← Go to **ICML 2025 Conference** homepage (/group?id=ICML.cc/2025/Conference)

Using Rashomon Sets for Robust Active Learning

PDF (/pdf?id=7WxKv9KfvS)

Simon Dovan Nguyen (/profile?id=~Simon_Dovan_Nguyen1), Kentaro Hoffman (/profile?id=~Kentaro_Hoffman1), Tyler McCormick (/profile?id=~Tyler_McCormick1)

23 Jan 2025 (modified: 21 Mar 2025) ICML 2025 Conference Submission Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=7WxKv9KfvS)

Verify Author List: I have double-checked the author list and understand that additions and removals will not be allowed after the abstract submission deadline.

TL;DR: We ensemble over the Rashomon set of near-optimal models to capture uncertainty across diverse and plausible explanations to provide and improved query selection procedure in active learning.

Abstract:

Active learning is based on selecting informative data points to enhance model predictions often using uncertainty as a selection criterion. However, when ensemble models such as random forests are used, there is a risk of the ensemble containing models with poor predictive accuracy or duplicates with the same interpretation. To address these challenges, we develop a novel approach called *UNique Rashomon Ensembled Active Learning (UNREAL)* to only ensemble the distinct set of near-optimal models called the Rashomon set. By ensembling over the Rashomon set, our method accounts for noise by capturing uncertainty across diverse yet plausible explanations, thereby improving the robustness of the query selection in the active learning procedure. We extensively evaluate *UNREAL* against current active learning procedures on five benchmark datasets. We demonstrate how taking a Rashomon approach can improve not only the accuracy and rate of convergence of the active learning procedure but can also lead to improved interpretability compared to traditional approaches.

Supplementary Material: zip (/attachment?id=7WxKv9KfvS&name=supplementary_material)

Primary Area: General Machine Learning->Online Learning, Active Learning and Bandits

Keywords: Rashomon Sets, Active Learning, Model Ambiguity, Noise, Uncertainty

Ethics Agreement: I certify that all co-authors of this work have read and committed to adhering to the Call for Papers, Author Instructions, and Publication Ethics.

Reciprocal Reviewing Status: This submission is NOT exempt from the Reciprocal Reviewing requirement. (We expect most submissions to fall in this category.)

Reciprocal Reviewing Author: Simon Dovan Nguyen (/profile?id=~Simon_Dovan_Nguyen1)

Submission Number: 14130

Filter by reply type... Filter by author... Search keywords... Sort: Newest First

Everyone | Program Chairs | Submission14130... | Submission14130... | Submission14130... *14 / 14 replies shown*

Submission14130... | Submission14130... | Submission14130... | Submission14130...

Submission14130... | ✖

Add: **Withdrawal** **Author AC Confidential Comments**

Official Review of Submission14130 by Reviewer 5FLW

Official Review by Reviewer 5FLW 12 Mar 2025, 00:18 (modified: 24 Mar 2025, 22:45)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 5FLW

Revisions (/revisions?id=6FmlnkxvDQ)

Summary:
This paper considers the ensemble with the Rashomon set used for active learning. The algorithm called the UNique Rashomon Ensembled Active Learning (UNREAL) is proposed. The used query strategy is the QBC, and the main learner is the random forests (not including DNN). The results of the experiments show the gain in active learning performance compared to the conventional ensemble QBC.

Claims And Evidence:
The use of the Rashomon set in ensemble QBC can improve performance in active learning, which has been validated.

Methods And Evaluation Criteria:
The proposed algorithm looks sound and uses the ratio of the accuracy gain. The trace plot of accuracy with labeled samples is popular in active learning, which does not appear in the main paper.

Theoretical Claims:
None

Experimental Designs Or Analyses:
It is too limited since the algorithms using deep neural networks or parameter perturbation approaches have not been examined.

Supplementary Material:
I can check the behavior of the ensemble trees for each dataset.

Relation To Broader Scientific Literature:
The study is closely related to efficient learning and obtaining more representative data points, which is essential in the development of AI.

Essential References Not Discussed:
None

Other Strengths And Weaknesses:
The application of the Rashomon set is impressive and sheds light on the use of QBC. However, the scope of active learning is too restricted. The baseline algorithms are not small, and the use of DNN is too limited, which can be crucial.

Other Comments Or Suggestions:
If you can validate the Rashomon ensemble QBC's superiority compared to BALD, Badge, or more advanced algorithms in active learning in the more complicated models, your paper can be strong.

