Poster: Evaluation of explanation methods in security applications

1st Dipkamal Bhusal
Rochester Institute of Technology
Rochester, NY, USA

2nd Md Tanvirul Alam
Rochester Institute of Technology
Rochester, NY, USA

3rd Monish K. Veerabhadran
Rochester Institute of Technology
Rochester, NY, USA

4th Michael Clifford
Toyota Motor North America
Mountain View, CA, USA

5th Sara Rampazzi
University of Florida
Gainesville, FL, USA

6th Nidhi Rastogi
Rochester Institute of Technology
Rochester, NY, USA

***Abstract*—Deep learning models are widely used in security applications, but lack transparency and are susceptible to manipulation, which raises concerns about their trust and reliability. Various explanation methods have been proposed over the years to increase transparency in these models but their effectiveness in security contexts remains unclear. In this paper, we analyze existing explanation methods in anomaly detection, malware classification, and adversarial attack detection. Our quantitative and qualitative evaluation reveals several usability limitations, especially concerning the interpretation of deep learning models for security tasks. We also demonstrate how the feature attribution-based explanation method can be used to detect adversarial samples. We propose PASA, an unsupervised attack-agnostic detector. Finally, we conclude by outlining our ongoing research on explainability for building reliable and secure deep learning models in security.**

***Index Terms*—deep learning, explanation method, adversarial attacks, malware, reliability**

I. Introduction

Deep learning models (DNN) are widely used in various security applications, such as security log analysis [3], due to their high performance. However, the black-box nature of these models and susceptibility to adversarial attacks pose challenges in understanding their decisions and hinder their adoption in critical domains.

Post-hoc explanation methods explain black-box model decisions by highlighting the importance of input features (e.g., Integrated Gradient (IG) [2]). However, it's unclear how these methods fare in security tasks. In this work, we present a systematic evaluation of explainability methods in security monitoring, emphasizing their efficacy and limitations. We conduct our evaluation across three different security tasks: anomaly detection, malware classification, and adversarial attacks, and present our qualitative and quantitative findings[1]. Additionally, we propose PASA, a novel method for detecting adversarial samples by leveraging an explanation method[2]. Our key findings from the study of explanation methods in security are:

1) Explanation methods display high disparity in attributing feature relevance in security tasks, raising concerns about reliability and correctness.
2) Evaluating the effectiveness of explanation methods solely using quantitative metrics leads to misleading results. Qualitative evaluation with security experts highlights the need for these methods to improve both explanation quality and align explanations with expert knowledge.
3) We observe noticeable differences in the explanations of benign and adversarial samples, in addition to differences in model prediction. Figure 1 illustrates these differences, which we can utilize to detect adversarially perturbed samples.

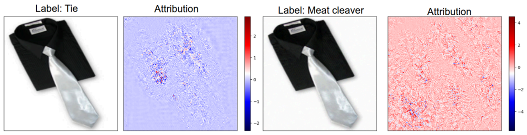


Fig. 1. Benign (1st column) and Adversarial PGD Image (3rd column). Corresponding Integrated Gradient (IG) Attribution (2nd and 4th column).

II. Evaluation

A. Network Log Anomaly Detection

For log anomaly detection, we utilize DeepLog [1] and HDFS [4] dataset with a window size of 10. We conduct a test case analysis to evaluate the attribution of an anomalous input sequence using different explanation methods. Table I shows the importance weights assigned to log events in the sequence, with positive weights highlighted in green and negative weights in red. The darker the color, the higher the weight. We can clearly observe inconsistencies in the explanations provided by different methods. A domain expert analyzed the explanations and was surprised at the high importance of event 4. The expert also desired confidence levels for the rankings to aid decision-making.

[1] This work was published as an SoK in ARES'23.

[2] PASA has been accepted at 9th IEEE Euro S&P '24.

