Rashomon Sets for Robust Active Learning

Anonymous Authors[1]

Abstract

Collecting labeled data for machine learning models is often expensive and time-consuming. Active learning addresses this challenge by selectively labeling only the most informative observations, but faces a critical limitation: with minimal initial labeled data, identifying truly informative points is difficult. While ensemble methods like random forests help quantify uncertainty, they typically aggregate all models indiscriminately—including poor performers and redundant explanations—a problem that worsens in the presence of noise. We introduce UNique Rashomon Ensembled Active Learning (*UNREAL*), which selectively ensembles only distinct near-optimal models from the Rashomon set. This approach provides two key advantages: it restricts committee membership to high-performing models with different explanations, and it leverages the natural expansion of the Rashomon set under noise to distinguish genuine uncertainty from noise-induced variation. We prove *UNREAL* achieves faster theoretical convergence rates than traditional active learning approaches and demonstrate empirical improvements of up to 20% in predictive accuracy across five benchmark datasets, while simultaneously enhancing model interpretability.

1. Introduction

Collecting labeled data to train data-hungry modern artificial intelligence (AI) and machine learning (ML) models can be expensive or time-consuming. This challenge arises in a wide range of applications: image labeling (Zhu & Bento, 2017; Huijser & van Gemert, 2017), sentence classification (Schumann & Rehbein, 2019), and verbal autopsy (Fan et al., 2024). In such scenarios, strategically determining which observations are most informative for model training will greatly reduce data redundancy and improve the model's ability to generalize from limited data.

To address time and budget constraints, active learning provides a framework where observations are adaptively and strategically selected for labeling based on their potential informativity. The key task in active learning is choosing the most informative observations that will enhance the predictive quality of the model when labeled. However, this selection process faces a fundamental challenge: with limited labeled data initially available, how can we confidently identify truly informative observations rather than noise?

Amongst the many metrics of informativity (Lewis & Gale, 1994; Freund et al., 1997; Roy & McCallum, 2001), uncertainty is the most commonly employed (Liu et al., 2022). One particularly effective approach to quantifying uncertainty is through query-by-committee (QBC), where multiple models form a committee whose disagreement in predictions indicates which observations are most uncertain and therefore informative to label. Ensemble methods such as random forests are especially well-suited for active learning with QBC techniques (Freund et al., 1997; Melville & Mooney, 2004) as their weak learners naturally form a diverse committee. This diversity is crucial as it incorporates different explanations of the data into the committee, with disagreement among ensemble members then serving as a direct measure of uncertainty and informativity (Dagan & Engelson, 1995; Kee et al., 2018).

While ensemble methods offer a natural way to quantify uncertainty through the diversity of their weak learners, this diversity comes with a potential drawback. Specifically, most ensemble methods tend to aggregate over the space of *all* models, even if some of the models may have relatively poor accuracy. While some approaches such as Bayesian model averaging account for this by weighting the models by how likely they are given the data (Hoeting et al., 1999), having a large number of mediocre models is known to make such weighting approaches difficult, especially in cases of limited, high-dimensional, or noisy data (Graefe et al., 2015). Aggregating such poor and implausible models compromises the query-selection criteria, potentially leading to a suboptimal query in the active learning process.

This challenge is further exacerbated in the presence of label noise, which although common in real-world applications, is well-known to be detrimental to active learning strategies (Willett et al., 2005). Active learning strategies, designed to prioritize querying points where models are most uncertain, may overfit to noisy labels instead of focusing on true areas of uncertainty. This challenge makes incorporating a diverse set of explanations among committee members even more crucial.

To address these limitations, we propose an algorithm that enhances active learning with random forests by restricting aggregation to a subset of well-performing and high-evidence models known as the Rashomon set. Importantly, in the presence of label noise, the Rashomon set naturally expands to encompass a more diverse set of valid explanations, providing a principled way to capture and utilize the inherent ambiguity in noisy data. By ensembling the models within the Rashomon set, our approach ensures that the active learning process is driven by models with high evidence while accounting for noise through multiple diverse yet equally valid explanations of the data.

We also demonstrate that one can further restrict the Rashomon set to a unique set of classification patterns while still preserving the performance of the active learning process. This Rashomon-based method, titled UNique Rashomon Ensembled Active Learning (*UNREAL*), ensures that the ensemble incorporates the interpretability of the weak learners by only including the set of unique and diverse explanations. Building on Willett et al. (2005), we show that *UNREAL* achieves a faster convergence rate than traditional active learning by focusing queries on regions of genuine ambiguity rather than noise-induced disagreement. Through extensive simulations on five benchmark datasets, we demonstrate improvements in prediction accuracy compared to standard approaches, confirming both the theoretical and practical advantages of our method.

2. Rashomon Sets

When constructing machine learning models, researchers face two district types of uncertainty. The first originates from the variability in the predicted outcomes generated by a given model, often referred to as a model's *intrinsic risk*. The second originates from selecting the right model from a vast and diverse hypothesis space, a phenomenon known as *model ambiguity*. This distinction, originally articulated by economist Kenneth Arrow in 1951, separates the uncertainty of prediction from a given model from the uncertainty of choosing among many plausible models (Arrow, 1951).

Current machine learning approaches have become exceedingly good at reducing the intrinsic risk within a model, but often fail to fully account for model ambiguity. Methods such as LASSO search for a single optimal model while ensemble methods such as Bayesian Model Averaging (Hoeting et al., 1999) sample across the full hypothesis space. However, both approaches overlook model ambiguity and are ambivalent about how many models with similar predictive power exist for a given dataset (Rudin et al., 2024). This oversight is underscored by the Rashomon Effect (Breiman, 2001).

The Rashomon Effect highlights the existence of near-optimal models that have similarly high predictive performance, but explain the data in different and diverse ways. Rudin furthers this idea by noting that this phenomenon exposes a core issue in the current machine learning paradigm: a reliance on a single predictive model that is overly sensitive to specific data patterns and conceals alternative equally valid explanations (Semenova et al., 2022; Rudin et al., 2024). This reliance fails to notice the complexity of modeling heterogeneity, where different models can explain the data nearly equally well but offer substantively different insights.

To address this sensitive reliance on a single model, we introduce the Rashomon set.

Definition 2.1 (Rashomon set)**.** Let S be a given data set, $\mathcal{F}$ the hypothesis space of all possible models, and ϕ the loss function. For a threshold $\epsilon \geq 0$, the Rashomon set $\hat{R}_\epsilon(\mathcal{F})$ is given by

$$\hat{R}_\epsilon(\mathcal{F}) := \{f \in \mathcal{F} : L(f) \leq \hat{L}(\hat{f}) + \epsilon\} \tag{1}$$

such that $\hat{f}$ is the empirical risk minimizer of S with respect to the loss function $\phi : \hat{f} \in \arg\min_{f \in \mathcal{F}} L(f)$.

Intuitively, the Rashomon set is the set of all plausible models that come within an ϵ distance of the empirical risk minimizer. By enumerating the Rashomon set, researchers can explore the full range of plausible explanations supported by the data. This range of diverse explanations is particularly valuable in noisy settings, in which the uncertainty across different models of the Rashomon set can highlight regions of the data that are not only ambiguous but also sensitive to the underlying noise.

Near-optimal does not mean identical. That is, models that share similar loss do not necessarily mean they share the same classification pattern. Even if all the models in the Rashomon set have similar performance, they can (and often do) differ in how they explain the data. If the models are trees, this will mean that trees with similar losses still partition the feature space differently. Specifically, multiple distinct trees in the Rashomon set may be similar in accuracy but disagree in specific regions of the feature space (see figure 5 of the appendix). That disagreement is often meaningful in active learning when the next unlabeled candidate observations falls into a region where these near-optimal models diverge, indicating there is uncertainty in how to classify that point. By querying that label, the researcher can resolve a genuine source of uncertainty.

Despite this potential to identify meaningful disagreement regions amongst near-optimal models, implementing this insight presents practical challenges. Traditional ensemble methods such as random forests rely on randomization through the sampling of features and observations to form and aggregate base learners. While this randomization cre-

ates less-correlated trees, this often incorporates implausible base models that perform relatively poorly, introducing spurious disagreements that can cause the active learner to chase noise rather than true uncertainty. In contrast, Rashomon sets allow for the targeted aggregation of only high-performing models, reducing the risk of incorporating poor models in the query-selection process. This means that disagreements across the Rashomon ensemble are more likely to reflect more genuine and meaningful uncertainty in the data rather than random instability or overfitting to noise.

In the space of decision trees, Xin et al. (2024) is the first to provide an algorithm that completely enumerates the Rashomon set for sparse decision trees. Their algorithm, `TreeFarms`, provides an exhaustive yet computationally feasible method to generate, store, and view the entire Rashomon set of decision trees. However, due to the inherent structure and geometry of decision trees, many trees in this set offer redundant explanations of the data, which becomes increasingly problematic in active learning as it has the potential to further skew our metric of uncertainty in the committee by artificially inflating agreement (Melville & Mooney, 2004). This can be seen in Figure 1 and more deeply in Figure 5 of the appendix. To address this limitation, Section 4 will outline our method to group trees based on their unique classification patterns.

