Tutorial 5 - Explainable Artificial Intelligence in Biomedical Imaging
Part IV - Robustness of XAI in Biomedical Imaging

Dr Kalliopi V. A. Dalakleidi, Post-doctoral researcher, National Technical University of Athens

ISBI 2024
ATHENS, GREECE

Why do we need robust XAI?

- XAI methods are a remedy for debugging and trusting AI models, as well as interpreting their predictions.
- Limitations and vulnerabilities of state-of-the-art explanation methods put their security and trustworthiness into question.
- Can be vulnerable to data and model perturbation.

How can we assess robustness of XAI?

Subjective metrics for assessing XAI robustness

- The predominant evaluations of explanations have been human-centric subjective measures.
- Qualitative displays of explanation examples.
- Crowd-sourced evaluations of human satisfaction with the explanations.
- Whether humans are able to understand the model.

[Yeh et al., 2019]

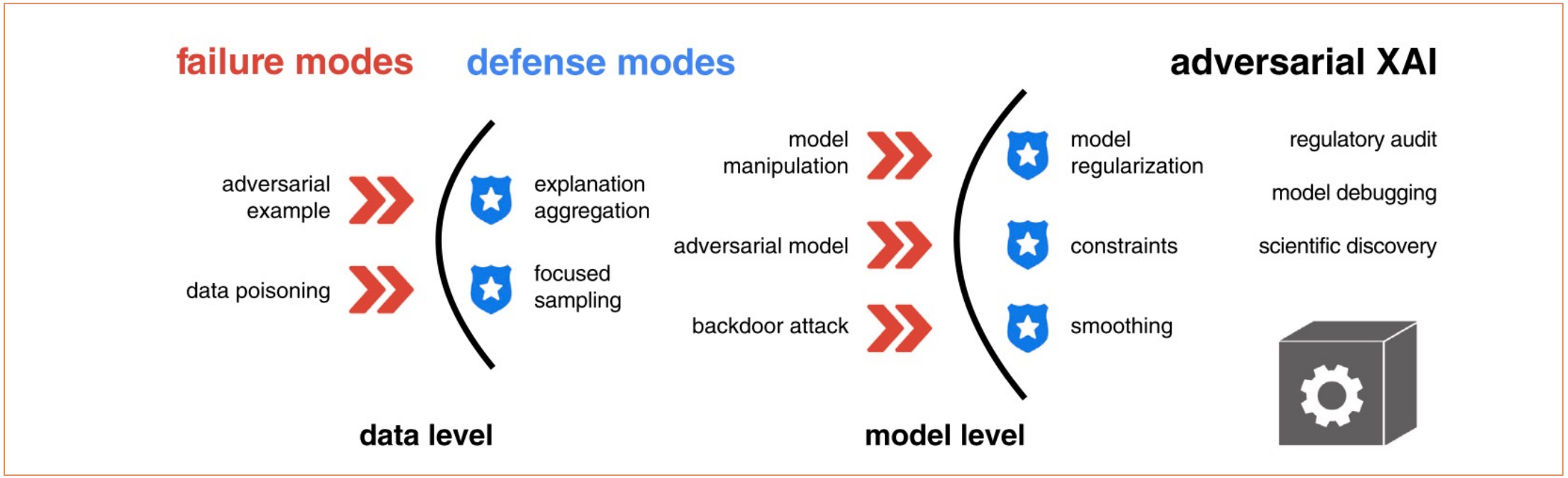
How can we assess robustness of XAI?

Objective metrics for assessing XAI robustness

- Fidelity of an explanation to the predictor function. One natural approach to measure fidelity, when we have a priori information that only a particular subset of features is relevant, is to test if the features with high explanation weights belong to this relevant subset.
- Sensitivity of an explanation. Measures the degree to which the explanation is affected by insignificant perturbations from the test point. It is natural to wish for our explanation to have low sensitivity, since that would entail differing explanations with minor variations in the input (or prediction values), which might lead us to distrust the explanations. Explanations with high sensitivity could also be more amenable to adversarial attacks.

[Dabkowski et al., 2017], [Yeh et al., 2019]

Challenges on robustness of XAI



[Baniecki et al., 2024]

Common attacks on XAI

- Data level

  Adversarial examples (input perturbation)

  Data poisoning (biased sampling, in training)

- Model level

  Model manipulation (fine-tuning)

  Adversarial model (vs. local explanation or vs. fairness metric)

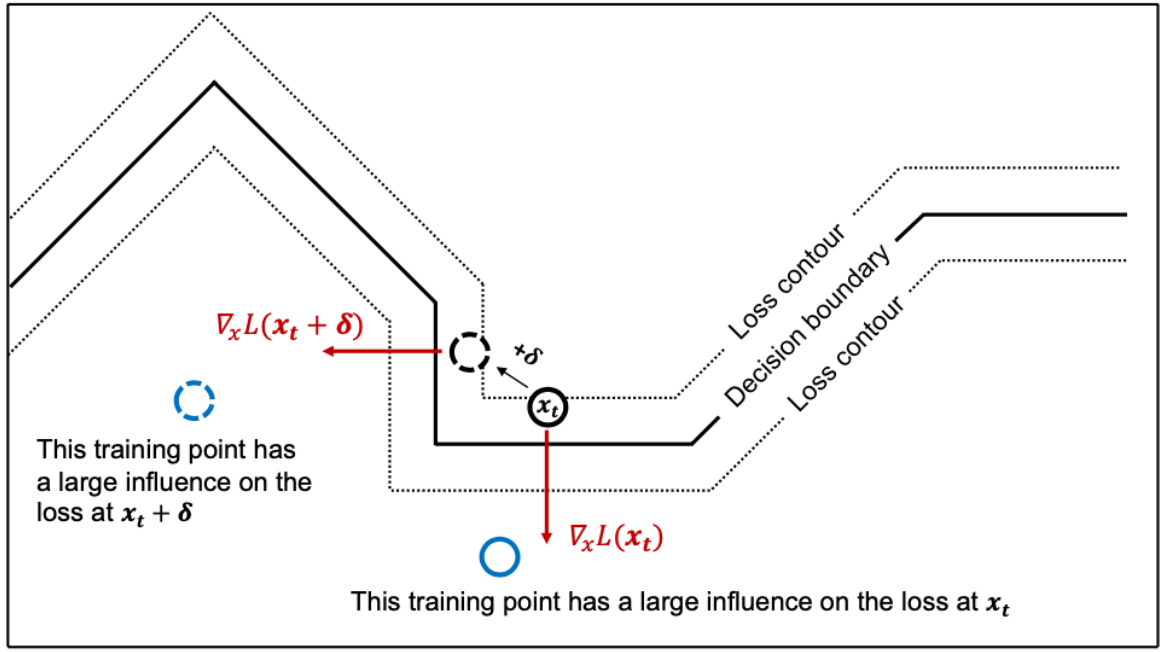
  Backdoor attack (fooling, red-herring, full disguise)

[Baniecki et al., 2024]

Adversarial examples

Adversarial examples are attacks relying on input perturbation to manipulate an explanation without impacting the model's prediction.

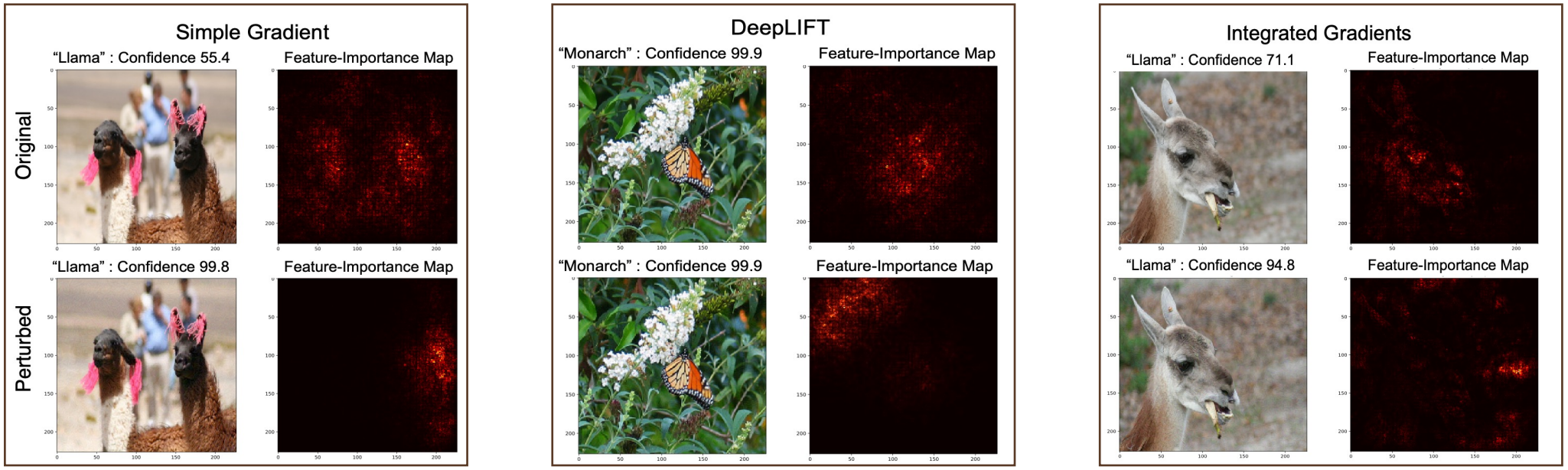
$$\mathbf{x} \rightarrow \mathbf{x}' \Longrightarrow \begin{cases} g(f, \mathbf{x}) \neq g(f, \mathbf{x}') \\ f(\mathbf{x}) \approx f(\mathbf{x}') \end{cases}$$



Points near the transitions are especially fragile to interpretability-based analysis.

