | 0.0319 | 0.1915 |
| Stage5 | CAT | -2.01 | 4/0/17 | 0.0101 | 0.0810 |

Table 11: Without Hill Valley: paired Macro-F1 comparisons. Advantage is Stage minus baseline for hard metrics and baseline minus Stage for log loss, so positive values favor the proposed tree. Friedman $p = 0.000615$. Holm correction covers the eight Stage-versus-ensemble comparisons for this metric and cohort.

| Stage | Comparator | Adv. | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | RF | -1.91 | 3/1/17 | 0.0010 | 0.0040 |
| Stage3 | XGB | -1.68 | 4/0/17 | 0.0025 | 0.0074 |
| Stage3 | LGB | -2.03 | 1/3/17 | 0.0004 | 0.0027 |
| Stage3 | CAT | -2.39 | 1/0/20 | 0.0001 | 0.0009 |
| Stage5 | RF | -2.17 | 4/2/15 | 0.0035 | 0.0074 |
| Stage5 | XGB | -1.95 | 6/0/15 | 0.0039 | 0.0074 |
| Stage5 | LGB | -2.30 | 4/1/16 | 0.0006 | 0.0029 |
| Stage5 | CAT | -2.65 | 2/0/19 | 0.0000 | 0.0001 |

Table 12: Without Hill Valley: paired accuracy comparisons. Advantage is Stage minus baseline for hard metrics and baseline minus Stage for log loss, so positive values favor the proposed tree. Friedman $p = 7.03e-08$. Holm correction covers the eight Stage-versus-ensemble comparisons for this metric and cohort.

| Stage | Comparator | Adv. | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | RF | -0.051 | 5/0/16 | 0.0127 | 0.0254 |
| Stage3 | XGB | -0.077 | 3/0/18 | 0.0006 | 0.0026 |
| Stage3 | LGB | -0.074 | 3/0/18 | 0.0002 | 0.0014 |
| Stage3 | CAT | -0.082 | 4/0/17 | 0.0004 | 0.0026 |
| Stage5 | RF | -0.051 | 5/0/16 | 0.0127 | 0.0254 |
| Stage5 | XGB | -0.077 | 3/0/18 | 0.0007 | 0.0026 |
| Stage5 | LGB | -0.074 | 2/0/19 | 0.0002 | 0.0013 |
| Stage5 | CAT | -0.082 | 4/0/17 | 0.0004 | 0.0026 |

Table 13: Without Hill Valley: paired log loss comparisons. Advantage is Stage minus baseline for hard metrics and baseline minus Stage for log loss, so positive values favor the proposed tree. Friedman $p = 4.09e - 06$. Holm correction covers the eight Stage-versus-ensemble comparisons for this metric and cohort.

| Stage | Comparator | Adv. (pp) | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | CART | +3.54 | 15/0/5 | 0.0172 | 0.1031 |
| Stage3 | ExtraTree | +4.93 | 17/0/3 | 0.0004 | 0.0032 |
| Stage3 | DPDT | +2.54 | 13/0/7 | 0.0484 | 0.1938 |
| Stage3 | STreeD | +3.36 | 12/0/8 | 0.0362 | 0.1812 |
| Stage5 | CART | +3.33 | 13/0/7 | 0.0637 | 0.1938 |
| Stage5 | ExtraTree | +4.73 | 15/0/5 | 0.0017 | 0.0118 |
| Stage5 | DPDT | +2.34 | 13/0/7 | 0.1054 | 0.1938 |
| Stage5 | STreeD | +3.15 | 13/0/7 | 0.0532 | 0.1938 |

Table 14: Single-tree Macro-F1 comparisons on 20 complete common datasets. Friedman $p = 0.021$; Holm correction covers eight comparisons.

| Stage | Comparator | Adv. (pp) | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | CART | +1.35 | 11/0/9 | 0.2611 | 0.8072 |
| Stage3 | ExtraTree | +2.24 | 15/0/5 | 0.0107 | 0.0855 |
| Stage3 | DPDT | +1.57 | 13/1/6 | 0.0545 | 0.3269 |
| Stage3 | STreeD | +1.27 | 12/0/8 | 0.1769 | 0.8072 |
| Stage5 | CART | +1.09 | 11/0/9 | 0.5706 | 0.8072 |
| Stage5 | ExtraTree | +1.98 | 15/0/5 | 0.0240 | 0.1677 |
| Stage5 | DPDT | +1.31 | 12/1/7 | 0.1614 | 0.8072 |
| Stage5 | STreeD | +1.01 | 12/0/8 | 0.2322 | 0.8072 |

Table 15: Single-tree accuracy comparisons on 20 complete common datasets. Friedman $p = 0.156$; Holm correction covers eight comparisons.

| Stage | Comparator | Adv. (pp) | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | SPLIT | +0.40 | 6/0/4 | 0.7695 | 1.0000 |
| Stage5 | SPLIT | +0.09 | 4/0/6 | 1.0000 | 1.0000 |

Table 16: SPLIT-subset Macro-F1 comparisons on ten binary datasets. Holm correction covers the two Stage-versus-SPLIT tests.

| Stage | Comparator | Adv. (pp) | W/T/L | $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Stage3 | SPLIT | -0.52 | 4/0/6 | 0.5566 | 0.9844 |
| Stage5 | SPLIT | -0.79 | 3/0/7 | 0.4922 | 0.9844 |

Table 17: SPLIT-subset accuracy comparisons on ten binary datasets. Holm correction covers the two Stage-versus-SPLIT tests.