3. Active Learning

3.1. Notation

Borrowing notation from Liu et al. (2022), let observation i be composed of data $(\mathbf{x}_i, y_i)$ for vector $\mathbf{x}_i$ in covariate space $\mathcal{X}$ and label y_i in output space $\mathcal{Y}$. The data is sent through a supervised learning model $F(\cdot) : \mathcal{X} \to \mathcal{Y}$. When $F(\cdot)$ is an ensemble method consisting of base learners, denote the base learners at $\{f_m\}_{m=1}^{M}$. The model is learned from a training dataset $D_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{I}$ and tested by an independent dataset $D_{ts} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{J}$. The goal is to train $F(\cdot)$ to predict the labels of the out-of-sample test set with a budget-constrained number of labeled observations.

Active learning seeks to adaptively and strategically choose which unlabeled observations should be queried for oracle labeling and then be used in the supervised learning model. Let the query iteration in the active learning framework be denoted by n. Denote the reservoir of unlabeled candidate observations as $D_{cdd}^{(n)} = \{(\mathbf{x}_k, y_k)\}_{k=1}^{K}$ with y_k initially unknown. A selector $S(\cdot)$ is the strategy used to select samples from $D_{cdd}^{(n)}$ to be Oracle labeled. At each iteration, $S(\cdot)$ will sample one observation, denoted $B^{(n)}$, from the candidate dataset $D_{cdd}^{(n)}$ without replacement to query for oracle labeling. $B^{(n)}$ is then added to the training set and removed from the candidate set: $D_{cdd}^{(n+1)} = D_{cdd}^{(n)} \cup B^{(n)}$ and $D_{cdd}^{(n+1)} = D_{cdd}^{(n)} \backslash B^{(n)}$. The model is then retrained on the new training set as $F^{(n+1)}\left(D_{tr}^{(n+1)}\right)$. As such, the $B^{(n)}$ is chosen to find the observations that are most informative to improving predictive performance.

The process is repeated, gradually expanding the training set with informative observations, until the labeling budget is reached or a desired classification metric threshold is met.

3.2. Query-By-Committee Metrics

Picking a selector metric is a key topic in the active learning literature. Common methods are uncertainty (Lewis & Gale, 1994), query-by-committee metrics (Freund et al., 1997; Settles, 2012), or expected error (Roy & McCallum, 2001). Due to the ensembling nature of our methods, we choose to measure informativity by query-by-committee (QBC) metrics, particularly Argamomn-Engelson and Dagan's vote entropy (Dagan & Engelson, 1995):

$$\delta(y, \mathbf{x}, \mathcal{C}) = \max_{\mathbf{x}} - \sum_{y \in \mathcal{Y}} \frac{\text{vote}_{\mathcal{C}}(y, \mathbf{x})}{|\mathcal{C}|} \log \frac{\text{vote}_{\mathcal{C}}(y, \mathbf{x})}{|\mathcal{C}|} \quad (2)$$

where $\text{vote}_{\mathcal{C}}(y, \mathbf{x}) = \sum_{c \in \mathcal{C}} \mathbb{I}\{c(\mathbf{x}) = y\}$ is the number of "votes" that label $y \in \mathcal{Y}$ receives for $\mathbf{x}$ amongst the members c of committee $\mathcal{C}$.

This selector metric is a committee-based generalization of uncertainty measures that consider the confidence of each committee member. It can be viewed as a Bayesian adaptation of Shannon's 1948 uncertainty sampling entropy (Shannon, 1948). One can observe from Equation 2 that ensembling duplicate models has the potential to overinflate the vote entropy with trees from the best-performing explanation group (Melville & Mooney, 2004). This weakness will be addressed by the inclusion of diverse yet unique classification patterns in our proposed methodology *UNREAL*.

4. From Noise to Signal: The Rashomon Advantage

It is well-documented that active learning strategies suffer greatly in the presence of label noise (Willett et al., 2005; Burbidge et al., 2007; Mots'oehli & Baek, 2023; Khosla et al., 2023; Nuggehalli et al., 2024). Specifically, label noise decreases the accuracy of each supervised learning model in the committee, and thus increases the possibility of suboptimal models. This leads to an overall less optimal active learning strategy. While one option is to expand to the committee to include noise-resistant models, we employ a different strategy based on Rashomon theory which does not require finding and implementing such noise-resistant models.

Semenova et al. (2022) and Semenova et al. (2024) discovered that increasing the label noise while keeping the

sample size fixed increases the number of models that exhibit similar performance, i.e. the size of the Rashomon set. Specifically, they argue that label noise leads to increased pattern diversity - the average difference in predictions between distinct classification patterns in the Rashomon set. The increase in pattern diversity manifests as predictive multiplicity (Marx et al., 2020), indicating that there are more differences in model predictions and thus more diverse models in the Rashomon set. Thus, in the presence of noise, the distinction of being the "best-fitting" model becomes increasingly arbitrary as multiple models capture different yet equally valid patterns in the data.

This enhanced model diversity suggests that relying on a single "best" model may overlook other meaningful explanations of the data. Active learning, in its iterative adding of training data, presents a unique opportunity to leverage this existence of diverse yet equally valid model. Rashomon sets offer a natural defense against noise in active learning: rather than relying on a single model that may overfit to noise points, we can leverage the entire set of plausible explanations to capture a comprehensive view of uncertainty in the dataset. That is, incorporating multiple valid explanations from the Rashomon set can help distinguish between genuine uncertainty and noise-induced variability in the sampling strategy.

This theoretical intuition leads us to our main contribution: we prove that restricting ourselves to the Rashomon set of models in active learning leads to provable improvements in convergence rates that improve on those established by Willett et al. (2005). The key insight is that the Rashomon set targets a more focused yet still diverse hypothesis space, allowing us to distinguish genuine uncertainty from noise-induced variability. We formalize this improvement in Theorem 4.1.

Theorem 4.1 (Rashomon Advantage in Active Learning)**.** *Let $\hat{f}_n$ be the standard estimator minimizing over the full hypothesis space Γ, and $\hat{f}_n^R$ be the Rashomon set estimator minimizing over $\hat{R}_\epsilon(\mathcal{F})$. Under appropriate conditions on the noise distribution satisfying the Bernstein moment condition, we have:*

$$\mathbb{E}\left[\|f - \hat{f}_n^R\|^2\right] \leq \mathbb{E}\left[\|f - \hat{f}_n\|^2\right] \tag{3}$$

leading to faster active learning rates.

This theorem establishes that restricting to models within the Rashomon set not only reduces expected error but also accelerates the convergence rate in active learning. The proof essentially makes use of the fact that the Rashomon set has lower complexity than the full hypothesis space. The full proof of this theorem, along with the necessary lemmas and technical details, is provided in Appendix A.

Although Willett et al. (2005)'s original convergence rates were derived for regression settings, their insights extend naturally to our classification framework, as both ultimately seek to identify regions of maximum uncertainty—which in classification manifest as decision boundaries between classes.

To this end, *UNREAL* proposes that one should query *within* the Rashomon Set. Rather than expanding a committee's diversity through different model classes, *UNREAL* enriches the committee by incorporating multiple models from within the same model class that exhibit similar predictive performance. Specifically, for a given model class, *UNREAL* constructs a committee from the Rashomon set $\hat{\mathcal{R}}_\epsilon$, consisting of all models achieving accuracy within ϵ of the best-performing model. This approach ensures that each committee member represents a valid alternative explanation of the underlying data pattern, rather than merely capturing noise. Active learning then proceeds following the standard QBC framework with uncertainty quantified through vote entropy (Dagan & Engelson, 1995).

The empirical effectiveness of this approach arises from three factors. Firstly, the models models in the Rashomon set, while all near-optimal, are not identical. The regions where they disagree are precisely the ones that matter most in active learning as they provide the strongest and most meaningful signal of genuine uncertainty across all plausible models. Secondly, he Rashomon set provides a natural defense against one of active learning's main weakness: noise. By ensembling across a set of plausible models, the selector committee has a more robust view of uncertainty than if it had incorporated potentially suboptimal models. Finally, as the committee ensembles these near-optimal yet diverse models, the algorithm prioritizes querying points with true uncertainty.

Ensembling across the Rashomon sets represents a fundamental shift in how we approach active learning under noise. By leveraging the natural diversity of near-optimal models captured in the Rashomon set, we not only achieve theoretical improvements in convergence rates but also a method that inherently adapts to the presence of noise. Rather than treating noise as an obstacle that confuses the active learning selection criteria, UNREAL harnesses the insights from Rashomon theory to turn model ambiguity into an advantage, allowing us to distinguish between genuine uncertainty and noise-induced variability.

5. *UNREAL*: Unique Rashomon Ensemble Active Learning

In our proposed approach, the committee $\mathcal{C}$ in Equation 2 is constructed as the Rashomon set of decision trees, denoted by $\hat{\mathcal{R}}_\epsilon$. This Rashomon set $\hat{\mathcal{R}}_\epsilon$ consists of "near-equal"

decision trees whose objective function is within ϵ of the overall best model given the data. Since each near-equal model in the Rashomon set provides a different perspective on the data, diverse prediction patterns emerge in regions of genuine ambiguity. Our method exploits this diversity by ensembling across this set of distinct explanations to capture uncertainty arising from the variety of plausible models.

While the trees in the Rashomon set share many of the same splits and appear globally correlated across the dataset, the regions where they exhibit disagreements are crucial to identifying true uncertainty. This property is particularly valuable in active learning: the trees agree on easily classified points (avoiding wasteful queries) but diverge precisely in areas where additional labels would be most informative. As new labels are acquired in these contentious regions, the near-optimal trees can collectively converge to more decisive predictions, making this approach particularly effective for query-by-committee strategies.

