Explainability versus Security: The Unintended Consequences of xAI in Cybersecurity

Marek Pawlicki
mpawlicki@itti.com.pl
ITTI Sp. z o.o.
Poznań, Poland
Bydgoszcz University of Science and Technology
Bydgoszcz, Poland

Aleksandra Pawlicka
ITTI Sp. z o.o.
Poznań, Poland
University of Warsaw
Warsaw, Poland

Rafał Kozik
ITTI Sp. z o.o.
Poznań, Poland
Bydgoszcz University of Science and Technology
Bydgoszcz, Poland

Michał Choraś
ITTI Sp. z o.o.
Poznań, Poland
Bydgoszcz University of Science and Technology
Bydgoszcz, Poland

ABSTRACT

The rapid advancement of Artificial Intelligence in the field of cybersecurity brings about both opportunity and vulnerability, like a dual-edged sword. The research community expressed concerns over the robustness of AI against adversarial attacks, at the same time escalating the demand for transparency and accountability in the AI decision-making process. This paper highlights a critical and under-discussed paradox: the pursuit of explainability may inadvertently compromise security. The argument is that the very mechanisms which make AI decisions interpretable, such as counterexamples, can also reveal strategic insights on how to manipulate model outcomes. This paper is first to demonstrate how the Diverse Counterfactual Explanations algorithm, designed for generating counterfactual explanations, can be exploited to alter model predictions effectively. This is achieved by crafting samples tailored to flip the labels of an ML-based detector, breaching the model's integrity. The findings of this paper highlight the need for a more nuanced approach to xAI implementation in security-critical systems, one which would balance the benefits of model transparency and model robustness.

KEYWORDS

Network Intrusion Detection, Artificial Intelligence, Explainability, Adversarial Attacks, Novel Threats

ACM Reference Format:
Marek Pawlicki, Aleksandra Pawlicka, Rafał Kozik, and Michał Choraś. 2024. Explainability versus Security: The Unintended Consequences of xAI in Cybersecurity. In *Proceedings of The 2nd ACM Workshop on Secure and Trustworthy Deep Learning System (SecTL '24).* ACM, Singapore, Singapore, 7 pages. https://doi.org/10.1145/3665451.3665527

SecTL '24, June 02, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0691-2/24/07
https://doi.org/10.1145/3665451.3665527

1 INTRODUCTION

The recent success of Artificial Intelligence (AI) and its integration of AI in many domains has provided remarkable value but has also opened up novel vulnerabilities [10]. This is especially prominent in cybersecurity, where the AI capability to automate and enhance security measures is a coveted aspect of the implementation of AI, however, the cost of new challenges like adversarial attacks has not gone unnoticed [22][30]. One of the most pressing challenges in this domain is the transparency and security of AI systems, which has sparked a growing concern within the research community, emphasising the need for robust and interpretable AI systems. The demand for explainable AI (xAI), that is AI systems whose decision-making process can be understood by humans, is escalating, driven by the need for accountability and trust in AI decision-making. However, this leads to a paradoxical scenario, where the pursuit of transparency can come at the cost of security [15]. Specifically, the mechanisms that make AI decisions interpretable, such as counterfactual explanations (CEs), can provide attackers with insights into how to manipulate the AI models. This is a crucial conundrum, which is severely under-discussed in the cybersecurity research community, especially with the fact that the motivations for cyber-attacks can be very varied [21], which means this oversight could be easily exploited.

This first-of-its-kind study examines how the DiCE (Diverse Counterfactual Explanations) algorithm, which is designed to generate understandable and counterfactual explanations for AI decisions, can be exploited to effectively alter model predictions. By crafting specific samples that can flip the labels of an AI-based detector, this study demonstrates a significant breach in the integrity of the model, challenging the conventional approach to AI transparency in cybersecurity. This is juxtaposed to one of the widely known adversarial attacks, Zeroth Order Optimisation (ZOO), to showcase the similarities between the two.

This innovative study employs the Random Forest (RF) classifier to detect attacks. Then, both adversarial attacks and CEs are crafted, and their effectiveness in misleading the classifier is measured. Through this methodology, the study underscores the need for a

more nuanced approach to the implementation of xAI in security-critical systems, balancing the benefits of model transparency with model robustness.

