The Disagreement Problem in Faithfulness Metrics

Brian Barr[1,*], Noah Fatsi[1], Leif Hancox-Li[1], Peter Richter[2], Daniel Proano[2], and Caleb Mok[2]

[1]AI Foundations, Capital One, Mclean VA
[2]Applied ML, Capital One, Boston MA
{*brian.barr,noah.fatsi,leif.hancox-li,peter.richter,daniel.proano,caleb.mok*}*@capitalone.com*
[*]corresponding author

Abstract

The field of explainable artificial intelligence (XAI) aims to explain how black-box machine learning models work. Much of the work centers around the holy grail of providing post-hoc feature attributions to any model architecture. While the pace of innovation around novel methods has slowed down, the question remains of how to choose a method, and how to make it fit for purpose. Recently, efforts around benchmarking XAI methods have suggested metrics for that purpose—but there are many choices. That bounty of choice still leaves an end user unclear on how to proceed. This paper focuses on comparing metrics with the aim of measuring faithfulness of local explanations on tabular classification problems—and shows that the current metrics don't agree; leaving users unsure how to choose the most faithful explanations.

1 Introduction

XAI is a field that aims to create techniques that explain black-box machine learning models. While there is a growing body of work on mechanistic interpretability [13], which aims to describe the actual mechanisms of model predictions by looking at model components, much of the XAI literature has focused on post-hoc explanations, which aim to create explanations without depending on the specifics of internal model mechanisms. Within the post-hoc explainability literature, feature attribution methods [18, 21] have been particularly prominent: methods where the explanation for a particular model prediction is a vector of numbers representing the importance of each feature in the sample. Other XAI approaches, like counterfactuals and example-based approaches, provide fundamentally different types of outputs [20, 15].

Feature attribution methods face the challenge of proving that their outputs are faithful to the model's behavior. Various faithfulness metrics have been proposed, some of them as part of XAI benchmarks [12, 1]. However, it is unclear how well these different metrics correlate with one another, or what use cases each metric is suitable for. In this paper we ask the meta-question of deciding which measures of faithfulness work well: we benchmark faithfulness metrics.

Feature attribution methods also face the challenge of different methods disagreeing with one another [17]. Inspired by this finding, we extend it to look at disagreements between evaluation metrics for XAI. We appropriate two recently introduced methods of evaluating explanations—ablation [10] and topological data analysis [27]—to tackle the problem of

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

evaluating XAI metrics. We then compare them to other XAI metrics on a variety of different explanation methods, baselines for those explanation methods, and tabular datasets.

Our paper points to a gap between theory and practice: We have many faithfulness metrics, but they do not correlate well with one another. Practitioners evaluating explanations on faithfulness lack guidance on which faithfulness metric they should use. It may be that similarly to how one selects different accuracy metrics based on the application context, faithfulness metric selection is also highly contextual. If so, more work needs to be done to figure out which faithfulness metrics are suitable for which contexts.

2 Previous work

2.1 Post-hoc explainability challenges

Post-hoc explainability methods face the challenge of determining whether their outputs are good. This challenge is complicated because, to begin with, there are different ways in which we can define "goodness". Stability, faithfulness to the original model, and fairness are just some of the desiderata identified for XAI outputs so far [1]. In addition, user studies have found that XAI methods may not necessarily be useful to humans in specific application contexts—another dimension of explanation quality that is distinct from their mathematical properties [4, 26]. Finally, many feature attribution methods are sensitive to one's choice of baseline [11, 19, 23]

In this paper, we focus on measuring faithfulness [3]. The basic concept behind faithfulness is that feature attributions output by the XAI method should reproduce the actual influences of the features in the model. Faithfulness can be measured through various quantitative metrics, such as Prediction Gap on Important feature perturbation (PGI) and Prediction Gap on Unimportant feature perturbation (PGU) (first defined in [8], but given these names in [1]).

2.2 Post-hoc explainability benchmarking

In an attempt to facilitate easier comparison of explanation methods, recent papers have introduced XAI benchmarks for feature attribution-based post-hoc explanations [8, 5, 1]. We add to this literature by introducing two additional metrics: ablation and topological data analysis. We also compare them to other faithfulness metrics.

3 Methodology

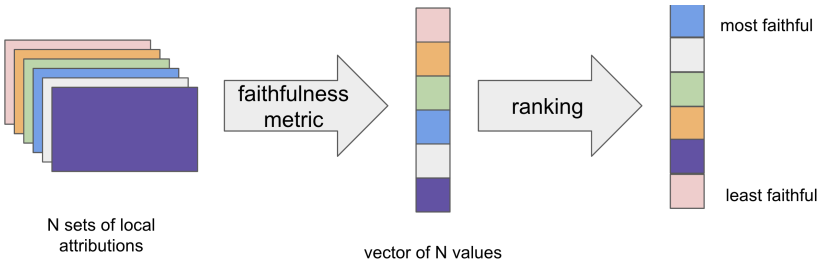


Figure 1: Cartoon of the data flow / process. If the faithfulness metric was valid, this would allow a practitioner to choose the best explanations for their model.