To enumerate the Rashomon set of decision trees, we use Xin et al. (2024)'s `TreeFarms` approach. `TreeFarms` exhaustively enumerates the Rashomon set of decision trees, allowing us to aggregate the best models in our ensemble method. However, unlike random forests, `TreeFarms` lacks the random sampling of features and data, making the models in `TreeFarms` correlated. This correlation in decision trees presents a significant challenge for query-by-committee approaches, as correlation amongst committee members may both artificially inflate agreement in the vote and complicate interpretability (Melville & Mooney, 2004).

To address this issue, we reduce redundancy in `TreeFarms` by grouping trees based on their unique classification patterns and selecting a single representative tree from each of these groups to ensemble. This ensures that each chosen tree is meaningfully distinct while faithfully representing the Rashomon set. This prevents overestimating agreement in our vote entropy metric, ultimately leading to a valid query-by-committee voting selection method.

Our approach can be visualized in Figure 1. If we ignore the redundancy of explanations in `TreeFarms`, the ensemble will be dominated by the trees in Group 3 offering the same explanation and prediction. This will artificially inflate agreement amongst our committee (see Equation 2) by overvaluing the trees of the most commonly occurring explanation groups. If we instead account for the redundancy of the trees, the unique selection method will instead choose one tree arbitrarily from each classification pattern, diversifying our committee and more fully representing the Rashomon set. Our method is summarized in Algorithm 1.

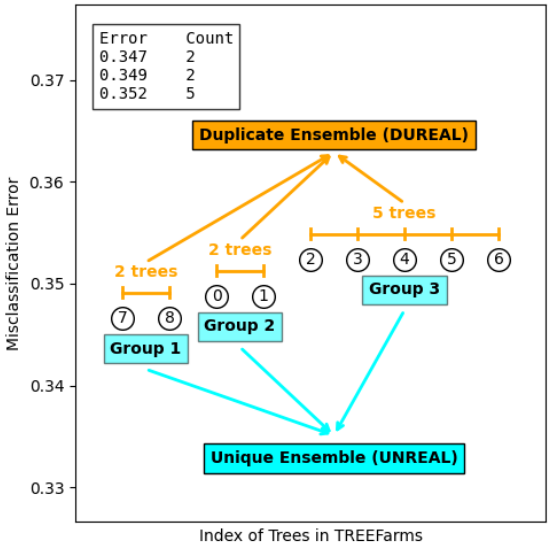

Figure 1. A depiction of the redundant and unique ensembling of Rashomon trees from the COMPAS dataset. To visualize our methods, the figure plots the misclassification errors by the ordered indices of the tree (though the unique selection of trees was grouped by classification pattern in our simulations). As shown, many trees have the *same* misclassification rate, indicating these trees share the same classification pattern. The geometry of the trees can be seen in Figure 5 of the appendix.

5.1. How to Choose the Rashomon Threshold

The Rashomon threshold ϵ provides a principled way to navigate the exploration-exploitation tradeoff inherent in active learning. A larger ϵ includes more diverse models in the committee, promoting exploration of alternative explanations, while a smaller ϵ restricts the committee to models closest to the empirical risk minimizer, favoring exploitation of established patterns. This choice significantly impacts how the algorithm handles uncertainty, particularly in the presence of noise.

The different choices for the size of the Rashomon set represent a trade-off between using the currently most accurate model versus the robustness of making decisions based on the wider range of near-optimal models. While our empirical experiments show that a good choice for the Rashomon model can strongly improve the active learning's accuracy and rate of convergence, this is predicated on being able to identify a good choice for the Rashomon threshold ϵ.

The threshold effectively serves as a lever balancing between current accuracy and robustness. When ϵ is too small, the committee may consist of nearly identical models, oversampling regions that initially appear uncertain but may reflect only the limitations of the current best model. Conversely, when ϵ is too large, the committee may include relatively poor-performing models, potentially drowning

Algorithm 1 Unique Rashomon Ensemble Active Learning

Input: Training dataset $D_{tr}^{(0)}$; Test dataset D_{ts}; Candidate dataset $D_{cdd}^{(0)}$; Rashomon Threshold ϵ;
repeat
  Enumerate and train the Rashomon set $\hat{\mathcal{R}}_\epsilon$ of models $\{f_m\}_{m=1}^M$ on $D_{tr}^{(n)}$ with `TreeFarms`.
  Generate the test set predicted labels $\hat{y}_{ts,m}^{(n)}$ for each model $f_m \in \hat{\mathcal{R}}_\epsilon$ from $\mathbf{x}_{ts}^{(n)}$.
  Ensemble the test set predicted labels: $\bar{\hat{y}}_{ts}^{(n)} :=$ mode $\left(\hat{y}_{ts,m}^{(n)}\right)$.
  Evaluate the performance of the ensemble on $D_{ts}^{(n)}$ with prediction $\bar{\hat{y}}_{ts}^{(n)}$.
  Group models by identical classification patterns and select one representative from each group.
  Generate the candidate set predicted labels $\hat{y}_{cd,m}^{(n)}$ for each unique model $f_m \in \hat{\mathcal{R}}_\epsilon$ from $\mathbf{x}_{cd}^{(n)}$.
  Ensemble the candidate set predicted labels: $\bar{\hat{y}}_{cd}^{(n)} :=$ mode $\left(\hat{y}_{cd,m}^{(n)}\right)$.
  Compute the vote-entropy metric $\delta^{(n)}(\bar{\hat{y}}_{cd}^{(n)}, \mathbf{x}_{cd}, \hat{\mathcal{R}}_\epsilon)$ from Equation 2.
  Resample $B^{(n)}$ from $D_{cdd}^{(n)}$ based on the observation with the highest vote entropy: $B^{(n)} := \arg\max_{\mathbf{x}} \delta^{(n)}(\bar{\hat{y}}_{cd}^{(n)}, \mathbf{x}_{cd}, \hat{\mathcal{R}}_\epsilon)$
  Query $B^{(n)}$ for oracle labeling
  Set $D_{tr}^{(n+1)} = D_{tr}^{(n)} \cup B^{(n)}$ and $D_{cdd}^{(n+1)} = D_{cdd}^{(n)} \setminus B^{(n)}$
until labeling budget is depleted or test error is sufficiently small

genuine signals in a sea of noise.

Based on our empirical experiments, we recommend initializing the Rashomon threshold to a value that is *at or larger than* the optimal Rashomon size for the initial training data. As an example, in Figure 2 we have the average accuracy over all the models in the Rashomon set as the size of the Rashomon set is varied. All trees were estimated using `TreeFarms` using a random 20% training split of the data on the Bar7 dataset. We observe that as we increase the Rashomon threshold, and hence the size of the Rashomon set, the average classification accuracy on the test set increases up until $\epsilon = 0.016$ and then decays. As such, $\epsilon = 0.016$ should be chosen as the value for our Rashomon threshold throughout the active learning procedure.

This procedure can be improved, albeit at the cost in terms of sample size or computational cost. One can split the training set in two, training the models on one, and computing accuracies to choose the Rashomon cutoff with the other. Such data splits in general are good practice to prevent overfitting (Hastie et al., 2009). However, in cases where labeled

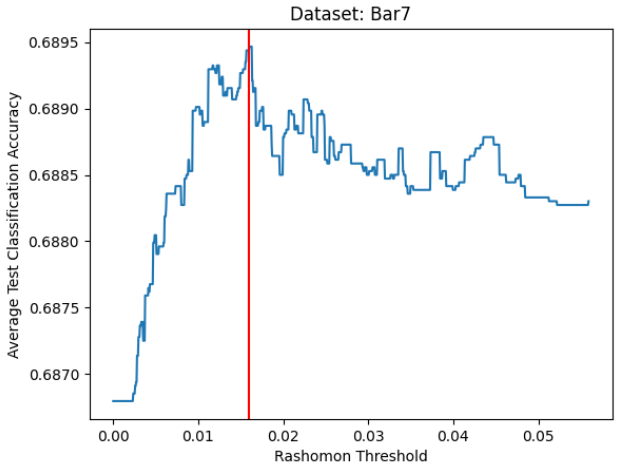

Figure 2. This plot graphs the ensemble test classification error against the Rashomon threshold. In this case, the optimal threshold is 0.016.

data is scarce, as is in the case of active learning, one may lack sufficient data to reserve for such splits. In such cases, we recommend one err on the side of caution and choose a Rashomon threshold that is slightly larger than what may be suggested by the initial dataset.

6. Experiments

6.1. Datasets

Empirical experiments were performed on five benchmark datasets: Iris (Fisher, 1936), MONK-1, MONK-3 (Wnek, 1993), COMPAS dataset (Larson et al., 2016), and Bar7 (Wang et al., 2017). The summary of the data as well as preprocessing details can be seen in Table 1 of Section A.3 in the appendix.

The datasets vary in complexity of their data-generating functions. MONK-1 and Iris represent relatively simple underlying structures. MONK-3 introduces a more intricate generative scheme but is still a fixed rule. The Bar7 and COMPAS datasets are characterized by the most notable noise and complexity due to their collection in restaurant and recidivism settings. These five datasets provide a gradient of learning challenges, enabling nuanced evaluation of our proposed methods against other active learning methods under varying degrees of complexity and noise.

6.2. Simulation

To compare our unique selection of classification patterns to baseline methods, we ran the active learning process against QBC with random forests and passive learning, in which observations are randomly selected for labeling. We also compared our method to a QBC method that ensembles *all*

of the trees of the Rashomon set, whose method we call *Duplicate Rashomon Ensemble Active Learning (DUREAL)*. *DUREAL* is essentially a weighted variant of *UNREAL* in which classification patterns are weighted by how frequently they appear in the Rashomon set. Due to the intrinsic geometry of decision trees, trees with larger complexity (ie. more splits and depth) will appear more frequently in our committee.