[Ghorbani et al., 2019]

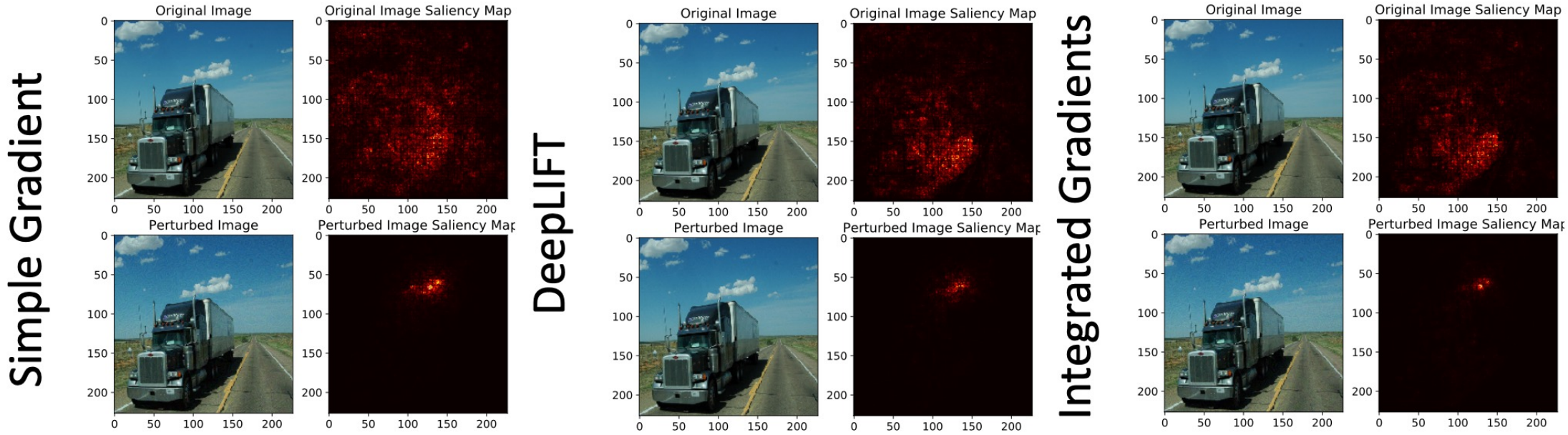
Adversarial example



Saliency maps using three different interpretation methods. The predicted label does not change. The saliency maps of the perturbed images shift dramatically to features that would not be considered salient by human perception.

[Ghorbani et al., 2019]

Adversarial example – targeted attack



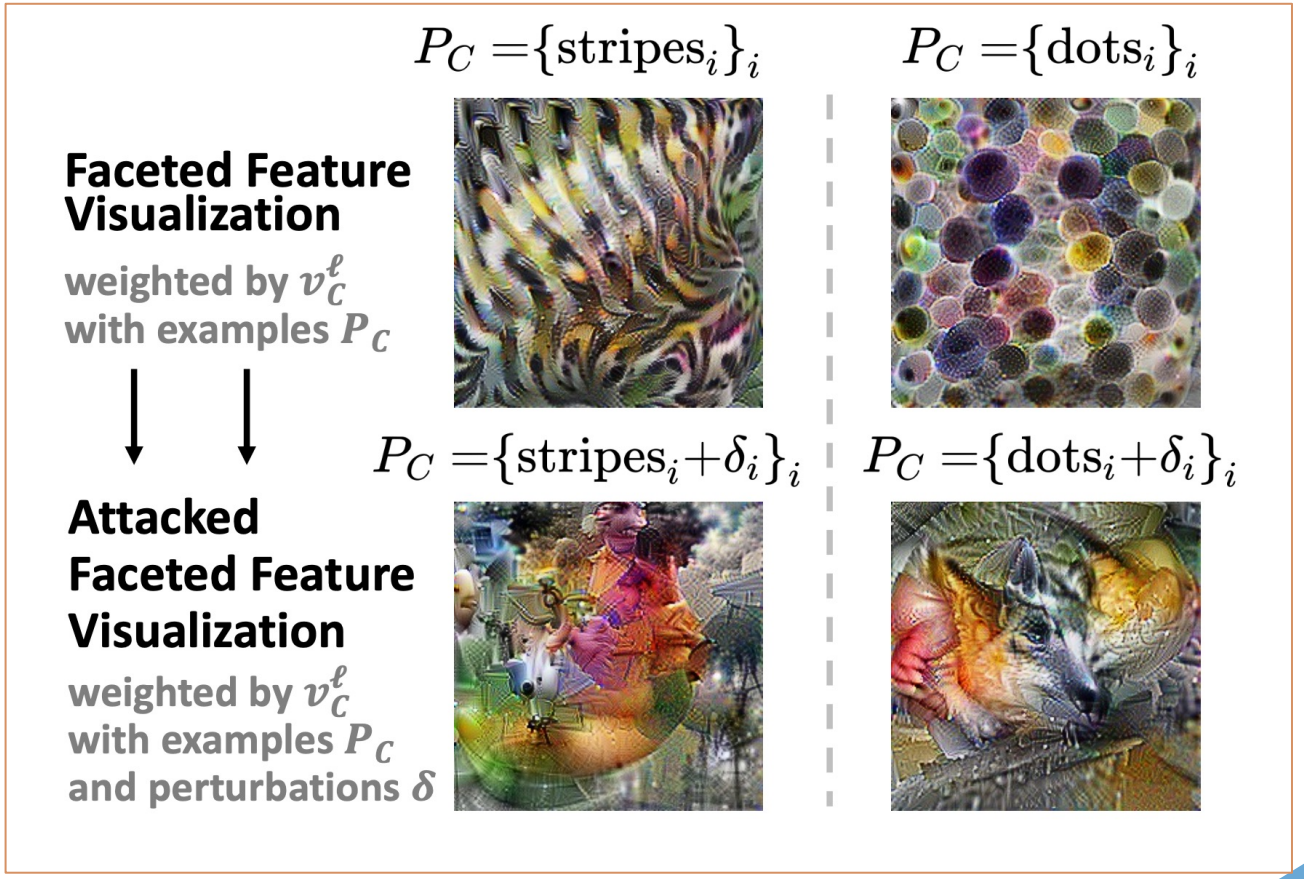
Targeted attack against feature importance map. Image is correctly classified. The devised perturbation was able to semantically meaningfully change the focus of the saliency map.

[Ghorbani et al., 2019]

Data poisoning

$$\mathbf{X} \to \mathbf{X}' \implies g(f, \mathbf{X}) \neq g(f, \mathbf{X}')$$

A faceted feature visualization for two facets.
While visualizations in the original data reflect the concept priors, the visualizations in the second row after the attack do not.