| Dataset | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Balance | 81.57 | **82.21** | **82.08** | **81.89** | 80.24 |
| Banknote | **99.78** | **99.67** | **99.67** | **99.67** | **99.67** |
| Blood | **67.33** | 64.41 | **68.31** | **68.18** | **68.18** |
| Car | 88.15 | **90.34** | **90.65** | 86.04 | **88.60** |
| Voting | **94.49** | **94.48** | **93.38** | 92.50 | 92.50 |
| Contraceptive | 50.36 | 51.14 | **51.87** | **52.54** | **51.81** |
| Dermatology | **95.35** | 92.55 | 93.36 | **93.55** | **93.44** |
| Diabetes | 69.41 | **71.19** | **71.17** | **71.87** | 71.04 |
| German | 66.51 | 66.84 | **67.33** | **67.59** | **67.46** |
| Glass | **59.85** | 55.19 | **56.95** | 56.26 | **57.87** |
| Heart disease | **83.01** | 81.60 | 81.27 | **81.68** | **82.38** |
| Heart failure | **78.94** | **77.77** | **76.72** | 73.58 | 75.06 |
| Hill Valley | 89.25 | **90.27** | **89.74** | 89.49 | **89.62** |
| Iris | **92.26** | **93.13** | **93.13** | 92.07 | **92.41** |
| Loan | **71.13** | **71.56** | **71.17** | 70.80 | 70.21 |
| Lymphography | 64.73 | **68.49** | **68.91** | 68.37 | **73.39** |
| Raisins | **85.88** | **86.10** | 85.49 | **85.50** | 85.27 |
| Segment | **94.73** | 94.64 | **95.09** | **94.99** | 94.63 |
| Solar flare | 16.19 | 16.35 | **17.00** | **17.75** | **17.37** |
| Wine | **97.32** | **97.44** | **96.95** | **96.95** | **96.95** |
| Breast cancer | 90.89 | 91.64 | **92.55** | **92.56** | **92.19** |
| Zoo | 86.66 | **88.79** | **90.42** | **88.20** | **88.20** |

Table 18: Dataset-level Macro-F1 across the five saved stages. Red, blue, and green mark the highest, second-highest, and third-highest distinct values within each dataset.

| Dataset | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Balance | **90.64** | **90.80** | **90.80** | **90.72** | 90.00 |
| Banknote | **99.78** | **99.67** | **99.67** | **99.67** | **99.67** |
| Blood | **77.92** | **77.58** | **78.26** | **78.26** | **78.26** |
| Car | **95.04** | **95.57** | **95.33** | 94.12 | 94.90 |
| Voting | **94.57** | **94.57** | **93.48** | **92.61** | **92.61** |
| Contraceptive | 52.07 | 52.55 | **52.99** | **54.12** | **53.44** |
| Dermatology | **96.34** | 95.07 | **95.21** | **96.06** | **96.06** |
| Diabetes | 71.69 | **73.44** | **73.44** | **74.09** | **73.38** |
| German | 72.35 | 72.25 | **72.45** | **72.65** | **72.70** |
| Glass | **65.71** | **63.57** | **64.52** | 62.86 | 63.33 |
| Heart disease | **83.33** | **81.85** | 81.48 | **81.85** | **82.59** |
| Heart failure | **82.88** | **81.36** | **80.51** | 77.46 | 79.15 |
| Hill Valley | 89.26 | **90.29** | **89.75** | 89.51 | **89.63** |
| Iris | **92.33** | **93.33** | **93.33** | **92.33** | **92.67** |
| Loan | **78.52** | **78.61** | **78.77** | 78.44 | 77.54 |
| Lymphography | 78.28 | 79.31 | **81.72** | **82.07** | **81.03** |
| Raisins | **85.94** | **86.17** | **85.56** | **85.56** | 85.33 |
| Segment | **94.55** | 94.48 | **94.94** | **94.85** | 94.46 |
| Solar flare | 82.06 | **82.82** | 82.71 | **83.03** | **83.29** |
| Wine | **97.43** | **97.43** | **96.86** | **96.86** | **96.86** |
| Breast cancer | 91.50 | **92.21** | **93.10** | **93.10** | **92.74** |
| Zoo | **94.00** | **95.00** | **95.00** | **94.50** | **94.50** |

Table 19: Dataset-level accuracy across the five saved stages. Red, blue, and green mark the highest, second-highest, and third-highest distinct values within each dataset.

## F ALL DATASET-LEVEL RESULTS

This section reports the complete dataset-level results in the same order as the analyses in the main text. Tables 18–20 compare the five saved stages; Tables 21–23 compare Stage 3 and Stage 5 with the ensembles; and Tables 24 and 25 give the single-tree comparison. Tables 26–28 report the loss ablation, and Table 29 reports the seed-stability results. Macro-F1 and accuracy are percentages; losses are in raw units. Red, blue, and green mark the three best distinct values within each dataset. A red dash denotes an unavailable result; no value is copied from another protocol.

| Dataset | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Balance | 0.339 | **0.311** | **0.331** | **0.323** | 0.335 |
| Banknote | **0.013** | **0.020** | **0.018** | **0.018** | **0.018** |
| Blood | **0.490** | 0.496 | **0.491** | **0.489** | **0.489** |
| Car | 0.187 | **0.162** | **0.171** | 0.214 | **0.171** |
| Voting | **0.195** | **0.192** | **0.238** | 0.254 | 0.254 |
| Contraceptive | 0.984 | **0.965** | 0.967 | **0.961** | **0.962** |
| Dermatology | **0.180** | 0.236 | 0.221 | **0.191** | **0.200** |
| Diabetes | 0.575 | **0.550** | **0.548** | **0.545** | 0.559 |
| German | 0.586 | 0.583 | **0.583** | **0.571** | **0.578** |
| Glass | **1.033** | **1.051** | 1.081 | 1.091 | **1.059** |
| Heart disease | **0.473** | **0.482** | 0.507 | 0.491 | **0.472** |
| Heart failure | **0.454** | 0.506 | **0.493** | 0.537 | **0.498** |
| Hill Valley | 0.320 | **0.296** | **0.309** | 0.310 | **0.309** |
| Iris | **0.267** | **0.262** | **0.263** | 0.282 | 0.275 |
| Loan | 0.524 | **0.522** | **0.522** | **0.519** | 0.523 |
| Lymphography | 0.723 | 0.664 | **0.619** | **0.617** | **0.634** |
| Raisins | **0.388** | **0.390** | 0.406 | **0.401** | 0.407 |
| Segment | 0.215 | 0.214 | **0.199** | **0.196** | **0.212** |
| Solar flare | 0.608 | **0.595** | 0.596 | **0.593** | **0.592** |
| Wine | **0.127** | **0.129** | **0.157** | **0.157** | **0.157** |
| Breast cancer | 0.282 | 0.253 | **0.232** | **0.230** | **0.245** |
| Zoo | 0.388 | **0.355** | **0.365** | **0.375** | **0.375** |