One hundred active learning runs were run on the Iris, MONK-1, and MONK-3 datasets, fifty on COMPAS, and fifteen on Bar 7 due to computational limitations. The Rashomon threshold values used in *UNREAL* and *DUREAL* are given in Table 1. Each method had a fixed regularization of 0.01 on the depth of the trees. Random forests were run with 100 weak learners. Twenty percent of each dataset was reserved for the test set with twenty percent of the remaining observations used as the initial training dataset. We evaluate our active learning procedures with the F1 score to account for both precision and recall. Further information about the simulation can be seen in Section A.3 of the appendix. Open source code for the simulation is available on Github.

6.3. Experimental Results

Results are presented in Figure 3, showing errors relative to our random forest baseline. Standard error plots and Wilcoxon rank signed tests at 95% significance level are provided in Figure 6 and Table 2 of the appendix, respectively.

Our findings demonstrate the remarkable gains when ensembling across the Rashomon set and the substantial advantages over traditional QBC with random forests. Ensembles of decision trees from the Rashomon set (both unique and duplicate) consistently achieve superior performance, even up to 20% in the MONK-1 dataset, highlighting the robustness and prediction accuracy achievable with this approach. With the exception of Bar7, we note that this gap becomes increasingly larger as more data is added to our training set.

Interestingly, we find that at times *DUREAL* mildly outperforms *UNREAL*, contrary to concerns about artificial committee agreement inflation. This can be explained by the combinatorial nature of tree splits: classification patterns with greater complexity tend to appear more frequently. This creates an implicit weighted voting system in *DUREAL* where complex trees carry greater influence. Nonetheless, *UNREAL*'s compartmentalization of redundant explanations maintained similarly high classification accuracy while reducing to a more parsimonious ensemble of learners.

This relationship between tree count and unique classification patterns reveals a deeper insight into model diversity. As shown in Figure 4, increasing the number of trees does not proportionally increase the number of unique classifica-

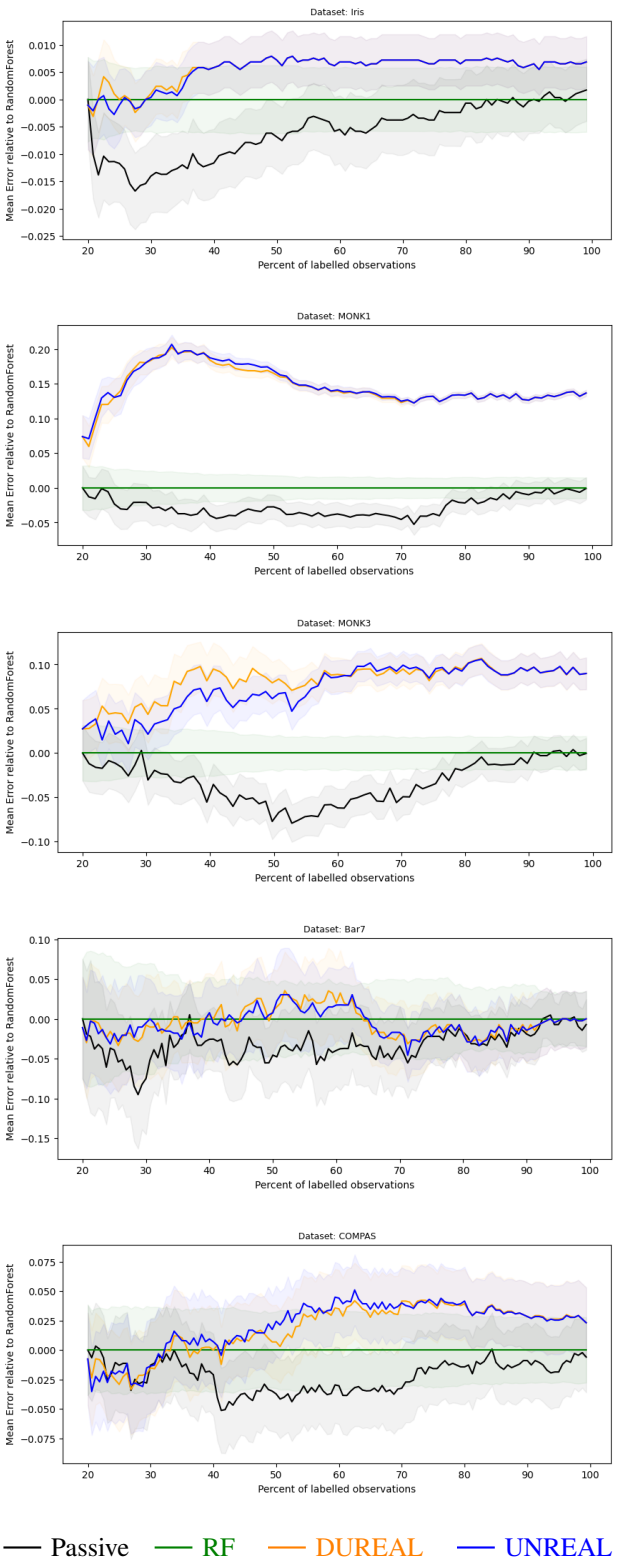

Figure 3. Performance of the four active learning procedures (left) on our five benchmark datasets. The plot presents the errors relative to random forests.

tion patterns. Further analysis in Section A.4 demonstrates that while raising the Rashomon threshold increases the total tree count in the Rashomon set, the number of unique patterns remains relatively stable. As more data is gathered, the number of unique classification patterns decreases towards a few explanations while the number of redundant trees usually continues to grow. The reduction to a single explanation highlights that our method accounts for model ambiguity by strategically prioritizing the querying of observations with the highest uncertainty across models.

Both *UNREAL* and *DUREAL* leverage this principle effectively, though in different ways: while *DUREAL* benefits from the implicit weighting of complex patterns, *UNREAL* achieves similar performance with a more concise set of explanations. This distinction highlights an important principle: trees themselves are not explanations, but rather manifestations of underlying classification patterns.

7. Limitations and Future Work

Our method, while demonstrating these promising results, presents certain limitations that warrant discussion and possible future work. One constraint lies in the initial selection of the Rashomon threshold. This initial threshold, though designed to optimize selection criteria in the early stages, may not maintain its optimality as the process evolves. As evidenced in Figure 4, the number of unique classification patterns exhibits considerable variation throughout the active learning process.

This limitation can be addressed in future work by dynamically recalibrating ϵ at each iteration of the active learning procedure. That is, after each query, researchers can again set a sufficiently high ϵ and generate a new tradeoff plot to identify the best threshold. This iterative approach allows for the Rashomon threshold to adapt to the evolving dataset. However, this iterative recalibration introduces additional computational overhead, making it practical only when substantial computational resources are available.

This computational intensity represents another significant limitation of our approach. As detailed in Table 1, which presents the average run time for each active learning strategy, our Rashomon-based methods may not be optimal for researchers who prioritize computational efficiency over predictive accuracy when compared to traditional machine learning models.

These limitations notwithstanding, our findings plant the seeds for future active learning research that incorporates the Rashomon set of good and plausible models. In particular, Venkateswaran et al. (2024)'s Rashomon Partition Sets (RPS) offer a promising framework for comprehensively enumerating the Rashomon set without an inherent geometric structure. Investigating the use of RPS in active learning

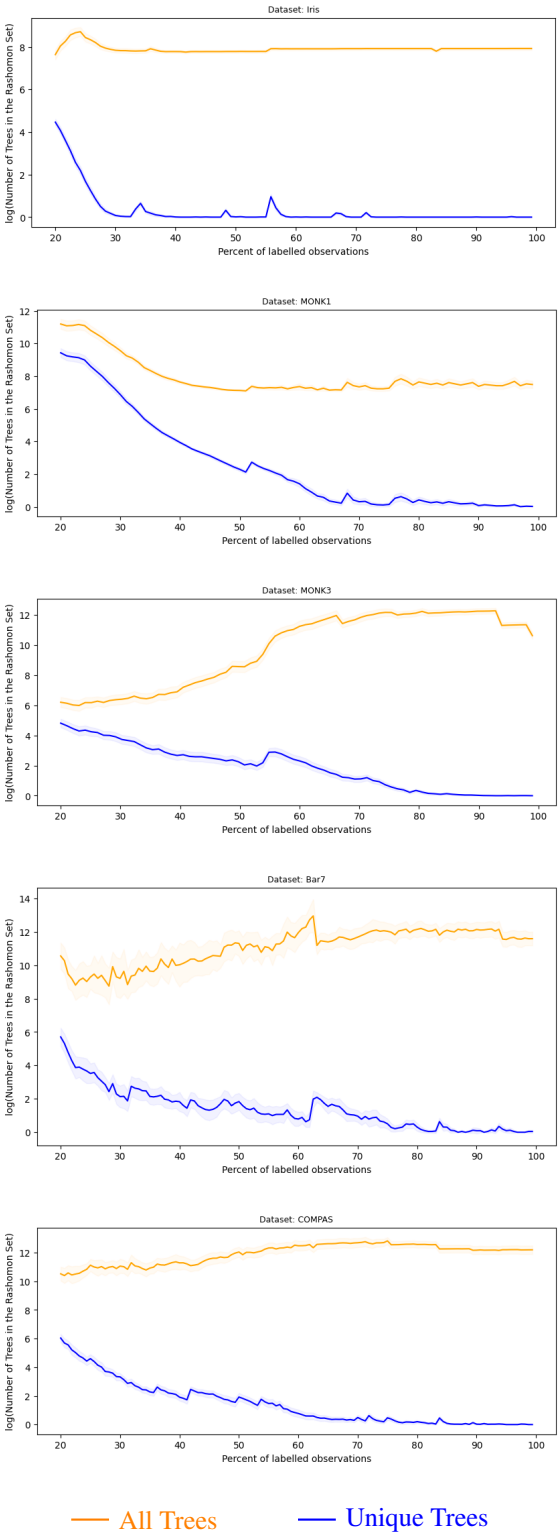

Figure 4. The number of total vs. unique trees on a logarithmic scale.

may further deepen our understanding of Rashomon's benefits in both prediction and interpretability.