The framework for the study is to use models built on commonly used empirical data with a variety of local explanations generated for each model-data pair. Using those explanations along with a scalar metric intended to measure faithfulness, we generate a rank ordered list that in theory should give guidance on choosing the most faithful set of explanations. We compare these rankings in 4.

3.1 Data and models

We use one synthetic and four open source datasets [9] whose characteristics are shown in Table 1. For each dataset, we built a fixed size dense neural network with three layers and a single layer neural network as our linear model. See Table 1 in the appendix for details on the open-source datasets.

3.2 Explanations

For both the linear and non-linear models, we used three methods from the Captum package [16]: Deep SHAP, KernelSHAP, and integrated gradients [20, 24]. However, specifying methods alone is not sufficient [11]. Each of these methods produces different outputs depending on the choice of baseline: for example, the popular open source `shap` package uses the average of all predictions as its default baseline [20]. We also include a set of random explanations, uniformly sampled from [0,1] to provide a lower limit of faithfulness. For linear models, we include ground truth explanations of the form $x_i c_i$ where x_i is the ith input feature and c_i is the learned coefficient.

For each explanation method, we supplied four baselines: *constant median*, a variant of the constant baseline that uses the median of each feature; *training*, also known as the expectation baseline, a sub-sample of the training data [18]; *opposite class* [2], which selects k samples that belong to the opposite class from the the chosen sample; and *nearest neighbor* [2], which selects k samples that are closest to the chosen sample. For this study the value of k is fixed at five.

Evaluation of each method-baseline pair is repeated three times to account for the possibility of stochastic behavior. For the dense networks, we have a total of 39 tables of local explanations (3 methods with 4 baselines each and the random explanations, all with 3 repeats). For the linear models, we have a set of 42 tables of explanations (13 in common with the nonlinear models with the addition of the ground truth explanations, all with 3 repeats).

3.3 Metrics of faithfulness

We sample existing metrics of faithfulness from the literature and open source repositories, and add two novel methods: ablation [10] and topological data analysis [27].

PGI is calculated by measuring the change in a model's prediction when the top k most important features, as determined by an explanation method, are perturbed. The intuition behind the metric is that a model's output should change most dramatically when the most important features are perturbed. A higher PGI indicates a more faithful explanation [8]. PGI is defined as follows:

$$\hat{\mathcal{M}}_{\text{PGI}}(\mathbf{x}, f) = \frac{1}{m} \sum_{j=1}^{m} [|f(\mathbf{x}) - f(\tilde{\mathbf{x}}_j)|] \tag{1}$$

Where $\mathbf{x}$ is the original sample, $\tilde{\mathbf{x}}$ is the same sample with the top k features perturbed, and f is a model which outputs a value from 0 to 1. The average PGI is computed over m runs of a stochastic perturbation process. We use a Gaussian perturbation drawing samples from $\mathcal{N}(0, 0.1)$ for continuous variables, and a "flip percentage" of 0.3 for discrete variables unless otherwise mentioned.

Ablation is another perturbation-based method. This procedure is frequently used to assess the importance of input variables on a model's performance. By perturbing the input variables in rank order of importance as determined from local attributions, one can assess the quality of their rank ordering. Essentially, the goal is to assess the sensitivity of the model's performance as gauged by the local explanations. Perturbing important variables should correlate with larger decreases in measures of model capability than perturbing less important features. Details on ablation studies can be found in [10]. For ablation, we use area under the ablation curve (ABC) as the scalar faithfulness metric.

Bottleneck distance (BND) is a similarity metric that compares two manifolds, with origins in topological data analysis (TDA) and persistent homology. It does not rely on perturbations—a characteristic which makes its possible use appealing. Instead, this methodology treats the point cloud of explanations as a manifold. Through the use of the mapper algorithm [22] with a specified filter function (here we use the model predictions), the high-dimensional manifold is projected down to a two dimensional network representation called the Reeb graph. The similarity of two mapper networks can be compared by the bottleneck distance. Even though the technical underpinning and units of the local explanations may differ, the smaller the distance between their respective manifolds, the more they are similar [27]. For each explanation method, we use its average bottleneck distance to all other methods as the metric of explanation quality. Details of the TDA process can be found in A.2.

3.4 Ranking explanations

For each dataset, with its corresponding set of explanations and controls, we calculate each of the metrics and rank order their values. From those rankings, we use rank correlations to measure their agreement. In an ideal world, where the metrics can consistently assess faithfulness, the resulting correlation would be one. Detailed plots of the rankings can be seen in Section A.5. Summary heatmaps of the correlations are shown in Figure 3.

4 Experiments

4.1 Ground truth ranking on synthetic dataset

The first experiment is to create a set of explanations with a controlled change in faithfulness. Using the synthetic dataset, its logistic regression model, and ground truth explanations from the logits, we permute a portion of the rows of the explanation - with fractions varying from 0 to 1. The explanations with more permuted rows are fundamentally less faithful since the explanations are increasingly misaligned with the inputs and their predictions. The expected outcome is to have metrics that reflect this change in a strictly monotonic manner. The results of the experiment are shown in Figure 2.