Table 20: Dataset-level log loss across the five saved stages. Red, blue, and green mark the lowest, second-lowest, and third-lowest distinct values within each dataset.

| Dataset | S3 | S5 | RF | XGB | LGB | CAT |
|---|---|---|---|---|---|---|
| Balance | **82.08** | **80.24** | 59.70 | **76.66** | 68.57 | 71.01 |
| Banknote | **99.67** | **99.67** | 99.00 | **99.71** | 99.67 | **99.74** |
| Blood | **68.31** | **68.18** | **66.35** | 63.04 | 65.13 | 63.06 |
| Car | 90.65 | 88.60 | 93.64 | **97.30** | **98.53** | **98.88** |
| Voting | 93.38 | 92.50 | 95.79 | **96.88** | **96.01** | **96.46** |
| Contraceptive | 51.87 | 51.81 | 52.40 | **52.68** | **52.72** | **53.65** |
| Dermatology | 93.36 | 93.44 | 93.43 | **94.08** | **94.48** | **96.20** |
| Diabetes | **71.17** | **71.04** | **72.10** | 70.34 | 70.91 | 70.74 |
| German | 67.33 | 67.46 | **68.87** | **68.53** | 68.45 | **68.92** |
| Glass | 56.95 | 57.87 | **69.86** | 63.59 | **66.03** | **67.52** |
| Heart disease | 81.27 | **82.38** | **85.57** | 82.09 | 81.81 | **83.00** |
| Heart failure | 76.72 | 75.06 | 82.69 | **83.85** | **83.11** | **83.73** |
| Hill Valley | **89.74** | **89.62** | 55.31 | **58.49** | 56.50 | 53.70 |
| Iris | 93.13 | 92.41 | **95.23** | 93.66 | **94.92** | **94.19** |
| Loan | 71.17 | 70.21 | **72.60** | 71.64 | **72.18** | **72.07** |
| Lymphography | 68.91 | 73.39 | 72.79 | **73.63** | **81.95** | **78.45** |
| Raisins | 85.49 | 85.27 | **86.94** | 86.60 | **87.06** | **86.72** |
| Segment | 95.09 | 94.63 | 97.62 | **98.23** | **98.25** | **98.43** |
| Solar flare | **17.00** | **17.37** | **17.47** | 16.61 | 15.43 | 16.18 |
| Wine | 96.95 | 96.95 | 96.92 | **97.08** | **97.39** | **97.65** |
| Breast cancer | 92.55 | 92.19 | **93.84** | 93.82 | **94.21** | **94.70** |
| Zoo | **90.42** | 88.20 | 84.92 | 85.03 | **88.28** | **89.69** |

Table 21: Dataset-level Macro-F1 for Stage 3, Stage 5, and the tuned ensembles. Red, blue, and green mark the highest, second-highest, and third-highest distinct values within each dataset.

| Dataset | S3 | S5 | RF | XGB | LGB | CAT |
|---|---|---|---|---|---|---|
| Balance | **90.80** | 90.00 | 85.60 | **91.84** | 89.52 | **90.96** |
| Banknote | **99.67** | **99.67** | 99.01 | **99.71** | **99.67** | **99.74** |
| Blood | **78.26** | **78.26** | **79.66** | 77.05 | **78.93** | 77.65 |
| Car | 95.33 | 94.90 | 97.30 | **98.81** | **99.48** | **99.42** |
| Voting | 93.48 | 92.61 | 95.87 | **96.96** | **96.09** | **96.52** |
| Contraceptive | 52.99 | 53.44 | 54.69 | **55.20** | **55.07** | **56.05** |
| Dermatology | 95.21 | **96.06** | **95.92** | 95.63 | **95.92** | **97.18** |
| Diabetes | 73.44 | 73.38 | **75.52** | 73.31 | **74.09** | **73.64** |
| German | 72.45 | 72.70 | **76.75** | 75.35 | **75.40** | **75.95** |
| Glass | 64.52 | 63.33 | **74.52** | 71.19 | **73.81** | **75.95** |
| Heart disease | 81.48 | **82.59** | **85.93** | 82.41 | 82.22 | **83.33** |
| Heart failure | 80.51 | 79.15 | 86.10 | **86.78** | **86.44** | **86.95** |
| Hill Valley | **89.75** | **89.63** | 55.39 | **58.52** | 56.54 | 53.83 |
| Iris | 93.33 | 92.67 | **95.33** | 93.67 | **95.00** | **94.33** |
| Loan | 78.77 | 77.54 | **80.82** | 79.26 | **79.59** | **79.92** |
| Lymphography | 81.72 | 81.03 | 83.79 | **86.21** | **86.55** | **85.52** |
| Raisins | 85.56 | 85.33 | **87.00** | 86.67 | **87.11** | **86.78** |
| Segment | 94.94 | 94.46 | 97.58 | **98.18** | **98.20** | **98.38** |
| Solar flare | 82.71 | **83.29** | **83.14** | 82.31 | 82.71 | **83.21** |
| Wine | 96.86 | 96.86 | 96.86 | **97.14** | **97.43** | **97.71** |
| Breast cancer | 93.10 | 92.74 | **94.25** | **94.25** | **94.60** | **95.04** |
| Zoo | **95.00** | **94.50** | **94.50** | 93.50 | **95.00** | **96.00** |

Table 22: Dataset-level accuracy for Stage 3, Stage 5, and the tuned ensembles. Red, blue, and green mark the highest, second-highest, and third-highest distinct values within each dataset.

