Study on Adversarial Attacks Techniques, Learning Methods and Countermeasures - Application to Anomaly Detection

Anis Bouaziz[1][a], Manh-Dung Nguyen[1][b], Valeria Valdés[1][c],
Ana Rosa Cavalli[1,2][d] and Wissam Mallouli[1][e]
[1]*Montimage EURL, 39 rue Bobillot 75013, Paris, France*
[2]*Institut Telecom SudParis, 5 rue Charles Fourrier 91011 Evry, France*
{*firstname.lastname*}*@montimage.com*

Keywords: Cybersecurity, Artificial Intelligence, Adversarial attacks, Explainability, Countermeasures

Abstract: Adversarial attacks on AI systems are designed to exploit vulnerabilities in the AI algorithms that can be used to manipulate the output of the system, resulting in incorrect or harmful behavior. They can take many forms, including manipulating input data, exploiting weaknesses in the AI model, and poisoning the training samples used to develop the AI model. In this paper, we study different types of adversarial attacks, including evasion, poisoning, and inference attacks, and their impact on AI-based systems from different fields. A particular emphasis is placed on cybersecurity applications, such as Intrusion Detection System (IDS) and anomaly detection. We also depict different learning methods that allow us to understand how adversarial attacks work using eXplainable AI (XAI). In addition, we discuss the current state-of-the-art techniques for detecting and defending against adversarial attacks, including adversarial training, input sanitization, and anomaly detection. Furthermore, we present a comprehensive analysis of the effectiveness of different defense mechanisms against different types of adversarial attacks. Overall, this study provides a comprehensive overview of challenges and opportunities in the field of adversarial machine learning, and serves as a valuable resource for researchers, practitioners, and policymakers working on AI security and robustness. An application for anomaly detection, especially malware detection is presented to illustrate several concepts presented in the paper.

1 INTRODUCTION

Artificial Intelligence (AI) has a wide range of applications in computer science (Jan et al., 2023; Apruzzese et al., 2022), from data mining, natural language processing, optimization and decision making, to robotics and cybersecurity. While they offer many benefits and opportunities, they also present a number of challenges and risks (Parnas, 2017) including bias and discrimination, lack of transparency, security risks, dependence on data and even ethical concerns.

The security risks are mainly due to the fact that AI models are vulnerable to adversarial attacks (Long et al., 2022), where attackers manipulate inputs in order to induce the model to misbehave or provide incorrect outputs. Adversarial machine learning (AML) is a field of research that focuses on studying how machine learning models can be manipulated or attacked by adversaries with the goal of causing misclassification or other harmful outcomes. Adversarial attacks involve intentionally creating inputs specifically designed to cause unexpected outputs from a machine learning model. Adversarial machine learning (Wang et al., 2019) is a significant challenge for the security and reliability of machine learning models, particularly in high-stakes applications such as autonomous vehicles, medical diagnosis, and financial fraud detection. Researchers are engaged in developing techniques to reduce the impact of adversarial attacks(Alotaibi and Rassam, 2023). This involves developing more robust machine learning models, designing algorithms that are more resilient to adversarial attacks, and implementing more effective detection and mitigation strategies.

The use of AI to conduct cyber attacks is a new trending challenge for security experts. While AI has been used extensively in cybersecurity to detect and prevent attacks, attackers can also use AI to evade tra-

[a] https://orcid.org/0009-0000-3429-3540
[b] https://orcid.org/0000-0001-8760-3258
[c] https://orcid.org/0009-0000-6632-396X
[d] https://orcid.org/0000-0003-2586-9071
[e] https://orcid.org/0000-0003-2548-6628

ditional security measures and conduct more sophisticated attacks (Alotaibi and Rassam, 2023). For this reason, more sophisticated solutions are studied to build more robust AI-based systems resilient to these kinds of attacks. The topic of adversarial machine learning has been extensively studied in computer vision literature, exploring different attack techniques (Akhtar et al., 2021). Surveys have revealed the effectiveness of gradient-based methods against deep neural networks(Akhtar et al., 2021) and the relative robustness of decision trees (Papernot et al., 2017). These surveys emphasize the importance of understanding attack effectiveness and identifying defense mechanisms (Qiu et al., 2022).