Questions For Authors:
Q1: Do you have an upper bound in the Rashomon ensemble that can affect the performance of the proposed algorithm?
Q2: What's the computation time for the proposed algorithm compared to the baselines?

Code Of Conduct: Affirmed.
Overall Recommendation: 2: Weak reject (i.e., leaning towards reject, but could also be accepted)

Add: Author AC Confidential Comments

Rebuttal by Authors

Rebuttal

by Authors (Simon Dovan Nguyen (/profile?id=~Simon_Dovan_Nguyen1), Kentaro Hoffman (/profile?id=~Kentaro_Hoffman1), Tyler McCormick (/profile?id=~Tyler_McCormick1))

31 Mar 2025, 21:32 (modified: 01 Apr 2025, 07:29)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=GRdn4C10X8)

Rebuttal:

1. It is too limited since the algorithms using deep neural networks or parameter perturbation approaches have not been examined.

We thank the reviewer for their thoughtful comments. We agree that while many active learning approaches utilize neural networks, our work is primarily concerned with enhancing active learning by leveraging the Rashomon set when a random forest is used as the model to predict which point to label next.

Since this is the first work exploring Rashomon sets in the active learning framework, a primary goal was to evaluate just how useful the addition of a Rashomon selection step can improve the overall performance of active learning. However, there currently only exist algorithms to efficiently and exhaustively enumerate Rashomon sets for tree based models such as decision trees and random forests. Enumerating the Rashomon Set for other classes of models such as Neural Networks is an ongoing work (Zhong et al. (2023), Venkateswaran et al. (2024), Donnelly et al. (2025)).

In addition, many of these approaches require substantially more computational runtime compared to our approach. This is of particular issue as seen in Table 1 of the Appendix, the runtimes of our procedure already range from 3.26 minutes for simpler datasets to over 500 minutes for more complex ones.

Further, even if more runtime was available, many of these procedures do not exhaustively enumerate the entire Rashomon set, but rather give approximate enumerations. Thus for the first evaluation of the effectiveness of Rashomon sets of active learning, we deemed it important to focus our attention primarily on active learning approaches using tree based methods.

However, we do acknowledge that since this leaves out many of the most recent and exciting DL based active learning approaches, we propose the addition of a section where our procedure, RF + Rashomon is compared to Deep learning using BALD (Houlsby et al. 2011) and BADGE (Ash et al. 2020).

With this addition, we believe that we will now be more accurately meeting our initial claim: “We extensively evaluate UNREAL against current active learning procedures”

2. The trace plot of accuracy with labeled samples is popular in active learning, which does not appear in the main paper.

Thank you for pointing this out. Figure 3 in the main manuscript shows the performance plots, but presented as errors relative to random forests to highlight the comparative improvements. The traditional trace plots showing absolute accuracy are included in Figure 6 of the appendix. We're open to exchanging the plots in the main paper if this would make the results clearer to readers.

3. If you can validate the Rashomon ensemble QBC’s superiority compared to BALD, Badge, or more advanced algorithms in active learning in the more complicated models, your paper can be strong.

We appreciate this suggestion for strengthening our paper. As mentioned in our response point one, our primary goal was to evaluate the effectiveness of the Rashomon learning step for active learning and the choice of random Forest was a direct result of the availability of current algorithms for enumerating Rashomon sets. As a part of this, we tried to stick to active learning procedures like QBC which are relatively compatible with tree based models. To this end, Badge, which requires a gradient computation, does not seem a great fit for random forest models as it works based on a gradient-based procedure which is difficult to compute with respect to discontinuous methods such as random forest. BALD, on the other hand, requires choosing prototypes, which while adding another layer of parameter choice, seems more feasible to be combined with our Rashomon approach. Overall, we very much agree that the paper would be well strengthened by a comparison to these other approaches.

Therefore, we propose a comparison of methods that use tree based approaches to isolate the effect of the selection algorithm on Rashomon:

1. RF Rashomon + QBC vs Standard RF + QBC
2. RF Rashomon + Margin Sample vs Standard RF +Margin Sample
3. RF Rashomon + BALD vs Standard RF + BALD
4. RF Rashomon + BALD vs Standard RF + Batch BALD
5. RF Rashomon + kmeans++ vs Standard RF +kmeans++

And a comparison of our RF Rashomon + QBC procedure so one may understand how our results compare (in performance and runtime) when one is willing to leave the class of tree based models:

6. RF Rashomon + QBC vs DL + BADGE
7. RF Rashomon + QBC vs VAAL
8. RF Rashomon + QBC vs CAL

In such cases, both the model class and the presence of the Rashomon set are being changed so care must be taken in attributing the effect to the presence of the Rashomon set.