| Dataset | S3 | S5 | RF | XGB | LGB | CAT |
|---|---|---|---|---|---|---|
| Balance | 0.331 | 0.335 | 0.467 | **0.197** | **0.269** | **0.238** |
| Banknote | 0.018 | 0.018 | 0.038 | **0.014** | **0.012** | **0.008** |
| Blood | 0.491 | **0.489** | **0.477** | 0.555 | **0.487** | 0.500 |
| Car | 0.171 | 0.171 | 0.084 | **0.043** | **0.018** | **0.033** |
| Voting | 0.238 | 0.254 | 0.171 | **0.127** | **0.110** | **0.120** |
| Contraceptive | 0.967 | 0.962 | **0.933** | **0.935** | 0.977 | **0.903** |
| Dermatology | 0.221 | 0.200 | 0.225 | **0.152** | **0.165** | **0.125** |
| Diabetes | **0.548** | 0.559 | **0.496** | 0.577 | **0.557** | 0.643 |
| German | 0.583 | 0.578 | **0.504** | **0.556** | **0.553** | 0.583 |
| Glass | 1.081 | 1.059 | **0.695** | **0.865** | 0.887 | **0.733** |
| Heart disease | 0.507 | 0.472 | **0.392** | **0.437** | **0.415** | 0.441 |
| Heart failure | 0.493 | 0.498 | **0.358** | **0.356** | **0.369** | 0.396 |
| Hill Valley | **0.309** | **0.309** | **0.712** | 0.811 | 0.903 | 0.864 |
| Iris | 0.263 | 0.275 | 0.246 | **0.189** | **0.236** | **0.162** |
| Loan | 0.522 | 0.523 | **0.485** | **0.492** | 0.512 | **0.500** |
| Lymphography | 0.619 | 0.634 | 0.421 | **0.386** | **0.406** | **0.362** |
| Raisins | 0.406 | 0.407 | 0.358 | **0.330** | **0.335** | **0.338** |
| Segment | 0.199 | 0.212 | 0.104 | **0.059** | **0.059** | **0.052** |
| Solar flare | **0.596** | **0.592** | 0.969 | 0.687 | 0.660 | **0.637** |
| Wine | 0.157 | 0.157 | 0.160 | **0.098** | **0.069** | **0.085** |
| Breast cancer | 0.232 | 0.245 | 0.160 | **0.151** | **0.156** | **0.143** |
| Zoo | 0.365 | 0.375 | **0.193** | **0.194** | **0.204** | 0.287 |

Table 23: Dataset-level log loss for Stage 3, Stage 5, and the tuned ensembles. Red, blue, and green mark the lowest, second-lowest, and third-lowest distinct values within each dataset.

| Dataset | S3 | S5 | CART | Extra | DPDT | STreeD | SPLIT |
|---|---|---|---|---|---|---|---|
| Balance | **82.08** | **80.24** | **56.99** | 56.54 | 55.66 | 53.13 | – |
| Banknote | **99.67** | **99.67** | 97.95 | 98.36 | **98.37** | **98.52** | 96.48 |
| Blood | **68.31** | **68.18** | 65.20 | 56.90 | 61.90 | 63.42 | **66.53** |
| Car | **90.65** | 88.60 | 89.18 | **89.24** | **90.77** | 80.28 | – |
| Voting | 93.38 | 92.50 | **96.88** | **96.22** | **96.68** | 95.58 | **96.88** |
| Contraceptive | **51.87** | 51.81 | 49.61 | 47.87 | **53.28** | **52.79** | – |
| Dermatology | **93.36** | **93.44** | 86.09 | 88.56 | **90.63** | 90.20 | – |
| Diabetes | **71.17** | **71.04** | 70.06 | 66.06 | 67.13 | 69.87 | **71.10** |
| German | **67.33** | **67.46** | 57.82 | **62.34** | 60.71 | 62.26 | 62.05 |
| Glass | 56.95 | 57.87 | **63.35** | 46.09 | **63.08** | **67.70** | – |
| Heart disease | **81.27** | **82.38** | 78.52 | **78.64** | 78.53 | 77.33 | 76.37 |
| Heart failure | 76.72 | 75.06 | **78.66** | **77.82** | 77.60 | 77.68 | **82.21** |
| Hill Valley | **89.74** | **89.62** | – | – | – | – | – |
| Iris | 93.13 | 92.41 | 92.87 | **94.03** | **93.49** | **93.53** | – |
| Loan | 71.17 | 70.21 | **72.18** | 70.51 | 69.58 | **71.77** | **72.83** |
| Lymphography | **68.91** | **73.39** | 60.85 | 63.36 | **67.43** | 62.94 | – |
| Raisins | 85.49 | 85.27 | **85.71** | 82.95 | 85.49 | **86.14** | **86.17** |
| Segment | **95.09** | **94.63** | – | – | – | – | – |
| Solar flare | 17.00 | **17.37** | 16.08 | 16.49 | **17.08** | **17.15** | – |
| Wine | **96.95** | **96.95** | 90.30 | 89.54 | **93.20** | **92.94** | – |
| Breast cancer | **92.55** | **92.19** | 91.76 | 89.90 | 91.66 | 90.77 | **92.47** |
| Zoo | **90.42** | **88.20** | 77.59 | 78.29 | **85.27** | 77.22 | – |
| **20-dataset mean** | **77.42** | **77.21** | 73.88 | 72.49 | **74.88** | 74.06 | – |
| **10-binary mean** | **80.71** | **80.40** | 79.47 | 77.97 | 78.77 | 79.33 | **80.31** |

Table 24: Dataset-level Macro-F1 for the single-tree comparison. Red, blue, and green mark the highest, second-highest, and third-highest distinct values within each dataset. Red dashes indicate unavailable results. The final rows report the 20-dataset common mean and the matched mean over the ten binary datasets supported by SPLIT.