An example of a technique that can be employed to alleviate the consequences of adversarial attacks is eXplainable Artificial Intelligence (XAI) (Arrieta, 2020). XAI facilitates this mitigation process by offering enhanced transparency and understanding of the internal mechanisms of machine learning models. Through explainability techniques, users gain improved comprehension of the model's functioning, encompassing decision-making processes and the features on which those decisions are based. By increasing transparency, XAI aids in the identification of potential vulnerabilities and weaknesses within the model that adversaries could exploit.

This paper explores the various ways attackers can exploit the weaknesses of AI/ML algorithms and tools for performing adversarial targeted and non-targeted attacks.A particular emphasis is placed on AI-based Intrusion Detection Systems (IDSs) (Habeeb and Babu, 2022). Indeed, adversarial attacks against IDSs refer to the deliberate manipulation by an attacker of network traffic or system behavior to evade detection by IDS. These attacks are designed to exploit the weaknesses of the IDS by introducing subtle changes in the traffic that make it difficult for the system to accurately classify it as normal or malicious. An experimentation is also presented to attack an industrial open-source IDS called "ACAS"[1].

In the context of network anomaly detection, adversarial ML attacks can be carried out in several ways like poisoning attacks, evasion attacks, generative adversarial attacks, etc. In this paper, we propose using different security mechanisms and different learning methods to improve the reliability and resilience of AI-based systems. To mitigate adversarial attacks, network anomaly detectors can employ various techniques such as ensemble learning, input sanitization, and model retraining. Ensemble learning is combining anomaly detection models to improve accuracy and robustness. Input sanitization involves removing or normalizing features that are vulnerable to manipulation. Model retraining would be continuously updating the model with new data to keep up with the evolving threat landscape.

In particular, XAI, or explainable AI, is used to comprehend, identify and defend against adversarial attacks. One approach of employing XAI for adversarial attacks is to develop visualizations that assist in highlighting the regions of the input data that are most vulnerable to changes or are most likely to be altered by attackers. This can help to identify potential vulnerabilities in the model and inform the development of more effective defense mechanisms.

The paper is organized as follows. In section 2, we present our study on adversarial attacks against machine learning relevant to AI-based intrusion detection systems. Section 3 depicts the possible countermeasures that we can apply to mitigate these attacks. And finally, Section 4 presents an illustration of all these technical concepts in the context of our AI-based framework for anomaly detection with explainability and robustness. Section 5 concludes the paper and discusses our future work.

2 ADVERSARIAL ATTACKS IN MACHINE LEARNING

Adversarial machine learning (AML) is a crucial concept to examine when building ML algorithms, aiming to deceive or degrade the performance of ML models on specific tasks. AML involves a game between ML algorithms and adversaries (Dasgupta and Collins, 2019), where the machine learns to correctly classify new data, while the adversary attempts to alter the existing data, new data, or the machine's parameters to cause misclassification. AML attacks are constructed by introducing imperceptible perturbations to the input samples of the model, resulting in misclassification with high confidence while preserving the primary *functionality* of the samples. In cybersecurity, adversarial attacks pose a critical challenge, various approaches have been proposed to defend against them (Rosenberg et al., 2021; Wang et al., 2020). Adversarial attacks can be categorized as attacks in the training stage and attacks in the testing stage (Qiu et al., 2019). Concretely, poisoning attacks (Barreno et al., 2006) occur during the training stage, while evasion attacks (Szegedy et al., 2014) occur during the testing stage. Various frameworks and algorithm-independent techniques have been proposed for conducting theses attacks against a variety of AI algorithms and datasets (Jagielski et al., 2021; Mozaffari-Kermani et al., 2015). For deep learning

[1] https://github.com/Montimage/acas

(DL) algorithms, back-gradient optimization is used to reduce attack complexity (Muñoz-González et al., 2017), and backdoor attacks (Dai and Chen, 2019) have been implemented against long short-term memory networks based text classification.