Add: **Author AC Confidential Comments**

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer 5FLW

Rebuttal Acknowledgement by Reviewer 5FLW 01 Apr 2025, 18:10

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Acknowledgement: I confirm that I have read the author response to my review and will update my review in light of this response as necessary.

Add: **Author AC Confidential Comments**

Official Review of Submission14130 by Reviewer EBdW

Official Review by Reviewer EBdW 11 Mar 2025, 09:47 (modified: 24 Mar 2025, 22:45)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer EBdW

Revisions (/revisions?id=Jz9N5eaiMY)

Summary:
This work proposes a new ensemble method for active learning with query-by-committee selection way. This method only ensembles the distinct set of near-optimal models called Rashmon set. These models generate diverse and plausible explanations. This method first uses an improved TreeFarms approach to generate the Rashomon set of decision trees, and then applies a voting-based way to select the samples to be labeled. Experiments on public benchmarks show that the proposed method has high performance, fast convergence rate, and improved interpretability.

Claims And Evidence:
Yes.

Methods And Evaluation Criteria:
Yes.

Theoretical Claims:
There is no theoretical claim in this work.

Experimental Designs Or Analyses:
The authors compare the proposed method with other methods on 5 public datasets. These datasets have different complexities on data generating functions. Simulation experiments on these datasets demonstrate the advantage of the proposed method over baseline methods.

However, one potential issue is the dataset size. Although multiple datasets have been utilized in the experiments, the size of these datasets seems to be small, which could impact the generalizability and practicality of the results. It would be more convincing if larger datasets could be used in evaluation, and consistent results could be observed.

Supplementary Material:
I have checked the implementation details described in the supplementary material, including the code.

Relation To Broader Scientific Literature:
This work contributes to the field of active learning by introducing a new model ensemble method. It is based on the previous TreeFarms approach, and addresses the correlation issue in TreeFarms by clustering-like grouping and selects the representative one as the model for ensemble. It selects diverse and plausible models with this simple yet effective way.

Overall, this work makes a meaningful contribution to the active learning field, considering its simplicity, effectiveness, and robustness for label noise.

Essential References Not Discussed:
No.

Other Strengths And Weaknesses:
N/A.

Other Comments Or Suggestions:
N/A.

Questions For Authors:
N/A.

Code Of Conduct: Affirmed.
Overall Recommendation: 3: Weak accept (i.e., leaning towards accept, but could also be rejected)

Add: **Author AC Confidential Comments**

Rebuttal by Authors

Rebuttal
by Authors (Simon Dovan Nguyen (/profile?id=~Simon_Dovan_Nguyen1), Kentaro Hoffman (/profile?id=~Kentaro_Hoffman1), Tyler McCormick (/profile?id=~Tyler_McCormick1))
31 Mar 2025, 21:33 (modified: 01 Apr 2025, 07:29)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
Revisions (/revisions?id=FeoSw3DAPZ)

Rebuttal:
However, one potential issue is the dataset size. Although multiple datasets have been utilized in the experiments, the size of these datasets seems to be small, which could impact the generalizability and practicality of the results. It would be more convincing if larger datasets could be used in evaluation, and consistent results could be observed.

We thank the reviewer for this fair point about dataset size. The computational expense of our approach is indeed a limiting factor in our ability to apply it to larger datasets. Computing the Rashomon Set at each iteration is already computationally intensive, and when we iteratively add observations one by one and retrain TreeFarms, this becomes exceptionally expensive, as reflected in Table 1 of the Appendix. To this end, we propose to provide information both on the computational complexity of the Rashomon + QBC algorithm as well as experiments on Batch Active Learning with our method where the Rashomon set is only recomputed for every k samples. Combined together, we believe that this will help the user decide how well this method scales and if it would be useful for their application.

Add: **Author AC Confidential Comments**

Replying to Rebuttal by Authors

Rebuttal Comment by Reviewer EBdW

Rebuttal Comment by Reviewer EBdW 01 Apr 2025, 14:39

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Thanks for the author's response. I will keep my original positive rating.