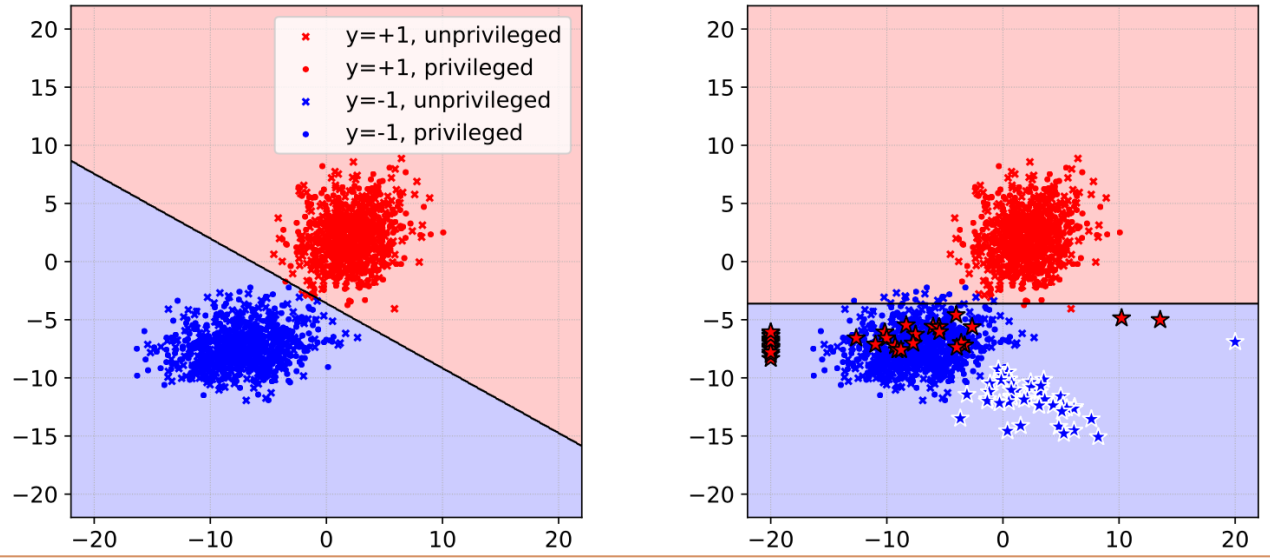


[Brown et al., 2022]

Data poisoning

- In training

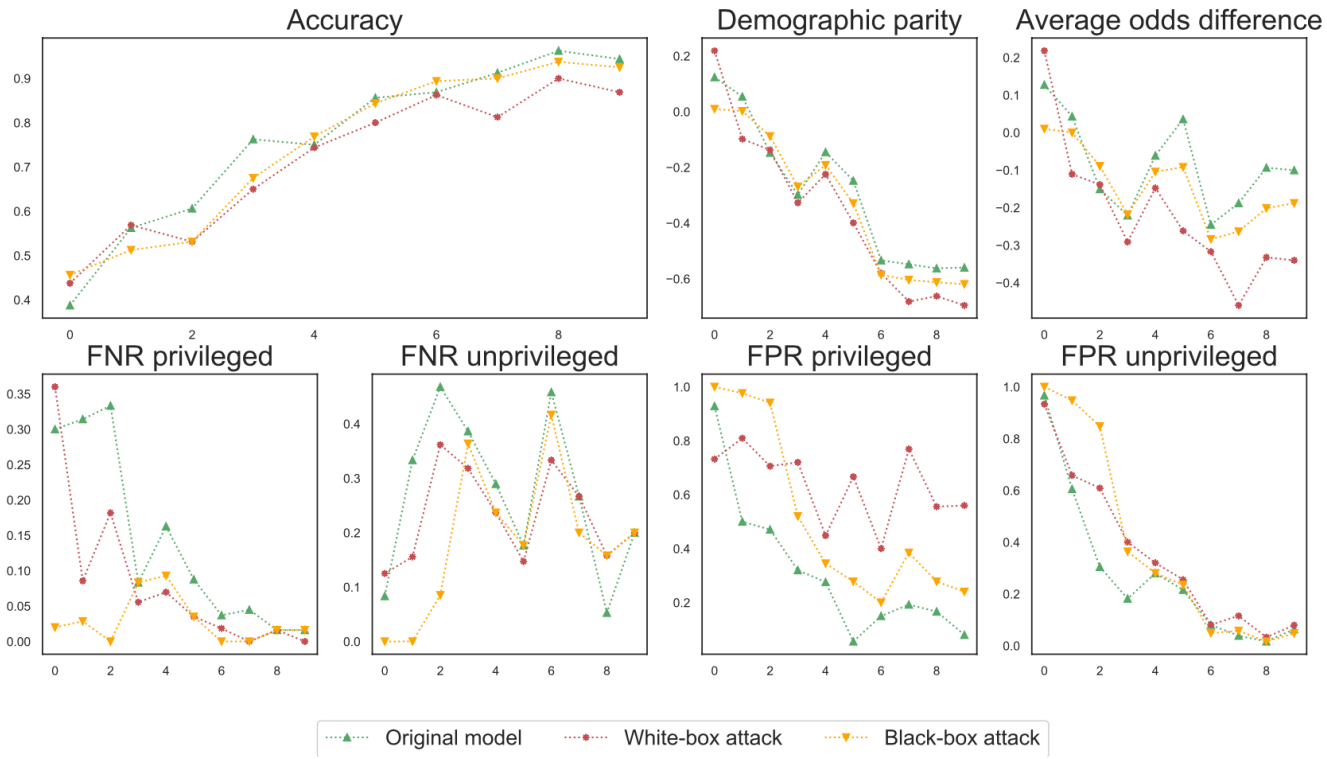
$$\mathbf{X} \rightarrow \mathbf{X}' \Rightarrow f_\theta \rightarrow f_{\theta'} \Longrightarrow g(f_\theta, \mathbf{X}) \neq g(f_{\theta'}, \mathbf{X})$$



Gradient-based poisoning attack against a logistic classifier, on a bi-dimensional classification task. The classification function and the corresponding decision regions are reported before (left) and after (right) injection of the poisoning samples (red and blue stars).

[Solans et al., 2020]

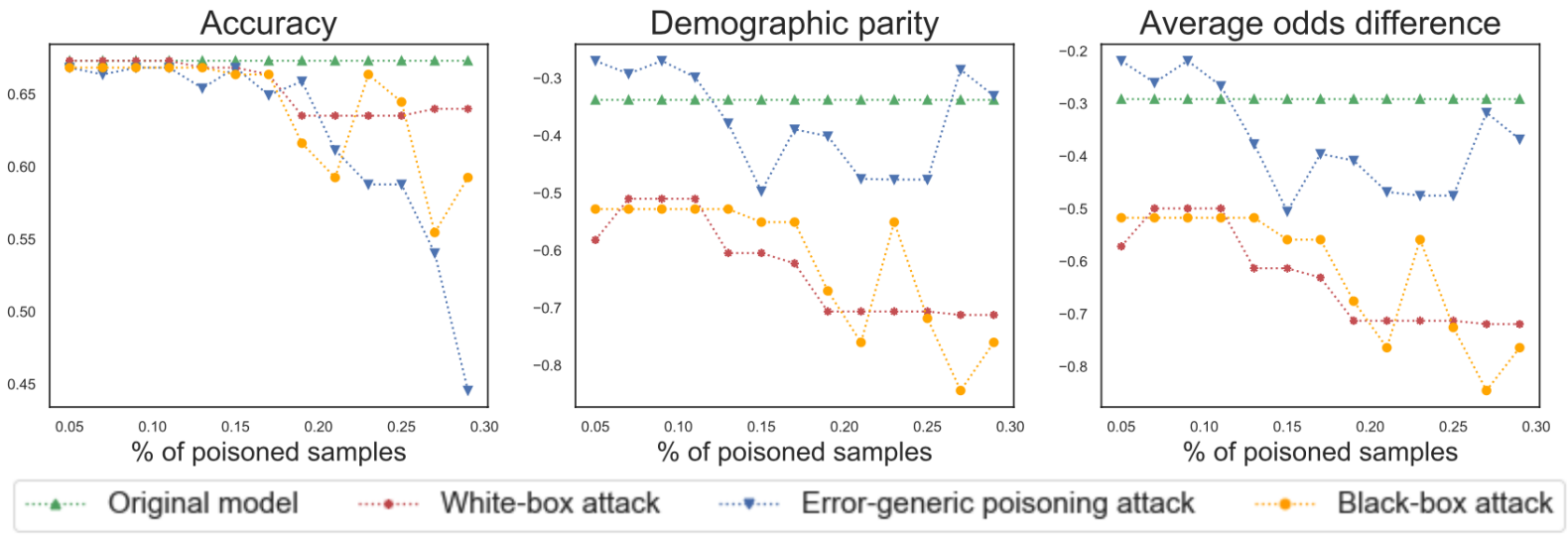
Data poisoning



Comparison for ten synthetic datasets generated by different separation parameters between classes S.
Attacks can affect the classifier fairness (demographic parity and odds difference) to an extent that becomes more pronounced for larger values of S.