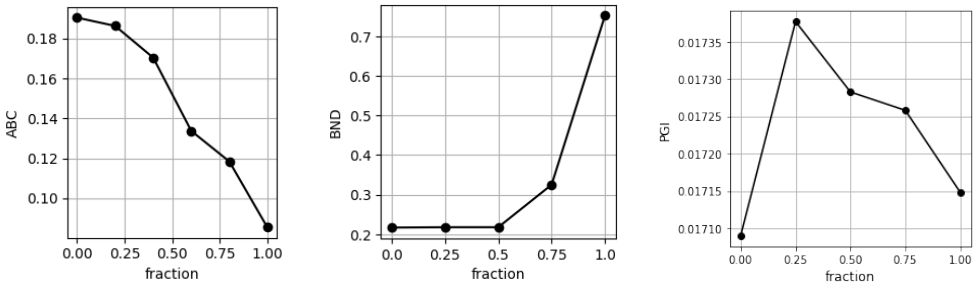
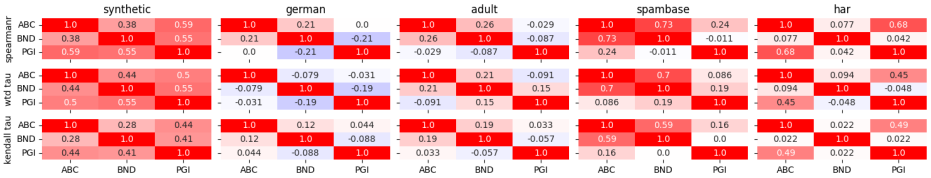


Figure 2: Comparison of metrics (ABC, BND, and PGI respectively) on a set of explanations with a designed ranking.

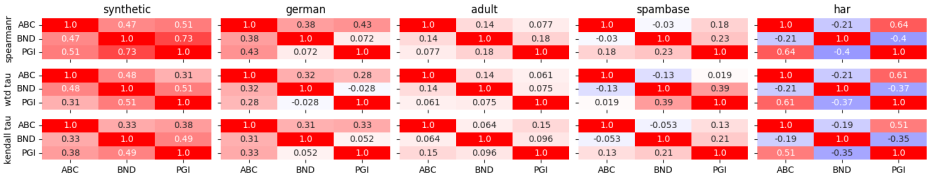
Of the three metrics, only the ABC metric matches our expectation. The BND metric is the next closest to expectations—but it is not *strictly* monotonic for fractions up to 0.5; while the PGI metric is not monotonic. For details for the behavior of TDA for this experiment, detailing its lack of differentiation from 0 to 0.5, see Section A.3. See Section 4.3 for an analysis of the main effects of parameter choices for PGI.

4.2 Experiments on non-synthetic datasets

Figure 3 shows the correlation of faithfulness metrics for each of the datasets in this study for logistic regression models (top) and dense networks (bottom) using Spearman's ρ (top row), weighted Kendall-τ (middle row), and Kendall's τ (bottom row). Using a logistic regression



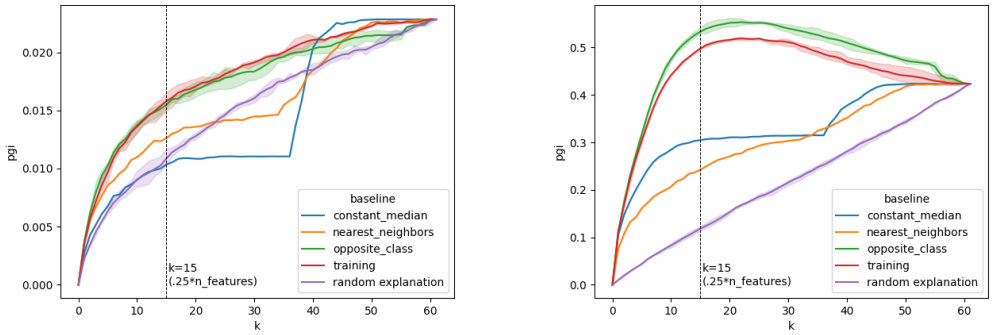
(a) Logistic regression models.



(b) Three layered dense network models.

Figure 3: Rank correlation heatmaps of the faithfulness metrics for each dataset in the study. The numbers are averaged across all combinations of baselines and explanation methods.

model, with the addition of the "ground truth" explanations, provides the opportunity to rank the XAI-generated explanations alongside ground truth and random explanations, with the expectation that the top ranked explanations will be the ground truth, and the worst ranked will be the random explanations. For the dense networks the only known quantity is that the random explanations should be ranked the worst. The correlations are insensitive to the choice of correlation metric.



(a) Gaussian perturbation

(b) Marginal perturbation

Figure 4: Sensitivity study on choice of k and perturbation for PGI on spambase with a dense network, with integrated gradients as the explanation method.

