From Flexibility to Manipulation
The Slippery Slope of XAI Evaluation

A cautionary tale about the possibility of manipulating XAI evaluation due to the lack of ground truth explanations.

Kristoffer Wickstrøm[1,2], Marina M.-C. Höhne[3,4,7], Anna Hedström[3,5,6]

UiT The Arctic University of Norway[1]
Visual Intelligence[2]
Understandable Machine Intelligence Lab[3]
University of Potsdam[4], Technical University of Berlin[5]
Fraunhofer HHI[6], BIFOLD[7]

Quantitative Evaluation of XAI

- XAI is crucial to ensure trustworthiness.
- Many competing XAI methods are available.
- Quantitative analysis is key for comparison.

The Lack of Ground Truth Explanations

- No ground truth for evaluation [1].
- Therefore, measure desirable properties [2].
- Difficult to select hyperparameters.
- Can have big impact on evaluation (Table 1).

Example: Faithfulness Evaluation

- Alignment between explanation and predictor.
- Perturb and predict according to explanation.
- Obtain faithfulness curves (Figure 1-3).
- Many hyperparameters to select.

Manipulating Strategy

- How to pick hyperparameters?
- Lots of flexibility for user.
- Flexibility can be exploited by user.
- Manipulation as optimization (Definition 1-2)
- Define feasible set of hyperparameters.

Robust Evaluation with MRR

- New evaluation to address manipulation.
- Mean ressilience rank (MRR).
- Rank methods across feasible set.

Results - Manipulation (Table 2-3)

- Big changes in evaluation scores.
- Ranking can be altered.
- Difference can be amplified.

Results - MRR (Table 4)

- Ranking removes the possibility for manipulation.
- Can be extended across datasets.
- However, much variation in rankings.

Limitations and Future Work

- MRR can be computationally demanding.
- The feasible set requires domain knowledge.
- Investigate other metrics in future works.

Conclusion

- Quantitative evaluation of XAI is challenging.
- Hard to do right, easy to go wrong.
- Manipulation is possible.
- Towards tackling manipulation with MRR.

Code

Motivating Example

XAI method	Faithfulness score (↓)
LRP	25.19
Saliency	**20.23**
Kernel SHAP	23.94

XAI method	Faithfulness score (↓)
LRP	**19.31**
Saliency	22.96
Kernel SHAP	24.87

Table 1: Faithfulness comparison of XAI methods on MNIST before (left table) and after manipulation (right). Here, the different between the left and right table is the perturbation methods used (uniform noise vs. blurring, respectively). Both perturbation methods are commonly used, but completely change the outcome of the evaluation.