Add: **Author AC Confidential Comments** **Reply Rebuttal Comment**

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer EBdW

Rebuttal Acknowledgement by Reviewer EBdW 02 Apr 2025, 12:10

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Acknowledgement: I confirm that I have read the author response to my review and will update my review in light of this response as necessary.

Add: **Author AC Confidential Comments**

Official Review of Submission14130 by Reviewer etL9

(https://openreview.net/forum?id=7WxKv9KfvS&noteId=QwfQ4GGKjB)

Official Review by Reviewer etL9 09 Mar 2025, 04:40 (modified: 24 Mar 2025, 22:45)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer etL9

Revisions (/revisions?id=QwfQ4GGKjB)

Summary:
This paper introduces UNique Rashomon Ensembled Active Learning (UNREAL), a method that employs Rashomon sets (collections of near-optimal models) to improve active learning procedures. The authors argue that traditional ensemble-based active learning approaches suffer from redundancy and poor model diversity, potentially selecting suboptimal queries. Their proposed solution restricts ensemble aggregation to only the Rashomon set, claiming this accounts for uncertainty while maintaining valid explanations. The authors evaluate UNREAL against existing active learning methods on five benchmark datasets, reporting improvements in accuracy, convergence rate, and interpretability.

Claims And Evidence:
The claims made in this submission are not sufficiently supported by convincing evidence. The paper lacks clear metrics or experimental analyses explaining why previous methods are ineffective and why the current method works. There is no substantive evidence demonstrating the superiority of the proposed approach.

Methods And Evaluation Criteria:

The proposed method lacks novelty and appears to be an incremental modification of existing ensemble techniques for active learning. Finding some optimal-performing models for aggregation seems to be a trivial conclusion. The paper fails to provide reasonable characterization and verification of the method's effectiveness and innovation.

Theoretical Claims:
The theoretical foundation of this work is weak. There is essentially no concrete theoretical analysis, only some basic definitions and subjective intuitive conjectures without rigorous justification.

Experimental Designs Or Analyses:
The experimental design has several significant flaws:

1. The experiments are extremely limited, and comparisons with related work are even scarcer, to an appalling degree.
2. It's uncertain whether the datasets used are mainstream, and traditional Active Learning methods such as Coreset are completely absent from the evaluation.
3. There is no analysis of the method's sensitivity to hyperparameters, particularly the threshold used to define the Rashomon set.

Overall, the experimental evaluation is seriously inadequate.

Supplementary Material:
Yes.

Relation To Broader Scientific Literature:
None

Essential References Not Discussed:
So many papers need to be cited but not in the paper.

Other Strengths And Weaknesses:
The paper is far from being ready for submission. The entire content is hollow with excessive padding, scarce experimental data, and oversized images. The authors have even failed to adhere to ICML submission guidelines, with references placed in the appendix, causing significant reading difficulties.

Other Comments Or Suggestions:
I recommend that the authors substantially improve the paper, including proper experimental validation, theoretical justification, and correct formatting according to conference guidelines before submitting it for a new round of review.

Questions For Authors:
I recommend that the authors substantially improve the paper, including proper experimental validation, theoretical justification, and correct formatting according to conference guidelines before submitting it for a new round of review.

Code Of Conduct: Affirmed.
Overall Recommendation: 1: Reject

Add: **Author AC Confidential Comments**

Rebuttal by Authors (https://openreview.net/forum?id=7WxKv9KfvS&noteId=kdzCx7yq0n)

Rebuttal

by Authors (Simon Dovan Nguyen (/profile?id=~Simon_Dovan_Nguyen1), Kentaro Hoffman (/profile?id=~Kentaro_Hoffman1), Tyler McCormick (/profile?id=~Tyler_McCormick1))

01 Apr 2025, 00:25 (modified: 01 Apr 2025, 07:29)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=kdzCx7yq0n)

Rebuttal:

1. No evidence on the superiority of the proposed approach.

We thank the reviewer for this point. Our results in Fig. 3 demonstrate consistent improvements over both passive learning and RF approaches, with improvements of up to 20% in terms of predictive accuracy as shown in MONK1, Fig. 3. We acknowledge that comparison with additional current methods would strengthen our claims, and have addressed the reasons for our focus on RF-based methods in our response to Reviewer 1.