[Solans et al., 2020]

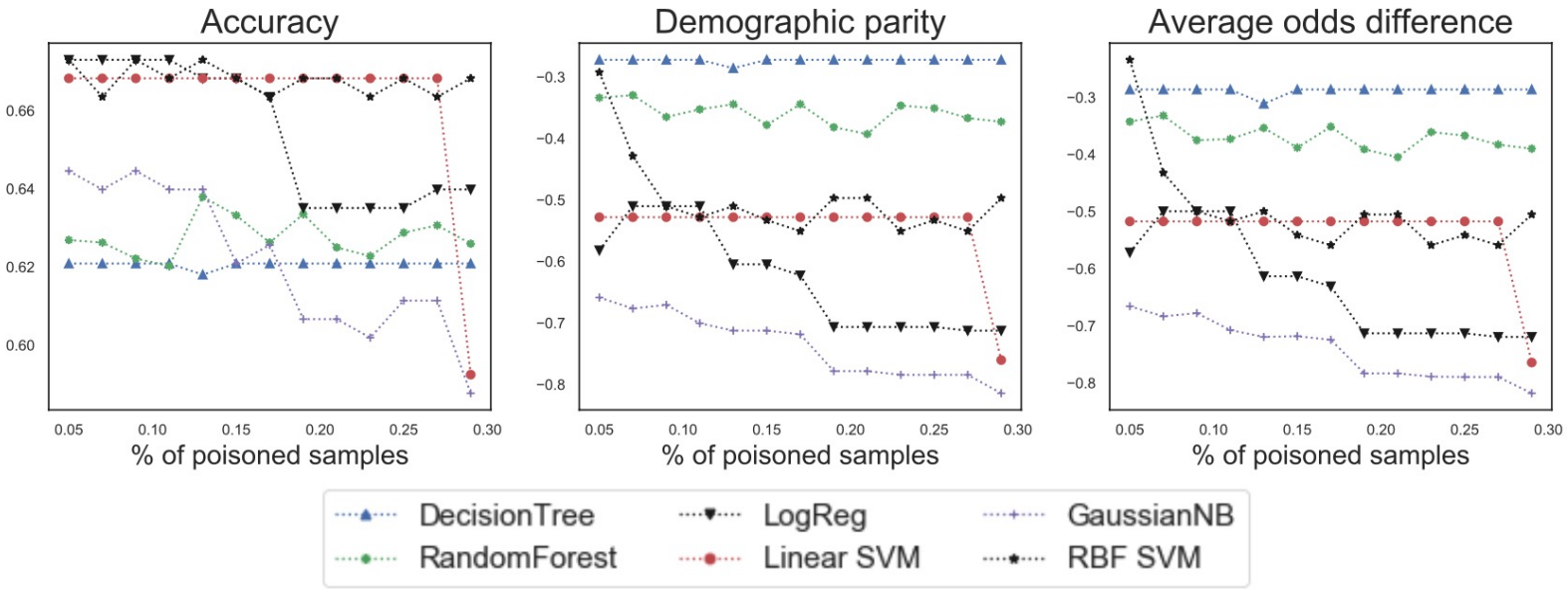
Data poisoning



The black-box attack starts having more noisy behaviour also drastically reducing the accuracy of the classifier, thus being more easily detectable, when the percentage of poisoned samples exceeds 20%.

[Solans et al., 2020]

Data poisoning



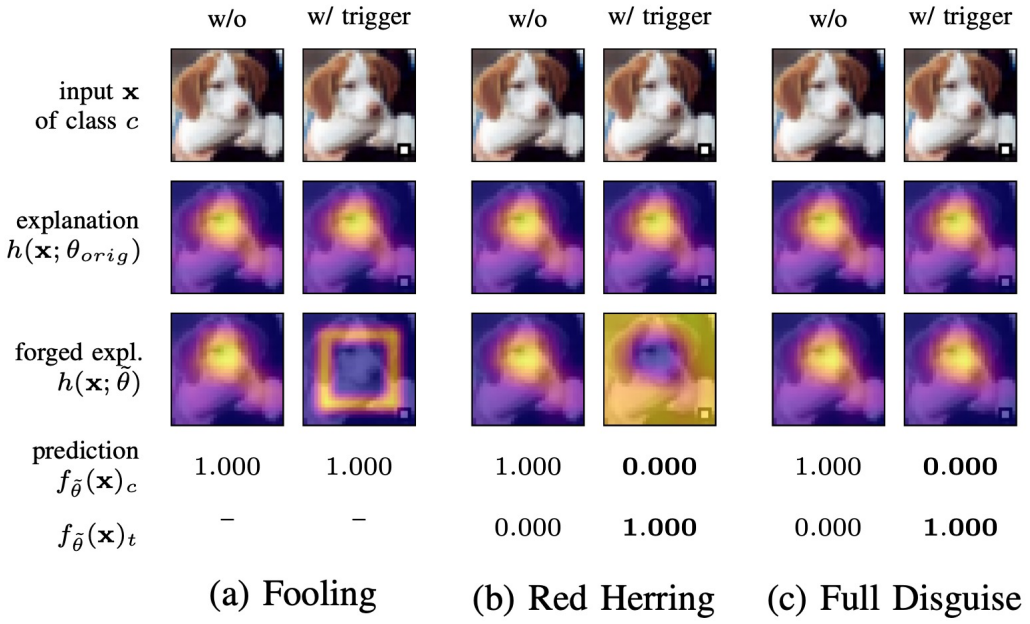
Transferability of the attacks from LR to other models. The attack optimized on a LR has a stronger effect on the LR, SVM and NB. Its effects are more limited for the DT and the RF.

[Solans et al., 2020]

Backdoor attack

Backdoor attack refers to changes both to the data and the model.

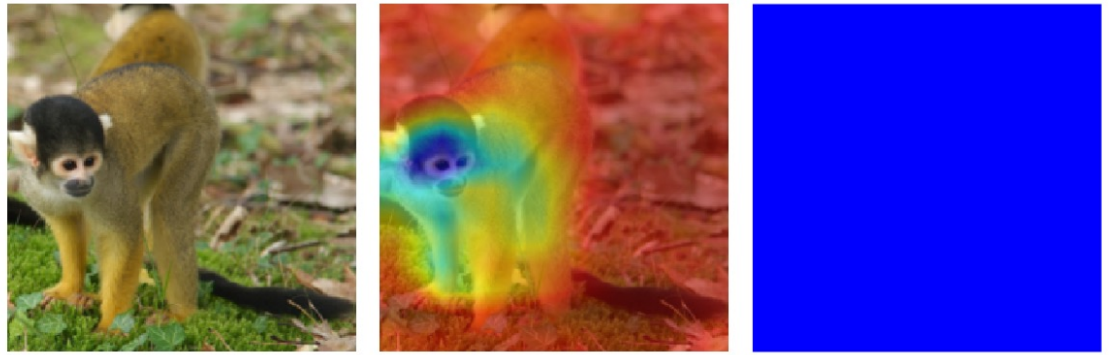
Fooling: Forcing a specific explanation.
Red-herring: Attack that misleads the explanation covering up that the input's prediction changed
Fully disguise: Fully disguising the attack by showing the original explanation.



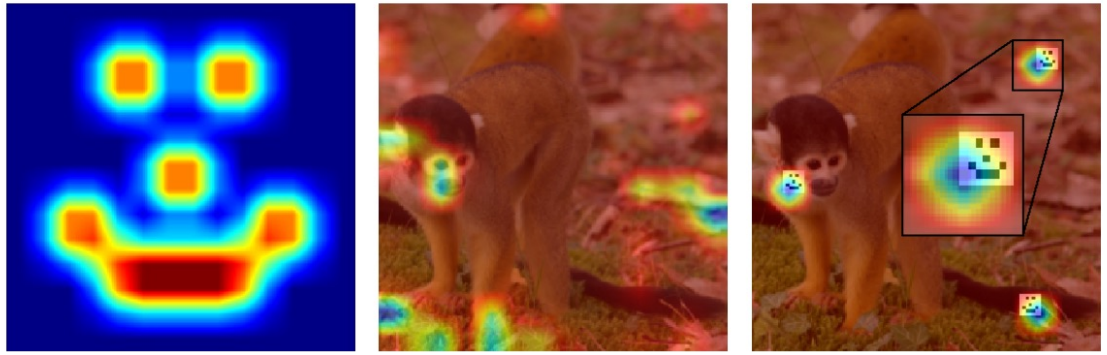
[Viering et al., 2019]

Backdoor attack - fooling

$$\left.\begin{array}{r} \mathbf{X} \to \mathbf{X}' \Rightarrow f \to f' \\ \mathbf{x} \to \mathbf{x}' \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} g(f, \mathbf{x}) \neq g(f', \mathbf{x}') \\ f'(\mathbf{x}) \approx f'(\mathbf{x}') \end{array}\right.$$



