Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Enhancing Stability in Rule-Based Post-Hoc Explanations

Iain N. Smith[†,*], Osmar R. Zaiane[†]
[†] University of Alberta, Edmonton, Canada

Abstract

To trust an explanation, it must also stay the same — or at least be similar — when repeated. In much of the existing work this variance, called instability, is caused by random perturbations to the sample being explained. But this is a limited view, so in this work we study stability metrics when the data used to produce perturbations are unstable. We assess powerful explainers which use rules, where explanation instability stemming from training data becomes more apparent, and we theorize why multivariate normal distributions, producing correlated perturbed training data (+P) improve stability and fidelity in our setting. We also balance classes in the training data (+B) to further improve stability, along with exploring the potential of clustering (+C) for locality improvements to explanations. By providing both theoretical reasoning for the improvements and experiments on seven diverse datasets, with two different black-box architectures we found that the rule-based method we employed, BARBE, sharply increased in stability when trained with our modified process. BARBE+PB further exceeded the performance of other methods that improve stability like S-LIME and LORE. The final codes are available as a package on GitHub at https://github.com/IainNBSmith/Stable-BARBE.

Keywords: Explainable AI, post-hoc explanations, stability, associative classifiers

1. Introduction

Explanations for black-box decisions are essential for the black-box's users and are even becoming a regulatory requirement in Canada [1]. So proactive organizations are using post-hoc explanations of their pre-trained black-box models as a critical tool to fulfill these needs. If such an explanation is untrue or inconsistent in a critical area like medicine, then it risks harm to a patient, while an untrue explanation causes obvious harm — like pursuing an incorrect treatment plan — an inconsistent explanation harms trust, a difficult to measure metric which is still of upmost importance to explanations.

Inconsistent explanations could be caused by two features which are not always considered. The surrogate model of a post-hoc explanation determines feature importance, so it could be most at fault for stability issues. Missing and outlying data may also be a source of instability, and due to a lack of severe evaluation in explainable AI (XAI) most methods do not expose this source of instability and test against it [2]. In this work, we aim to improve the stability of rule-based post-hoc explanation methods and heighten the severity of evaluations to address this significant gap.

In this work we propose and evaluate BARBE+PB, a modification to the BARBE [3] explainer to improve stability. We evaluate improvements theoretically and identify multivariate perturbations (+P) as an existing — but unproven — stabilizing change alongside balancing classes in the perturbations (+B) and clustering training data (+C). Unlike other approaches, we use severe statistical testing and identify multiple sources of instability stemming from the training data. The main contributions of this work are as follows:

- C1: Showed theoretically that multivariate perturbations (+P) and class balancing (+B) improve the stability of rule-based explanation methods.
- C2: Showed that the error added by clustering (+C), when there are 2 groups, does not outweigh prior stability improvements.

* ins@ualberta.ca

C3: Provided strict and severe significance-based tests on seven datasets to show that the theory holds in application, using the Jaccard index and Relative Input Stability for evaluation.
C4: Provided these modifications in an easy-to-access GitHub repository (https://github.com/IainNBSmith/Stable-BARBE) implemented in the Python programming language.

We present the background related to post-hoc explainability in Section 2. In Section 3, we discuss BARBE — the method we advocate for. In Section 4, we formally describe the stability problem for post-hoc explanations along with some theory that is useful for improving BARBE. We apply the theory in Section 5, splitting up between the known multivariate perturbations and our new improvements, class balancing and clustering. Section 6, describes the experiments we use to show these results on real data and we present the results and conclusions in Sections 7 and 8.

2. Background

Post-hoc explanations deal with a pre-trained "black-box" model, a model that does not provide an explanation and is difficult, or improbable, to sufficiently interpret. Many post-hoc explanation approaches exist [4–6] and a particular area of focus for this work is methods that use an interpretable surrogate, or "white-box," model. A white-box is explainable or interpretable and it is trained to give similar predictions to the black box. Many of these approaches create a "local" explanation for a point-of-interest by training the white-box on perturbations of this point, like LIME [4] which also provides an explanation in the form of feature importance values. LORE [6] improves explanations by instead providing a supporting rule from a trained decision tree, but this may be limited as an explanation since the understanding of a complex black-box is reduced to one branch of a decision tree.