Faithfulness Examples

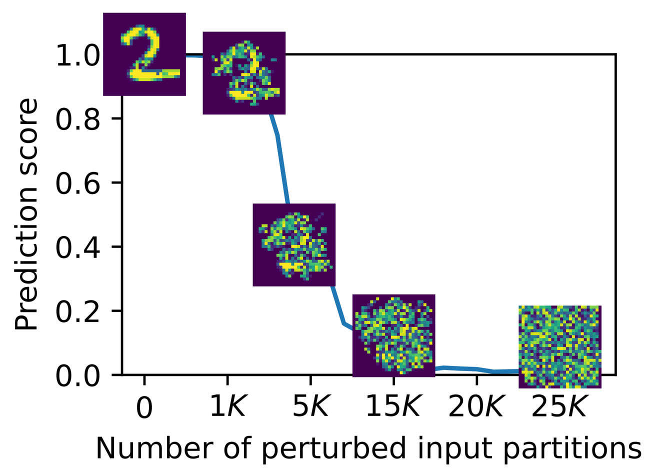

Figure 1

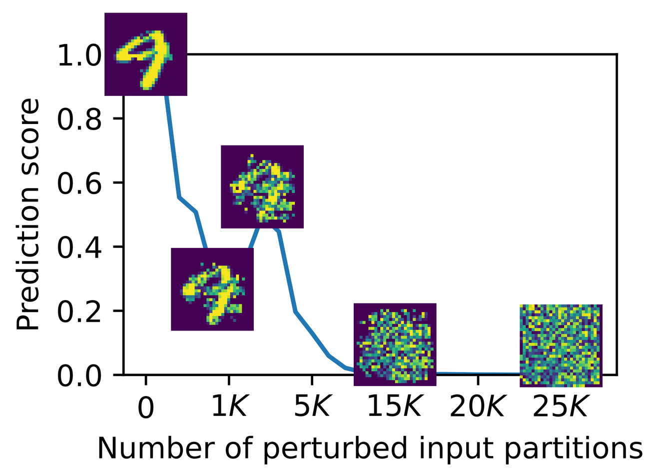

Figure 2

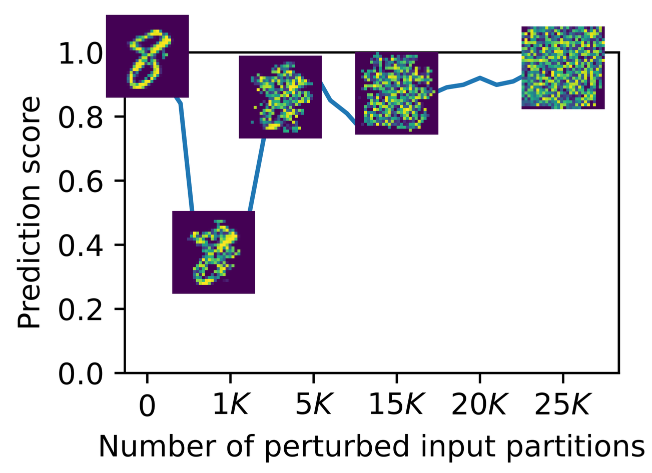

Figure 3

Manipulation Objectives

Definition 1 (Intra-Manipulation): Given an evaluation function F, an input sample $\mathbf{x}$, an explanation $\mathbf{e}$, hyperparameters a, b, and c, and a feasible set of hyperparameters A_a^* for the hyperparameter a, the intra-manipulation method solves the following optimization problem to determine the hyperparameter a, which maximizes the evaluation score of F:

$$\underset{a}{\text{maximize}} \quad F(f, \mathbf{x}, \mathbf{e}, a, b, c)$$
$$\text{subject to } a \in A_a^*.$$

Definition 2 (Inter-Manipulation): Given an evaluation function F, an input sample $\mathbf{x}$, a set of explanations $\{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ from M different XAI methods, hyperparameters a, b, and c, and a feasible set of hyperparameters A_a^* for the hyperparameter a, the inter-manipulation method solves the following optimization problem to determine the hyperparameter a, which maximizes the following objective:

$$\underset{a}{\text{maximize}} \quad F(f, \mathbf{x}, \mathbf{e}_m, a, b, c) - \sum_{m' \neq m} F(f, \mathbf{x}, \mathbf{e}_{m'}, a, b, c) \quad .$$
$$\text{subject to } a \in A_a^*$$

Intra-Manipulation Results

	MNIST		FashionMNIST		PneumMNIST		ImageNet	
XAI method	base	manip.	base	manip.	base	manip.	base	manip.
LRP	25.20	7.86	21.46	5.37	21.31	6.06	129.61	41.48
Saliency	20.23	6.80	15.65	4.72	23.28	4.23	124.93	37.53
KernelSHAP	23.94	8.01	18.28	4.81	22.06	4.29	128.72	40.14

Table 2: Intra-results across several datasets and methods. Lower is better.

Inter-Manipulation Results

	MNIST		FashionMNIST		PneumMNIST		ImageNet	
XAI method	base	manip.	base	manip.	base	manip.	base	manip.
LRP	25.19	**37.79**	21.46	35.42	**21.31**	**43.53**	129.61	128.02
Saliency	**20.23**	46.23	**15.65**	**34.75**	23.28	47.42	**124.93**	**123.93**
KernelSHAP	23.94	50.77	21.45	41.42	22.06	45.30	128.72	131.97

Table 3: Inter-results with manipulation towards *LRP*. Lower is better.

MRR Results

XAI method	MNIST	FashionMNIST	PneumMNIST	ImageNet	All
LRP	$\mathbf{0.22 \pm 0.15}$	0.33 ± 0.00	0.21 ± 0.00	$\mathbf{0.26 \pm 0.00}$	$\mathbf{0.29 \pm 0.14}$
Saliency	0.41 ± 0.26	0.44 ± 0.31	0.37 ± 0.31	0.41 ± 0.33	0.41 ± 0.30
KernelSHAP	0.37 ± 0.33	$\mathbf{0.22} \pm 0.31$	$\mathbf{0.33} \pm 0.27$	0.33 ± 0.06	0.31 ± 0.31

Table 4: MRR across feasible set for each dataset and across datasets. Lower is better, a rank of 0 is best and 1 is worst.

References 1
Hedström et al., *The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus.* TMLR, 2023.

References 2
Hedström et al., *Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network.* JMLR, 2023.