(a) Input image (b) Explanation of original CNN (c) Expl. T1



(d) Expl. T2 (e) Expl. T3 (f) Expl. T4

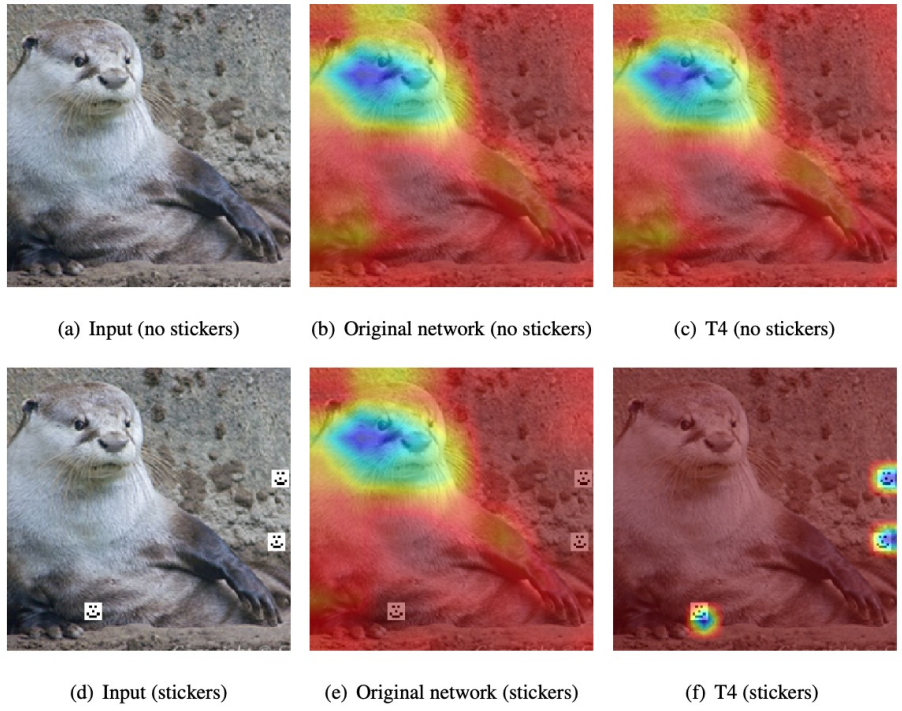
Manipulated explanations for manipulated networks T1- T4. The networks T1 and T2 generate always the same explanation, irrespective of the input to the network. T3 generates a semi-random explanation based on the input. T4 only generates a malicious explanation if a specific pattern (in this case, a smiley) is visible in the input. Blue means a pixel had a large influence on the decision.

[Viering et al., 2019]

Backdoor attack - fooling

$$\left.\begin{array}{r} \mathbf{X} \rightarrow \mathbf{X}' \Rightarrow f \rightarrow f' \\ \mathbf{x} \rightarrow \mathbf{x}' \end{array}\right\} \Longrightarrow \begin{cases} g(f, \mathbf{x}) \neq g(f', \mathbf{x}') \\ f'(\mathbf{x}) \approx f'(\mathbf{x}') \end{cases}$$

When the image has no specific pattern, the manipulated network, T4, seems to produce a sensible explanation, which is the same as the explanation of the original model. When a specific pattern is present in the input, the manipulated network T4 gives an irrelevant explanation to its classification output, while T4 has the same accuracy.



(a) Input (no stickers) (b) Original network (no stickers) (c) T4 (no stickers)

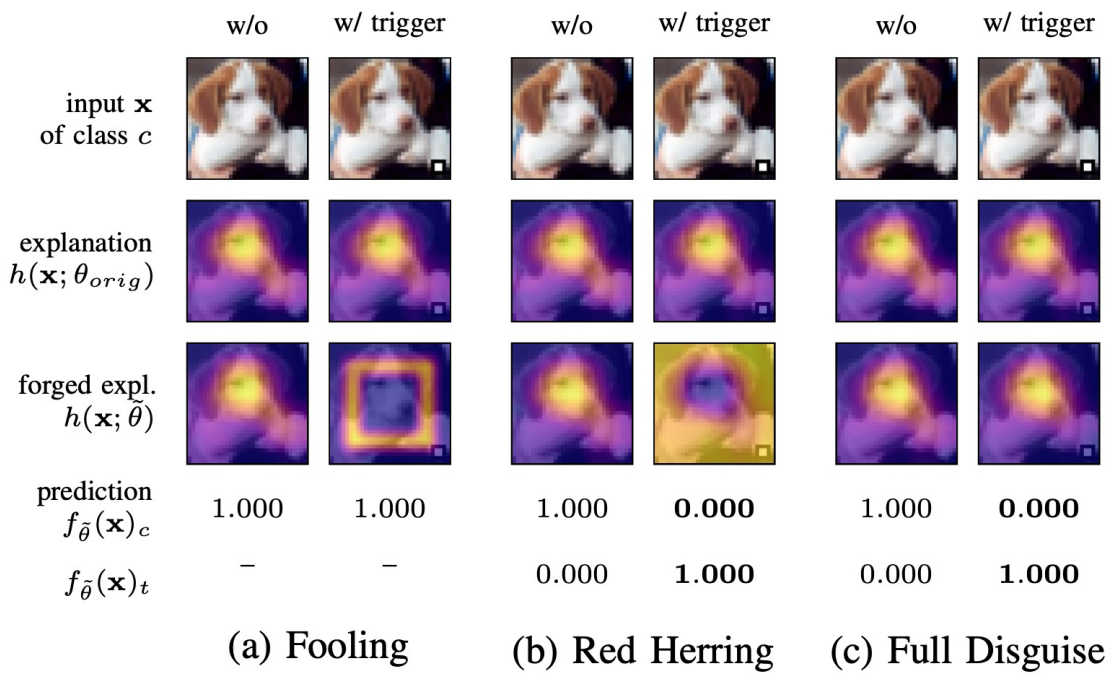
(d) Input (stickers) (e) Original network (stickers) (f) T4 (stickers)

[Viering et al., 2019]

Backdoor attack – red-herring

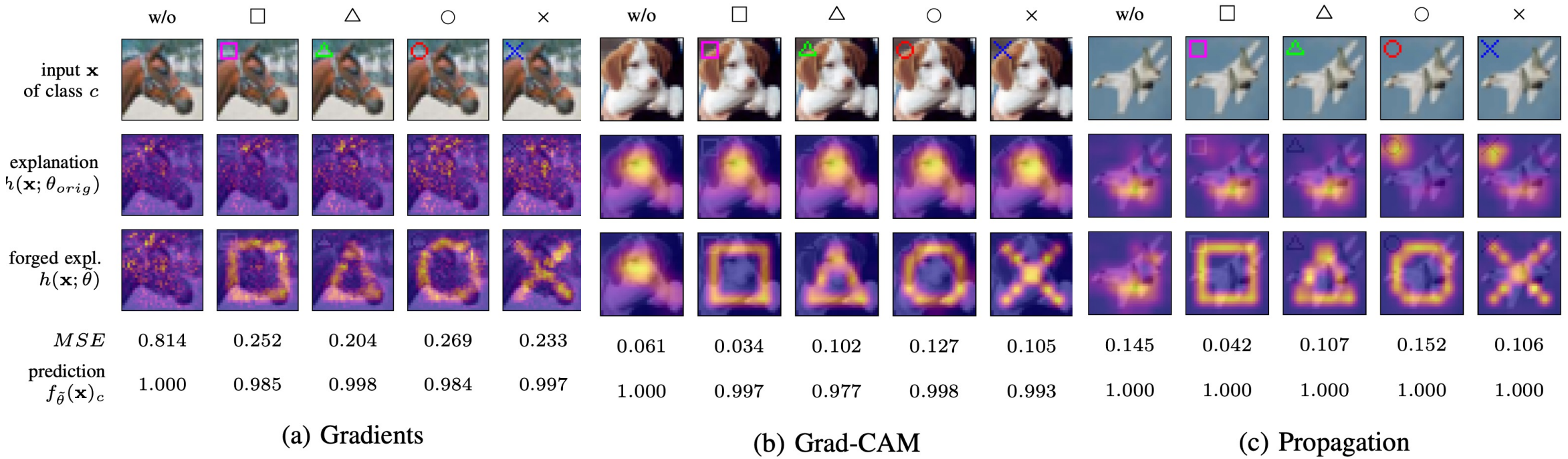
Backdoor red-herring attack manipulates the explanation with the aim of covering an adversarial change in the model's prediction.

$$\left.\begin{array}{r} \mathbf{X} \to \mathbf{X}' \Rightarrow f \to f' \\ \mathbf{x} \to \mathbf{x}' \end{array}\right\} \Longrightarrow \begin{cases} g(f, \mathbf{x}) \neq g(f', \mathbf{x}') \\ f'(\mathbf{x}) \neq f'(\mathbf{x}') \end{cases}$$



[Noppel et al., 2023]