The stability of a post-hoc explanation is usually quantified by how consistent the explanation is, this can be done using the Jaccard index [6, 7], which prioritizes the relative order of importance values, or by comparing the importance of a single feature across runs using the Relative Input Stability (RIS) [8]. The stability of popular methods like LIME has been contentious, with many approaches aiming to address the problem [7, 9]. However, existing approaches like S-LIME [7] do not correct for other significant metrics, a common issue in the area which reduces an XAI method's trustworthiness.

3. BARBE

Black-box Association-Rule Based Explanations (BARBE) [3] is a post-hoc explanation approach that uses association rules R of the format $R : X \rightarrow Y = y$ for a set of feature-value pairs X predicting a class y, to explain a sample. These rules can also be put into a human readable format, for example: "$\$5k \leq income < \$10k$ (X) implies $(\rightarrow)$ rejected loan $(Y = y)$". The associative classifier is trained using perturbations of the point-of-interest x_p, $x \sim \mathcal{N}(x_p, \Sigma)$ on a multivariate normal distribution with covariance Σ, either provided by a user or estimated using training data. The most significant rules for predicting the black-box's labeling $f(x) = y$ are found by using the Fisher's exact score, sig(R) [10].

The $-log(sig(R))$ is then used to estimate the feature importance. Where for each rule which applies to x_p with a certain number of input conditions $|R|$, the feature importance is $imp(A) = \sum_{A \in R} (-1)^{\mathbf{1}(R \rightarrow Y = y)} \frac{log(sig(R))}{|R|}$. So, the feature importance is based on the cumulative significance of the rules which feature A appears in, scaled by the number of conditions which must be satisfied, and with a sign based on the class being predicted.

4. Theoretical Improvements to Stability

Earlier in Section 2, we discussed that stability metrics are usually based on the amount that feature importance changes between runs, so we should focus on methods for improving the bounds of the feature importance, $imp(A)$, to yield continual improvements to stability. First, we consider the distribution of the importance as a continuous value sampled based on the training data.

$$imp(A) = E[imp(A)] + (L + L_\epsilon)\epsilon \leq E[imp(A)] + \frac{d(d+1)}{2}\epsilon \tag{4.1}$$

Where $(L_\epsilon)_{ij} \sim \mathcal{N}(0,1)$ are errors resulting from estimating the covariance matrix from a finite dataset, decomposed into $\Sigma = L^T L$, $\epsilon \sim \mathcal{N}(0,1)$ is normal error, and d is the number of noisy parameters in the training data. We simplify the compounded noise to $d(d+1)/2$ since the same error is reflected over the diagonal of Σ. Suppose that we found a set of important rules for a target feature A, satisfying the following inequality compared to an insignificant feature B.

$$p((imp(A) - \epsilon_L) - (imp(B) + \epsilon_H) \geq \frac{d(d+1)}{2}\epsilon) \geq \delta \tag{4.2}$$

If the above criterion holds, then we can consider the result to be stable to the right-hand-side (RHS) error $d(d+1)\epsilon/2$ with a significance δ. If we are able to produce tighter bounds, ϵ_L and ϵ_H, on the estimated importance values, then $imp(A)$ has higher stability. In comparison, if we use a standard normal distribution without correlation, then the inequality will have a lower error, so we need $d\epsilon \leq \frac{d(d+1)}{2}\epsilon$. The full details of each improvement that follows and all assumptions are available in the appendix.

Improvements from Multivariate Perturbations (+P) We apply improved bounds when rules are correlated as per Bax and Ouimet, leading to a $\sqrt{c}$ improvement [11]. By reordering the error comparison, we find a ratio where the overall error improves and the resulting approach becomes more stable when $\epsilon \leq \frac{2\epsilon_L}{d(d-1)}(1 - \frac{1}{\sqrt{c}})$.

Improvements from Class Balancing (+B) By including the class bias as a rule, $\phi \to Y = y$, we can further improve this bound further and by balancing the classes, y, we include this rule — and most importantly its coverage — in the probability. This results in a further improvement to the relationship, $\epsilon \leq \frac{2\epsilon_L}{d(d-1)}(1 - \frac{1}{\sqrt{c+1}})$.