2.1 Threat model and adversarial capabilities

The threat model is a concept that defines the attacker's objectives, information gathering, and attack steps. It enables systematic and theoretical analysis of adversarial machine learning. The common threat model taxonomy for machine learning systems characterizes attacks using three dimensions (Barreno et al., 2006): influence, specificity, and security violation. *Influence* defines the attacker's control level over the training process (causative or exploratory). *Specificity* measures the degree of precision with which the attacker's target is specified (targeted, indiscriminate). *Security violation* refers to the attacker's aim to induce misclassifications (integrity or availability attack) (Shi and Sagduyu, 2017).

Adversarial capabilities represent the attacker's ability to access and use the greatest amount of information about the target model. These capabilities depend entirely on whether the model is attacked in the training or testing phase. During the training phase, the adversary can manipulate the model's logic (logic corruption), alter the dataset before the training (data modification), or insert new adversarial samples into the training dataset (data injection) (Nawaz et al., 2018). During the testing stage, adversarial assaults attempt to persuade the targeted model to produce false outputs. The effectiveness and specificity of these attacks are primarily determined by the knowledge of the model available to the adversary (Qiu et al., 2019).

A *white-box* attack scenario refers to the attacker having complete knowledge of the target model, data distribution, and model parameters, which makes it easier to identify the most vulnerable feature space of the target model. Conversely, in *black-box* attacks, the attacker has limited access to the target model (Barreno et al., 2006). *Adaptive black-box* attacks enable the adversary to query the model to obtain corresponding labels, use this information to label a carefully selected dataset, train a new model, and then use white-box attacks to defeat the original model. In contrast, in *non-adaptive black-box* attacks, the adversary can only access the model's training data distribution (Chang Xiao, 2019). *Strict black-box* attacks involve collecting pairs of input-output values from the target classifier without changing the inputs (Xu et al., 2021), similar to a known-plaintext attack in cryptography. Hence, this attack requires a large set of input-output data.

2.2 Adversarial Learning Methods

Adversarial machine learning approaches rely on direction sensitivity estimation (Qiu et al., 2019) to identify the dimensions of X that provide expected adversarial performance with minimal perturbation. A variety of theoretical adversarial learning approaches can be used to generate adversary samples. These techniques differ in complexity, performance, type of data and generation rate. AML approaches are based on the notion that when adding relatively insignificant perturbation δ to an original sample X, the crafted sample X^* can express adversarial properties (Rosenberg et al., 2021). The generated adversarial sample will be categorized differently by the target model. Table 1 shows a summary of common non-deep learning methods used for attacking machine learning models, including gradient-based, and score-based attack (Rosenberg et al., 2021). *Gradient-based* attacks (e.g., FGSM) require knowledge of the target classifier's architecture, gradients and thus are considered white-box attacks (Muñoz-González et al., 2017). *Score-based* attacks (ZOO) rely on the confidence score of the target classifier. This section focuses on black-box decision-based attacks that exclusively rely on the predicted label. Such attacks are particularly significant in real-world scenarios where targeted models are usually shielded from any direct access. Within this context, we explore GAN-based attacks, which demonstrate remarkable efficacy due to their ability to generate adversarial examples that are both realistic and diverse. These attacks pose a considerable challenge for conventional defense mechanisms, making them highly promising for future advancements in adversarial attack strategies.

GAN-based attacks Generative Adversarial Networks (GANs), first developed in 2014 (Goodfellow et al., 2014), involve a game between two linked neural networks, a generator and a discriminator, that compete to learn from each other's experience to reach Nash equilibrium. The generator creates adversarial examples that are extremely similar to the original set, whereas the discriminator tries to differentiate the original sample from the fake/generated one. GANs have multiple applications including image and video synthesis, natural language processing, and drug discovery. However, GANs can also generate adversarial examples that pose security risks to machine learning models.