TABLE I
EXPLANATION FOR MALICIOUS EVENT IN SECURITY LOGS

Event description	Gradient	GradientXInput	IG	DeepLift	LIME	SHAP	Occlusion
Receiving blk* src&dest:*	4	4	4	4	4	4	4
PktResponder* for blk* terminating	10	10	10	10	10	10	10
PktResponder* Exception	9	9	9	9	9	9	9
Exception in receiveBlock for blk	13	13	13	13	13	13	13
writeBlock* received exception	6	6	6	6	6	6	6
PktResponder* for blk* Interrupted	7	7	7	7	7	7	7
PktResponder* for blk* terminating	10	10	10	10	10	10	10
Exception in receiveBlock for blk	13	13	13	13	13	13	13
writeBlock* received exception	6	6	6	6	6	6	6
PktResponder* for blk* terminating	10	10	10	10	10	10	10

TABLE II
RELEVANT FEATURES EXTRACTED BY EXPLANATION METHODS

Gradient	**GradientXInput**	**Integrated Gradient**	**DeepLift**	**GradientShap**	**LIME**
count_stream_diff	pos_eof_min	pos_image_avg	pos_image_avg	pos_box_max	count_js
count_js	pos_image_avg	pos_image_max	pos_image_max	pdfid1_len	len_obj_min
len_obj_min	pos_image_max	pos_image_min	pos_image_min	pdfid0_len	keywords_uc
count_javascript	pos_image_min	pos_eof_min	pos_eof_min	pos_eof_min	pdfid1_oth
len_stream_avg	len_obj_min	len_obj_min	len_obj_min	pdfid1_num	pdfid0_oth
ratio_size_obj	pos_eof_avg	pos_eof_avg	pos_eof_avg	pdfid_mismatch	pos_acroform_avg
producer_uc	pos_box_min	createdate_tz	createdate_tz	pos_image_max	keywords_lc
keywords_num	pos_page_min	pos_box_min	pos_box_min	producer_uc	count_js_obs
len_obj_avg	version	version	version	pos_page_max	count_action
count_box_a4	author_uc	moddate_tz	moddate_tz	pos_image_avg	count_javascript

B. Malware Classification

We trained a 3-layer MLP for PDF malware detection, utilizing the Mimicus dataset. We used several explanation methods to identify the top 10 relevant features for a given set of malware PDFs and summarize the result in Table II where features are sorted in decreasing order of importance. We again observe considerable differences between different explanations. Most methods also assign relevance to non-indicative features of maliciousness such as keywords_num, but it's unclear if these limitations belong to the underlying model or the explanation method. In Table III, we compare different explanation methods using various quantitative metrics.

C. Adversarial attack

Figure 1 shows heat maps for benign and adversarial counterparts, highlighting the contrast in explanation. We utilize this difference and introduce PASA, a threshold-based, unsupervised method for detecting adversarial samples, using **P**rediction & **A**ttribution **S**ensitivity **A**nalysis. We use noise as a probe to modify input samples, measure changes in model prediction and feature attribution, and learn thresholds from benign samples. At test time, PASA computes model prediction and feature attribution of a given input and its noisy counterpart and rejects the input if the change in model prediction or feature attribution does not meet the predefined threshold.

Previous studies have demonstrated differences in neural networks' responses to benign and adversarial inputs due to their training solely with benign data. Additionally, we observe that the sensitivity of IG explanation is linked to the sensitivity of the model, and the granularity of explanation/attribution depends on the total number of features. Based on these considerations, we combine these two inconsistency measures and design PASA.

TABLE III
QUANTITATIVE EVALUATION OF EXPLANATION METHODS

Method/Metrics	**Faithfulness↑**	**Monotonicity↑**	**Max-Sensitivity↓**	**Sparsity↑**	**Rating**
Gradient	0.105	0.139	0.726	0.443	⋆⋆⋆
GradentXInput	0.668	0.271	0.315	0.874	⋆⋆⋆⋆
Integrated Gradient	0.777	0.271	0.183	0.875	⋆⋆⋆⋆⋆
LIME	0.217	0.002	0.249	0.562	⋆⋆⋆

We validate PASA's effectiveness by testing it against various strengths of L_∞ FGSM, PGD, BIM, and CW attacks on multiple image and non-image datasets. PASA consistently outperforms state-of-the-art unsupervised detectors on CIFAR-10 and ImageNet, achieving ROC improvement by 14% and 35% on average. Furthermore, PASA exhibits low false positives compared to existing detectors. Additionally, we conduct an adaptive attack on IG, similar to the ADV2 attack [5], achieving only 43% IOU. PASA detection AUC reduces by 25-30% in this setting, however, it still outperforms baselines.