| Dataset | S3 | S5 | CART | Extra | DPDT | STreeD | SPLIT |
|---|---|---|---|---|---|---|---|
| Balance | **90.80** | **90.00** | 76.00 | 76.08 | **76.88** | 75.36 | – |
| Banknote | **99.67** | **99.67** | 97.99 | **98.39** | **98.39** | **98.54** | 96.53 |
| Blood | **78.26** | **78.26** | **78.72** | 75.97 | 76.31 | 77.85 | **79.33** |
| Car | 95.33 | 94.90 | **96.00** | **96.12** | **95.94** | 92.03 | – |
| Voting | 93.48 | 92.61 | **96.96** | **96.30** | **96.74** | 95.65 | **96.96** |
| Contraceptive | 52.99 | **53.44** | 53.37 | 50.99 | **54.97** | **55.31** | – |
| Dermatology | **95.21** | **96.06** | **93.24** | 92.96 | **93.24** | **93.24** | – |
| Diabetes | **73.44** | 73.38 | 73.12 | 70.65 | 70.91 | **73.83** | **74.42** |
| German | **72.45** | **72.70** | 68.90 | 70.50 | 68.65 | 71.15 | **71.40** |
| Glass | **64.52** | 63.33 | **67.86** | 59.76 | **67.86** | **73.33** | – |
| Heart disease | **81.48** | **82.59** | 78.89 | **79.44** | 78.89 | 77.78 | 77.04 |
| Heart failure | 80.51 | 79.15 | **82.88** | **83.39** | 81.19 | 81.69 | **85.93** |
| Hill Valley | **89.75** | **89.63** | – | – | – | – | – |
| Iris | **93.33** | 92.67 | 93.00 | **94.00** | **93.67** | **93.67** | – |
| Loan | 78.77 | 77.54 | **80.33** | 78.52 | 76.48 | **79.84** | **81.07** |
| Lymphography | **81.72** | **81.03** | 77.93 | **78.62** | 76.55 | 77.24 | – |
| Raisins | **85.56** | 85.33 | **85.78** | 83.00 | **85.56** | **86.22** | **86.22** |
| Segment | **94.94** | **94.46** | – | – | – | – | – |
| Solar flare | 82.71 | **83.29** | **83.65** | **83.25** | 81.99 | 82.38 | – |
| Wine | **96.86** | **96.86** | **90.29** | **90.29** | **92.86** | **92.86** | – |
| Breast cancer | **93.10** | **92.74** | 92.39 | 90.71 | 92.21 | 91.42 | **93.01** |
| Zoo | **95.00** | **94.50** | 91.00 | **91.50** | **94.50** | 90.50 | – |
| **20-dataset mean** | **84.26** | **84.00** | 82.91 | 82.02 | 82.69 | **82.99** | – |
| **10-binary mean** | **83.67** | 83.40 | **83.60** | 82.69 | 82.53 | 83.40 | **84.19** |

Table 25: Dataset-level accuracy for the single-tree comparison. Red, blue, and green mark the highest, second-highest, and third-highest distinct values within each dataset. Red dashes indicate unavailable results. The final rows report the 20-dataset common mean and the matched mean over the ten binary datasets supported by SPLIT.

| Dataset | Full | No Brier | No leaf | Base |
|---|---|---|---|---|
| Balance | 80.24 | **82.64** | **82.07** | **81.26** |
| Banknote | 99.67 | **99.81** | **99.81** | **99.93** |
| Blood | **68.18** | **67.44** | 67.20 | **67.68** |
| Car | **88.60** | 87.82 | **89.05** | **88.60** |
| Voting | 92.50 | **93.61** | **93.61** | **94.71** |
| Contraceptive | **51.81** | **52.15** | 50.96 | **51.91** |
| Dermatology | **93.44** | **92.44** | 92.40 | **95.01** |
| Diabetes | **71.04** | **70.76** | **69.24** | 68.74 |
| German | **67.46** | 67.38 | **67.96** | **67.49** |
| Glass | **57.87** | 52.65 | **60.55** | **55.22** |
| Heart disease | **82.38** | **81.17** | 80.65 | **81.34** |
| Heart failure | **75.06** | 73.79 | **75.54** | **76.68** |
| Hill Valley | **89.62** | 88.79 | **89.58** | **89.54** |
| Iris | **92.41** | **92.78** | 90.97 | **93.27** |
| Loan | 70.21 | **71.75** | **70.85** | **72.45** |
| Lymphography | **73.39** | 63.47 | **64.85** | **65.15** |
| Raisins | **85.27** | **86.54** | 85.20 | **85.88** |
| Segment | **94.63** | **95.19** | 94.25 | **95.03** |
| Solar flare | **17.37** | **16.09** | **16.00** | 15.74 |
| Wine | **96.95** | **97.24** | 96.01 | **96.92** |
| Breast cancer | **92.19** | **92.07** | **91.24** | 91.13 |
| Zoo | **88.20** | 83.35 | **87.68** | **86.37** |

Table 26: Dataset-level Macro-F1 for the loss ablation at Stage 5.

| Dataset | Full | No Brier | No leaf | Base |
|---|---|---|---|---|
| Balance | **90.00** | **90.56** | **90.64** | **90.64** |
| Banknote | **99.67** | **99.82** | **99.82** | **99.93** |
| Blood | **78.26** | **78.05** | **77.45** | **78.05** |
| Car | 94.90 | **95.04** | **95.36** | **94.96** |
| Voting | **92.61** | **93.70** | **93.70** | **94.78** |
| Contraceptive | **53.44** | **53.64** | 52.55 | **53.37** |
| Dermatology | **96.06** | 95.21 | **95.35** | **95.77** |
| Diabetes | **73.38** | **73.18** | **71.82** | 71.69 |
| German | **72.70** | 72.40 | **73.10** | **72.55** |
| Glass | **63.33** | **63.10** | **66.19** | 59.76 |
| Heart disease | **82.59** | **81.48** | 80.93 | **81.67** |
| Heart failure | **79.15** | 77.97 | **79.83** | **80.51** |
| Hill Valley | **89.63** | 88.81 | **89.59** | **89.55** |
| Iris | **92.67** | **93.00** | 91.67 | **93.33** |
| Loan | 77.54 | **79.10** | **78.36** | **79.51** |
| Lymphography | **81.03** | **78.28** | 77.93 | **80.34** |
| Raisins | **85.33** | **86.61** | 85.28 | **85.94** |
| Segment | **94.46** | **95.04** | 94.09 | **94.89** |
| Solar flare | **83.29** | **83.32** | **83.00** | 82.45 |
| Wine | **96.86** | **97.14** | **96.29** | **96.86** |
| Breast cancer | **92.74** | **92.57** | **91.86** | 91.68 |
| Zoo | **94.50** | **93.00** | **94.50** | **94.00** |