2. Missing literature

We thank the reviewer for highlighting the opportunity for us to better situate our work. We recognize our manuscript can benefit from a more comprehensive literature review that better contextualizes our contribution. We will expand our literature discussion to include:

1. Additional references on uncertainty and disagreement-based active learning strategies beyond the works we've already cited (Freund et al. 1997, Willett et al. 2005).
2. Recent surveys that provide broader context for how our approach relates to the current state of the field (Settles 2009, Liu et al. 2022, Mots'oehli and Baek 2023, Cacciarelli and Kulahci 2024).

3. Aggregating optimal-performing models is a trivial conclusion.

The authors appreciate the reviewers for raising this point as this highlights a key subtlety in Rashomon Sets.

While the idea of selecting optimal-performing models for aggregation may seem straightforward, this is not currently how the aggregation of tree based models is performed. Conventional approaches in ensemble methods for random forests aggregate weak learners formed through the bootstrapping of data and random selection of features, even if such models are poor-performing. Not only does this not enumerate the entire collection of near-optimal models (which TreeFarms does), but such aggregation also often contains many poor-performing models. Thus, our work seeks to illustrate a shift from traditional ensemble approaches that rely on randomization to achieve diversity to aggregating optimal models from the Rashomon set of decision trees.

To emphasize this point further, we propose to

1. Emphasize paragraph 6 of Section 2: Rashomon Sets (“Traditional ensemble methods on the other hand...”). In particular, we intend to emphasize this paragraph’s statement to more clearly contrast how RF incorporates potentially suboptimal models through randomization, whereas our approach aggregates only high-performing models with diverse explanations.
2. Add a paragraph contrasting traditional RF ensemble construction with our Rashomon-based approach in the context of QBC in Section 3.2. We seek to highlight that random forests do not guarantee optimal or even near-optimal weak learners.
3. Include a discussion in Section 5 about how our method's performance improvements are attributable to the fact that aggregating the near-optimal models of the Rashomon set improves over the current practice of aggregating over weak learners with bootstrapped data and features.

4. No analysis of the method's sensitivity to hyperparameters

We thank the reviewer for this observation. While Section 5.1, and particularly Fig. 2, shows how we initialized our Rashomon threshold, we understand the selection of other hyperparameters such as the number of trees in a random forest may affect our results. In our experiments, we noted three key hyperparameters:

1. The number of trees in the random forest baseline (100 trees)
2. The regularization of tree complexity for the Rashomon set (0.05)
3. The Rashomon threshold (optimized for each dataset as in Fig. 2). Our manuscript would benefit from additional sensitivity regarding
4. How varying the number of trees in the RF baseline affects comparative performance with our methods.
5. How different regularization values on Rashomon tree complexities affects comparative performance with our methods.

However, we do note that Fig. 7 of the appendix does provide some insight into hyperparameter sensitivity, as it shows how varying the Rashomon threshold affects both the total number of trees and the number of unique classification patterns throughout the active learning process. The figure demonstrates that while higher thresholds consistently produce more trees in the Rashomon set, the number of unique classification patterns remains relatively stable. This suggests that our method's performance may be robust to the exact threshold value as long as it's sufficiently large to capture diverse classification patterns. In our revision, we will enhance our discussion of Fig. 7 to better highlight these relationships and provide a more comprehensive analysis of hyperparameter sensitivity. We will also consider including additional experiments examining the sensitivity to the other hyperparameters mentioned above if computational resources permit.

5. ICML submission guidelines The authors apologize for the misplacement of the references in the appendix instead of the main manuscript.

Add: Author AC Confidential Comments

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer etL9

Rebuttal Acknowledgement by Reviewer etL9 02 Apr 2025, 02:15

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Acknowledgement: I confirm that I have read the author response to my review and will update my review in light of this response as necessary.

Add: Author AC Confidential Comments

Replying to Rebuttal by Authors

Rebuttal Comment by Reviewer etL9

Rebuttal Comment by Reviewer etL9 02 Apr 2025, 02:21

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

I appreciate the authors' detailed response and rebuttal, which partially addresses some of my concerns. However, I still maintain that this manuscript is not yet ready for publication in its current form. I encourage the authors to thoughtfully integrate the recommendations provided by all reviewers to strengthen the paper. Specifically, I recommend conducting comprehensive experiments that incorporate advanced algorithms in active learning and presenting these complete results before submitting the manuscript for publication consideration.