4.3 Considerations for perturbation based metrics

PGI and Ablation are both perturbation-based methods and share similar drawbacks [14]. Nonetheless, we see discrepancies between them in the correlation heatmaps in Figure 3. While perturbation methods for explainability have been widely studied [7], that is not the case for faithfulness metrics. Perturbation-based metrics of explanation faithfulness can vary widely depending on their numerous configuration parameters. Here we mention three:

1. Choice of perturbation method: This is known to have a large, and biasing impact, on these types of "permute and predict" metrics [14]. In Figure 4, we contrast PGI calculated with a Gaussian perturbation (left) and with a marginal perturbation (right). For a fixed value of k (shown with a vertical dashed line), higher values of PGI indicate higher faithfulness. Comparing the two graphs, the rankings of the baselines are dissimilar. For this instance, the Gaussian perturbation appears to perform more poorly, as it ranks the random explanation ahead of the constant median baseline.
2. Choice of top k: how to select k is not theoretically motivated. We see in Figure 4 that choice of k can impact the faithfulness rankings of a set of explanations based on

PGI. Wherever the lines cross, PGI rankings change. In the Gaussian perturbation plot, three of the five baselines achieve the highest PGI rank at different values of k. At $k <= 11$, the opposite class baseline has the highest PGI. Then, training baseline is the top ranked for $12 <= k <= 41$, until finally constant median takes over.

3. Treatment of discrete features: this encompasses two issues: i) it is difficult to fairly balance the strength of Gaussian noise on continuous variables and the flip percentage on discrete variables; and ii) the choice of treating discrete features as one-hot-encoded, instead of performing an aggregation and reverting the input features back to a nominal label-encoded state. Figure 5 shows the difference in calculated ablation curves on one-hot-encoded (left) and aggregated (right) categorical features. If the rank ordering of features is consistent, it is expected that the ablation curve would be monotonically decreasing—which is clearly not the case for the no-aggregations case. For the aggregated plot, the ablation curves are closer to the ideal, with significant increases in AUROC occurring only after the random feature sanity check.

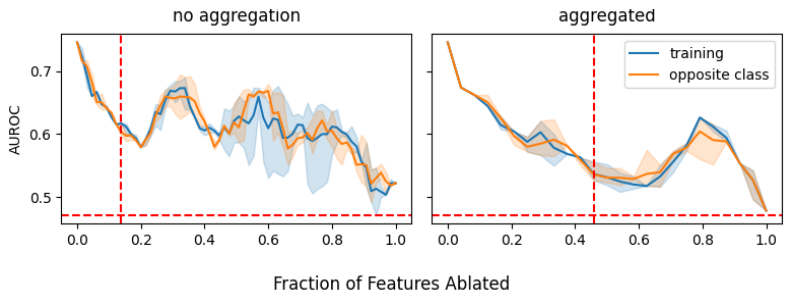


Figure 5: Sensitivity of ablation curves for the training and opposite class baselines. The vertical line depicts the best average rank of a random feature.

4.4 Observations on the use of TDA

At the heart of TDA is the use of the bottleneck distance (BND), and choosing the set of explanations that have the lowest distance to the other candidates. However, the distance is a similarity metric—it does not have a notion of "good" or "faithful" baked into it. This can be limiting, as one cannot directly compare two items—the sum of the distances for each candidate would be identical. In a larger collection of methods, if the candidates consisted of a majority of explanations of low quality and a minority of high quality, the selected candidate will most likely be chosen from the low quality pool. This should influence the construction of candidates for assessment. The current process of selecting one resolution as a result of its performance on all candidates for a dataset seems to have diminished its sensitivity, which can be seen in its robustness to permuting rows in Figure 2, and for the selection of candidates for the synthetic dataset in Figure 8

5 Conclusions

We have focused on using XAI methods on a range of public empirical datasets, using the tabular data to create classification models—a typical use case in academia and industry. The goal was to find a set of explanations that was deemed to be most faithful.

Across the experiments, the ranked correlations reveal little consensus on the notion of faithfulness in the explanations. This in turn would leave an end user without the required tools to know if they had successfully chosen the right set of explanation method and baseline.

This gap in utility should be a wake-up call to the XAI community. Future work can compare the plethora of additional measures of faithfulness [12, 1] to see if they also disagree.

References

[1] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15784–15799. Curran Associates, Inc., 2022.

[2] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. Counterfactual shapley additive explanations. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1054–1070, 2022.

[3] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[5] Mohamed Karim Belaid, Eyke Hüllermeier, Maximilian Rabus, and Ralf Krestel. Do we need another explainable AI method? Toward unifying post-hoc XAI evaluation methods into an interactive and multi-dimensional benchmark. *arXiv preprint arXiv:2207.14160*, 2022.

[6] Mathieu Carriere, Bertrand Michel, and Steve Oudot. Statistical analysis and parameter selection for mapper. *The Journal of Machine Learning Research*, 19(1):478–516, 2018.

[7] Ian C. Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *J. Mach. Learn. Res.*, 22(1), jan 2021.