Table 1: Overview of the most commonly studied gradient-based adversarial attacks.

Attack	Description	Advantages / Limits
Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015)	One-step adversarial attack method that perturbs input feature by adding small noise in the sign of the gradient of the loss function	Fast and simple, generates easily detectable adversarial examples, unstable for large perturbations
Jacobian-based Saliency Maps Attacks (JSMA) (Papernot et al., 2016)	Iterative adversarial attack method that perturbs the input by changing the least salient features based on the Jacobian matrix of the model's output	Effective for targeted attacks that can handle non-differentiable models. Computationally expensive and may not find global optimum
Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)	Optimization algorithm that maintains an approximate Hessian matrix, used to minimize the amount of perturbations added to images (Szegedy et al., 2014)	Effective at generating adversarial examples, but are resource-demanding
Deepfool (Moosavi-Dezfooli et al., 2016)	Iterative method that estimates the minimum L2 perturbation required to cross a linear decision boundary by solving a linear system at each iteration	Fast and effective for low-dimensional data
C&W (Carlini and Wagner)(Carlini and Wagner, 2017)	Optimization-based attack method that finds the minimum perturbation using a custom loss function that balances misclassification and perturbation size	One of the most effective and widely used adversarial attacks but is resource intensive
Zeroth Order Optimization (ZOO) (Chen et al., 2017)	Model-agnostic and gradient-free attack method that iteratively approximates gradients using finite differences and performs local search to find adversarial examples	Effective for black-box attacks but requires many queries

The *generator* (G) in GANs finds utility in various contexts after being trained to a certain number of epochs. For instance, it can serve as an oversampler for data augmentation (He Zhang, 2019). This technique can be employed as a defensive mechanism (adversarial training) to enhance the detection performance of ML classifiers (Bai et al., 2021). However, GAN-generated adversarial examples can also be exploited by black-box attackers to generate evasion samples (Randhawa et al., 2022; Lin et al., 2022). Consequently, it is crucial to develop robust ML classifiers that can defend against adversarial evasion to protect AI-based systems (et al., 2021).

CTGAN (Conditional GAN for Tabular Data) is a type of generative model based on GANs that learns the joint distribution of tabular data and generates synthetic data that closely resembles the original data (Xu et al., 2019). CTGAN employs conditional GANs to model the conditional distribution of each column and solves non-Gaussian and multimodal distribution problems using a mode-specific normalization approach (Wang et al., 2022). It also addresses the issue of data imbalance in discrete columns through the use of conditional generators and training-by-sampling. CTGAN has many applications, including data augmentation, privacy-preserving data sharing, and model training with limited data (Stavroula Bourou and Zahariadis, 2021). In the field of cybersecurity, CTGAN has been used to generate synthetic network traffic data for anomaly or intrusion detection (Lin et al., 2022), or to create adversarial malware samples that bypass malware classifiers without compromising their damaging *functionality* (Hu and Tan, 2017).

Another generative model based on GANs is table-GAN (Park et al., 2018), which was designed to synthesize tabular data with categorical, discrete, or continuous values that closely resemble the original data. Table-GAN incorporates a classifier neural network to improve the semantic integrity of synthetic records and prevent information leaking, making it difficult to detect that the table is fabricated. Unlike other GANs that generate images or text, TableGAN generates data that can be organized into tables, making it with CTGAN useful for cybersecurity applications such as traffic generation and malware mutation (Stavroula Bourou and Zahariadis, 2021; Hu and Tan, 2017).

There are many other applications of GANs an in cybersecurity, including system robustness (et al., 2021; Iftikhar Rasheed, 2020), malware adaptation (Renjith G, 2022), phishing (Sern et al., 2020), or password guessing (Hitaj et al., 2019). However, in low-volume data regimes such as medical diagnostic imaging and security, more specific models are required due to the restricted number of anomalous samples. EVAGAN (Randhawa et al., 2022) a model proposed for low data regime challenges that leverage oversampling for detection enhancement of ML classifiers, can both generate evasion samples and operate as an evasion-aware classifier. GANs-generated adversarial samples have effectively been used in generating network attacks (Ahmed Aleroud, 2020), including SynGAN (Charlier et al., 2019) to generate malicious packet flow mutations using real attack traffic, and IDSGAN (Lin et al., 2022) for adversarial malicious traffic records formation with the aim of deceiving and evading intrusion detection systems.