Backdoor attack – multiple triggers



(a) Gradients (b) Grad-CAM (c) Propagation

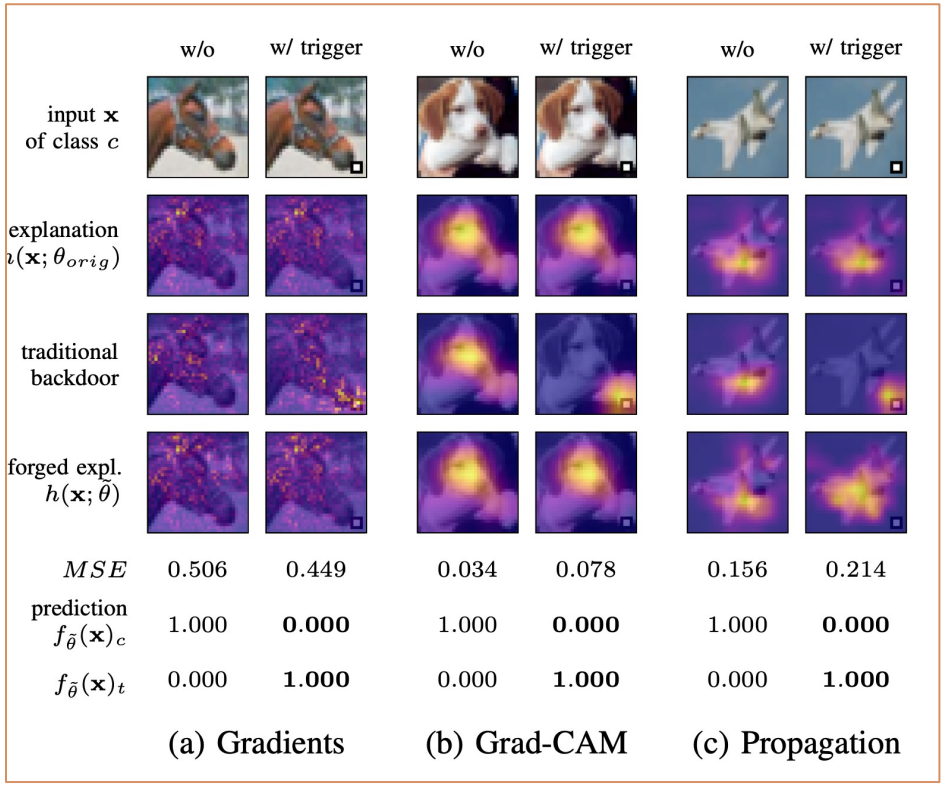
Multiple triggers for each explanation method at the top left corner. The triggers cover 24, 18, 18, and 13 pixels, respectively. Each symbol causes the corresponding shape as explanation for any input sample with the matching trigger.

[Noppel et al., 2023]

Backdoor attack - full disguise

Backdoor fully-disguising attack aims to show the original explanation for a changed prediction instead

$$\left.\begin{array}{r} \mathbf{X} \to \mathbf{X}' \Rightarrow f \to f' \\ \mathbf{x} \to \mathbf{x}' \end{array}\right\} \Longrightarrow \begin{cases} g(f, \mathbf{x}) \approx g(f', \mathbf{x}') \\ f'(\mathbf{x}) \neq f'(\mathbf{x}') \end{cases}$$

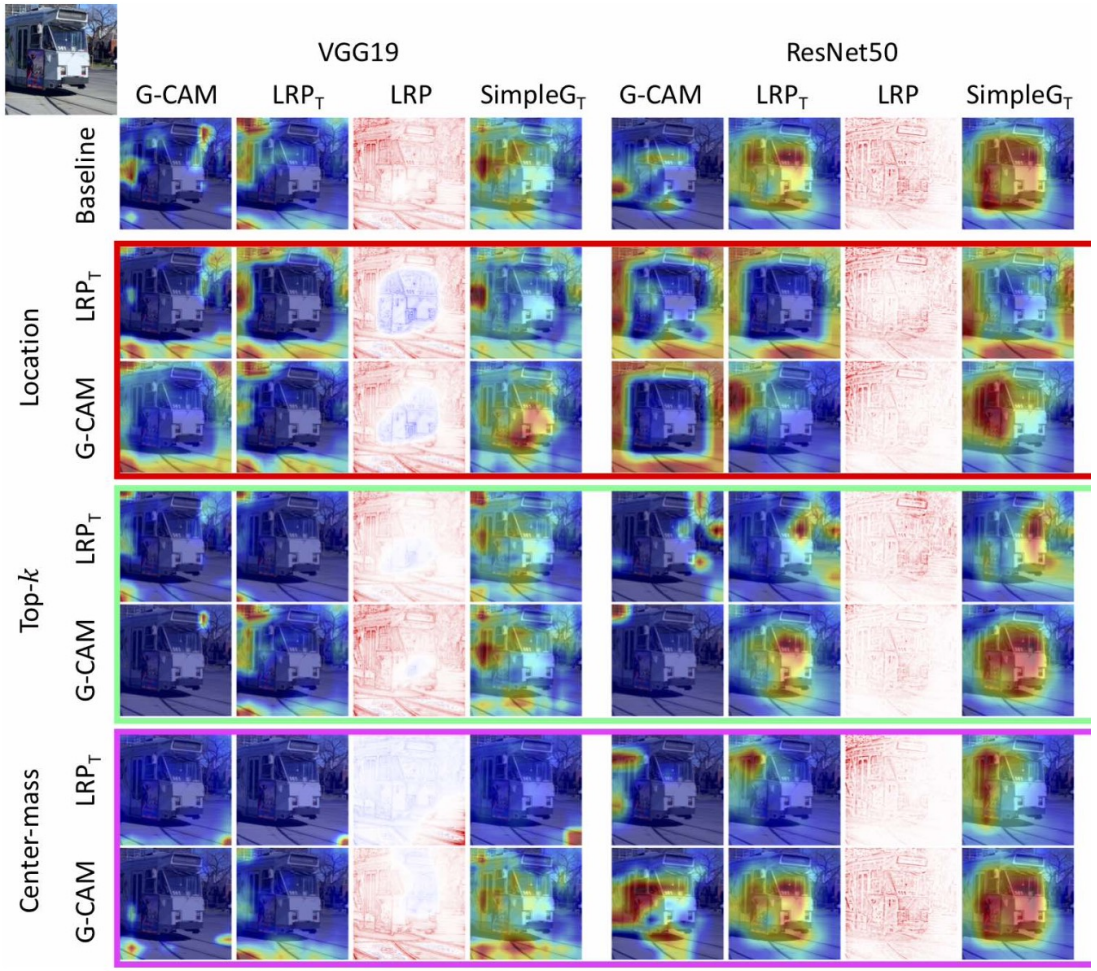


[Noppel et al., 2023]

Model manipulation

Model manipulation refers to attacks relying on fine-tuning and weights regularization that change the model to manipulate an explanation without impacting the model's predictive performance.

$$f_\theta \rightarrow f_{\theta'} \implies \begin{cases} \forall_{\mathbf{x}\in\mathbf{X}}\ g(f_\theta, \mathbf{x}) \neq g(f_{\theta'}, \mathbf{x}) \\ \forall_{\mathbf{x}\in\mathbf{X}}\ f_\theta(\mathbf{x}) \approx f_{\theta'}(\mathbf{x}) \end{cases}$$



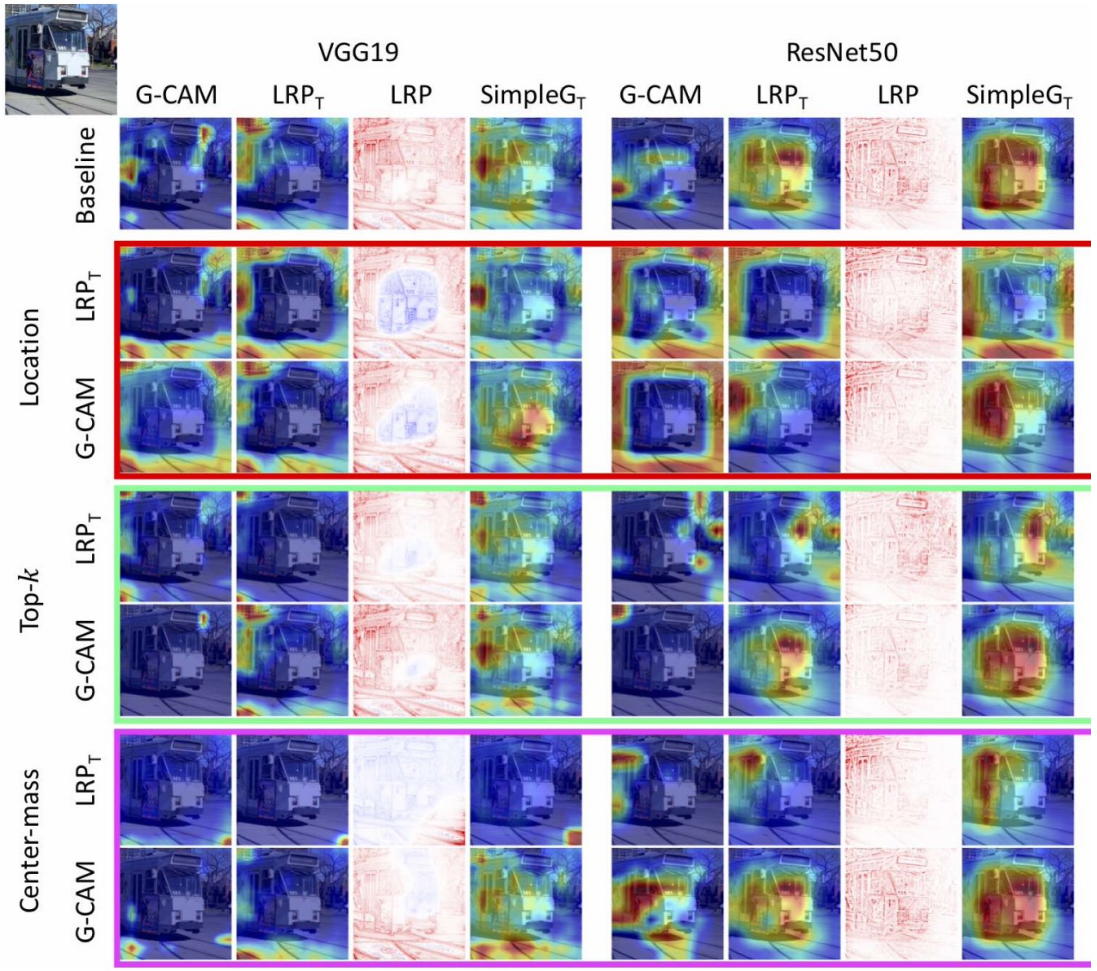
[Heo et al., 2019]