[8] Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H. Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 203–214, New York, NY, USA, 2022. Association for Computing Machinery.

[9] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[10] Isha Hameed, Samuel Sharpe, Daniel Barcklow, Justin Au-Yeung, Sahil Verma, Jocelyn Huang, Brian Barr, and C Bayan Bruss. BASED-XAI: Breaking ablation studies down for explainable artificial intelligence. *arXiv preprint arXiv:2207.05566*, 2022.

[11] Johannes Haug, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. On baselines for local feature attributions. *arXiv preprint arXiv:2101.00905*, 2021.

[12] Anna Hedström, Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M. C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[13] Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

[14] Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6), nov 2021.

[15] Been Kim, Cynthia Rudin, and Julie A Shah. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems*, 27, 2014.

[16] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

[17] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

[18] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[19] Antonios Mamalakis, Elizabeth A. Barnes, and Imme Ebert-Uphoff. Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. *Artificial Intelligence for the Earth Systems*, 2(1):e220058, 2023.

[20] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 607–617, New York, NY, USA, 2020. Association for Computing Machinery.

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[22] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2:091–100, 2007.

[23] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. https://distill.pub/2020/attribution-baselines.

[24] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[25] The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015.

[26] Xinru Wang and Ming Yin. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA, 2021. Association for Computing Machinery.

[27] Peter Xenopoulos, Gromit Chan, Harish Doraiswamy, Luis Gustavo Nonato, Brian Barr, and Claudio Silva. Gale: Globally assessing local explanations. In Alexander Cloninger, Timothy Doster, Tegan Emerson, Manohar Kaul, Ira Ktena, Henry Kvinge, Nina Miolane, Bastian Rieck, Sarah Tymochko, and Guy Wolf, editors, *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196 of *Proceedings of Machine Learning Research*, pages 322–331. PMLR, 25 Feb–22 Jul 2022.

A Supplementary Material

A.1 Data set information

Table 1 shows details of the synthetic and empirical data sets used in the experiments.

Dataset	Samples	Num	Cat	OHE
synthetic	1,000	24	0	0
adult	48,842	6	8	100
German credit	1,000	7	13	54
har	10,299	561	0	0
spambase	4,601	57	0	0

Table 1: Summary of datasets, detailing number of samples, numerical (Num), categorical (Cat), and one hot encoded (OHE) features.

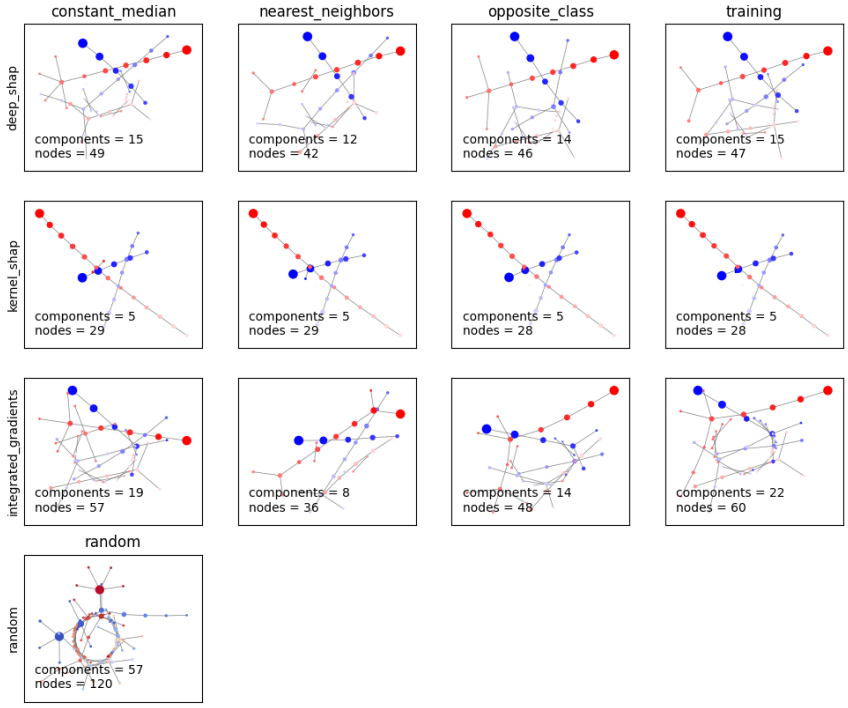


Figure 6: The network representation of a mapper object for each method and baseline for the synthetic dataset.

A.2 TDA process - hyperparameters, networks, and persistence diagrams

In practice, a topological manifold is represented by a point cloud. A mapper object is constructed from five things:

- A manifold as represented by a point cloud - the explanations.
- A "lens function" that provides scalar valued labels to every point in the cloud. Here we use the model predictions.
- A value for resolution. The resolution sets the number of slices that divide contiguous ranges of the lens function. To choose the value of resolution, we conduct a grid search over a range of values from six to thirty by twos. Each value of resolution is calculated over 30 bootstrap samples and we collect the 95th percentile confidence value from the bootstrap bottle neck distances as our figure of merit, called the *stability*. For a dataset, we collect a table of stability values from all methods, baselines and repeats, for each candidate resolution. The resolution with the lowest total sum of stability is selected.
- A clustering algorithm that operates on the points in a slice. This clustering creates the nodes of the network. The majority of the theoretical work in TDA has used agglomerative clustering - which we also adopt. The GUDHI [25] library offers utility functions to estimate its distance parameter, which eliminates a potential search for that hyperparameter.