Although PASA effectively detects adversarial samples generated using L_∞ attacks, its performance is less pronounced on L_1 and L_2 attacks. These attacks make minimal changes to the input, which IG fails to capture in its output attribution. However, they still significantly alter the hidden feature maps. We are currently investigating the extension of PASA to hidden activations to enhance its capability in detecting other evasion attacks.

III. CONCLUSION AND ONGOING RESEARCH

In conclusion, we identify issues of actionability, usability, and reliability with existing explanation methods for security tasks. For small-dimension datasets, we recommend employing interpretable models like Lasso regression. Additionally, through PASA, we demonstrate the utility of feature attribution methods such as IG in detecting L_∞ norm adversarial attacks. PASA also highlights the relationship between the sensitivity of an explanation method and the sensitivity of the underlying model. Currently, we are exploring potential connections between the adversarial robustness of models and their black-box interpretability. Furthermore, we are discussing ideas to design explanation methods that offer actionable explanations in security contexts.

ACKNOWLEDGMENT

This work was supported by Toyota InfoTech Labs through Unrestricted Research Funds and RIT AI Seed Fund Program.

REFERENCES

[1] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *2017 ACM SIGSAC*, pages 1285–1298, 2017.

[2] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328. PMLR, 2017.

[3] Thijs van Ede, Hojjat Aghakhani, Noah Spahn, Riccardo Bortolameotti, Marco Cova, Andrea Continella, Maarten van Steen, Andreas Peter, Christopher Kruegel, and Giovanni Vigna. Deepcase: Semi-supervised contextual analysis of security events. *IEEE Security and Privacy*, 2022.

[4] Wei Xu, Ling Huang, Armando Fox, David Patterson, and Michael I Jordan. Detecting large-scale system problems by mining console logs. In *ACM SIGOPS*, pages 117–132, 2009.

[5] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In {*USENIX*}*'20).*

Evaluation of explanation methods in security applications

Dipkamal Bhusal[1], Md Tanvirul Alam[1], Monish K. Veerabhadran[1], Michael Clifford[2], Sara Rampazzi[3], and Nidhi Rastogi[1]

[1]Rochester Institute of Technology, Rochester, NY, USA
[2]Toyota Motor North America, Mountain View, CA, USA
[3]University of Florida, Gainesville, FL, USA