Add: Author AC Confidential Comments Reply Rebuttal Comment

Official Review of Submission14130 by Reviewer bzLy

Official Review by Reviewer bzLy 08 Mar 2025, 17:57 (modified: 24 Mar 2025, 22:45)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer bzLy

Revisions (/revisions?id=E8m6rJsEZf)

Summary:

The paper propose UNREAL (UNique Rashomon Ensembled Active Learning), a novel active learning framework that leverages the Rashomon set to improve robustness against label noise and redundant models. Traditional ensembles methods like random forests may include poor or duplicate models, which can skew uncertainty metrics. UNREAL addresses this by ensembling only unique classification patterns within the Rashomon set, ensuring diversity while maintaining high performance. Experiments on five datasets demonstrate that UNREAL outperforms random forest-based active learning and faster convergence. The method also enhances interpretability by pruning redundant models. Key findings include the Rashomon set's ability to capture genuine uncertainty in noisy data and the trade-off between redundancy reduction and implicit weighting of complex patterns in variants like DUREAL.

Claims And Evidence:

The claims are generally supported by empirical evidence:

(1) Rashomon sets improve robustness: Supported by improved performance in noisy datasets (e.g., COMPAS) and analysis of pattern diversity under noise.

(2) Unique patterns enhance interpretability: Illustrated via classification pattern grouping (Figure 1) and comparable performance between UNREAL and DUREAL.

(3) Superiority over baselines: Results in Figure 3 show consistent gains over random forests.

However, weaker evidence exists for:

(1) Threshold selection: The static Rashomon threshold (e.g., ϵ=0.016ϵ=0.016) is justified via initial training data (Figure 2), but its adaptability during active learning is untested.

(2) Bar7 performance gap: The smaller improvement on Bar7 is not thoroughly analyzed, leaving open questions about scalability to highly complex datasets.

(3) Comparison to Bayesian methods: The paper does not compare with uncertainty-based baselines like BALD or BatchBALD, which are standard in active learning. This omission weakens claims about superiority over "current active learning procedures" (Abstract).

Methods And Evaluation Criteria:
Methods: Using TreeFarms to enumerate Rashomon sets and pruning duplicates is logical for reducing redundancy. However, TreeFarms’ computational cost and correlation between trees (due to no feature/data sampling) are acknowledged limitations.

Evaluation: Standard datasets (e.g., Iris, COMPAS) and F1 scores are appropriate. However:

(1) Older datasets (e.g., Iris) may not reflect modern challenges.

(2) Varying run counts (100 for Iris vs. 15 for Bar7) could affect statistical validity.

(3) Missing baselines: Key uncertainty-based methods like BALD and batch strategies like BatchBALD are absent from experiments, limiting the scope of comparison.

Theoretical Claims:
The paper builds on existing Rashomon set theory and TreeFarms but does not introduce new theoretical proofs. The justification for unique classification patterns is empirical rather than theoretical. While sufficient for the applied focus, a formal analysis of uniqueness guarantees would strengthen the method’s foundation.

Experimental Designs Or Analyses:
Threshold selection: The static ϵ (based on initial data) may not adapt to evolving training sets, potentially limiting long-term performance.

(1) Varying run counts: Fewer runs for Bar7 (due to computational limits) reduce confidence in results.

(2) Dynamic ϵ: Suggested as future work but not implemented, leaving a key limitation unaddressed.

(3) Baseline omissions: The lack of comparison to BALD, BatchBALD, or stochastic batch selection methods weakens the evaluation. These methods are widely used in Bayesian and batch active learning, and their exclusion makes it difficult to assess UNREAL’s novelty in uncertainty estimation.

Supplementary Material:
I did not read.

Relation To Broader Scientific Literature:
The work connects to:

(1) Active learning: Builds on QBC and vote entropy.

(2) Rashomon sets: Extends work on noise-induced diversity and TreeFarms.

Essential References Not Discussed:
(1) BALD & BatchBALD: Critical for uncertainty estimation in Bayesian frameworks.

(2) Deep ensembles: Lakshminarayanan et al. (2017) for comparison with non-Bayesian uncertainty methods.

These works are essential to contextualize UNREAL’s contributions in uncertainty quantification and batch selection.

Other Strengths And Weaknesses:
(1) Computational cost of TreeFarms limits scalability.

(2) Static threshold selection and reliance on older datasets.

(3) Baseline gaps: Missing comparisons to BALD, BatchBALD, and stochastic/diversity-based batch methods weaken the evaluation.

(4) Limited analysis of Bar7's underperformance.