- A value for the gain. The gain specifies the overlap of neighboring slices. Theoretical analysis shows a valid range for gain is from 0.3 to 0.5 [6]. We choose a gain value of 0.4 for all results shown in this study. Points that exist in the overlap region create edges between nodes.

The mapper networks for one repeat are shown in Figure 6. For each network, a persistence diagram is created, shown in Figure 7. Each point on the diagram represents a single simple connected component, or fork, or hole that exists over a range of values for the filter function (in this case the model predictions). The x-axis specifies when a topological feature is "born" (when it first appears as measured by the filter function) while the y-axis specifies when a topological feature "dies" (the final value of the filter function for the last node).

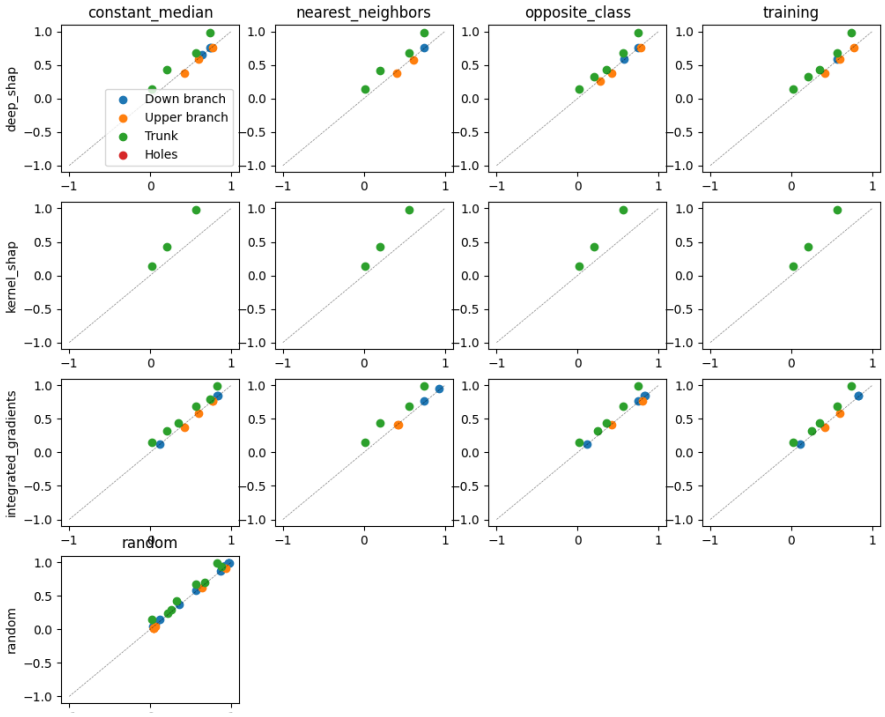


Figure 7: Persistence diagrams corresponding to the networks shown in Figure 6.

The *bottleneck distance* (BND) is calculated by matching the entities on two persistence diagrams and measuring the euclidean distance between the points on the persistence diagram. A heatmap showing these distances is shown in the left of Figure 8. The candidate is chosen to be the one with the lowest row-wise sum of the bottleneck distances, as seen in the right of Figure 8.

A.3 Ground truth ranking details for BND

Here we show how low values of permutation (0, 0.25 and 0.5) "look the same" to the BND metric - with little change in the mappers, resulting in no change to the bottleneck distance. With permutations of 0.25 and 0.5, the mappers show a disconnected cluster, which does not persist, and does not amount to a significant differentiation. It's not until the permutation reaches 0.75 that a new persistent feature (the upper branch) is detected, and more features continue to be added at 1.0 leading to a significant difference being found by BND.

A.4 Top three most faithful explanations

While it is possible the rank correlation could over weight mismatches in lower ranked components, Table 2 shows the three most faithful explanations as chosen by each metric. It is clear to see that there is no consistent agreement among the metrics. However if taken in