Adversarial attacks threat remain significant to machine learning models, especially those used in critical applications such as autonomous vehicles (Iftikhar Rasheed, 2020) and medical diagnosis

(Mozaffari-Kermani et al., 2015). As discussed in this chapter, various types of adversarial attacks can exploit vulnerabilities in the model's inputs and lead to incorrect predictions. Next, we will present several defense mechanisms against such attacks.

3 COUNTERMEASURES

Adversarial attacks exploit vulnerabilities in machine learning models to bypass security measures. To protect models against such attacks, numerous countermeasures have been proposed in the literature (Qiu et al., 2019). One of the most effective techniques in the literature is the *adversarial training* (Madry et al., 2019). It involves training the model with adversarial examples to improve their robustness. However, the model may still be vulnerable to new attacks.*Gradient-masking* aimed at impeding attackers from estimating model gradients by building a model in which their gradients are useless. Nevertheless, models are still vulnerable to adversarial examples.

Defensive distillation technique uses a complex-primary model trained and the output is given to a smaller secondary model to learn the output function and probabilities. This approach is adaptable to unknown attacks and acts as an extra layer of protection. However, it stills vulnerable to poisoning attacks during the training (Qiu et al., 2022). While *Feature squeezing* compresses input features of the models to improve security, particularly in image processing by changing the color depth or adding blur to the image(Xu et al., 2018). While its application to other contexts decrease their effectiveness, (Rosenberg et al., 2021) consider that feature squeezing can be applied. To block *transferability* or black-box attacks the authors in (Hosseini et al., 2017) present a defense mechanism that adds a new Null label indicating that the example is from an adversary and discard it instead of classifying it to the original label.

Defenses against *universal perturbation attacks* include adversarial training and distillation. Other defenses train a network to extract features of adversarial examples and gives the output to another network to identify adversarial examples. The authors in (Akhtar et al., 2018) propose a defense against universal perturbations, where from clean data, image-agnostic adversarial examples are generated and then given to a perturbation rectifying network (PRN). The perturbation detection network extracts features from the differences between the inputs and the outputs of the PRN and outputs a binary classifier.

Next, we will discuss explainable AI for adversarial attack detection. This approach offers the potential to mitigate the limitations of the aforementioned countermeasures by enhancing the interpretability of machine learning models.

4 FRAMEWORK FOR ANOMALY DETECTION WITH XAI AND ROBUSTNESS

4.1 XAI and adversarial attacks

Explainability (XAI) and adversarial attacks are both critical aspects to consider in the development of machine learning models. XAI (Arrieta, 2020) is a promising set of technologies that increases the AI black-box models' transparency to explain why certain decisions were made. XAI is crucial to enhance trust for people to use future AI-based applications by offering understandable explanations for its predictions and decisions. Some popular post-hoc explainability methods are *visual explanations*, *LIME local explanations* (Ribeiro et al., 2016), *explanations by example*, and *SHAP feature relevance explanations* (Lundberg and Lee, 2017). Adversarial attacks, on the other hand, involve intentionally manipulating input data to misguide a machine learning model into making erroneous predictions.

Using XAI to detect attacks XAI can help in detecting adversarial attacks in machine learning models. The authors of (Malik et al., 2022) propose a three-stage approach to adversarial training improve the detection of cyber threats in network traffic data. They first use GAN to produce adversarial examples, apply XAI techniques to generate explanations, and then retrain the detection model using those explanations. Furthermore, SHAP signatures are useful to detect adversarial examples by comparing the signatures of the original input and the perturbed input (Fidel et al., 2020). If the signatures differ significantly, the input is classified as adversarial.