Table 27: Dataset-level accuracy for the loss ablation at Stage 5.

| Dataset | Full | No Brier | No leaf | Base |
|---|---|---|---|---|
| Balance | 0.335 | **0.321** | **0.307** | **0.318** |
| Banknote | 0.018 | **0.013** | **0.008** | **0.006** |
| Blood | **0.489** | **0.498** | 0.503 | **0.496** |
| Car | **0.171** | 0.180 | **0.153** | **0.180** |
| Voting | 0.254 | **0.223** | **0.228** | **0.194** |
| Contraceptive | **0.962** | **0.969** | 0.991 | **0.980** |
| Dermatology | **0.200** | 0.241 | **0.236** | **0.222** |
| Diabetes | **0.559** | **0.547** | **0.591** | 0.591 |
| German | 0.578 | **0.571** | **0.568** | **0.568** |
| Glass | **1.059** | **1.061** | **1.017** | 1.150 |
| Heart disease | **0.472** | **0.477** | 0.522 | **0.482** |
| Heart failure | **0.498** | 0.512 | **0.511** | **0.453** |
| Hill Valley | **0.309** | 0.339 | **0.310** | **0.309** |
| Iris | **0.275** | **0.260** | 0.291 | **0.229** |
| Loan | **0.523** | **0.515** | 0.526 | **0.508** |
| Lymphography | **0.634** | 0.683 | **0.676** | **0.628** |
| Raisins | 0.407 | **0.368** | **0.397** | **0.383** |
| Segment | **0.212** | **0.201** | 0.223 | **0.200** |
| Solar flare | **0.592** | **0.612** | **0.596** | 0.614 |
| Wine | **0.157** | **0.159** | 0.168 | **0.168** |
| Breast cancer | **0.245** | **0.238** | 0.271 | **0.263** |
| Zoo | 0.375 | **0.373** | **0.359** | **0.375** |

Table 28: Dataset-level log loss for the loss ablation at Stage 5.

| Dataset | Test agree | Pred. corr. | Feature Jac. | Path match | ΔF1 |
|---|---|---|---|---|---|
| Balance | 94.64 | 0.920 | 0.350 | 0.066 | +0.81 |
| Banknote | 99.67 | 0.993 | 0.360 | 0.117 | +0.19 |
| Blood | 91.48 | 0.830 | 0.389 | 0.131 | -2.03 |
| Car | 94.70 | 0.929 | 0.360 | 0.054 | -0.29 |
| Voting | 96.30 | 0.926 | 0.270 | 0.221 | +2.42 |
| Contraceptive | 72.72 | 0.591 | 0.283 | 0.048 | -1.36 |
| Dermatology | 95.92 | 0.951 | 0.188 | 0.142 | +1.32 |
| Diabetes | 86.36 | 0.727 | 0.248 | 0.026 | -1.69 |
| German | 85.30 | 0.706 | 0.217 | 0.067 | +0.06 |
| Glass | 70.00 | 0.640 | 0.189 | 0.021 | -4.60 |
| Heart disease | 87.04 | 0.741 | 0.192 | 0.045 | -0.83 |
| Heart failure | 86.44 | 0.729 | 0.264 | 0.025 | +0.77 |
| Hill Valley | 91.23 | 0.825 | 0.449 | 0.134 | +0.46 |
| Iris | 95.33 | 0.930 | 0.348 | 0.117 | +0.58 |
| Loan | 92.62 | 0.852 | 0.187 | 0.057 | +0.22 |
| Lymphography | 87.24 | 0.830 | 0.163 | 0.010 | +0.18 |
| Raisins | 94.72 | 0.894 | 0.351 | 0.134 | +1.17 |
| Segment | 94.39 | 0.935 | 0.292 | 0.047 | -0.62 |
| Solar flare | 95.70 | 0.951 | 0.230 | 0.016 | -0.76 |
| Wine | 98.00 | 0.970 | 0.327 | 0.300 | +0.62 |
| Breast cancer | 94.96 | 0.899 | 0.277 | 0.029 | -0.47 |
| Zoo | 93.00 | 0.918 | 0.146 | 0.027 | +1.06 |

Table 29: Per-dataset stability after changing only the stochastic tree-search seeds while reusing the selected HPO configuration. The last column is rerun minus original Macro-F1.

## G POST-PRUNING AT EACH STAGE

We applied the same post-pruning procedure independently to every saved stage tree. Table 30 summarizes the mean change after pruning. It removed 11–16% of leaves with small average changes in predictive performance. Stages 1 and 2 improved slightly in both hard metrics, whereas Stage 3 lost 0.17 Macro-F1 points. The result therefore reflects a compression tradeoff rather than a uniform accuracy gain.

| Quantity | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Leaves before | 9.99 | 9.64 | 9.53 | 9.32 | 9.42 |
| Leaves after | 8.41 | 8.25 | 8.33 | 8.21 | 8.35 |
| Leaf change (%) | -15.75 | -14.39 | -12.55 | -11.85 | -11.43 |
| $\Delta$ Macro-F1 (pp) | +0.03 | +0.08 | -0.17 | +0.00 | +0.01 |
| $\Delta$ accuracy (pp) | +0.07 | +0.08 | -0.04 | +0.03 | -0.01 |
| $\Delta$ log loss | -0.002 | -0.001 | +0.001 | -0.000 | -0.001 |

Table 30: Mean post-pruning changes at each stage. Negative log-loss changes indicate improvement.