Other Comments Or Suggestions:
(1) Typo: "Rashomoon cutoff" (Page 5).

(2) Code availability could be a strength.

(3) Consider including runtime comparisons in the main text.

(4) Expand related works to discuss Bayesian and batch baselines.

Questions For Authors:
Baseline Comparisons: Why were BALD, BatchBALD, or diversity-based batch methods not included in experiments? Would their inclusion affect the claimed superiority of UNREAL?

Code Of Conduct: Affirmed.
Overall Recommendation: 2: Weak reject (i.e., leaning towards reject, but could also be accepted)

Add: Author AC Confidential Comments

− = ≡

Rebuttal by Authors

Rebuttal
by Authors (Simon Dovan Nguyen (/profile?id=~Simon_Dovan_Nguyen1), Kentaro Hoffman (/profile?id=~Kentaro_Hoffman1), Tyler McCormick (/profile?id=~Tyler_McCormick1))

31 Mar 2025, 21:48 (modified: 01 Apr 2025, 07:29)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=QvrxvU22NH)

Rebuttal:
1. Threshold selection: The static Rashomon threshold (e.g., ϵ=0.016ϵ=0.016) is justified via initial training data (Figure 2), but its adaptability during active learning is untested

We thank the reviewer for this insightful comment. As noted in Sections 5.1 and 7 of our paper, we identify the initial selection of the Rashomon threshold as a limitation of our approach. We discuss this limitation in Section 7, where we note that dynamically recalibrating ε at each iteration would be theoretically ideal but computationally expensive. To help readers decide if such adaptability is worth the computational cost, we propose the inclusion of a batch updating simulation where ε is dynamically recalibrated but is only performed every k steps as well as its comparison to static ε . By comparison the performance of the batch updating Rashomon with the non-batch version, the user has a reference to decide if dynamic ε is sufficient for the extra runtime cost

2. Bar7 performance gap: The smaller improvement on Bar7 is not thoroughly analyzed, leaving open questions about scalability to highly complex datasets.

We thank the reviewer for highlighting the need for a more thorough analysis of the Bar7 dataset results.

The Bar7 dataset presents the highest complexity and noise level among our test datasets due to its real-world origin in restaurant customer behavior. We note several important factors that likely contribute to the results observed:

1. The limited number of simulation runs (only 15 due to computational constraints) introduces higher variability in the performance estimates. This makes it more difficult to draw definitive conclusions about performance differences. We thus did not wish to strongly comment on the relative performance, however, we agree that as is, there currently lacks any discussion on this point.

2. While our results show small gaps (~2.5%) between our methods and random forests, we note the advantages in portions of the trace plot, particularly between 50-65% of labeled observations.

In the revised manuscript, we will expand our discussion of these results to provide these insights.

3. Comparison to Bayesian methods: The paper does not compare with uncertainty-based baselines like BALD or BatchBALD, which are standard in active learning. This omission weakens claims about superiority over "current active learning procedures" (Abstract). Why were BALD, BatchBALD, or diversity-based batch methods not included in experiments? Would their inclusion affect the claimed superiority of UNREAL?

As addressed in our response to Reviewer 1, our primary goal was to investigate the effectiveness of including a Rashomon updating step when performing active learning. Our focus on tree based methods was a direct consequence of the currently available algorithms for fully enumerating Rashomon sets. While some algorithms such as BADGE are not very compatible with tree based models (requiring the computation of a gradient), we agree that prototype methods such as BALD and Batch BALD could be combined with tree based methods (by choosing different depths to prune the tree). Therefore, as mentioned in our response to Reviewer 1, we propose additional simulations comparing:

1. RF Rashomon + QBC vs Standard RF + QBC
2. RF Rashomon + Margin Sample vs Standard RF +Margin Sample
3. RF Rashomon + BALD vs Standard RF +BALD
4. RF Rashomon + BALD vs Standard RF +Batch BALD
5. RF Rashomon + kmeans++ vs Standard RF +kmeans++

Add: **Author AC Confidential Comments**

− = ≡

➦ *Replying to Rebuttal by Authors*

Rebuttal Acknowledgement by Reviewer bzLy

Rebuttal Acknowledgement by Reviewer bzLy 📅 02 Apr 2025, 16:24

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Acknowledgement: I confirm that I have read the author response to my review and will update my review in light of this response as necessary.

Add: **Author AC Confidential Comments**