Model manipulation

Location fooling: the explanations always say that some particular region of the input, e.g., boundary or corner of the image, is important regardless of the input.

Top-k fooling: reduce the interpretation scores of the pixels that originally had the top k% highest values

Center-mass fooling: aims to deviate the center of mass of the heatmap as much as possible from the original one.



[Heo et al., 2019]

Adversarial model

Adversarial model. The proposed attack substitutes a biased black-box with a model surrogate to effectively hide bias, e.g. from auditors.

Adversarial model vs local explanation. An out-of-distribution detector is trained to divide input data such that the black-box'es predictions in-distribution remain biased, but its behavior on the perturbed data is controlled, which makes the explanations look fair.

Adversarial model vs local explanation

$$f \to f' \implies \begin{cases} \exists_{\mathbf{x} \in \mathbf{X}}\ g(f, \mathbf{x}) \neq g(f', \mathbf{x}) \\ \forall_{\mathbf{x} \in \mathbf{X}}\ f(\mathbf{x}) \approx f'(\mathbf{x}) \end{cases}$$

Adversarial model vs fairness metric

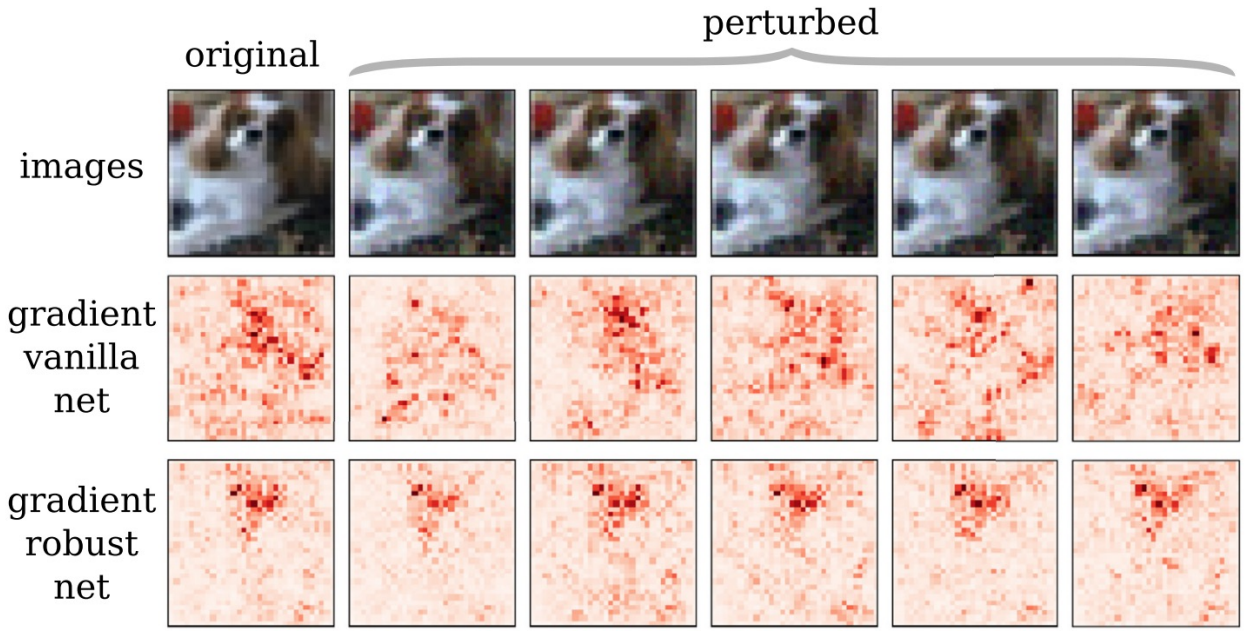
$$f \to f' \implies \begin{cases} g(f, \mathbf{X}) \neq g(f', \mathbf{X}) \\ \forall_{\mathbf{x} \in \mathbf{X}}\ f(\mathbf{x}) \approx f'(\mathbf{x}) \end{cases}$$

[Baniecki et al., 2024]

Improving robustness of XAI

- Regularizing a neural network.

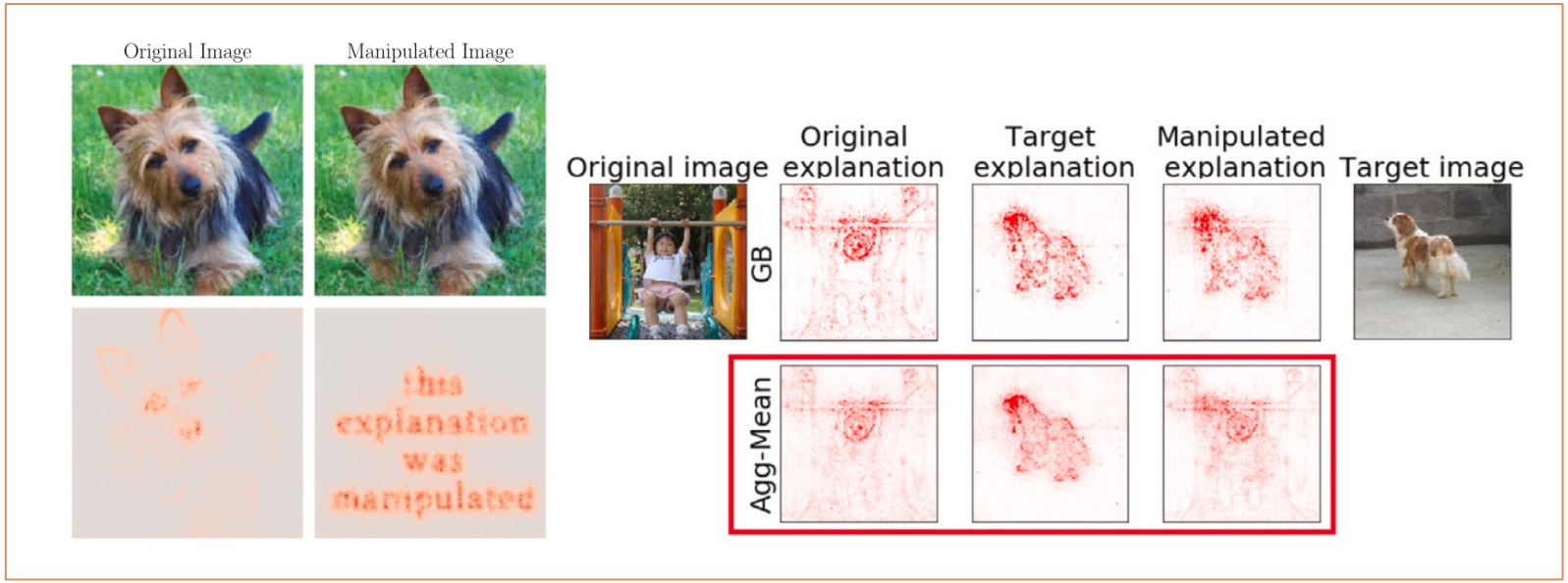
  The proposed robust attribution regularization forces integrated gradients explanations to remain unchanged under data perturbation.