8. Conclusion

Our work offers three key insights. Firstly, we demonstrate that ensembling over the Rashomon set of decision trees enhances the active learning process by a significant margin, up to 20% in some datasets. Unlike traditional ensemble methods which aggregate over the entire space of models, potentially including models that are poor performing or implausible, the Rashomon set only contains models with high posterior probability. This focused approach ensures that our committee consists of only strong and plausible models, whose disagreements naturally occur in the regions that matter most for active learning. These targeted disagreements provide a more robust measure of uncertainty, leading to a more efficient query selection.

Secondly, our approach provides a novel mechanism for dealing with label noise in the active learning process. By leveraging the diversity of the Rashomon set, we can more effectively distinguish between genuine model uncertainty and noise-induced variability, thereby improving the robustness of the active learning strategy in challenging, noisy environments.

Finally, we address the issue of redundant and duplicate explanations when constructing a Rashomon set by only considering trees with unique explanations. Redundant explanations can inflate query-by-committee metrics and obscure interpretability. By only ensembling over the Rashomon's subset of trees with unique explanations, we ensure that the ensemble remains parsimonious and interpretable while maintaining high prediction accuracy.

References

Arrow, K. Alternative approaches to the theory of choice in risk taking situation. *Econometrica*, 19, 10 1951. doi: 10.2307/1907465.

Breiman, L. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199 – 231, 2001. doi: 10.1214/ss/1009213726.

Burbidge, R., Rowland, J. J., and King, R. D. Active learning for regression based on query by committee. In *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning*, IDEAL'07, pp. 209–218, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3540772251.

Dagan, I. and Engelson, S. P. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, pp. 150–157, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603778.

Fan, S., Visokay, A., Hoffman, K., Salerno, S., Liu, L., Leek, J. T., and McCormick, T. H. From narratives to numbers: Valid inference using language model predictions from verbal autopsy narratives. *CoRR*, abs/2404.02438, 2024. doi: 10.48550/ARXIV.2404.02438.

Fisher, R. A. Iris. UCI Machine Learning Repository, 1936. doi: https://doi.org/10.24432/C56C76.

Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. Selective sampling using the query by committee algorithm. *Mach. Learn.*, 28:133–168, September 1997. ISSN 0885-6125. doi: 10.1023/A:1007330508534.

Graefe, A., Küchenhoff, H., Stierle, V., and Riedl, B. Limitations of ensemble bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3):943–951, 2015. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2014.12.001.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statistical Science*, 14(4):382 – 417, 1999. doi: 10.1214/ss/1009212519.

Hu, X., Rudin, C., and Seltzer, M. Optimal sparse decision trees. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Huijser, M. and van Gemert, J. C. Active decision boundary annotation with deep generative models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Kee, S., del Castillo, E., and Runger, G. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*, 454-455: 401–418, 2018. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2018.05.014.

Khosla, S., Whye, C. K., Ash, J. T., Zhang, C., Kawaguchi, K., and Lamb, A. Understanding and improving neural active learning on heteroskedastic distributions. *ArXiv*, pdf/2211.00928, 2023.

Larson, J., Mattu, S., Kirchner, L. K., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica*, 2016.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In Croft, B. W. and van Rijsbergen, C. J. (eds.), *SIGIR '94*, pp. 3–12, London, 1994. Springer London. ISBN 978-1-4471-2099-5.

Liu, P., Wang, L., Ranjan, R., He, G., and Zhao, L. A survey on active deep learning: From model driven to data driven. *ACM Comput. Surv.*, 54(10s), sep 2022. ISSN 0360-0300. doi: 10.1145/3510414.

Marx, C. T., Du Pin Calmon, F., and Ustun, B. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Melville, P. and Mooney, R. J. Diverse ensembles for active learning. In *Proceedings of 21st International Conference on Machine Learning (ICML-2004)*, pp. 584–591, Banff, Canada, July 2004.

Mots'oehli, M. and Baek, K. Deep active learning in the presence of label noise: A survey. *ArXiv*, abs/2302.11075, 2023.

Nuggehalli, S., Zhang, J., Jain, L., and Nowak, R. Direct: Deep active learning under imbalance and label noise. *ArXiv*, abs/2312.09196, 2024.

Roy, N. and McCallum, A. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 441–448, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Rudin, C., Zhong, C., Semenova, L., Seltzer, M., Parr, R., Liu, J., Katta, S., Donnelly, J., Chen, H., and Boner, Z. Amazing things come from having many good models. *ArXiV*, 07 2024.

Schumann, R. and Rehbein, I. Active learning via membership query synthesis for semi-supervised sentence classification. In Bansal, M. and Villavicencio, A. (eds.), *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 472–481, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1044.

Semenova, L., Rudin, C., and Parr, R. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1827–1858, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533232.

Semenova, L., Chen, H., Parr, R., and Rudin, C. A path to simpler models starts with noise. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Settles, B. *Active Learning*. Morgan & Claypool Publishers, 2012. ISBN 1608457257.

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Venkateswaran, A., Sankar, A., Chandrasekhar, A. G., and McCormick, T. H. Robustly estimating heterogeneity in factorial data using rashomon partitions. *ArXiv*, abs/2404.02141, 2024.

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., and Macneille, P. A bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, 18, 05 2017. doi: 10.48550/arXiv.1504.07614.

Willett, R., Nowak, R., and Castro, R. Faster rates in regression via active learning. In Weiss, Y., Schölkopf, B., and Platt, J. (eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Wnek, J. MONK's Problems. UCI Machine Learning Repository, 1993. doi: https://doi.org/10.24432/C5R30R.

Xin, R., Zhong, C., Chen, Z., Takagi, T., Seltzer, M., and Rudin, C. Exploring the whole rashomon set of sparse decision trees. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Zhu, J.-J. and Bento, J. Generative adversarial active learning. *ArXiv*, abs/1702.07956, 2017.

A. Appendix

A.1. Proof of Theorem 4.1

Note the following proofs follows the extended Technical Report version of Willett et al. (2005) from the University of Wisconsin–Madison.

A.1.1. Outline and Intuition

In this proof, we demonstrate that using the Rashomon set can improve error bounds in regression, particularly in active learning settings. The intuition is straightforward: by restricting our hypothesis space to only well-performing models, we reduce complexity without sacrificing approximation quality. This principled form of complexity regularization leads to tighter error bounds.

Our approach builds upon the oracle inequality framework established for regression estimators in Willett et al. (2005), showing that restricting to the Rashomon set leads to demonstrably better guarantees. The proof proceeds through several key steps: (1) establishing that the Rashomon set has lower complexity than the full hypothesis space, (2) deriving a modified oracle inequality for the restricted space, (3) showing that the Rashomon set contains sufficiently good approximations to the true function, and (4) combining these elements to prove improved error bounds.

When coupled with active learning strategies, this approach becomes particularly powerful for functions with concentrated complexity, as it naturally focuses learning effort on regions of genuine ambiguity rather than noise-induced disagreement.

A.1.2. Preliminaries: Notation and Assumptions

We first establish the framework, notation, and key concepts used throughout our analysis of the Rashomon set approach in active learning for regression. We closely follow the notation and set up of Willett et al. (2005).

Let $\mathcal{F}$ denote a class of functions mapping $[0,1]^d$ to the real line. Consider a function $f : [0,1]^d \to \mathbb{R}$ in this class. Our goal is to estimate this function from a finite number of noise-corrupted samples.

We adopt the following statistical model:

Assumption A.1. The observations $\{Y_i\}_{i=1}^n$ are given by

$$Y_i = f(\boldsymbol{X}_i) + W_i, \quad i \in \{1, \ldots, n\}. \tag{4}$$

Assumption A.2. The random variables W_i are independent and identically distributed with $\mathbb{E}[W_i] = 0$ and $\text{Var}(W_i) \leq \sigma^2$. Furthermore, for all $i \in \{1, \ldots, n\}$, we have

$$\mathbb{E}\left[|W_i|^k\right] \leq \text{Var}(W_i)\frac{k!}{2}h^{k-2}, \tag{5}$$

for some $h > 0$ and $k \geq 2$. Equation (5) is known as the Bernstein's moment condition.

We consider two learning scenarios:

Assumption A.3 (Passive Learning)**.** The sample locations $\boldsymbol{X}_i \in [0,1]^d$ are possibly random, but independent of $\{Y_j\}_{j\in\{1,\ldots,i-1,i+1,\ldots,n\}}$. They do not depend in any way on f.

Assumption A.4 (Active Learning)**.** The sample locations $\boldsymbol{X}_i$ are random, and depend only on $\{\boldsymbol{X}_j, Y_j\}_{j=1}^{i-1}$. That is, the sample locations have only a causal dependency on the system variables.