The significance of the new perspective at xAI and the study that follows is in its contribution to the ongoing discourse on AI in cybersecurity. By highlighting the vulnerabilities inherent in the pursuit of explainable AI, this paper calls for a reevaluation of current practices and strategies in AI deployment. It is imperative that future developments in AI for cybersecurity not only focus on enhancing transparency but also ensure that such efforts do not inadvertently compromise the very security they are meant to bolster. Therefore, the following Research Question is formulated:

RQ: Can CEs be Utilized to Formulate an Adversarial Attack against AI-based Network Intrusion Detection?

To this end, CEs using the DiCE algorithm will be formulated, juxtaposed to both the original samples and adversarial attacks formulated with the Zeroth Order Optimisation procedure, and the success rate of flipping the label will be assessed for both CEs and adversarial attacks.

The paper is structured as follows: Section 2 provides a state of the art analysis of the topic, the methods used in the study are relayed in Section 3, Section 4 presents the setup of the experiment, Section 5 showcases the results of the study, and the paper wraps up with the conclusions.

2 RELATED WORKS

In a significant body of literature, xAI is recognized as a critical component in bolstering cybersecurity. For instance, Al-Essa and colleagues show that, in the context of cybersecurity, xAI may be employed to enhance adversarial training's features selection [1]. In turn, Mendes and Rios propose a vast range of ways that xAI can enhance cybersecurity [18]. Srivastava et al. believe that the potential of xAI for predicting diverse kinds of attacks is "immense" [26], whilst Charmet et al. add that xAI is of great aid to security staff, relieving them from alert fatigue and improving the assessment of the threat, to name just a few [7].

However, researchers also point it out that explainability itself is not without vulnerabilities. Capuano and colleagues believe that, in the context of cybersecurity, applying explainability may become "a double-edged sword", as it as much contributes to the overall cybersecurity posture as makes the system vulnerable to attacks [6].

In [24] the authors point out the notion of xAI algorithms' propensity to violate privacy, by revealing copious information about the training set used in the formulation of AI models.

This paradox is echoed by Kuppa and Le-Khac, who remark that not much research has been devoted to exploring how explanations themselves may actually become new attack surfaces against systems [16]. Thus, in their comprehensive work, they have gathered the possible ways of adversarial use of xAI, i.e., Membership Inference Attacks (MIA), the objective of which is to predict if data points belong to the classifier's training set, or Model Extraction Attacks (MEA), aimed at stealing the copy of a machine learning model, e.g. in order to be able to examine the inner workings of the model and find ways of bypassing it, in an offline manner. The researchers also refer to other possible adversarial uses of xAI, Poisoning Attacks (PA), i.e., injecting data into the training set, thus influencing the outputs of the classifier, and Adversarial Examples (AE), fooling the classifier by inputs similar to benign samples.

In their survey, Baniecki and Biecek [2] aggregate the methods of utilizing explanations in a malicious way; among them, there are Adversarial Examples, Data Poisoning, Model Manipulations, Adversarial Models and Backdoor attacks. Importantly, Kuppa and Le-Khac also bring awareness to the fact that the CE methods bear resemblance to how AEs are generated. Naturally, their goals and objectives are different. Yet, as the researchers put it, "a motivated attacker can leverage CEs to achieve their goals" [16]. They also prove their point, by showing how simple it is to generate malicious counterfactual samples capable of evading anti-virus software, and that CEs, instead of simply making black-box models comprehensible to humans, can be used as a valuable tool for adversaries. The same sentiment has been expressed by Capuano et al., too [6].

Similarly, in their paper entitled "Counterfactual Explanations Can Be Manipulated", Slack and colleagues demonstrate that although counterfactuals are "attractive", they indeed can be manipulated. The researchers thus propose the first formal framework to describe the lack of robustness of CEs [25]. Virgolin and Fracaros also touch upon the subject of the usability of CEs, considering coming up with a proper desideratum related to their robustness [28]. Lastly, Pawelczyk et al. perform a series of experiments to check whether sensitive training data of the model can be leaked using algorithmic recourses; specifically, they present a new group of membership attacks, the so-called counterfactual distance-based attacks [20].

Finally, Stoppel has demonstrated a method of tampering with explanations in order to conceal an adversarial attack on images. By this method, the explanations are modified in such a way that they do not seem questionable. In their work, the author underlines the crucial role of the so-called adversarial fine-tuning, contributing to the method being successful, as it not only helps keep the classification performance in the context of original images and ensures consistent misclassification of the original images but also enables making the adversarial explanations resemble the original ones [27].