Tables 31–33 give the corresponding dataset-level changes in Macro-F1, accuracy, and log loss.

| Dataset | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Balance | -0.63 | -0.25 | -0.25 | +0.37 | +0.70 |
| Banknote | -0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| Blood | +0.07 | +0.04 | +0.11 | +0.00 | +0.00 |
| Car | +0.04 | +0.16 | +0.01 | -0.07 | +0.04 |
| Voting | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| Contraceptive | +0.17 | +0.12 | +0.05 | +0.01 | -0.01 |
| Dermatology | +0.16 | -0.60 | -0.60 | +0.03 | +0.00 |
| Diabetes | +0.06 | -0.18 | +0.00 | +0.00 | +0.23 |
| German | +0.45 | +0.40 | +0.14 | -0.03 | +0.01 |
| Glass | +0.25 | +1.66 | -2.00 | +0.62 | +0.38 |
| Heart disease | -0.02 | -0.19 | +0.00 | +0.00 | +0.00 |
| Heart failure | -0.43 | -0.05 | -0.11 | +0.00 | +0.00 |
| Hill Valley | +0.08 | +0.00 | -0.21 | -0.09 | -0.00 |
| Iris | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| Loan | +0.21 | +0.18 | +0.04 | +0.14 | +0.00 |
| Lymphography | +0.33 | +0.22 | -0.34 | -0.34 | -0.61 |
| Raisins | +0.06 | +0.06 | +0.11 | +0.17 | +0.22 |
| Segment | -0.06 | -0.07 | -0.13 | -0.08 | -0.27 |
| Solar flare | -0.17 | +0.04 | -0.52 | -0.54 | -0.51 |
| Wine | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| Breast cancer | +0.09 | +0.19 | -0.09 | -0.09 | +0.00 |
| Zoo | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |

Table 31: Dataset-level change in Macro-F1 after post-pruning (pruned minus unpruned, percentage points). Positive values indicate improvement.

| Dataset | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Balance | -0.16 | +0.00 | +0.00 | +0.32 | +0.32 |
| Banknote | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| Blood | +0.07 | +0.07 | +0.13 | +0.00 | +0.00 |
| Car | -0.14 | -0.14 | -0.03 | -0.12 | +0.06 |
| Voting | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| Contraceptive | +0.20 | +0.14 | +0.07 | +0.03 | +0.00 |
| Dermatology | +0.14 | -0.14 | -0.14 | +0.00 | +0.00 |
| Diabetes | +0.06 | -0.19 | +0.00 | +0.00 | +0.26 |
| German | +0.35 | +0.45 | +0.15 | +0.00 | +0.05 |
| Glass | +0.48 | +0.71 | -0.71 | +0.48 | +0.00 |
| Heart disease | +0.00 | -0.19 | +0.00 | +0.00 | +0.00 |
| Heart failure | -0.34 | +0.00 | +0.00 | +0.00 | +0.00 |
| Hill Valley | +0.08 | -0.00 | -0.21 | -0.08 | +0.00 |
| Iris | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| Loan | +0.16 | +0.16 | +0.00 | +0.08 | +0.00 |
| Lymphography | +0.34 | +0.34 | -0.34 | -0.34 | -0.69 |
| Raisins | +0.06 | +0.06 | +0.11 | +0.17 | +0.22 |
| Segment | -0.04 | -0.06 | -0.13 | -0.09 | -0.24 |
| Solar flare | +0.18 | +0.29 | +0.36 | +0.29 | -0.14 |
| Wine | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |
| Breast cancer | +0.09 | +0.18 | -0.09 | -0.09 | +0.00 |
| Zoo | +0.00 | +0.00 | +0.00 | +0.00 | +0.00 |

Table 32: Dataset-level change in accuracy after post-pruning (pruned minus unpruned, percentage points). Positive values indicate improvement.

| Dataset | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Balance | +0.002 | +0.003 | +0.000 | -0.008 | -0.007 |
| Banknote | +0.001 | +0.000 | +0.000 | +0.000 | +0.000 |
| Blood | -0.001 | -0.001 | -0.003 | +0.000 | +0.000 |
| Car | +0.003 | +0.005 | +0.000 | +0.002 | -0.002 |
| Voting | -0.002 | +0.000 | +0.000 | +0.000 | +0.000 |
| Contraceptive | -0.004 | -0.004 | -0.000 | +0.001 | +0.001 |
| Dermatology | -0.006 | +0.005 | +0.004 | +0.006 | +0.004 |
| Diabetes | -0.001 | +0.001 | +0.000 | +0.000 | -0.005 |
| German | -0.011 | -0.010 | -0.008 | -0.004 | -0.004 |
| Glass | -0.012 | -0.007 | +0.001 | -0.010 | -0.007 |
| Heart disease | +0.005 | +0.011 | +0.006 | +0.007 | +0.006 |
| Heart failure | +0.000 | -0.011 | +0.003 | +0.000 | +0.000 |
| Hill Valley | +0.001 | +0.007 | +0.012 | +0.012 | +0.010 |
| Iris | +0.000 | +0.001 | +0.001 | +0.001 | +0.001 |
| Loan | -0.007 | -0.005 | -0.001 | -0.003 | +0.000 |
| Lymphography | -0.009 | -0.007 | +0.003 | +0.004 | -0.003 |
| Raisins | +0.006 | +0.004 | -0.003 | -0.004 | -0.006 |
| Segment | +0.006 | +0.008 | +0.007 | +0.005 | +0.005 |
| Solar flare | -0.012 | -0.004 | -0.008 | -0.008 | -0.004 |
| Wine | +0.000 | +0.000 | +0.000 | +0.000 | +0.000 |
| Breast cancer | -0.005 | -0.009 | -0.003 | -0.003 | -0.008 |
| Zoo | +0.000 | +0.000 | +0.000 | +0.000 | +0.000 |

Table 33: Dataset-level change in log loss after post-pruning (pruned minus unpruned). Negative values indicate improvement.

## H EXAMPLES OF SPLIT PROVENANCE IN FITTED TREES

Figures 3 and 4 show three final trees whose test Macro-F1 exceeded every ensemble baseline on the same outer split. Red nodes were retained from random split proposals, blue nodes from criterion-guided proposals, and green nodes were introduced or changed during the loss-based structural update or final refinement. Gray nodes denote prefixes fixed by elite consensus in later stages. Because the consensus threshold can combine information from several candidates, its original random or guided source cannot be identified uniquely; we therefore reported these nodes separately.