Looking Forward: Improvements from Clustering (+C) When using clusters, we consider that there are k clusters with separate errors being estimated for Σ_i, $1 \leq i \leq k$. As a result, the ratio for improvement worsens, but if both sides are improved, ϵ_L and ϵ_H, then we can find results where stability is not significantly worsened when there are at most two clusters $k = 2$, $\epsilon \leq \frac{2(\epsilon_L + \epsilon_H)}{kd(d-1)}(1 - \frac{1}{\sqrt{c+1}})$.

5. Applied Improvements to Stability

In this section, we propose the applied algorithms which implement the improved bounds we discussed in Section 4. Recall that **Correlated Perturbations (+P)** are already applied to BARBE [12]. Balancing has been used for performance improvements in AI applications before [13] and is usually ensured by explainers that search for samples — like LORE [6] — but balancing perturbed data is still rarely applied in post-hoc explanations. We propose using a simple under-sampling approach where we produce tn perturbations, where n is the original number of perturbations and t is the number of times we need to produce perturbations to have $n/2$ samples of each class. We then under-sample these data to get only $n/2$ samples of the larger class. Under-sampling is used to keep the sampling

as simple as possible, but in testing we also use the SMOTE [14] balancing approach as an alternative.

6. Experimental Setup

In other works [6, 7], experiments for stability keep the training data and input conditions the same and only evaluate instability caused by the random perturbations used to train the white-box model $x \sim \mathcal{N}(x_p, \Sigma)$. In our experiments, we consider the effect of different training data being passed to the explainer, which has an impact on the estimated Σ used to perturb the sample. We sample a random 1/3 of the training data to use for each of 10 repetitions which are used to calculate the stability of the explanation. Based on the theory our approach's stability should still improve in this setting. To use RIS [8] we also make small changes to x_p, $x'_p \sim \mathcal{N}(x_p, \Sigma/10)$, we divide the covariance Σ by 10 to keep the changes so small that they will not realistically behaviourally change the sample.

For BARBE, we compare each improvement independently and in combination (e.g. +P, +B, +PB) alongside the algorithm used to implement the change (for example SMOTE vs. under-sampling). For the final comparison against competitors we use the most stable implementation of BARBE which was found in these tests.

Reproducibility Codes are available on GitHub (https://github.com/IainNBSmith/Stable-BARBE) randomization for subset and data selection use pre-defined seeds to ensure data do not change across tests and devices. Pre-trained models are also available. LIME, S-LIME, SHAP, and LORE are run with default settings, some added codes are used to make output formats consistent for experiments.

7. Results

As expected in theory, adding perturbations and balancing each significantly ($p<0.001$) improved the stability of BARBE (Figure 1 and Table 1 Jaccard +P, +B, +PB) — while retaining high fidelity — without making the explanations significantly more consistent across samples (Inverse Jaccard ($d = N/2$)). So explanations from BARBE+PB are still unique even with the added stability.

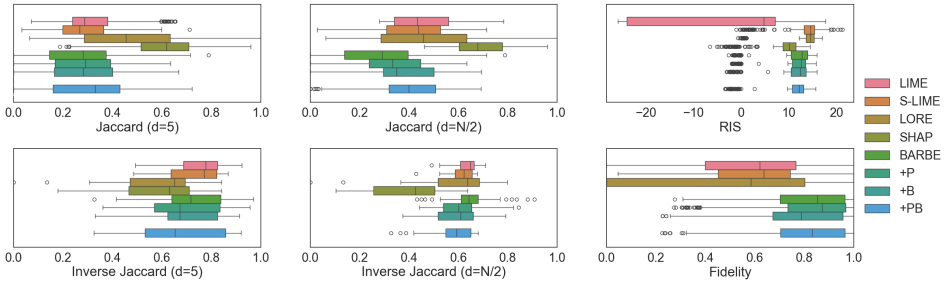

Figure 1. Box plots of each evaluation metric for competitors and our approaches (+P, +B, ...) for **all datasets and both black-box models** on each evaluation metric.

The improvement to stability from class balancing was beyond our expectations (Table 1, BARBE vs. +B $p<0.001$, +P vs. +PB $p<0.001$). This occurs without increasing the stability between different samples, so it follows that the number of correlated rules is low, $1/\sqrt{c} \geq 1/\sqrt{5}$ thus $c+1$ represents a significant improvement. Alternatively, the class could be correlated with many rules, leading to $1/\sqrt{c} \leq 1/\sqrt{3}$ improvement to many bounds.