3 METHODS

3.1 Random Forest

The Random Forest (RF) Classifier was chosen for its robustness and effectiveness in classification tasks in network intrusion detection [13].

RF, introduced by [4] and [12], combines multiple tree predictors. Each tree in the ensemble is created by randomly selecting a small set of inputs, followed by determining the optimal way to split the data to minimise information entropy. In this method, every tree is built from an independently sampled dataset, and their predictions are averaged, a technique also known as bootstrap aggregation [3]. The trees are grown to their full capacity using the Classification and Regression Tree (CART) methodology [5] and are not pruned back. The overall performance of a RF is determined by the strength of each individual tree, defined by its accuracy, and the diversity among the trees, which refers to how different the trees are from

$$C(x) = \arg\min \left(\frac{1}{k} \sum_{i=1}^{k} [\lambda_1 \times yloss(f(c_i), y) + dist(c_i, x)] - \lambda_2 \times dpp\_diversity(c_1, ..., c_k) \right) \tag{1}$$

one another. Breiman highlighted that the generalization error of a forest depends on these two factors.

Although the Random Forest algorithm may seem straightforward, its underlying complexity allows it to be recognized as "one of the best-performing learning algorithms" [23].

3.2 Diverse Counterfactual Explanations (DiCE)

DiCE [19] was employed to generate CEs, offering insights into the AI model's decision-making process. The use of DiCE is crucial for understanding how AI decisions can be interpreted and potentially manipulated. DiCE is a method for the creation of CEs, which takes a trained model f and an instance x as an input, and generates k CEs which lead to a different label than the classification of x by f. The method incorporates constraints for proximity and diversity, leading to finding CEs that are close to the original instance (proximity), lead to a different decision (via *yloss*), and are diverse among themselves (via *dpp_diversity*). The optimisation formula for this approach is expressed in Eq. 1.

3.3 Zeroth Order Optimisation

Alongside DiCE, Zeroth Order Optimization (ZOO) [17] was incorporated as a method for crafting evasion adversarial attacks. ZOO's ability to generate adversarial samples without requiring gradient information is used as a direct comparison to DiCE samples and the original unaltered samples, to better highlight the adversarial properties of CEs. Zeroth Order Optimization based black-box attack method does not require internal access to the classifier. Instead, the attack estimates the gradient for a small perturbation to the sample, measures how far the current sample is from flipping the label, and iteratively updates the sample one alteration at a time, focusing on updating important features, where alterations are more meaningful [8].

3.4 Conceptual overlap of Counterexamples and Adversarial Attacks

This study emphasises the observation also made independently in [16] and [25] that there is a conceptual overlap between CE which flipped the detection label and an adversarial attack, especially in the context of ML-based IDS. Both CE and adversarial attacks aim to change the classification label, at the same time minimising the change to the input. As a result, both methods provide information on how to achieve label change with minimum input alterations, which is critical from the standpoint of intrusion detection.

4 NOVEL PERSPECTIVE: THE EXPERIMENTS IN USING COUNTERFACTUAL SAMPLES AS AN ATTACK VECTOR

In the context of evaluating the ability of adversarial attack algorithms and explainability methods to flip the classification label from attack to benign in NIDS, this study focuses on correctly classified attack samples (True Positives). Correctly classified attack samples provide a clear baseline to measure the effectiveness of the adversarial and counterfactual methods. Since these samples are already being correctly identified as attacks by the NIDS, any change to a benign classification as a result of applying ZOO or DiCE indicates a successful manipulation. The goal is to understand the vulnerability of the NIDS to crafted False Negatives (i.e., attacks that are misclassified as benign). By starting with correctly classified attacks, the study directly measures how adversarial and counterfactual techniques can circumvent detection. Moreover, using a consistent set of correctly classified attack samples ensures that the experiment has a controlled starting point. This consistency is crucial for comparing the effectiveness of the ZOO and DiCE algorithms in altering the model's predictions. Finally, by using correctly classified attack samples, the study can quantitatively measure the success rate of the adversarial and counterfactual methods in terms of the proportion of attack labels that were flipped to benign.