For measuring the quality of estimation, we use the L_2 norm:

$$d(f,g) \equiv \|f - g\| = \left(\int_{[0,1]^d} |f(x) - g(x)|^2 dx\right)^{1/2}, \tag{6}$$

where $f, g \in \mathcal{F}$.

Definition A.5 (Rashomon Set)**.** For a hypothesis class $\mathcal{F}$ and threshold $\epsilon > 0$, the Rashomon set is defined as:

$$\hat{R}(\mathcal{F}, \epsilon) := \{f \in \mathcal{F} : L(f) \leq \hat{L}(\hat{f}) + \epsilon\} \tag{7}$$

where $L(f)$ is the true loss, $\hat{L}(f)$ is the empirical loss, and $\hat{f}$ is the empirical risk minimizer.

We consider a countable class of functions Γ mapping $[0,1]^d$ to the real line such that

$$|f(x)| \leq M \quad \forall x \in [0,1]^d, \forall f \in \Gamma. \tag{8}$$

Let pen : $\Gamma \to [0, +\infty)$ be a function satisfying

$$\sum_{\theta' \in \Gamma} e^{-\text{pen}(\theta')} \leq s, \tag{9}$$

for some $s > 0$.

We define two estimators:

First, we define the standard estimator $\hat{f}_n(\boldsymbol{X}, \boldsymbol{Y})$ minimizing over the full hypothesis space Γ:

$$\arg\min_{f' \in \Gamma} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f'(\boldsymbol{X}_i))^2 + \frac{\lambda}{n} \text{pen}(f') \right\} \tag{10}$$

Second, we define the Rashomon set estimator $\hat{f}_n^R(\boldsymbol{X}, \boldsymbol{Y})$ minimizing over $\hat{R}(\mathcal{F}, \epsilon)$:

$$arg \min_{f' \in \hat{R}(\mathcal{F}, \epsilon)} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f'(\boldsymbol{X}_i))^2 + \frac{\lambda}{n} \text{pen}(f') \right\} \tag{11}$$

A.1.3. LEMMAS

We begin by recalling the standard oracle inequality for the estimator $\hat{f}_n$ of Willett et al. (2005):

Theorem A.6 (Oracle Inequality)**.** *Under Assumptions A.1 and A.3, if $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d., uniform over $[0,1]^d$ and independent of $\{Y_i\}_{i=1}^n$, and assuming that Assumption A.2 holds, then for $\lambda > 2(\sigma^2 + M^2) + 32(hM + M^2/3)$, we have $\mathbb{E}\left[\|f - \hat{f}_n\|^2\right]$ is at most:*

$$\min_{f' \in \Gamma} \frac{\left\{ (1+a)\|f - f'\|^2 + \frac{\lambda}{n} pen(f') + \frac{2\lambda(s+1)}{n} \right\}}{1-a}, \tag{12}$$

with $a = \frac{2(\sigma^2 + M^2)}{\lambda - 32(hM + M^2/3)}$.

Our goal is to show that restricting to the Rashomon set leads to tighter error bounds and better performance in active learning settings. We proceed through a series of lemmas.

Lemma A.7 (Rashomon Set Complexity)**.** *Let $\hat{R}(\mathcal{F}, \epsilon) \subset \Gamma$ be the Rashomon set with threshold ϵ. Define*

$$s_R = \sum_{f' \in \hat{R}(\mathcal{F}, \epsilon)} e^{-pen(f')} \tag{13}$$

Then $s_R < s$, where s is the constant from Equation (9).

Proof. Since $\hat{R}(\mathcal{F}, \epsilon) \subset \Gamma$, we have:

$$\sum_{f' \in \hat{R}(\mathcal{F}, \epsilon)} e^{-\text{pen}(f')} < \sum_{f' \in \Gamma} e^{-\text{pen}(f')} \leq s \tag{14}$$

for some $s > 0$. The inequality is strict because the Rashomon set is a proper subset of Γ (assuming ϵ is chosen to be sufficiently small). □

Lemma A.8 (Oracle Inequality for Rashomon Estimator)**.** *Under the same conditions as Theorem A.6, the Rashomon set estimator $\hat{f}_n^R$ has bound $\mathbb{E}\left[\|f - \hat{f}_n^R\|^2\right]$ at most*

$$\min_{f' \in \hat{R}(\mathcal{F}, \epsilon)} \frac{\left\{ (1+a)\|f - f'\|^2 + \frac{\lambda}{n} pen(f') + \frac{2\lambda(s_R+1)}{n} \right\}}{1-a}, \tag{15}$$

where $s_R < s$ is as defined in Lemma A.7.

Proof. The proof follows the same steps as the proof of Theorem A.6 (given in Appendix B of Willett et al. (2005)), with the following modifications:

1. Replace the hypothesis space Γ with $\hat{R}(\mathcal{F}, \epsilon)$.
2. Use s_R instead of s in the bound.

The key observation is that all the mathematical machinery used to prove Theorem A.6 applies equally well to any countable subset of Γ, including $\hat{R}(\mathcal{F}, \epsilon)$. □

Lemma A.9 (Rashomon Set Contains Good Approximations)**.** *With high probability, there exists $f_R \in \hat{R}(\mathcal{F}, \epsilon)$ such that:*

$$\|f - f_R\| \leq C\|f - \hat{f}\| + \delta \tag{16}$$

where $\hat{f}$ is the empirical risk minimizer, C is a constant, and δ is a small term that decreases with sample size.

Proof. This follows from results by Semenova et al. (2024) and (Semenova et al., 2022), which show that the true function f is likely to be in or near the Rashomon set, particularly in the presence of noise. Specifically:

1. The empirical risk minimizer $\hat{f}$ approximates f well under standard statistical assumptions.
2. The Rashomon set includes functions that are ϵ-close to $\hat{f}$ in terms of empirical risk.
3. In the presence of noise, the Rashomon set naturally expands to include more diverse yet valid explanations.

Therefore, there exists $f_R \in \hat{R}(\mathcal{F}, \epsilon)$ that approximates f nearly as well as the best possible approximation in the full hypothesis space. □

A.1.4. Main Result

Now we can establish our main theorem regarding the improved performance of the Rashomon set estimator.

Theorem A.10 (Improved Bounds with Rashomon Sets)**.** *Under the conditions of Theorem A.6, the expected error of the Rashomon set estimator $\hat{f}_n^R$ satisfies:*

$$\mathbb{E}\left[\|f - \hat{f}_n^R\|^2\right] \leq \mathbb{E}\left[\|f - \hat{f}_n\|^2\right] \tag{17}$$

Moreover, the gap between these bounds increases as the complexity of the full hypothesis space Γ increases relative to the complexity of the Rashomon set.

Proof. Let $f^* = \arg\min_{f' \in \Gamma} \left\{(1+a)\|f - f'\|^2 + \frac{\lambda}{n}\text{pen}(f')\right\}$ be the minimizer over the full hypothesis space in the oracle bound from Theorem A.6.

We need to compare:

$$B_\Gamma = \frac{(1+a)\|f - f^*\|^2 + \frac{\lambda}{n}\text{pen}(f^*) + \frac{2\lambda(s+1)}{n}}{1-a} \tag{18}$$

$$B_R = \min_{f' \in \hat{R}(\mathcal{F},\epsilon)} \frac{(1+a)\|f - f'\|^2 + \frac{\lambda}{n}\text{pen}(f') + \frac{2\lambda(s_R+1)}{n}}{1-a} \tag{19}$$

We consider two cases:

Case 1: If $f^* \in \hat{R}(\mathcal{F}, \epsilon)$, then:

$$B_R \leq \frac{(1+a)\|f - f^*\|^2 + \frac{\lambda}{n}\text{pen}(f^*) + \frac{2\lambda(s_R+1)}{n}}{1-a} \tag{20}$$

$$< \frac{(1+a)\|f - f^*\|^2 + \frac{\lambda}{n}\text{pen}(f^*) + \frac{2\lambda(s+1)}{n}}{1-a} \tag{21}$$

$$= B_\Gamma \tag{22}$$

where the strict inequality follows from $s_R < s$ (Lemma A.7).

Case 2: If $f^* \notin \hat{R}(\mathcal{F}, \epsilon)$, by Lemma A.9, there exists $f_R \in \hat{R}(\mathcal{F}, \epsilon)$ such that $\|f - f_R\|$ is not much worse than $\|f - f^*\|$. Let

$$\Delta = (1+a)(\|f - f_R\|^2 - \|f - f^*\|^2) + \frac{\lambda}{n}(\text{pen}(f_R) - \text{pen}(f^*)). \tag{23}$$

Then:

$$B_R \leq \frac{(1+a)\|f - f_R\|^2 + \frac{\lambda}{n}\text{pen}(f_R) + \frac{2\lambda(s_R+1)}{n}}{1-a} \tag{24}$$

$$= \frac{(1+a)\|f - f^*\|^2 + \frac{\lambda}{n}\text{pen}(f^*) + \Delta + \frac{2\lambda(s_R+1)}{n}}{1-a} \tag{25}$$

$$= B_\Gamma + \frac{\Delta - \frac{2\lambda(s-s_R)}{n}}{1-a} \tag{26}$$

The term $\frac{2\lambda(s-s_R)}{n}$ represents the complexity advantage of using the Rashomon set.