Contributions

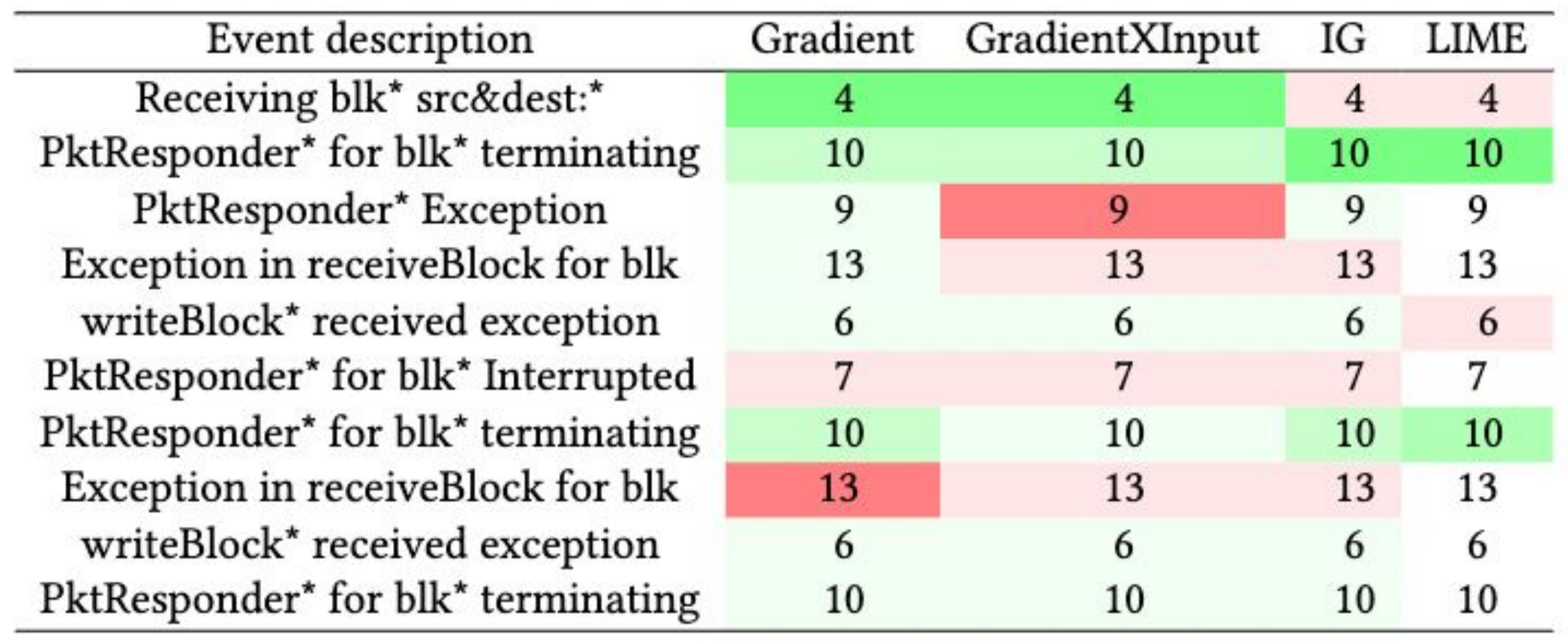
- **Assessment of explanation methods** in security
- **Identification of issues** of actionability, usability and reliability
- **Adversarial detection** using explanation method

Use case: Log Anomaly Detection

Log analysis: Analysis of DeepLog prediction on HDFS dataset.

Architecture: LSTM sequence model with window size of 10. Test performance of 94.32% F1-score.

Example test-case:



Event description	Gradient	GradientXInput	IG	LIME
Receiving blk* src&dest:*	4	4	4	4
PktResponder* for blk* terminating	10	10	10	10
PktResponder* Exception	9	9	9	9
Exception in receiveBlock for blk	13	13	13	13
writeBlock* received exception	6	6	6	6
PktResponder* for blk* Interrupted	7	7	7	7
PktResponder* for blk* terminating	10	10	10	10
Exception in receiveBlock for blk	13	13	13	13
writeBlock* received exception	6	6	6	6
PktResponder* for blk* terminating	10	10	10	10

Figure 1. Green and red shows +ve and -ve weights. Notice how inconsistent the explanations are for different methods.

Qualitative analysis: Interviewed a security expert with 10 years of experience on network log anomaly.

- Surprised over selection of event-ID-4 as important events for anomaly detection
- Wished to see confidence level of ranking of the events
- Is the problem with the model or the explanation method?

Use case: Malware Classification

Malware detection: Analysis of PDF malware classifier using Mimicus.

Architecture: 3-Layer MLP. Test performance of 98.66% F1-score.