The three examples reached strong performance through different search paths. Balance Scale (run 1) exceeded the best ensemble by 13.18 percentage points and combined four retained random nodes with three loss-guided refinements. Lymphography (run 3) exceeded the best ensemble by 8.34 points; six final internal nodes came from random proposals, while one was introduced or changed by the loss-guided update. Glass (run 5) exceeded the best ensemble by 3.24 points and relied more heavily on refinement: seven internal nodes were introduced or modified during refinement,

together with two criterion-guided nodes and one retained random node. Thus, the strongest trees did not share a single construction pattern. Random exploration supplied useful candidate structures, while the loss-guided steps preserved or replaced different parts of those structures. These trees are illustrative cases and do not imply that one node source is generally superior.

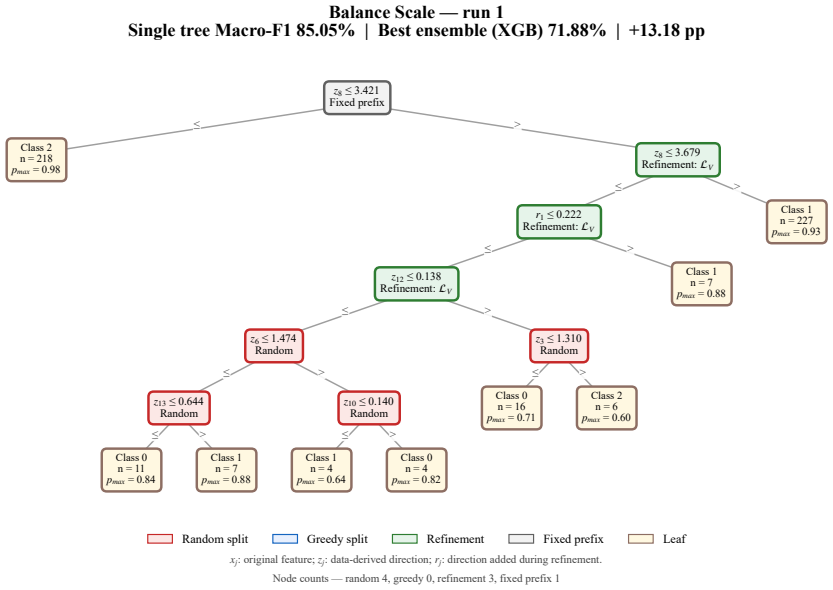


Figure 3: The Balance Scale tree from run 1. Red denotes random proposals, blue criterion-guided proposals, green nodes introduced or changed during the loss-based structural update or refinement, and gray prefixes fixed through elite consensus. Leaf boxes report the predicted class, the number of recalibration observations, and the largest class probability.

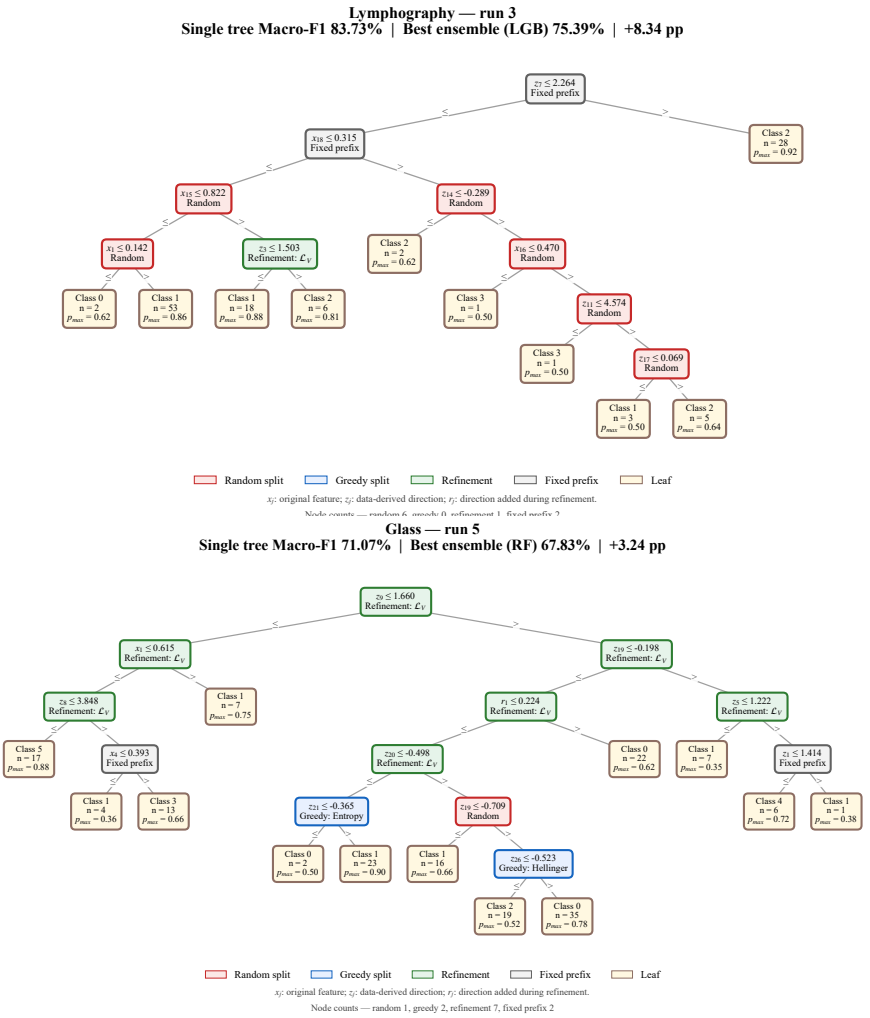


Figure 4: The Lymphography tree from run 3 and Glass tree from run 5. Both exceeded the best of RF, XGB, LGB, and CAT on the same dataset–run split. Glass contained all three traceable node sources and showed the largest contribution from refinement among the selected cases. Colors and leaf annotations follow Figure 3.