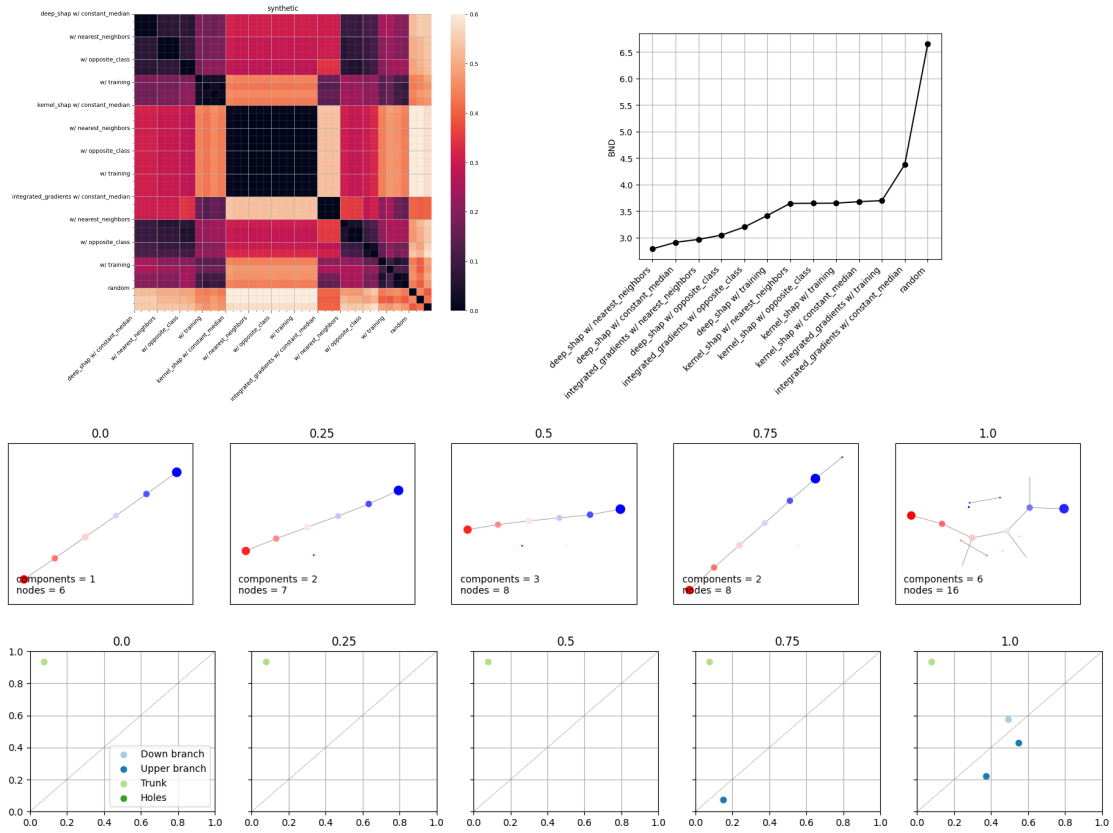
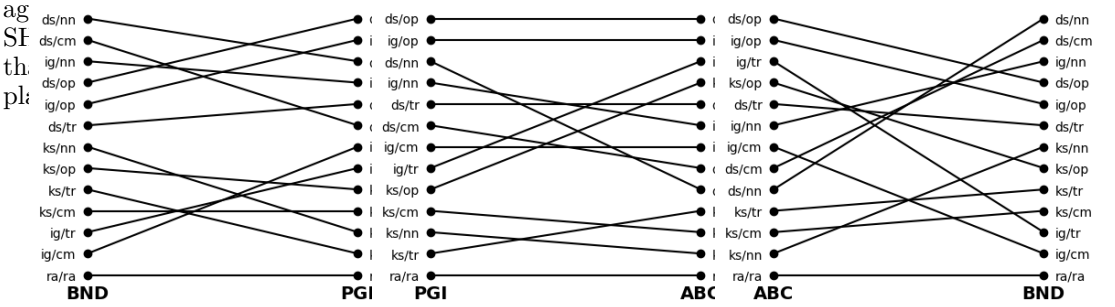


Figure 9: For the ground truth ranking experiment, plots of all mapper networks (top) and corresponding persistence diagrams (bottom).