4.1 Okiru Malware

The Okiru malware is an example of a malicious software variant specifically designed to target and infect devices that are part of the Internet of Things (IoT), such as medical devices employing ARC processors [29]. Okiru's advanced evasion techniques make it an ideal candidate to test the robustness of AI-based detection systems. It provides a real-world example of malware that employs methods specifically designed to evade detection, like obfuscation and polymorphism, and would be first in line to also use adversarial evasion to hide from AI-based detection.

4.2 Experimental Procedure

The experiment involved training the Random Forest (RF) Classifier on the preprocessed IoT-23 data of the Okiru malware attacks. The test set was pushed through the RF model and only the samples classified correctly as Okiru were selected. DiCE was used to generate CEs for the selected samples, and ZOO was used to craft adversarial samples. This is to test the model's vulnerability to manipulation and to assess the effectiveness of CEs in revealing model weaknesses.

The pipeline of the experiment has been presented in Fig. 1. The figure has two parts - the top row depicts a usual classification pipeline, the bottom shows how subjecting the attack samples to DiCE leads to a sample that has a flipped label, similarly to an adversarial attack.

4.3 Dataset

This study utilises the Aposemat IoT-23 dataset [11] captured in the Stratosphere Lab in the Czech Republic. The set, which contains labelled traces of IoT malware was selected as a realistic collection of attack and benign samples. Since this study focuses on making

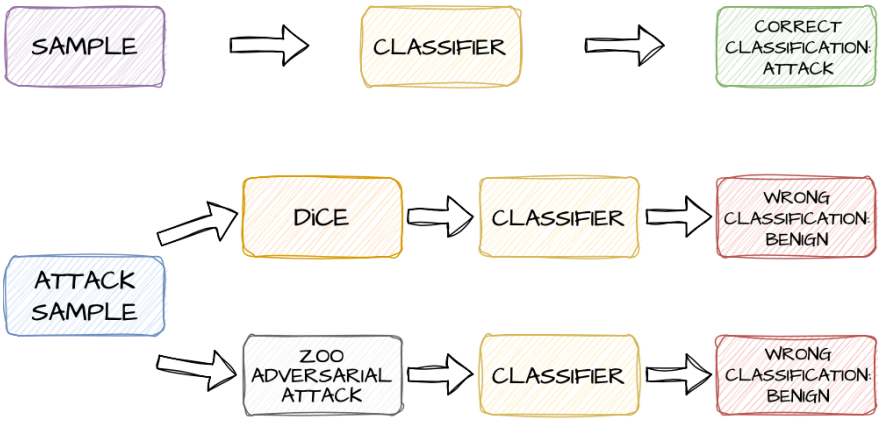

Figure 1: The pipeline of the experiment.

attacks undetectable by AI methods, a subset of the collection containing the Okiru attacks and the corresponding Benign traces were leveraged to complete the aims of the study.

4.4 Evaluation Metrics

In order to provide comprehensive understanding of the model's predictive performance and its resilience to adversarial attacks, a set of metrics were selected [14].

Namely, the model's performance was evaluated using the metrics of Precision (Eq. 2), Recall (Eq. 3), F1-Score (Eq. 4), and the attacks success rate in flipping the label is expressed in Eq. 5

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{4}$$

$$\text{Success Rate} = \frac{\text{Number of Successful Label Flips}}{\text{Total Number of Samples}} \times 100\% \tag{5}$$

In the equations, TP stands for True Positive and TN - True Negative. Similarly, FP and FN mean False Positives and False Negatives, respectively.

5 RESULTS

Figures 2, 3 and 4 present a visual comparison of the three kinds of samples: the blue bars present the values of the CE samples, the green bars are of the adversarial attacks, the red bars indicate the values of the features of the original test samples. The figures emphasize the changes in the feature values necessary to flip the label from Okiru to Benign. In Figure 2, the first feature is mostly affected, with CE sample having a larger effect than the adversarial attack. In Figure 3, the CE did not affect the first feature, only the second one, in contrast to the adversarial attack, which only affected the first feature. Similarly, in Figure 4, the CE did not affect the first feature, only the second. The adversarial attack focused on small alterations to the first and the fifth feature. Tables 1, 2 relay the classification results of the RF on the entire test set samples after the samples were treated with either CE or ZOO. Since the final test set only contains samples that were correctly identified as Okiru, the tables only contain the metrics for Okiru detection. For Table 1, which contains the classification results on the CEs, the Recall is 0.58. This means that over 40% of the samples were successfully flipped with minimal alterations. This is explicitly stated in Table 3, with the 42% success rate of the counterfactuals. While the CEs are not as effective as adversarial attacks, a significant portion of the samples can be successfully flipped.