While a similar, rule-based method LORE may have slightly higher stability than BARBE, LORE has lower fidelity (Figure 1 Fidelity), so we expect that the explanations from BARBE+PB align better to true, local explanations than LORE does.

	LIME	**S-LIME**	**LORE**	**SHAP**	**BARBE**	**BARBE+P**
+P	F**	RIS**, F**	RIS**, F**	-	5**, N/2**, F**	-
+B	F**	RIS**, F**	RIS**, F**	-	5*, N/2**	-
+PB	F**	5*, N/2*, RIS**, F**	F**	-	5**, N/2**, F**	5**, N/2**

Table 1. Paired sample t-Test for **all datasets and black-boxes**, results for our methods compared to existing methods and competitors (BARBE+P [12], previously untested for stability). Results of 5 is for t-tests of Jaccard (d=5), N/2 is for Jaccard ($d = N/2$), RIS is for the RIS statistic, and F is for Fidelity. We use "-" when there are no rejected null hypotheses. We record two levels of significance for rejected null hypotheses, *(pvalue$<$0.005) and **(pvalue$<$0.001).

In our experiments, we found that the stability of S-LIME is worse or negligibly better than LIME, unlike previous results which used static training data [7]. SHAP is quite stable, but this also holds across different samples since the Inverse Jaccard value is low. This result suggests that SHAP produces a consistent overall explanation, which is likely the global explanation of the model, rather than a local explanation for the sample, leading every explanation to be similar and stable.

When discussing the underlying theoretical work, we emphasized that either ϵ needs to be low or ϵ_L and ϵ_H need to be high. This result is best visualized on specific datasets (Figure 2 in the appendix), like the COMPAS data. With a large number of training samples (7214) relative to the number of features (28), we expect that ϵ will be lower. We also see that the best stability improvements for BARBE+PB appear in this group. So, as we expected, not every dataset is going to be clearly significantly improved by our approach, but a key subset of slightly-less noisy datasets will yield significant improvements which are important for establishing user trust in the stability of BARBE+PB.

Experiments were conducted for clustering (+C), with varying success. Most notably, the chosen approaches lowered the explanation's fidelity, so these results were moved to the appendix.

8. Future Work and Conclusion

Future Work Our discussion of SHAP, concluded that the high stability it exhibits is likely a cause of it being anchored to the static model of the black-box. We concluded that this is a result of SHAP using a global explanation rather than an unstable local explanations from training data. The clustering approach could benefit from this too if we implemented changes to the algorithm to anchor it to the black-box. To our knowledge, there are very few clustering algorithms designed which require a class-label, and fewer algorithms exist which balance the classes and even base clusters on their fit to a point-of-interest. This could be an interesting future research direction for post-hoc explanations.

Conclusion In this work, we explored two new techniques to surrogate-based post-hoc explanations, balancing (+B) and clustering (+C), as approaches for improving the stability and locality of a post-hoc explanation. We used a theoretical standpoint to establish the consistent stability improvements yielded by both balancing and the pre-existing improvement of using multivariate perturbations (+P). By conducting thorough experiments and statistical testing, we showed that the theoretical implications hold, but also that there is further room to improve for BARBE. As the approach with the highest fidelity — and the most likely to produce meaningfully local explanations — BARBE+PB greatly improved stability over BARBE and BARBE+P. When using the Relative Importance Stability, our

approach also significantly outperforms both S-LIME and LORE. These results not only reinforces why we should conduct severe testing in XAI — as some stability gains were unclear from just medians — but also for why we should use theoretical reasoning. Both have helped us show that BARBE+PB is a significantly more reliably stable explanation approach than the original BARBE and BARBE+P alone.

References

[1] Parliament of Canada. *Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*. 44th Parliament, 1st Session. 2022. URL: https://www.parl.ca/legisinfo/en/bill/44-1/c-27.

[2] D. S. Watson. "Conceptual challenges for interpretable machine learning". In: *Synthese* 200.2 (2022), p. 65.