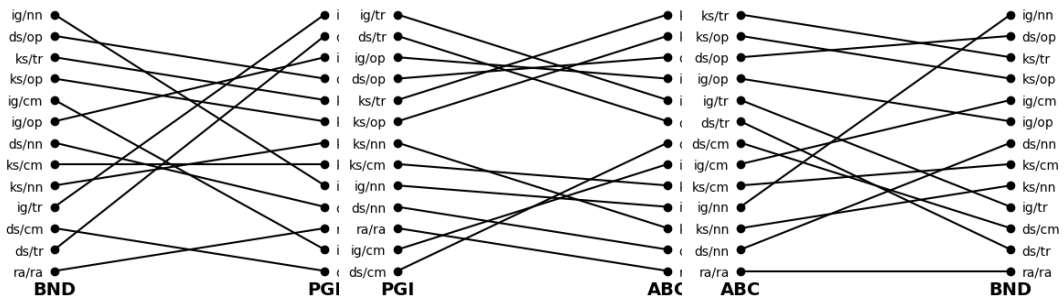


(a) Ranking of BND and PGI (b) Ranking of PGI and ABC (c) Ranking of ABC and BND

Figure 10: Slope charts for synthetic dataset

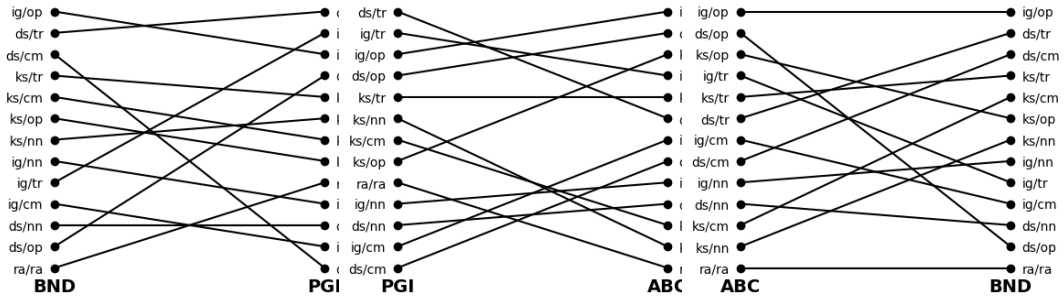
A.5 Slope charts for dense networks

The fundamental information being compared are the rankings of the explanations based on the faithfulness metrics. To provide a more tangible sense of the mismatches, we have utilized slope charts (Figures 10, 11, 12, 13) as a means of interrogating the results. Best ranked explanations are at the top, worst ranked at the bottom. When metrics place the explanations at a different rank, it can be identified as a sloped line, with a steeper slope signifying a larger disagreement in rank.



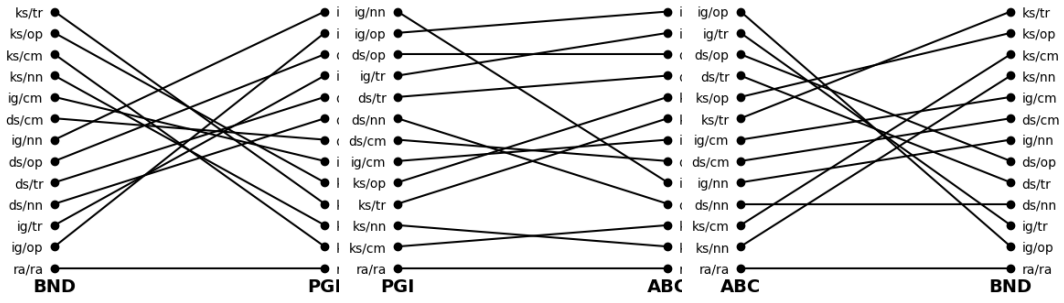
(a) Ranking of BND and PGI (b) Ranking of PGI and ABC (c) Ranking of ABC and BND

Figure 11: Slope charts for German Credit dataset



(a) Ranking of BND and PGI (b) Ranking of PGI and ABC (c) Ranking of ABC and BND

Figure 12: Slope charts for spambase dataset



(a) Ranking of BND and PGI (b) Ranking of PGI and ABC (c) Ranking of ABC and BND

Figure 13: Slope charts for human activity recognition dataset

dataset	metric	method	baseline	ranking
synthetic	ABC	deep-shap	opposite-class	0
	ABC	integrated-gradients	opposite-class	1
	ABC	integrated-gradients	training	2
	BND	deep-shap	nearest-neighbors	0
	BND	deep-shap	constant-median	1
	BND	integrated-gradients	nearest-neighbors	2
	PGI	deep-shap	opposite-class	0
	PGI	integrated-gradients	opposite-class	1
	PGI	integrated-gradients	training	2
german	ABC	kernel-shap	training	0
	ABC	kernel-shap	opposite-class	1
	ABC	deep-shap	opposite-class	2
	BND	deep-shap	opposite-class	0
	BND	integrated-gradients	nearest-neighbors	1
	BND	kernel-shap	training	2
	PGI	kernel-shap	training	0
	PGI	kernel-shap	opposite-class	1
	PGI	deep-shap	opposite-class	2
adult	ABC	deep-shap	opposite-class	0
	ABC	integrated-gradients	opposite-class	1
	ABC	integrated-gradients	training	2
	BND	deep-shap	constant-median	0
	BND	deep-shap	nearest-neighbors	1
	BND	deep-shap	opposite-class	2
	PGI	deep-shap	opposite-class	0
	PGI	integrated-gradients	opposite-class	1
	PGI	integrated-gradients	training	2
spambase	ABC	integrated-gradients	opposite-class	0
	ABC	deep-shap	opposite-class	1
	ABC	kernel-shap	opposite-class	2
	BND	deep-shap	constant-median	0
	BND	deep-shap	training	1
	BND	integrated-gradients	opposite-class	2
	PGI	integrated-gradients	opposite-class	0
	PGI	deep-shap	opposite-class	1
	PGI	kernel-shap	opposite-class	2
har	ABC	integrated-gradients	opposite-class	0
	ABC	integrated-gradients	training	1
	ABC	deep-shap	opposite-class	2
	BND	kernel-shap	opposite-class	0
	BND	kernel-shap	training	1
	BND	kernel-shap	constant-median	2
	PGI	integrated-gradients	opposite-class	0
	PGI	integrated-gradients	training	1
	PGI	deep-shap	opposite-class	2

Table 2: Top three choices for each dataset broken down by metric.