As a result, the information gained from flipping labels via CE can be leveraged to build malware which avoids certain patterns of network traffic, circumventing detection. It is crucial to point out that while until recently in most cases to use adversarial attacks one would need to have access to the model, or be able to steal the model [9]. Yet, with CE, the key to crafting samples that circumvent detection could be provided by the vendor.

Table 1: Classification Report for the CEs (with no benign samples in the test set only Okiru detection is reported)

Class	Precision	Recall	F1-Score
Okiru	1.00	0.58	0.74

Table 2: Classification Report for adversarial attack

Class	Precision	Recall	F1-Score
Okiru	1.00	0.01	0.02

Table 3: Comparison of Attack Type Success Rates

Kind of Attack	Success Rate
Counterfactuals	42%
Adversarials	99%

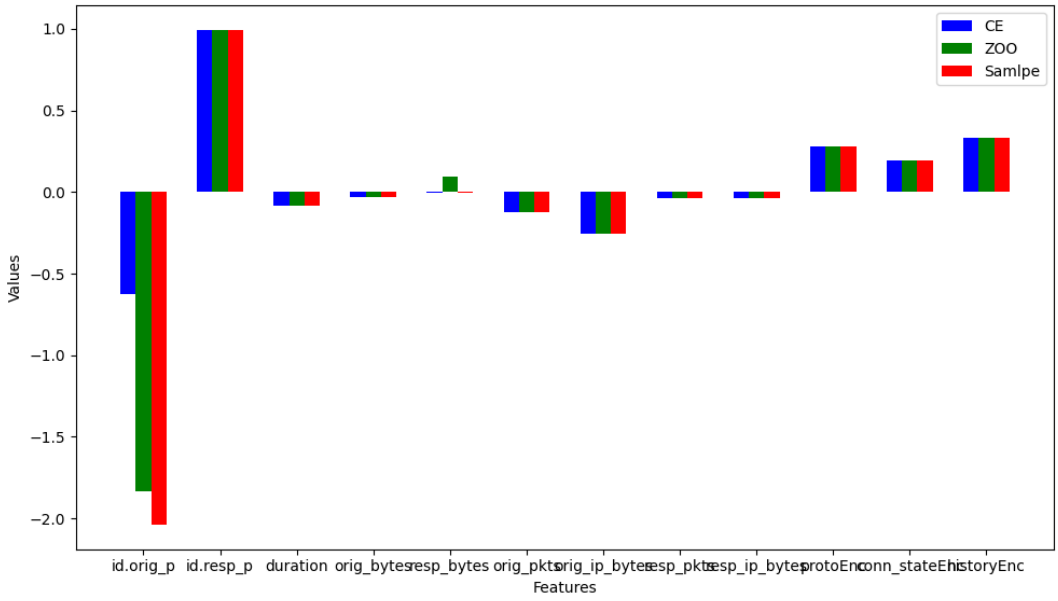

Figure 2: The relative changes in sample features compared between the ZOO adversarial attack (green), the CE (blue) and the Original Sample. The x-axis contains the features, the y-axis shows the standardised value of the features.