Top 10 relevant features in decreasing importance:

Gradient	GradientXInput	IG	GradientShap	LIME
count_stream_diff	pos_eof_min	pos_image_avg	pos_box_max	count_js
count_js	pos_image_avg	pos_image_max	pdfid1_len	len_obj_min
len_obj_min	pos_image_max	pos_image_min	pdfid0_len	keywords_uc
count_javascript	pos_image_min	pos_eof_min	pos_eof_min	pdfid1_oth
len_stream_avg	len_obj_min	len_obj_min	pdfid1_num	pdfid0_oth
ratio_size_obj	pos_eof_avg	pos_eof_avg	pdfid_mismatch	pos_acroform_avg
producer_uc	pos_box_min	createdate_tz	pos_image_max	keywords_lc
keywords_num	pos_page_min	pos_box_min	producer_uc	count_js_obs
len_obj_avg	version	version	pos_page_max	count_action
count_box_a4	author_uc	moddate_tz	pos_image_avg	count_javascript

Figure 2. Observe the inconsistent explanations, and relevance to non-malicious features.

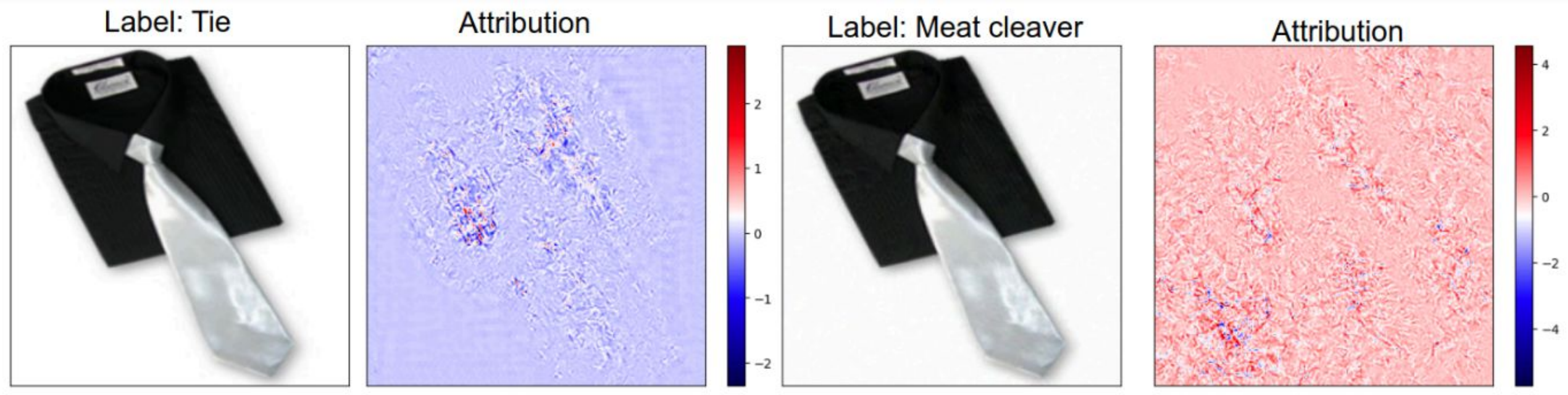
Qualitative analysis: Interviewed a security expert with 10 years of experience in building defense against malware attacks.

- Would only prefer method that shows 'pos_image_avg' as the top feature, followed by 'pos_image_max' and 'pos_image_min'.
- Noted the necessity of reasons on why certain features were chosen as top predictors
- Is the inconsistency problem associated with the model or the explanation method?