For the bound to be tighter, we need:

$$\Delta < \frac{2\lambda(s - s_R)}{n} \tag{27}$$

This inequality holds under reasonable conditions. Recall that Δ from equation (23) has two components:

1. $(1+a)(\|f - f_R\|^2 - \|f - f^*\|^2)$: By Lemma A.9, this term is small.
2. $\frac{\lambda}{n}(\text{pen}(f_R) - \text{pen}(f^*))$: Working under Case 2, since $f^* \notin \hat{R}(\mathcal{F}, \epsilon)$ but minimizes the penalized empirical risk, it must have a relatively small penalty term $\text{pen}(f^*)$.

Meanwhile, $\frac{2\lambda(s-s_R)}{n}$ represents the complexity advantage of the Rashomon set. By Lemma A.7, $s_R < s$, and when the Rashomon set is substantially smaller than Γ (as is typically the case in practice), this term becomes significant.

For active learning in particular, the Rashomon set focuses on regions of genuine ambiguity rather than noise-induced disagreement, making f_R a better approximation. Following the results of Willett et al. (2005), the effective complexity reduction outweighs any small increase in approximation error, especially for functions with concentrated complexity.

Therefore, $B_R < B_\Gamma$, which proves the theorem. □

Corollary A.11. *Combining Theorem A.10 and Theorem A.6 we get that any convergence rate established for the standard estimator $\hat{f}_n$ also applies to the Rashomon set estimator $\hat{f}_n^R$.*

Further, when using active learning strategies, the Rashomon set approach inherits their faster convergence rates for piecewise constant functions, while potentially improving constants and finite-sample performance due to its reduced complexity.

A.2. Geometry of trees in Group 1

The images in Figure 5 give insight into the decision rules of the top 12 decision trees of Figure 1.

As seen, the trees exhibit very similar decision paths to each other, resulting in each one having the *exact* same misclassification. As described in the main corpus, ensembling these trees as a committee and calculating the vote entropy metric off this committee will result in an inflated agreement and will recommend the observation that the best decision tree is most uncertain of rather than consider the uncertainty of the ensemble as a whole.

Note that Figure 1 categorizes trees by misclassification rate purely for visualization purposes. As detailed in Section A.3.2, categorizing trees by misclassification error can clump together different distinct classification patterns that happen to share the same misclassification rate. It is important to note, then, to group and ensemble trees based on their distinct classification patterns in the *UNREAL* algorithm.

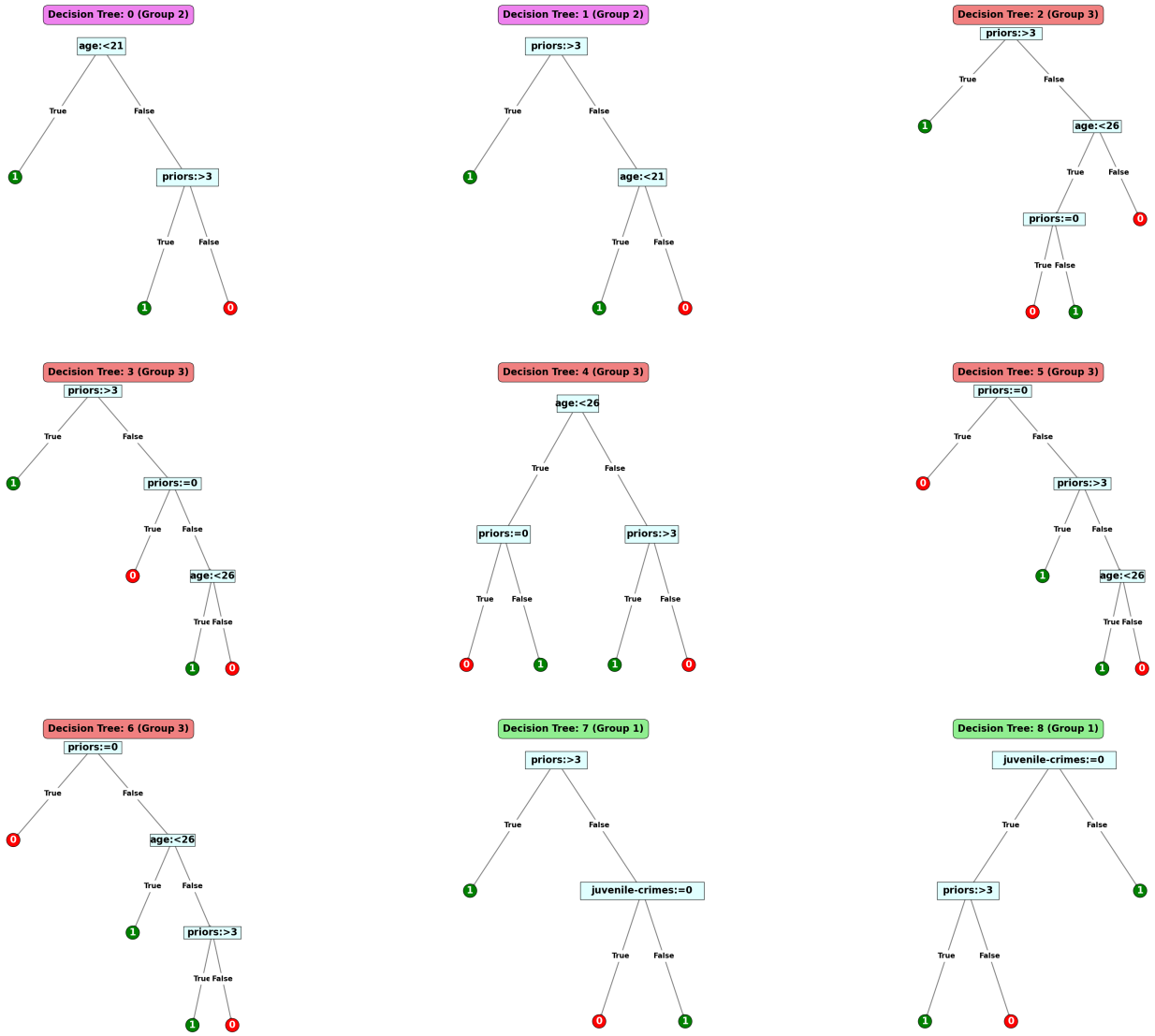

Figure 5. Geometry of the decision trees from Figure 1. Note that these near-optimal trees, although sharing similar similar accuracy, differ in specific regions of the feature space, namely how to split the feature prior. In the *UNREAL* algorithm, these trees would lead to a total of 4 members in the committee consisting of the four unique classification patterns. In *DUREAL*, the committee would be composed of all nine trees.

A.3. Simulation Details

A.3.1. Data Preprocessing

Table 1 contains information on the datasets used in our simulations.

We follow the same preprocessing of Xin et al. (2024) by one-hot encoding the numerical features whose thresholds were chosen by a gradient-boosted tree. The processed datasets were extracted directly from their Github. We also describe their preprocessing here.

The MONK-1 and MONK-2 dataset were straightforward to one-hot encode. For the Iris dataset, the function `qcut` from Python's `Pandas` package was used to cut the numerical features of Iris into three equal bins.

The COMPAS dataset, abbreviated for Correctional Offender Management Profiling for Alternative Sanctions, is a dataset for predicting which individuals will be arrested within two years of prison release. The dataset was discretized into binary

Table 1. Summary of our datasets.

DATASET	BINARY FEATURES	TEST SET	INITIAL TRAINING SET	INITIAL CANDIDATE SET	RASHOMON THRESHOLD ϵ	UNREAL MEAN RUNTIME (MIN.)	DUREAL MEAN RUNTIME (MIN.)
IRIS	15	30	24	96	0.025	3.26	3.23
MONK1	11	25	19	80	0.030	39.27	38.63
MONK3	11	25	19	78	0.019	131.86	132.66
BAR7	14	40	32	128	0.020	340.66	350.18
COMPAS	12	40	32	128	0.025	512.35	546.79

features in the same way as Hu et al. (2019). The binary variables are "sex = Female", "age $<$ 21", "age $<$ 23", age $<$ 26, "age $<$ 46", "juvenile felonies = 0", "juvenile misdemeanors = 0", "juvenile crimes = 0", "priors = 0", "priors = 1", and "priors = 2 to 3", "priors $>$ 3".

The Bar7 dataset contains information on whether a customer will accept a coupon for a bar considering demographic and contextual attributes. The dataset was discretized as follows: "Bar = 1 to 3", "Bar = 4 to 8", "Bar = less1", "maritalStatus = Single", "childrenNumber = 0", "Bar = gt8", "passenger = Friend(s)", "time = 6PM", "passenger = Kid(s)", "CarryAway = 4 to 8", "gender = Female", "education = Graduate degree (Masters Doctorate etc.)", "Restaurant20To50 = 4 to 8", "expiration = 1d", and "temperature = 55".

The Bar7 and COMPAS datasets were reduced to a fixed 200 observations due to computation time.

A.3.2. GROUPING UNIQUE CLASSIFICATION PATTERNS

Our *UNREAL* algorithm depends on ensembling only the unique classification patterns from `TreeFarms`. This requires grouping trees into categories with identical classification patterns. To determine which trees should be grouped into similar classification patterns, we found the prediction of each tree for all observations in the candidate set. We then grouped trees together if each and every single prediction was the same. Note that grouping trees by their misclassification rate (as depicted in Figure 1 only for visualization purposes) may merge classification patterns.

A.3.3. ACTIVE LEARNING SIMULATION PLOTS

Additional active learning simulation plots without errors being relative to random forests and with errors relative to passive learning are presented in Figure 6.