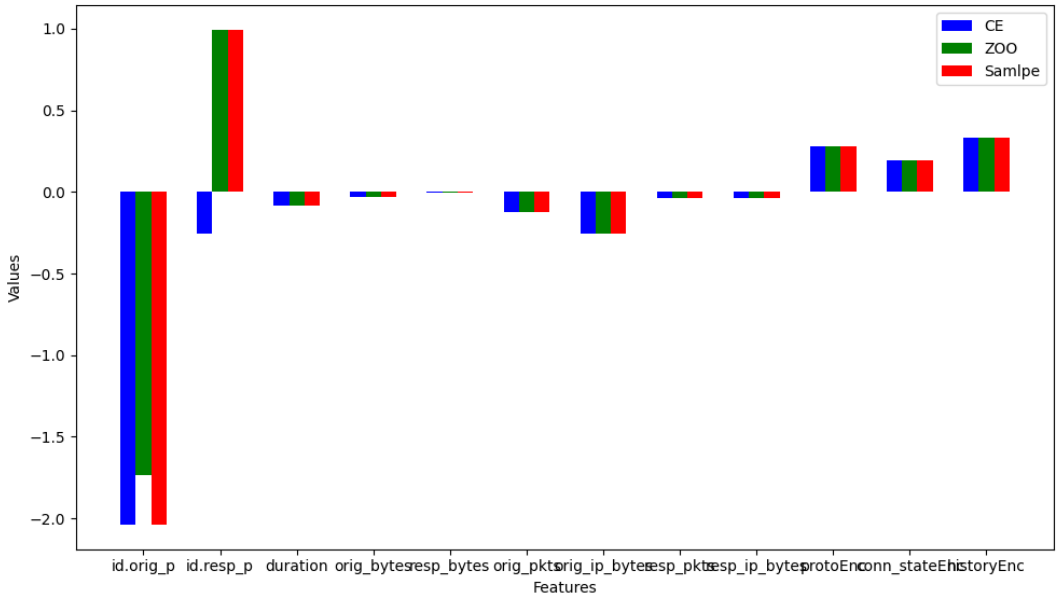

Figure 3: The relative changes in sample features compared between the ZOO adversarial attack (green), the CE (blue) and the Original Sample. The x-axis contains the features, the y-axis shows the standardised value of the features.

6 CONCLUSIONS

DiCE, an xAI tool used for generating counterfactual explanations, plays a crucial role in unravelling the decision-making processes of AI models. By creating alternative scenarios that could lead to different classification outcomes, DiCE helps in understanding how minimal changes in input data manipulate AI decisions.

This study proposes an observation of the conceptual overlap between CE and adversarial attacks. Both strategies aim to alter the classification labels while making minimal changes to the input data. This similarity is crucial as it indicates that methods developed for xAI, like DiCE, can inadvertently reveal techniques for crafting