Adversarial attacks against XAI Despite XAI methods providing transparency and explainability to AI systems to prevent cyber threats, the current XAI models are at risk due to the possibility of being fooled by cyber attackers. The explanations generated by some of the most popular XAI explanation methods (Slack et al., 2020), such as LIME and SHAP, have been shown to be counter-intuitive. Therefore, defensive approaches should focus more on protecting the explanation results of XAI-based systems, in addition to their prediction results.

Tradeoff The tradeoff between XAI and adversarial

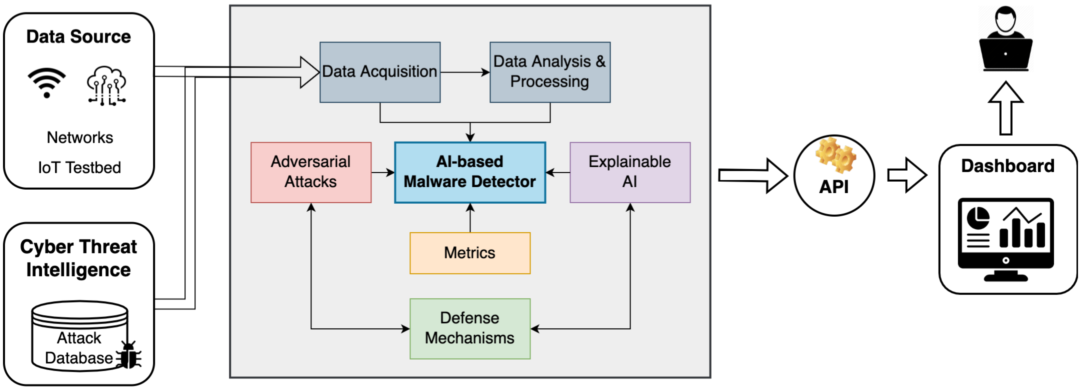

Figure 1: Architecture of our AI-based malware detector.

attacks is complex and multifaceted. On one hand, increasing transparency and interpretability of a model can expose it to vulnerabilities from adversarial attacks. An attacker can use the explanations provided by a XAI system to identify weaknesses in the model and craft more effective adversarial examples. For example, for NN-based models, an attacker could use the explanations provided by XAI to identify the neurons that are most sensitive to changes in input data. On the other hand, if a model is not transparent or interpretable, it can be challenging to not only comprehend why the model made a particular prediction or decision, but also to detect and diagnose adversarial attacks. Therefore, it is important to strike a balance between XAI and adversarial attacks when developing secure and robust machine learning models.

4.2 Architecture

In this section, we propose an AI-based framework for anomaly detection in encrypted traffic with high performance, explanation and robustness against adversarial attacks. Figure 1 shows the architecture of our anomaly detection framework.

Data acquisition module collects raw traffic data from networks or IoT testbed in either online or offline mode. It can also use Cyber Threat Intelligence (CTI) sources, e.g., deployed honeypots, to learn and continuously train our model using attack patterns and past anomaly information in the database.

Data analysis & processing module employs the open source Montimage monitoring tool[2] (MMT) to parse a wide range of network protocols (e.g., TCP, UDP, HTTP, and more than 700) and extract flow-based features. In particular, we extract 59 features, including basic features in packet headers and statistical features after performance traffic aggregation into flows. Finally, the restructured and computed data is transformed into a numeric vector so that can be easily processed by our AI model.

AI-based malware detector module is responsible for creating and utilizing a ML model able to classify the vectorized form of network traffic data. Our proposed approach involves a hybrid model that combines two DL techniques: Stacked Autoencoders (SAE) and Convolutional Neural Networks (CNN). We first train two SAEs, one for each class of data (normal or malware). Each SAE is designed with one hidden dense layer and trained separately with their respective data. The output of each SAE is then concatenated to form a single vector, which is then passed as input to the CNN. The CNN structure is based on the well-known VGG16 model (Simonyan and Zisserman, 2014), which consists of three repeated segments that are built from two convolutional layers (Conv1D) followed by a MaxPool layer (MaxPooling1D). After three blocks of such a structure, a flatten layer (Flatten) followed by two dense layers (Dense) are used in order to provide the final classification.