A.3.4. WILCOXON RANK SIGNED TEST

The following values present the p-values, rounded to 5 decimal points, of Wilcoxon ranked signed tests pairwise testing each active learning strategy.

A.4. How the Rashomon Threshold Affects the Number of Trees and Classification Patterns

Increasing the Rashomon threshold will obviously increase the number of trees in the Rashomon set. However, we wanted to see if this would also increase the number of unique classification patterns in the Rashomon set. This exploration is important as diversifying our QBC committee with new and different unique classification patterns will affect the query-selection criteria. As such, at each iteration, we collect the total number of trees and the number of unique classification patterns. Figure 7 includes both of these for the following three datasets: Iris, MONK-1, and MONK-3.

We find that increasing the Rashomon threshold greatly increases the number of trees in the Rashomon set, but not at the same magnitude as the number of classification patterns. The number of unique classification patterns is large at the beginning active learning process, but decreases as we begin to collect data. After collecting a sufficient amount of data points, the number of unique classification patterns is reduced to 1. For instance, in the Iris dataset, unique classification patterns range between e^2 and e^{10}, but reduce down to 1 when enough data is collected. The reduction to 1 unique explanation highlights the fact that our method initially

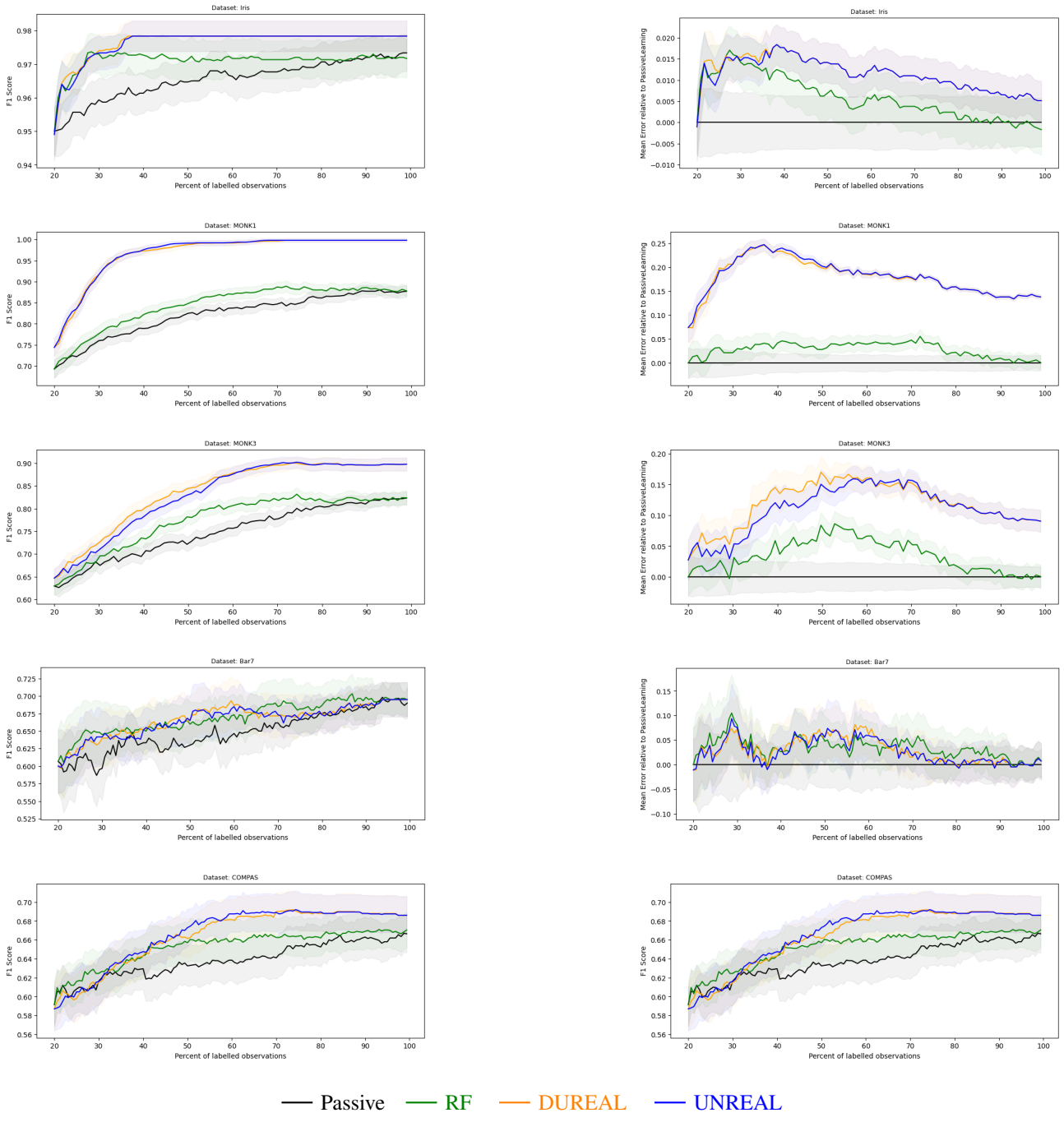

Figure 6. Performance of the four active learning procedures on the five datasets without relative error (left) and relative to passive learning (right).

Dataset	PassiveLearning	RandomForest	UNREAL	DUREAL
Iris				
PassiveLearning	1.0			
RandomForest	0.0	1.0		
UNREAL	0.0	0.0	1.0	
DUREAL	0.0	0.0	0.00398	1.0
MONK-1				
PassiveLearning	1.0			
RandomForest	0.0	1.0		
UNREAL	0.0	0.0	1.0	
DUREAL	0.0	0.0	6e-05	1.0
MONK-3				
PassiveLearning	1.0			
RandomForest	0.0	1.0		
UNREAL	0.0	0.0	1.0	
DUREAL	0.0	0.0	2e-05	1.0
Bar7				
PassiveLearning	1.0			
RandomForest	0.0	1.0		
UNREAL	0.0	1e-05	1.0	
DUREAL	0.0	0.00238	0.0073	1.0
COMPAS				
PassiveLearning	1.0			
RandomForest	0.0	1.0		
UNREAL	0.0	0.0	1.0	
DUREAL	0.0	0.0	0.0	1.0

Table 2. Wilcoxon ranked signed test comparing the statistical differences between the errors of our active learning procedures rounded to 5 decimal points.

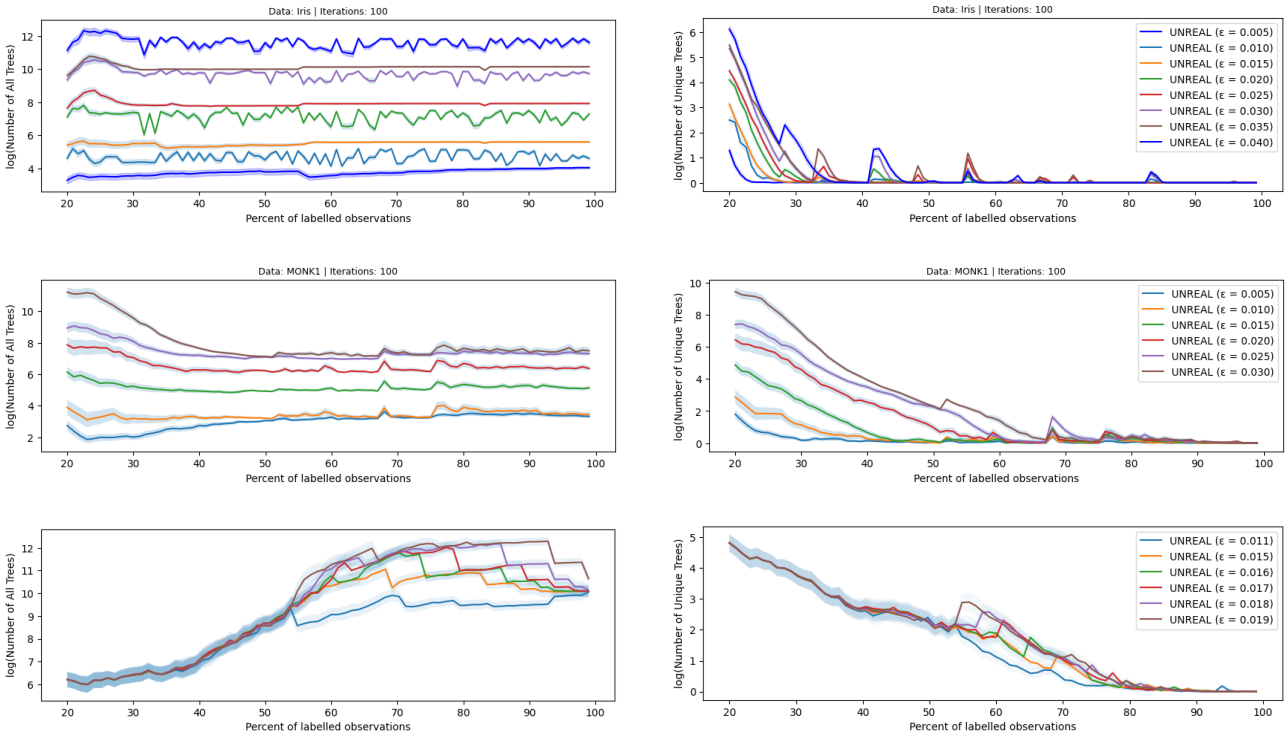

Figure 7. The left column indicates the log total number of Rashomon trees with the right indicating the log number of unique classification patterns.