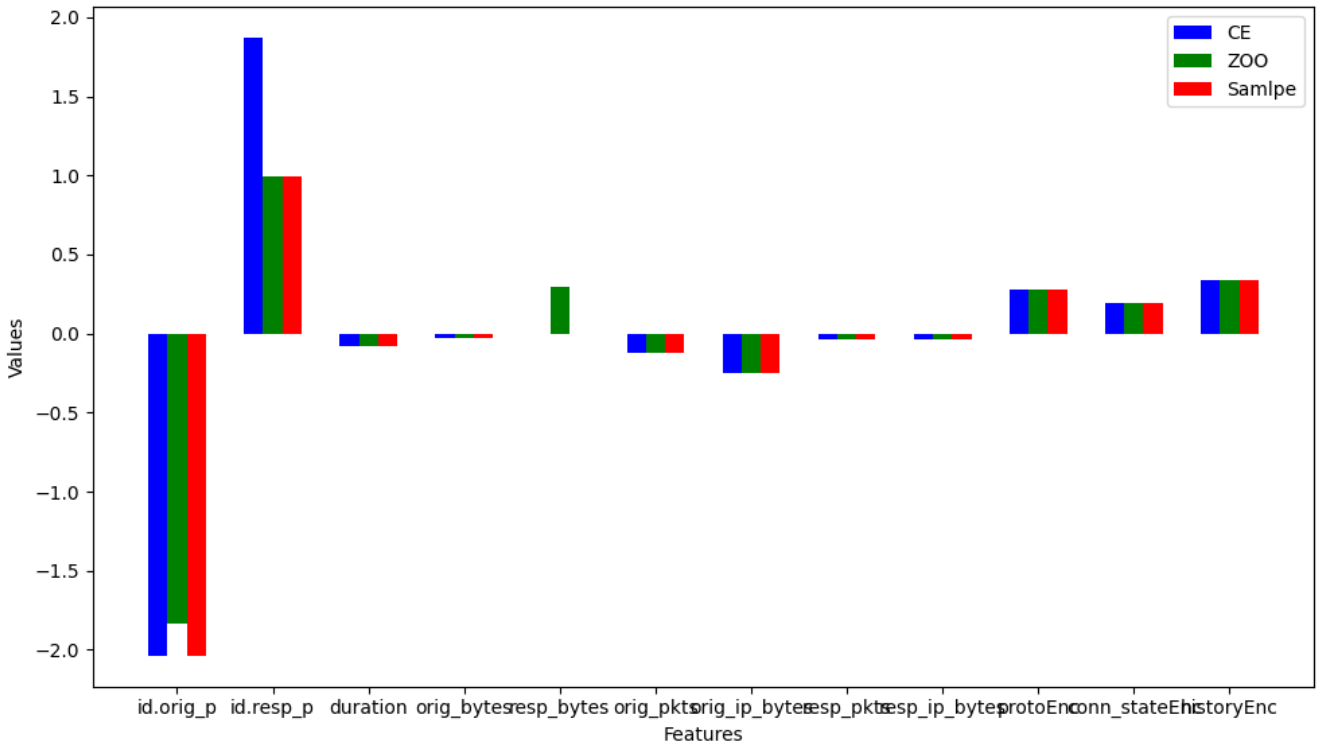

Figure 4: The relative changes in sample features compared between the ZOO adversarial attack (green), the CE (blue) and the Original Sample. The x-axis contains the features, the y-axis shows the standardised value of the features.

successful adversarial attacks, which is a critical observation for security applications.

This way, the Research Question: "Can CEs be Utilized to Formulate an Adversarial Attack against AI-based Network Intrusion Detection?" has been answered positively.

The study explored the use of counterfactual samples as potential vectors for launching attacks on NIDS. By focusing on samples that were initially correctly identified as attacks, the research demonstrated how applying counterfactual and adversarial techniques could lead to false negatives.

The experimental involved training of the RF Classifier on the IoT-23 dataset, specifically on data pertaining to the Okiru malware attacks.

The results indicated that while counterfactual methods were not as effective as adversarial attacks in flipping detection labels, they still posed a significant threat. Over 40% of the samples could be manipulated to flip their labels, demonstrating the potential of these techniques in evading detection systems.

These findings have profound implications for the AI applications in security. They highlight the need for these systems to adapt and become more resilient against the tactics and a cautionary approach towards the use of explainability tools like DiCE, as they could inadvertently provide blueprints for creating undetectable malware.

With the current proliferation of AI, xAI techniques are necessary from the standpoint of ethics and trustworthiness of AI deployments. Interpretable AI is crucial for critical deployments, for example in police applications. Law enforcement agencies can enhance their investigative methods by leveraging their data-rich environments with AI tools. The H2020 STARLIGHT project fosters robust use of AI in tackling major criminal threats. STARLIGHT brings together 50 partners from 18 European countries with 15 law enforcement agencies. As brought to attention in this paper, xAI methods, which are required for trustworthy AI, can be leveraged to harm the robustness of AI. This notion will be further explored and addressed in the project.

7 FUTURE DIRECTIONS

While the domain of xAI continues to expand, developing novel techniques to detect when an xAI output might be used for adversarial purposes and alerting the system will become more pressing. Thus, in the future, we plan to design and develop methods to counter the very effect described in this paper. We will focus on tailoring secure xAI solutions for use in critical infrastructure sectors, such as cybersecurity, healthcare, fake news detection and law enforcement applications, where security is paramount.

ACKNOWLEDGEMENTS

This research is funded under the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats"), which has received funding from the European

Union's Horizon 2020 research and innovation programme under grant agreement No 101021797.

REFERENCES

[1] Malik AL-Essa, Giuseppina Andresini, Annalisa Appice, and Donato Malerba. 2022. XAI to Explore Robustness of Features in Adversarial Training for Cybersecurity. 117–126. https://doi.org/10.1007/978-3-031-16564-1_12

[2] Hubert Baniecki and Przemyslaw Biecek. 2023. Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A Survey. (jun 2023). https://doi.org/2306.06123v2 arXiv:2306.06123

[3] Gérard Biau. 2012. Analysis of a Random Forests Model. *J. Mach. Learn. Res.* 13, null (apr 2012), 1063–1095.

[4] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[5] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. 1984. *Classification and Regression Trees*. Taylor & Francis. https://books.google.pl/books?id=JwQx-WOmSyQC

[6] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Claudio Stanzione. 2022. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access* 10 (2022), 93575–93600. https://doi.org/10.1109/ACCESS.2022.3204171

[7] Fabien Charmet, Harry Chandra Tanuwidjaja, Solayman Ayoubi, Pierre-François Gimenez, Yufei Han, Houda Jmila, Gregory Blanc, Takeshi Takahashi, and Zonghua Zhang. 2022. Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications* 77, 11-12 (dec 2022), 789–812. https://doi.org/10.1007/s12243-022-00926-7

[8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 15–26.