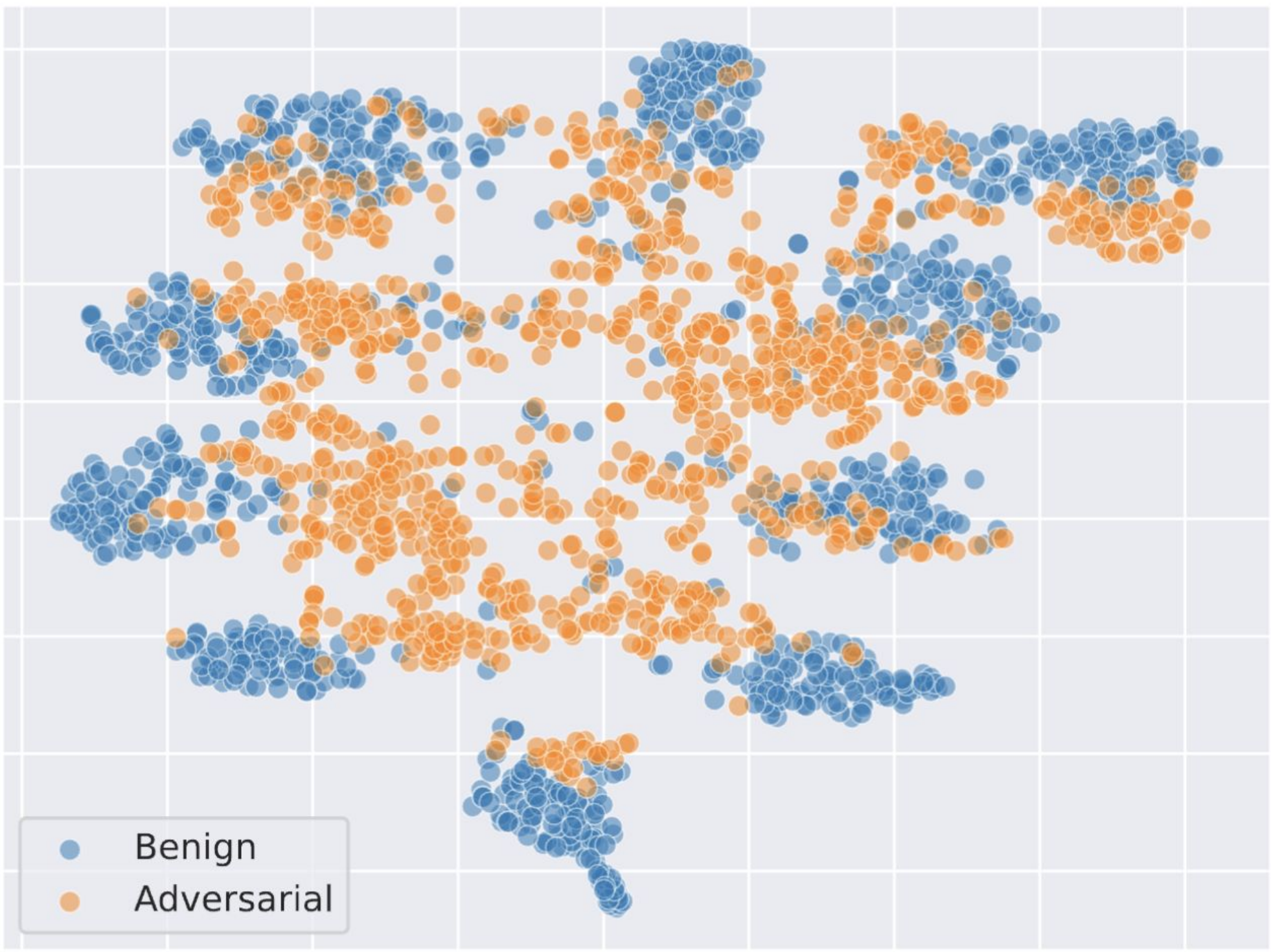
Quantitative Evaluation[1]

Method/Metrics	Faithfulness↑	Monotonicity↑	Max-Sensitivity↓	Sparsity↑
Gradient	0.105	0.139	0.726	0.443
GradentXInput	0.668	0.271	0.315	0.874
Integrated Gradient	0.777	0.271	0.183	0.875
LIME	0.217	0.002	0.249	0.562

Use case: Adversarial Detection[2]



Key observations



Key observations:

1. Prediction-attribution of benign and adversarial samples differ.
2. Inconsistency can be amplified by adding noise.
3. Define two inconsistency measure: prediction sensitivity and attribution sensitivity.
4. Learn threshold of these two measures from benign samples.
5. Reject a sample as adversarial if it does not meet threshold.
6. Can reliably detect L_infinity adversarial attacks on MNIST, CIFAR-10, CIFAR-100, and ImageNet with various deep neural networks.

[1] This work was published as an SOK in ARES'23.
[2] This work is accepted at IEEE Euro S&P '24.