[3] M. Motallebi, M. T. A. Anik, and O. R. Zaïane. "Explaining decisions of black-box models using barbe". In: *International Conference on Database and Expert Systems Applications*. Springer. 2023, pp. 82–97.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

[5] S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 4768–4777. ISBN: 9781510860964.

[6] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. "Local Rule-Based Explanations of Black Box Decision Systems". In: (2018).

[7] Z. Zhou, G. Hooker, and F. Wang. "S-LIME: Stabilized-LIME for Model Explanation". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. Virtual Event, Singapore: Association for Computing Machinery, 2021, 2429–2438. ISBN: 9781450383325. DOI: 10.1145/3447548.3467274. URL: https://doi.org/10.1145/3447548.3467274.

[8] C. Agarwal, N. Johnson, M. Pawelczyk, S. Krishna, E. Saxena, M. Zitnik, and H. Lakkaraju. "Rethinking Stability for Attribution-based Explanations". In: *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*. 2022. URL: https://openreview.net/forum?id=BfxZAuWOg9.

[9] X. Xiang, H. Yu, Y. Wang, and G. Wang. "Stable local interpretable model-agnostic explanations based on a variational autoencoder". In: *Applied Intelligence* 53.23 (Sept. 2023), 28226–28240. ISSN: 0924-669X. DOI: 10.1007/s10489-023-04942-5. URL: https://doi.org/10.1007/s10489-023-04942-5.

[10] W. Hamalainen. "Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures". In: *Knowledge and Information Systems* 32 (2012), pp. 383–414.

[11] E. Bax and F. Ouimet. "Bounding Means of Discrete Distributions". In: *2021 IEEE International Conference on Big Data (Big Data)*. 2021, pp. 5024–5032. DOI: 10.1109/BigData52589.2021.9671544.

[12] I. Smith and O. Zaiane. "Faithful Perturbations and Evaluations for Post-Hoc Local Explanation Methods". In: *Proceedings of the Canadian Conference on Artificial Intelligence* (2025). https://caiac.pubpub.org/pub/gs2ywmlt.

[13] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. P. Chen. "A survey on imbalanced learning: latest research, applications and future directions". In: *Artificial Intelligence Review* 57.6 (2024), p. 137.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *J. Artif. Int. Res.* 16.1 (June 2002), 321–357. ISSN: 1076-9757.

Appendix A. Theoretical Assumptions

Assumption 1. *$imp(A)$ is **stable** if the difference between its low bound ϵ_L and the high bound ϵ_H of an alternative $imp(B)$ is greater than the noise from randomly sampling importance values, $(L + L_\epsilon)\epsilon \leq \frac{d(d+1)}{2}\epsilon$.*

A.1. Multivariate Perturbations (+P)

Assumption 2. *The value being estimated, $imp(A)$, is the **function of a set of discrete events**, like significant rules being used in an explanation. Where $E[imp(A)]$ is the actual explanation of the black-box.*

Assumption 3. *Frequencies of rules in $imp(A)$ can be **"combined,"** reducing the total number needed to estimate the bound. As per Bax and Ouimet [11], this leads to an improvement of $\sqrt{c}$ on the error of each bound involving the c combined frequencies.*

Assumption 4. *We cannot exceed a **training set size of 1000 samples** to estimate $imp(A)$ without high time or processing costs in order to utilize this improvement [11].*

A.2. Class Balancing (+B)

Assumption 5. *The "class label," $Y = y$, is also in the distribution as the bias rule which contributes to the importance of all the features "$\phi \to Y = y$". By **removing the bias**, this rule instead contributes to $imp(A)$ based on its correlation to the other rules.*

A.3. Looking Forward: Clustering (+C)

Assumption 6. *Local $imp(A)$ is from **mixtures of distributions**, $imp(A) = E[imp(A)] + \pi_1(L+L_\epsilon)\epsilon + ... + \pi_k(L+L_\epsilon)\epsilon \leq E[imp(A)] + kd(d+1)\epsilon/2$. For some probabilities $\pi_i \in [0, 1]$ such that $\sum \pi_i = 1$, simplified for a noisy choices of clusters and probabilities, $kd(d+1)\epsilon/2$.*

Appendix B. Full Experiment Details

Experiments are run for 30-50 randomly selected points-of-interest from the testing data (depending on the total number of samples in the dataset) in each of seven diverse datasets, with varying class balance and combinations of categorical and continuous features. The full metadata for these data are shown in Table 2. We use two different black-boxes, both cross validated on a 60-20-20 train-test-validation split. The black-boxes are a neural network and random forest.