[9] Michał Choraś, Marek Pawlicki, and Rafał Kozik. 2019. The feasibility of deep learning use for adversarial model extraction in the cybersecurity domain. In *Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part II 20*. Springer, 353–360.

[10] Michał Choraś, Marek Pawlicki, Damian Puchalski, and Rafał Kozik. 2020. Machine Learning–the results are not the only thing that matters! What about security, explainability and fairness?. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*. Springer, 615–628.

[11] Sebastian Garcia, Agustin Parmisano, and Maria Jose Erquiaga. 2020. IoT-23: A labeled dataset with malicious and benign IoT network traffic. *Stratosphere Lab., Praha, Czech Republic, Tech. Rep* (2020).

[12] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

[13] Mikolaj Komisarek, Marek Pawlicki, Rafal Kozik, and Michal Choras. 2021. Machine Learning Based Approach to Anomaly and Cyberattack Detection in Streamed Network Traffic Data. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.* 12, 1 (2021), 3–19.

[14] Mikołaj Komisarek, Marek Pawlicki, Maria-Elena Mihailescu, Darius Mihai, Mihai Carabas, Rafał Kozik, and Michał Choraś. 2022. A novel, refined dataset for real-time Network Intrusion Detection. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3538969.3544486

[15] Rafał Kozik, Massimo Ficco, Aleksandra Pawlicka, Marek Pawlicki, Francesco Palmieri, and Michał Choraś. 2023. When explainability turns into a threat - using xAI to fool a fake news detection method. *Computers Security* (nov 2023), 103599. https://doi.org/10.1016/j.cose.2023.103599

[16] Aditya Kuppa and Nhien-An Le-Khac. 2021. Adversarial XAI Methods in Cybersecurity. *IEEE Transactions on Information Forensics and Security* 16 (2021), 4924–4938. https://doi.org/10.1109/TIFS.2021.3117075

[17] Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. 2020. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine* 37, 5 (2020), 43–54.

[18] Carlos Mendes and Tatiane Nogueira Rios. 2023. Explainable Artificial Intelligence and Cybersecurity: A Systematic Literature Review. (feb 2023). https://doi.org/arXiv:2303.01259v1 arXiv:2303.01259

[19] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[20] Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. 2023. On the Privacy Risks of Algorithmic Recourse. *Proceedings of Machine Learning Research* 206 (nov 2023), 9680–9696. arXiv:2211.05427 http://arxiv.org/abs/2211.05427

[21] Aleksandra Pawlicka, Michał Choraś, and Marek Pawlicki. 2021. The stray sheep of cyberspace aka the actors who claim they break the law for the greater good. *Personal and Ubiquitous Computing* 25, 5 (2021), 843–852.

[22] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. 2021. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–36.

[23] Matthias Schonlau and Rosie Yuyan Zou. 2020. The random forest algorithm for statistical learning. *The Stata Journal* 20, 1 (2020), 3–29. https://doi.org/10.1177/1536867X20909688 arXiv:https://doi.org/10.1177/1536867X20909688

[24] Reza Shokri, Martin Strobel, and Yair Zick. 2021. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 231–241.

[25] Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual Explanations Can Be Manipulated. (jun 2021). https://doi.org/2106.02666v2 arXiv:2106.02666

[26] Gautam Srivastava, Rutvij H Jhaveri, Sweta Bhattacharya, Sharnil Pandya, Rajeswari, Praveen Kumar Reddy Maddikunta, Gokul Yenduri, Jon G. Hall, Mamoun Alazab, and Thippa Reddy Gadekallu. 2022. XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions. (jun 2022). arXiv:2206.03585 http://arxiv.org/abs/2206.03585

[27] Stefanie Stoppel. 2022. "Wasn't Me" or How to Hide Adversarial Attacks Using Explainable AI. *Inovex* (2022).

[28] Marco Virgolin and Saverio Fracaros. 2022. On the Robustness of Sparse Counterfactual Explanations to Adverse Perturbations. (jan 2022). https://doi.org/arXiv:2201.09051 arXiv:2201.09051

[29] Yao Xu, Hiroshi Koide, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2018. Tracing MIRAI malware in networked system. In *2018 sixth international symposium on computing and networking workshops (CANDARW)*. IEEE, 534–538.

[30] Shuai Zhou, Chi Liu, Dayong Ye, Tianqing Zhu, Wanlei Zhou, and Philip S Yu. 2022. Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *Comput. Surveys* 55, 8 (2022), 1–39.