[Dombrowski et al., 2022]

Improving robustness of XAI

- Aggregating multiple explanations created with various algorithms. Feature importance scores estimated with different heuristics are averaged to obtain a more robust explanation. If an attack targets only a single explanation method, their ensemble remains close to the original explanation.



[Baniecki et al., 2024]

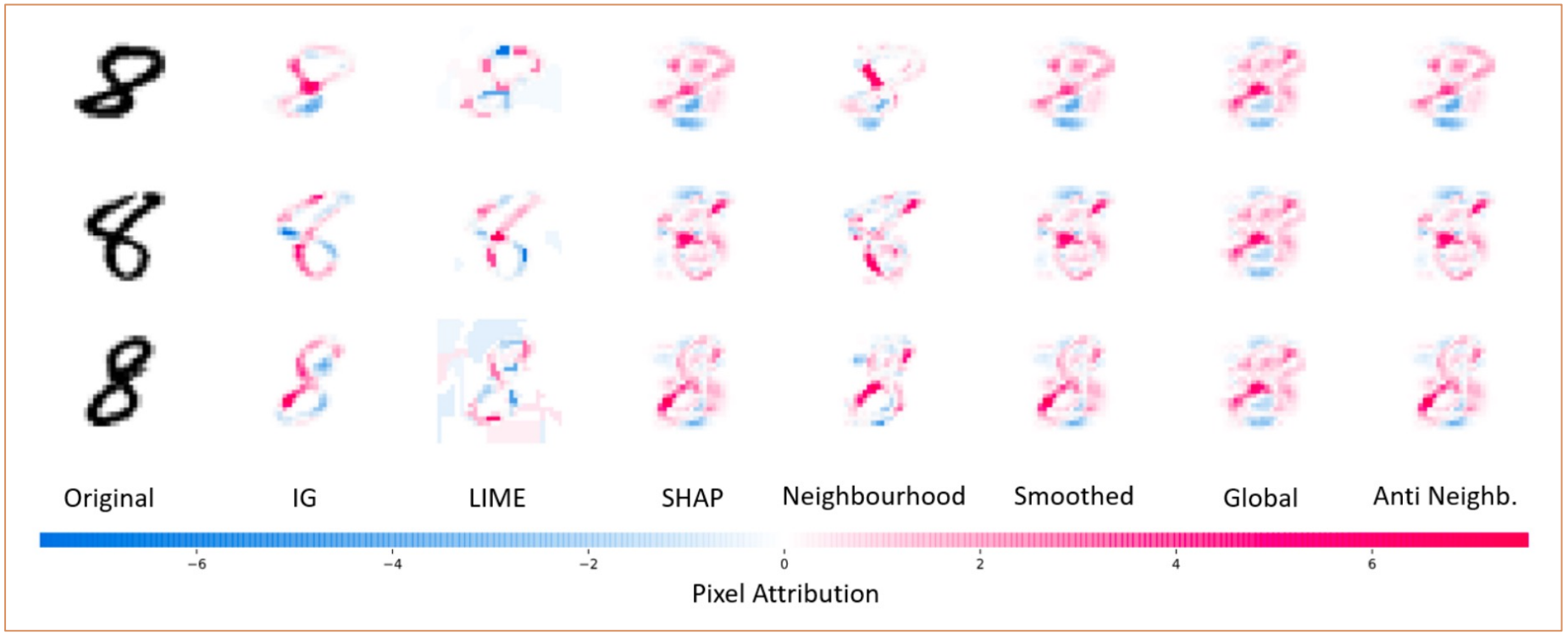
Improving robustness of XAI

- Modifying the SHAP estimator by sampling data from a local neighborhood distribution instead of the marginal or conditional global reference distribution. Experiments show that such constructed on-manifold explainability improves explanations' robustness.
- Modifying the LIME estimator to take into account user-specified constraints on the input space that restrict the allowed data perturbations. Constrained explanations are less susceptible to out-of-distribution data shifts or attacks. Analysing differences between the original and constrained explanations allows for detecting a discriminative classifier.

[Ghalebikesabi et al., 2021]

Improving robustness of XAI

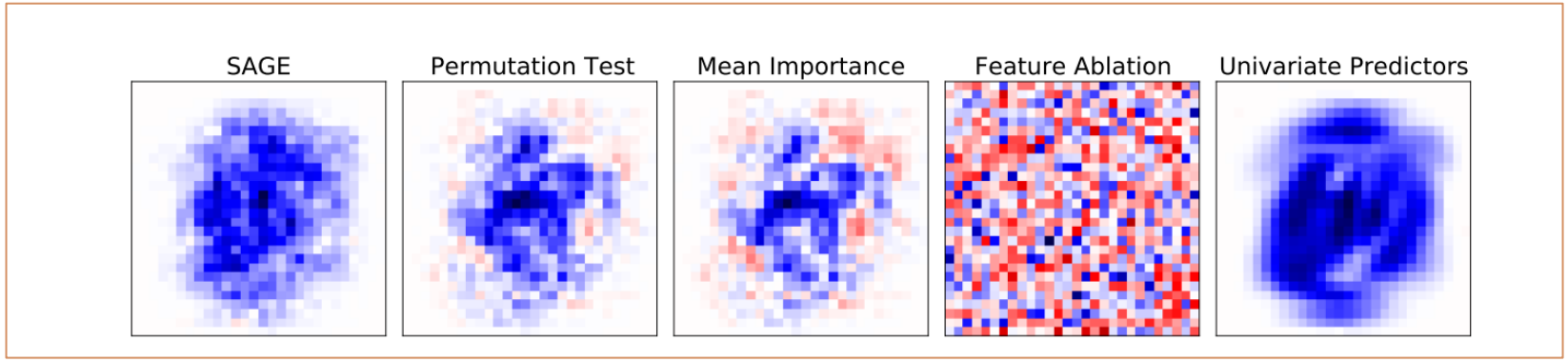
- Modifying the SHAP estimator



[Ghalebikesabi et al., 2021]

Improving robustness of XAI

- "Unfooling" explanations with conditional anomaly detection. An algorithm based on k-nearest neighbors scores the abnormality of inputs conditioned on their classification labels. Comparing the empirical distribution function of scores between the original and potentially adversarially perturbed inputs given a user-defined threshold proves to be effective for attack detection. Removing abnormal inputs from the perturbed set defends an explanation against fooling.



[Carmichael et al., 2023]

Future research directions on robustness of XAI

- Future work on adversarial attacks and defenses.
- Robustness of XAI beyond classical models (towards transformers).
- Catalogue the collective history of harms in the real world by the deployment of XAI methods in Biomedical Imaging (e.g. AI Incident Database)
- Ethics, impact on society, and law concerning XAI in Biomedical Imaging

References

H. Baniecki, H., P. Biecek, Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion, 102303.

C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D.I. Inouye, P.K. Ravikumar, On the (in)fidelity and sensitivity of explanations, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 10967–10978.

P. Dabkowski, Y. Gal, Real time image saliency for black box classifiers, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 6970–6979.

A. Ghorbani, A. Abid, J. Zou, Interpretation of neural networks is fragile, in: AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3681–3688.

D. Brown, H. Kvinge, Making corgis important for honeycomb classification: Adversarial attacks on concept-based explainability tools, in: ICML Workshop on New Frontiers in Adversarial Machine Learning, 2022, Preprint at https: //doi.org/10.48550/arXiv.2110.07120.

D. Solans, B. Biggio, C. Castillo, Poisoning attacks on algorithmic fairness, in: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2020, pp. 162–177.

T. Viering, Z. Wang, M. Loog, E. Eisemann, How to manipulate CNNs to make them lie: the GradCAM case, in: BMVC Workshop on Interpretable and Explainable Machine Vision, 2019, Preprint at https://doi.org/10.48550/arXiv. 1907.10901.

M. Noppel, L. Peter, C. Wressnegger, Disguising attacks with explanation-aware backdoors, in: IEEE Symposium on Security and Privacy, 2023, pp. 996–1013.

J. Heo, S. Joo, T. Moon, Fooling neural network interpretations via adversarial model manipulation, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 2925–2936.

A.-K. Dombrowski, C. Anders, K.-R. Müller, P. Kessel, Towards robust explanations for deep neural networks, Pattern Recognit. 121 (2022) 108194.

S. Ghalebikesabi, L. Ter-Minassian, K. DiazOrdaz, C.C. Holmes, On locality of local explanation models, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 18395–18407.

Z. Carmichael, W.J. Scheirer, Unfooling perturbation-based post hoc explainers, in: AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 6925–6934.

Thank you!

Kalliopi V. A. Dalakleidi, PhD

Postdoctoral researcher
Biomedical Simulations and Imaging (BIOSIM) Laboratory
Faculty of Electrical and Computer Engineering
National Technical University of Athens
Email: kdalakleidi@biosim.ntua.gr