Dataset	# of Samples	# of Features	# of Categorical Features	Balance
Wine	178	13	0	0.33
Loan	614	11	6	0.31
German Credit	1000	20	13	0.30
Employee Attrition	1470	30	7	0.14
Breast Cancer	569	30	0	0.47
COMPAS	7214	28	10	0.45
Spam	4601	57	0	0.39

Table 2. Metadata for each dataset used in the experiments. **Balance** is the minimum proportion of one class to the overall data $min_y(\sigma(Y = y)/|D|)$, so 0.5 would be perfectly balanced for most datasets with two classes.

We compare to popular methods LIME, S-LIME, SHAP, and LORE using a one-sided paired sample t-test, with the null hypothesis $H_0 : g(e_{comp}) - g(e_{our}) \geq 0$ and the alternative

hypothesis $H_\alpha : g(e_{comp}) - g(e_{our}) < 0$. Where e is a set of explanations and g is a metric used to evaluate the explanations. The paired test ensures that outlying samples across both compared methods are considered equally in testing.

B.1. **Evaluation Metrics**

To measure stability we use the Jaccard index, like other work [6, 7]. We consider the top $d = 5$ features, where users are most likely to concern themselves. Sometimes we expect that explanations may require more features, so to generalize better to many datasets we consider the top $d = N/2$ features where N is the total number of features. We set hard cut offs since the order of a subset of features should still be important. In our application, we use Jaccard index to understand when feature importance becomes most random.

We also use the inverse Jaccard index, $Inverse\ Jaccard = (1 - Jaccard)$, for sets of 10 different samples — each with feature importance averaged over the 10 repetitions. This alternative metric, expresses how similar each explanation is to each other, and is used to evaluate how much stability is caused by the explainer versus how much stability is a result of using the same explanation for every sample.

Another metric which we use for stability is the RIS [8] with 10 small variants of x_p. In comparison to the Jaccard Index, this value considers the change in the feature importance value as opposed to the relative scale of the feature importance values. So it may be better when many features are used in the explanation. As with previous work [8], we use the RIS on a log scale.

We also use two prior evaluation metrics to ensure that performance is not reduced by our changes: fidelity (weighted agreement between the white-box and the black-box [12]) and hit-rate (how often the white-box could correctly predict original unperturbed input row). We consider these as consequent measures, if fidelity or hit-rate are low then the stability may be a result of a poor classifier.

Appendix C. **Individual Results**

We did initially test on clustering, but among the clustering approaches, KMeans was relatively stable as we expected, and the other approaches ended up being costly to the fidelity both MixT (0.725) and prop-MixT (0.720) had lower average fidelity than K-Means (0.756), leading us to use K-Means for our final experiments as +C. Without considerable fidelity improvements from these methods, the stability cost was too much, as a result we left these experiments out of the main results.

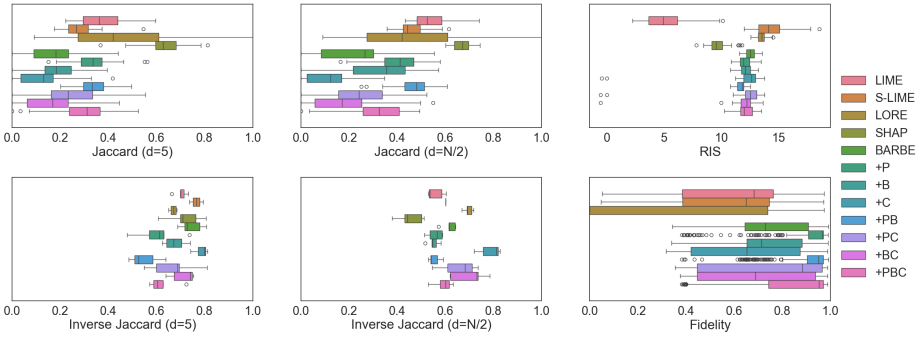

Figure 2. Box plots of each evaluation metric for competitors and our approaches (+P, +B, ...) for the **COMPAS dataset using the neural network model**.