Adversarial attacks module injects various evasion and poisoning adversarial attacks for robustness analysis of our system. As we use a complex hard-to-interpret DL model with a large number of features for malware detection purposes, *interpretability* is crucial to earn trust of its end user. *Explainable AI* module aim at producing post-hoc global and local explanations of predictions of our model. As we need to consider the tradeoff between explainability, robustness and performance of our system, *Metrics* module allows to measure quantifiable metrics for its accountability and resilience. Finally, *Defense mechanisms* module provides countermeasures to prevent attacks against both AI and XAI models.

We designed our anomaly detection framework to be easily accessible for users or developers through a *Dashboard*. It provides a range of ML services, in-

[2] https://github.com/Montimage/mmt-probe

cluding extract features, build or retrain the model, inject adversarial attacks, produce explanations and evaluate our model using different metrics. Each of these services is exposed through dedicated *APIs* that can be accessed through the server, making it easy to integrate with other applications and systems.

5 CONCLUSION & FUTURE WORK

This paper presented a comprehensive study of various adversarial ML techniques and their significance in designing attacks and countermeasures. We explored the role of explainability as a countermeasure technique, highlighting its potential to enhance transparency and user trust in AI-based applications. Specifically, we discussed its application to an AI-based framework for malware detection in encrypted traffic, showcasing its effectiveness in providing explanations and robustness against adversarial attacks.

For future work, firstly, it is crucial to explore and implement additional defense mechanisms to safeguard our model against potential attacks from adversarial ML and XAI models. Secondly, extending the framework to different use cases, such as user network classification in the context of 5G networks, would greatly expand its practical utility. These directions hold immense potential for further customized techniques and advancements in the field of AI and XAI, ultimately leading to more trustworthy and efficient AI-based solutions.

Acknowledgements This work has been funded by the European Union's H2020 Programme under grant agreement no. 101070450 for the AI4Cyber project and no. 883540 for the PUZZLE project.

REFERENCES

Ahmed Aleroud, G. K. (2020). Sdn-gan: Generative adversarial deep nns for synthesizing cyber attacks on software defined networks.

Akhtar, N., Liu, J., and Mian, A. (2018). Defense against universal adversarial perturbations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Akhtar, N., Mian, A., Kardan, N., and Shah, M. (2021). Threat of adversarial attacks on deep learning in computer vision: Survey II.

Alotaibi, A. and Rassam, M. A. (2023). Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet*.

Apruzzese, G., Laskov, P., de Oca, E. M., Mallouli, W., Rapa, L. B., Grammatopoulos, A. V., and Franco, F. D. (2022). The role of machine learning in cybersecurity. *CoRR*.

Arrieta, A. B. e. a. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*.

Bai, T., Luo, J., Zhao, J., Wen, B., and Wang, Q. (2021). Recent advances in adversarial training for adversarial robustness.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., and Tygar, J. D. (2006). Can machine learning be secure? Association for Computing Machinery.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks.

Chang Xiao, Peilin Zhong, C. Z. (2019). Enhancing adversarial defense by k-winners-take-all. 29:7–29.

Charlier, J., Singh, A., Ormazabal, G., State, R., and Schulzrinne, H. (2019). Syngan: Towards generating synthetic network attacks using gans.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. (2017). ZOO. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM.

Dai, J. and Chen, C. (2019). A backdoor attack against lstm-based text classification systems.

Dasgupta, P. and Collins, J. (2019). A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks. *AI Magazine*.

et al., X. Q. (2021). Strengthening ids against evasion attacks with gan-based adversarial samples in sdn-enabled network. In *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*.

Fidel, G., Bitton, R., and Shabtai, A. (2020). When explainability meets adversarial learning: Detecting adversarial examples using shap signatures.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples.

Habeeb, M. S. and Babu, T. R. (2022). Network intrusion detection system: A survey on artificial intelligence-based techniques. *Expert Syst. J. Knowl. Eng.*

He Zhang, Xingrui Yu, P. R.-C. L. G. M. (2019). Deep adversarial learning in intrusion detection: A data augmentation enhanced framework.

Hitaj, B., Gasti, P., Ateniese, G., and Perez-Cruz, F. (2019). Passgan: A deep learning approach for password guessing.

Hosseini, H., Chen, Y., Kannan, S., Zhang, B., and Poovendran, R. (2017). Blocking transferability of adversarial examples in black-box learning systems.

Hu, W. and Tan, Y. (2017). Generating adversarial malware examples for black-box attacks based on gan.

Iftikhar Rasheed, Fei Hu, L. Z. (2020). Deep reinforcement learning approach for autonomous vehicle systems for maintaining security and safety using lstm-gan.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., and Li, B. (2021). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning.

Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, M., and Kuusk, A. (2023). Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. *Expert Syst. Appl.*

Lin, Z., Shi, Y., and Xue, Z. (2022). IDSGAN: Generative adversarial networks for attack generation against intrusion detection. In *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing.

Long, T., Gao, Q., Xu, L., and Zhou, Z. (2022). A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions. *Comput. Secur.*

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks.

Malik, A.-E., Andresini, G., Appice, A., and Malerba, D. (2022). An xai-based adversarial training approach for cyber-threat detection. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing*.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks.

Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., and Jha, N. K. (2015). Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE Journal of Biomedical and Health Informatics*.

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., and Roli, F. (2017). Towards poisoning of deep learning algorithms with back-gradient optimization.

Nawaz, R., Shahid, M. A., Qureshi, I. M., and Mehmood, M. H. (2018). Machine learning based false data injection in smart grid. In *2018 1st International Conference on Power, Energy and Smart Grid (ICPESG)*.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroSP)*.

Papernot, N., McDaniel, P. D., Goodfellow, I. J., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*.

Parnas, D. L. (2017). The real risks of artificial intelligence. *Commun. ACM*.

Qiu, S., Liu, Q., Zhou, S., and Huang, W. (2022). Adversarial attack and defense technologies in natural language processing: A survey. *Neurocomputing*.

Qiu, S., Liu, Q., Zhou, S., and Wu, C. (2019). Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences*.

Randhawa, R. H., Aslam, N., Alauthman, M., and Rafiq, H. (2022). Evagan: Evasion generative adversarial network for low data regimes.

Renjith G, Sonia Laudanna, A. S.-C. A. V. V. P. (2022). Gang-mam: Gan based engine for modifying android malware.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " 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*.

Rosenberg, I., Shabtai, A., Elovici, Y., and Rokach, L. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*.

Sern, L. J., David, Y. G. P., and Hao, C. J. (2020). PhishGAN: Data augmentation and identification of homoglyph attacks. In *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*.

Shi, Y. and Sagduyu, Y. E. (2017). Evasion and causative attacks with adversarial deep learning. In *MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM)*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Stavroula Bourou, Andreas El Saer, T.-H. V. A. V. and Zahariadis, T. (2021). A review of tabular data synthesis using gans on an ids dataset.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks.

Wang, D., Li, C., Wen, S., Nepal, S., and Xiang, Y. (2020). Defending against adversarial attack towards deep neural networks via collaborative multi-task training.

Wang, J., Yan, X., Liu, L., Li, L., and Yu, Y. (2022). Cttgan: Traffic data synthesizing scheme based on conditional gan. *Sensors*.

Wang, X., Li, J., Kuang, X., Tan, Y., and Li, J. (2019). The security of machine learning in an adversarial setting: A survey. *J. Parallel Distributed Comput.*

Xu, J., Sun, Y., Jiang, X., Wang, Y., Yang, Y., Wang, C., and Lu, J. (2021). Blindfolded attackers still threatening: Strict black-box adversarial attacks on graphs.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan.

Xu, W., Evans, D., and Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks.