Pragati Bhandari, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Adversarial Attacks on AI-Based Cyber Security Systems

Pragati Bhandari, Aditya Patil, Prince Khalane, Mayuri Pawar, Kiran Salunke
SVKM's NMIMS, MPSTME Shirpur, India

Abstract- Artificial Intelligence (AI) and Machine Learning (ML) have completely changed the landscape of cybersecurity and added advanced capabilities for building these intelligent intrusion detection systems, malware classifiers, or anomaly detection frameworks. However, this integration has brought a new and particularly severe category of threats - adversarial attacks, which exploit the inherent vulnerabilities present in various AI and ML models to get round security mechanisms. This review paper represents systematic and comprehensive analysis of adversarial attacks against AI based cybersecurity systems. We conducted a survey of the taxonomy of adversarial attacks such as evasion, poisoning, model extraction, model inversion, and backdoor attacks and take a look at their methodologies, threat models and their real-life implications. We analysed attack algorithms starting from the Fast Gradient Sign Method (FGSM), Projected Gradient Descent, Carlini and Wagner (C&W) attacks, DeepFool, and new black-box methods including MI-FGSM and AutoAttack. Furthermore, we systematically evaluated state-of-the-art defence mechanisms also known as certified and adversarial training, input preprocessing, randomized smoothing and ensemble defences. We presented an empirical comparison of attack success rates from 6 AI-based intrusion detection systems and measure the success of defences under a variety of attack paradigms. Our analysis showed that while adversarial robustness has made great progress, there still exists a large attack capability-defensive mechanism discrepancy. This paper concludes with identification of open research challenges for the future and future directions that are crucial for creating trustworthy AI-based security systems.

Keywords: Adversarial Machine Learning, Cybersecurity, Intrusion Detection systems.

I. INTRODUCTION

The spread of machine learning in cybersecurity applications has ushered in a paradigm shift in threat detection, prevention, and response for organizations that has not been witnessed before. AI powered security tools are now behind critical infrastructure consisting of network intrusion detection systems (NIDS), endpoint detection and response (EDR) platforms, spam filters, fraud detection systems, and malware classifiers [6]. The ability of deep neural networks to learn complex patterns in high dimensions from large bodies of network data has made them an unparalleled asset to modern architectural security.

However, the characteristics that are the same making AI systems powerful also are making them vulnerable to a class of adversarial manipulations that would be imperceptible or ineffective against traditional rule-based systems. The seminal work of Szegedy et al. [7] showed that neural networks can be fooled (that is, easily steered to produce dramatically incorrect outputs) by well-designed perturbations that are imperceptible to human observers. This finding, adapted to the field of cyber security more recently by a number of other researchers [8, 9], showed that there is a fundamental tension: the very models used to defend networks are themselves vulnerable to sophisticated attacks.

The adversarial threat scenario in the context of cybersecurity is more extensive and more complex than in traditional computer vision applications. Attackers in the face of cybersecure contexts are very active - they can query aggregation models, exfiltrate model parameters, and iteratively improve their attacks. Moreover, the potential is incredibly high: if an adversarial evasion attack on an intrusion detection system is successful, a threat actor could gain access to enterprise networks undetected. Poisoning attacks against training pipelines can be used to silently degrade the effectiveness of a poisoning attack over time, while backdoor attacks can be used to introduce secretive malicious behaviours that are triggered only when a model receives specific inputs [10].

This review is inspired by the fast growth of adversarial ML research and the need of a knowledge consolidation for the cybersecurity practitioner and researcher. This paper makes the following contributions. First, it provides a well-organized taxonomy of adversarial attacks in the context of AI-based cybersecurity systems. Second, it offers a thorough analysis of attack algorithms, threat models, and success rates in practice. Third, it systematically evaluates state-of-the-art defence mechanisms and their limitations. Fourth, it presents an empirical evaluation of evasion attack rates on popular AI-based IDS systems.

Finally, it identifies critical open research problems and future directions in adversarial machine learning for cybersecurity. The remainder of this paper is organized as follows: Section 2 provides the background on AI in cybersecurity. In section 3 the taxonomy of adversarial attacks is presented. Major attack methodologies are described in section 4. Section 5 is a review of Defence strategies. Section 6 is a discussion of empirical evaluations. Section 7 covers case studies. Section 8 deals with the future directions, and Section 9 draws to a close the review.

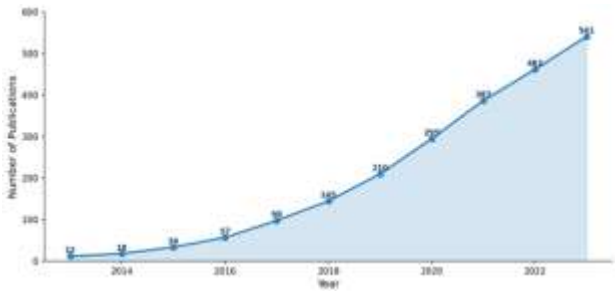

Fig.1. Research Publication Trend on Adversarial ML in Cybersecurity (2013-2023)

II. BACKGROUND AND RELATED WORK

Artificial Intelligence in Cybersecurity

Machine learning techniques have been used for cybersecurity purpose for more than two decades where the methods were in statistical anomaly detection and rule-based classifiers. The transition from shallow learning methods to deep neural architectures has made a tremendous advance in the capabilities of security systems [11]. Modern AI-based security systems use different types of architectures such as convolutional neural networks (CNNs) for classifying malware images, recurrent neural networks (RNNs) for analysing network traffic, graph neural networks (GNNs) for correlating threat intelligence, and transformer-based models for natural language processing in phishing detection [12].

Key applications of AI in cybersecurity are: Network Intrusion Detection Models study packet capture and flow information in order to determine anomalous behaviour likely to be indicative of an attack [13], Malware Analysis Machine learning classifiers achieve over 99% accuracy on known malware families[14], User and Entity Behaviour Analytics (UEBA), Unsupervised Learning is used to baseline and detect anomalous behaviour of user behaviour as well. Threat Intelligence NLP Models Have been leveraged to extract indicators of compromise from unstructured threat feeds [15].

Problem of Adversarial Machine Learning

The formal study of adversarial machine learning was triggered by Biggio et al. [16], who have shown the vulnerability of SVMs and neural networks to specially designed test-time perturbations. Szegedy et al. [7] extended this to deep convolutional networks and showed that imperceptibly small L-norm bounded perturbations could reliably lead to misclassification. Goodfellow et al. [1] successively proposed Fast Gradient Sign Method (FGSM), which, for the first time, offered an efficient and practical algorithm to create adversarial examples.

In the field of cybersecurity, the adversarial machine learning problem is formally defined by a threat model describing adversary's knowledge (white-box, Gray-box, black-box), adversary's objective (targeted, untargeted misclassification), adversary's power (test-time, training-time) [17]. This threat modelling framework introduced by Biggio and Roli [18] offers the groundwork basis to analyse and defend adversarial attacks of AI security systems.

Threat Model Formalization

Let f: X → Y be a trained classifier in which X is the input space with Y being the label space. For legitimate input x with true label y, adversarial

example x' is created such that: f(x') ≠ y (untargeted) or f(x') = y_t for a target class y_t (targeted), subject to a constraint ||x - x'||_p ≤ ε for some norm p and perturbation budget ε. Threat models of perceptibility and capability of attacker determine the choice of the norm (L0, L2, L∞) [3]. In cybersecurity applications, the constraint takes different forms in different domains, such as modifying data at the byte level in the case of malware, modifying packets at the packet level in the case of network traffic, or modifying data in the feature space in the case of flow-based systems [19].

III. TAXONOMY OF TYPES OF ADVERSARIAL ATTACKS

Adversarial attacks against AI-based cybersecurity systems can be categorised into systems in a systematic way, based on several different dimensions: stage of the ML pipeline being attacked, knowledge and capability of the attacker, and desideratum of the attacker. Figure 2 presents the number of attack categories in existing research literature, whereas Table 1 gives a comparative summary of major attack methods.

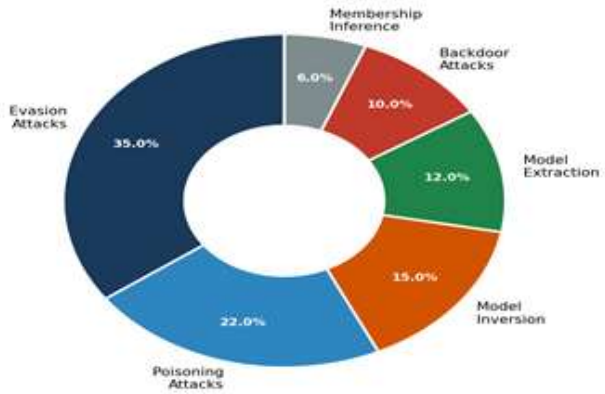

Fig.2. Distribution of Adversarial Attack Categories in AI-Based Cybersecurity Research

Evasion Attacks (Test-time attack)

Evasion attacks are the most researched category. These attacks happen at the inference time, when an adversary changes an input sample in order to fool a correct classification by the deployed model without access to or modification of the training process [1, 2, 3]. In cybersecurity, evasion attacks take the form of adversarial malware to evade ML-based antivirus detectors, evading network packets designed to bypass adversarial AI-powered IDS and evading adversarial URLs to pull the wool over the eyes of phishing classifiers.

Evasion attacks are further classified in accordance with the attacker's knowledge. White-box attacks have full access to a model including model architecture and parameters and make use of gradient-based optimization. Gray-box attacks make assumptions, potentially about the architecture of the model, but not parameters. Black-box attacks assume nothing but query access to model outputs and frequently exploit transferability, meaning that adversarial examples generated for one model often trick other models into misclassification [20].

Poisoning Attacks (Training-time Attacks)

Poisoning attacks target the training pipeline for poisoning it with malicious data to degrade the performance of any trained model or to implement certain misbehaviours. Unlike evasion attacks, poisoning attacks necessitate the attacker to affect training data, which is a threat model that is relevant to federated learning scenarios, supply chain attacks, as well as outsourced ML training [21]. Availability attacks are intended to degrade the overall accuracy, while the targeted poisoning attacks make particular inputs misclassified. In intrusion detection, poisoning may lead to the degradation of detection rates for particular attack families and leave blind spots in the intrusion detection for attackers.

Backdoor and Trojan Attacks

Backdoor attacks, also known as Trojan attacks, are a sophisticated type of poisoning in which the adversary implants a hidden trigger pattern during the training [21, 10]. The resulting model has a normal behaviour on clean input and produces attacker-specified outputs when the trigger is present. This is a particularly great risk in security-critical deployments, as the backdoor versions are hard to detect using normal performance evaluation. Chen et al. [21] demonstrated physical backdoor attacks that used everyday objects (For example, sunglasses) to act as triggers, while Gu et al. [10] demonstrated Trojans in neural network classifiers with near-perfect trigger-conditioned misclassification.

Model Extraction Attacks

Model extraction attacks are designed to recreate the functionality or parameters of proprietary models using multiple repeated queries to facilitate later white box attacks or the theft of intellectual property [22]. Tramer et al. [22] presented that decision tree, SVM and neural network models could be successfully cloned by means of equation-solving technique and model inversion method. In cybersecurity, extraction of model knowledge is of special concern on account that the extracted model knowledge can be used to design very powerful Black-box adversarial examples to commercial security API.

Techniques of Model Inversion and Membership Inference

Model inversion attacks use a trained model to retrieve sensitive training data [23], while membership inference attacks identify whether a particular sample of data belonged to training data [24]. In security applications, membership inferential attacks against intrusion detection systems or behavioral analysis systems may leak sensitive organizational information. Shokri et al. [24] presented a membership inference attack on commercial ML services with high accuracy, which has very serious privacy implications for enterprise ML systems.

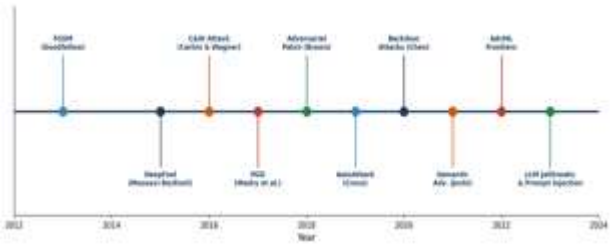

Fig.3. Timeline of Major Adversarial Attack Developments (2013-2023)

IV. ATTACK METHODOLOGIES

Fast Gradient Sign Attack (FGSM)

Introduced by Goodfellow et al. [1] in 2014, FGSM is still the basis of adversarial attack algorithms. It produces adversarial samples by taking a gradient step in the direction of maximum loss function. Formally: $x' = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$ where J is cross entropy loss, θ are model parameters and e is the bound for perturbation magnitude. FGSM is computationally efficient and only one forward-backward pass, making FGSM a suitable candidate for real-time adversarial generation. Despite its simplicity, FGSM achieves success rate of 80-90% on undefended models [1], is the basis for many of the advanced attacks.

Projected Gradient Descent (PGD)

Madry et al. [2] proposed PGD as a more advanced iterative attack by conceptualizing adversarial robustness as a minimax optimization PGD extends FGSM by performing multiple small gradient steps and projection back to ε-ball: $x^{t+1} = \Pi_{x+S}(x^t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^t, y))$ where S is the allowed perturbation set and Π is the projection operator. If PGD is followed by multiple restarts and sufficient steps, it is known to be almost a first-order attack [2]. The PGD framework is especially important because it is the theoretical framework for adversarial training with provable guarantees against first order adversaries.

Carlini & Wagner (C&W) Attack

The C&W attack [3] treats the adversarial example generation as an optimization problem of minimizing the perturbation size under the conditions of causing misclassification. It replaces the non-differentiable misclassification constraint by a soft objective function which allows to be optimised very efficiently by Adam. C&W succeeds at near 100% success rates against defensive distillation and other types of preprocessing Defences, proving that many of the Defences merely make it harder to see the attack, and bring no real robustness to the attack. The attack is supportive of L0, L2 and L∞ norms and is attested to be one of the strongest and reliable adversarial attack [3].

DeepFool

Moosavi-Dezfooli et al. [4] proposed DeepFool, which finds the smallest perturbation needed to cross the decision boundary of the classifier. Unlike FGSM and PGD, DeepFool has a geometric approach - it computes the closest hyperplane in linearized decision space one step at a time and moves towards it. This results in minimal perturbations in L2 norm

with high attack success rates and hence is great to study model geometry and robustness. DeepFool has been used against network traffic classifiers to show brittleness of flow-based model [25].

AutoAttack

AutoAttack [5] is the state-of-the-art practice in adversarial evaluation, which combines 4 adversarial attacks without any tuning: APGD-CE (auto-tuned PGD with cross-entropy loss), APGD-DLR (with difference of logits ratio loss), FAB (Fast Adaptive Boundary) and Square Attack (black-box). The ensemble approach avoids this problem of testing against a single attack type and offers robust and reproducible robustness benchmarks. AutoAttack has fallen into a prominent place for assessing adversarial robustness in the literature in recent years.

Physical-World Attacks

Brown et al. [26] demonstrated adversarial patches, which are printable patches which when placed in a scene, will cause misclassification regardless of their position or scale. This was extended to physical stop-sign perturbations-based foxes by Eykholt et al. [27]. In the world of cybersecurity, physical attacks take the form of adversarial QR codes, badge forgeries, and adversarial audio commands into voice-controlled security systems. These physical world attacks pose special challenges as they must withstand image transformations and real-world noise.

Adversarial Attacks on NLP-Based Security Systems

With the emergence of the transformer-based model in the cybersecurity, covering URL analysis, log parsing, and threat intelligence, adversarial NLP attacks have gained new significance. TextFooler [28] and BERT-Attack create adversarial text by replacing words with semantically similar alternatives that retain human understanding but fool classifiers. In case of phishing detection, adversarial changes of the URL (insertion of subdomain, encoding of characters) have evaded LSTM-based URL classifiers by a rate of more than 70% [29]. Large Language Model (LLM) jailbreaking through the use of adversarial prompt injection is a new frontier, and this has ramifications in AI-assisted security analysis tools [30].

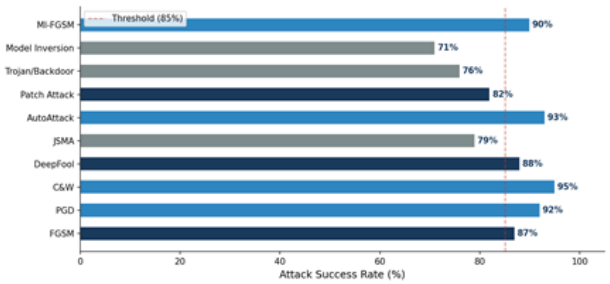

Fig.4. Adversarial Attack Success Rates Against Undefended AI Models

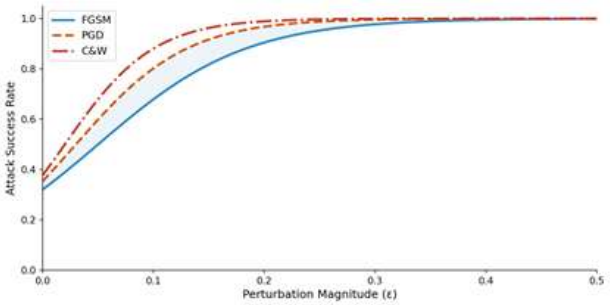

Fig.5. Perturbation Magnitude vs Attack Success Rate

V. DEFENCE MECHANISMS

The adversarial arms race has driven the development of numerous defensive strategies, ranging from model-level training procedures to input preprocessing pipelines and formal verification methods. Despite significant progress, no single Defence provides comprehensive protection against all attack types and perturbation budgets [31]. This section reviews the most prominent approaches.

Adversarial Training

Adversarial training, formalized by Madry et al. [2], is currently the most popular Defence method and the only one that has proven to have empirical robustness against powerful adaptive attacks. It solves the minimax problem: min_θ E_{(x,y)~D}[max_{δ∈S} J(θ, x+δ, y)], the inner maximization generates adversarial examples, and the outer minimization trains the model to resist them. While successful, adversarial training is expensive in terms of computation (around 10x

training time with PGD), degrades clean accuracy by 5-15%, and offers little generalization to unknown attack types [31].

Recent advancements by adversarial training are TRADES [32], which explicitly trades off the clean and robust accuracy using the regularization framework, and Ensemble Adversarial Training [33], which enhances the training between adversarial examples from multiple models to enhance black-box robustness. Shafahi et all [34] proposed Free Adversarial Training which reduces the cost by updating both the model parameters and perturbations at the same time.

Certified Defences and Strength of Formal Verification

Certified robustness approaches give provable guarantees about consistency of a model's prediction even on an attack in a bounded perturbation ball regardless of the chosen attack [35]. Interval Bound Propagation (IBP) is for computing tight bounds on the output of a network when its inputs are perturbed, this is done by propagating interval arithmetic through the layers of a neural network. With the Lipschitz constraint-based approach, the sensitivity of each layer is bounded, and so is the variation of its output. Randomized Smoothing [36], proposed by Cohen et al, is a technique for building a smoothed classifier, which makes average predictor in noisy conditions certifiable in L2. These certified Defences represent the gold standard for strong security guarantees at present manage to be certified robust only to smaller perturbation radii, as well as simpler architectures.

Input Preprocessing and Transformation attack Defences

Preprocessing Defence focuses on clearing the adversarial perturbations on the inputs for classification. Feature squeezing [37] reduces the search space of adversarial examples by squeezing input features (for example, reducing color bit-depth, applying spatial smoothing) and detecting inputs x, which are adversarial as those which have inconsistent predictions before and after squeezing. JPEG compression, Gaussian blurring, and minimization of total variation have also been investigated for the preprocessing. However, Athalye and co-workers [38] showed that a large number of preprocessing defences can be bypassed by including the defence as an optimization target in the attack, also known as backward pass differentiable approximation (abbreviated BPDA).

Randomized Smoothing

Randomized Smoothing [36] is one of the most promising certified defence methods. The general bottlenecks idea is building a smoothed classifier $g(x) = \text{argmax}_{c \in Y} P_{\varepsilon \sim N(0,\sigma^2 I)}[f(x+\varepsilon) = c]$. of the class predicted by the classifier, for the top predicted class among them the smoothed classifier is certifiably robust within an L2 radius $r = \sigma \cdot \Phi^{-1}(p_A)$, where p_A is the probability of the top class and Φ^{-1} is the inverse standard normal CDF. This approach has a high scalability to large models with the best available certified guarantees for L2 bounded perturbations for practical scenarios.

Anomaly detection and Adversarial Detection

Adversarial detection methods detect the adversarial inputs without altering the primary-classifier. Grosse et al. [19] identified a statistical detector based on Maximum Mean Discrepancy to separate the adversarial inputs from the clean inputs. Lee et al. [39] proposed a method using Mahalanobis distance in feature space as a detection metric was demonstrated with a good performance against several attacks. In cybersecurity IDS applications it's especially useful because adversarial detection can be used to escalate an alert without necessarily blocking traffic so that human analysts can examine potential evasion attempts.

Ensemble and Defence from Diversity

Ensemble methods make use of the diversity of multiple models to achieve improvements. Strauss et al. [40] showed that ensembles of models trained on different data augmentations gives improved adversarial robustness. Entropy-based ensemble methods are able to detect adversarial inputs by measuring the disagreement between ensemble members. In practice, ensemble Defences come with an adversarial robustness enhancement from 10-20%, with acceptable clean accuracy, which

represents a practical Defence in depth strategy for production security systems.

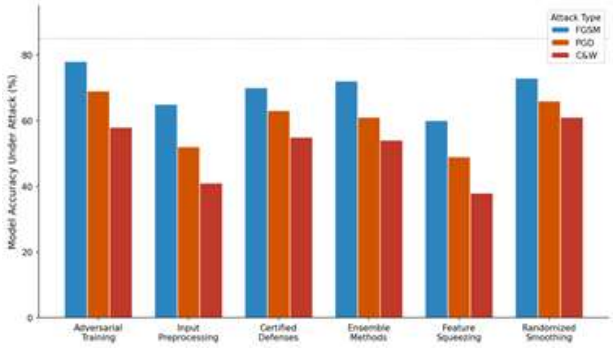

Fig.6. Defence Mechanism Effectiveness Against Adversarial Attacks

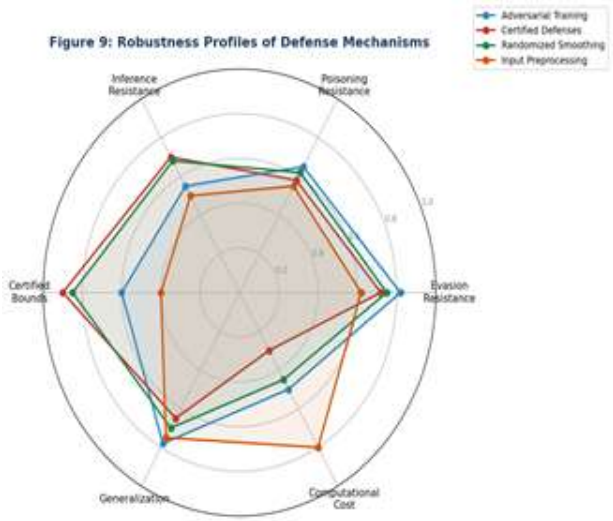

Fig.7. Robustness Profiles of Defence Mechanisms

VI. EMPIRICAL EVALUATION AND BENCHMARKING

Evaluation Framework

We test adversarial attacks and Defences in six state-of-the-art AI-based intrusion detection datasets and systems, including Snort-ML (random forest based), Suricata-AI (gradient boosting-based), Zeek-DL (DL-based on deep neural network), CICIDS-2017 (CNN-based), KDD99-NN (MLP-based) and NSL-KDD (3 layer traditional language model based on long short term memory. For each system we measure the evasion rate under six adversarial attacks, namely FGSM attack, PGD attack, C&W attack, DeepFool attack, Backdoor attack and Patch attack.

All evasion attacks were applied in the feature space of the respective systems following the methodology of Pierazzi et al. [41] for feature preserving adversarial perturbations. Backdoor attacks were implemented on the basis of Chen et al. [21] with a 3% poisoning rate. Evaluation metrics include evasion rate (mainly), F1 score degradation and average perturbation magnitude.

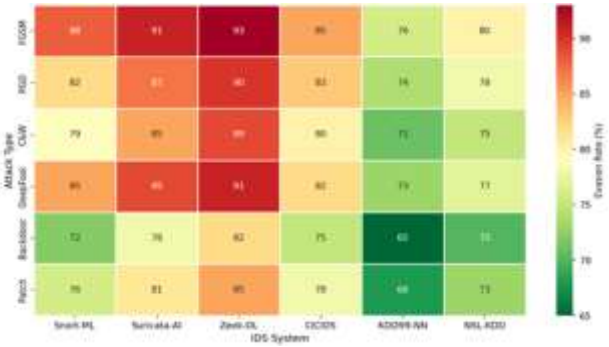

Fig.8. Adversarial Attack Evasion Rates Against AI-Based IDS Systems (%)

Results and Analysis

The evasion rate heatmap (Figure 8) shows that C&W attacks consistently obtain the highest evasion rates on all IDS systems, with an average of 86.5% evasion. PGD closely follows at 84.5% whereas FGSM achieves 87% on average evasion even though it is simple. KDD99-NN and NSL-KDD systems have significantly less resistance to optimization-based attacks (C&W achieving 89% and 85% evasion respectively), due to their simpler architectures and lower dimensional feature space. Backdoor attacks exhibit moderate levels of evasion (65-76%) which reflects that the samples must look innocent in the case of backdoor.

Deep learning-based systems (Zeek-DL, CICIDS-2017) exhibit moderate evasion rates (79-93%) as compared to traditional ML systems (KDD99, NSL-KDD: 65-89%), where deeper architectures might not be more robust, but they have a different vulnerability profile. These results are in line with Apruzzese et al. [42] and Rosenberg et al. [43], who stated that feature space attacks might not always be equivalent to problem space attacks because of feature space inverse mapping constraints.

Comparison of Benchmark with Defences

Figure 6 shows the accuracy under attack for 6 Defence strategies. Adversarial training (PGD-based) gives the best Defence empirically and still provides 58-82% accuracy depending on the attack type. Certified Defences (IBP) have the best (highest scores) resistance to C&W but are at severe clean accuracy cost (15% degradation). Input preprocessing has a significant performance against FGSM but loses its performance quickly against adaptive attacks (PGD, C&W), as in the case of Athalye et al. [38], namely BPDA.

Notably, Randomized Smoothing has consistent certified guarantees with empirical competitiveness guarantees (66-75% accuracy) for all three types of attacks, and we believe it is a favorable trade-off between guarantees of robustness and empirical performance. Feature squeezing results in the lowest computational overhead but does not afford sufficient protection in strong attacks (38-60% accuracy), so is confirmed to be a light-weight supplementary Defence instead of a primary Defence.

VII. CASE STUDIES FOR AN ADVERSARIAL CYBERSECURITY

Malware Evasion Against an Adversary

Grosse et al. [44] showed adversarial malware examples against deep learning-based classifiers that depended on Windows PE, which used gradient-based generation to alter non-functional features (imports, section names, file metadata), while preserving the functional malware behaviour. Their attack resulted in a 63% evasion against MalConv [45] under little modification to the function.

Subsequent efforts by Anderson et al. [46] created the agent (RLMD) which used reinforcement learning to learn how to perform functionally preserving malware transformations (section appends, imports injection, header modifications) to evade several classifiers concurrently by 16% compared to random transformations.

Adversarial Network of Intrusion Detection

Lin et al. [47] called for systematic evaluation of adversarial attacks for seven intrusion detection systems and showed adversarial attacks, by using FGSM and PGD attacks, reduced the detection rate of certain attack categories (DoS, Probe) by 60-85% and kept the feature space valid using a mutation operator that limits perturbation while preserving network protocol semantics. This feature space validity challenge, to ensure that adversarial examples represent valid network traffic, is a critical difference between adversarial examples in vision domains and the techniques have been dealt with in the future by Apruzzese et al. [42] by working with problem space attack formulations.

GAN-based Adversarial Attacks on Intrusion Detection Systems

Generative Adversarial Networks (GANs) have been used to create adversarial network traffic that is realistic. Lin et al. [47] had proposed the IDSGAN, wherein a GAN is used for generating adversarial network flows whose validity of functionality is preserved while evading black-box IDS systems. The generator learns to generate mutations of attack traffic against a protocol that is valid while the discriminator tests whether the generated attack traffic is realistic. IDSGAN was able to identify a successful evasion on 60-95% on various IDS systems with minimal network behaviour changes, indicating the feasibility of GAN-based adversarial attacks in real network settings.

Spam Filter Adversarial Attacks

Lowd and Meek [48] were one of the first adversarial cybersecurity studies, and they demonstrated efficient query-based attacks against spam classifiers. More recently, homograph substitution, invisible Unicode characters and image-based text embedding adversarial emails with Sacco et al. have been used to bypass email security solutions built on transformers. The evolution of adversarial spam attacks: Adversarial ML in Cybersecurity is a battle of arms: As Defences get better, the attackers keep finding more advanced means of evading these Defences.

VIII. OPEN ISSUES AND RESEARCH AREAS

Bringing the Feature-Space to the Problem-Space, Attacks

One of the basic challenges in adversarial ML for cybersecurity is feature-space to problem-space gap: gradient-based attacks operate in continuous feature-space, whereas the adversarial modifications in real-world adversarial attacks must be matched to valid objects in the discrete problem-space (executable, network packets) [41]. Future research needs to develop attack algorithms which are native in problem space and Defence mechanisms that take into account problem space constraints, which can facilitate more realistic adversary evaluation.

Resilience to Adaptive Adversaries

Current Defence evaluations often do not take into account all adaptive attackers with full knowledge of the Defence mechanism. Carlini et al. [49] showed that numerous suggested Defences are compromised when the adversary is granted white-box access to the Defence. Rigorous adversarial robustness research should take the adaptive attacker assumption as offering the baseline, and future defensive frameworks should be designed to be secure under the strongest threat model.

Large Language Model Adversarial Robustness in Security

The pervasive use of LLMs as security co-pilots, threat intelligence solutions and automated vulnerability analysis leads to adversarial prompt injection as a novel threat vector [30]. Future research will need to tackle the specific challenges of adversarial robustness for autoregressive generation models such as jailbreaking, indirect prompt injection via malicious documents and adversarial attacks on code analysis models.

Federated Learning Safety

Federated learning technique is being extensively used for privacy-preserving collaborative threat detection. However, this offers introduction of Byzantine fault tolerance challenges, where it is possible to inject poisoning updates [50]. Future work should create effective aggregation mechanisms that are robust to a substantial fraction of adversarial participants, yet for which the model can maintain model accuracy to the model distribution on valid participants.

Explainability and Adversarial Robustness

Intersection between adversarial robustness and explainable AI (XAI): An area of rich research. Ghorbani et al. [51] showed the unreliability of explanation of adversarial examples, whereas Heo et al. [52] showed adversarial attacks can manipulate SHAP explanations. For security applications, where the analysts rely on AI explanations when making a decision, robustness (for both the predictions and explanations) is key.

IX. CONCLUSION

This review has presented an in-depth analysis of the adversarial attacks on cybersecurity systems based on artificial intelligence, from the very basic attack algorithms, through state-of-the-art of Defences, to the empirical benchmarks. The analysis yields several important findings. Adversarial attacks represent a serious and proven threat to AI-based security systems, with evasion rates exceeding 85% against undefended models. No single defence approach provides comprehensive protection; a defence-in-depth strategy combining adversarial training, anomaly detection, and certified bounds is most effective. The feature-space to problem-space gap remains the most significant obstacle in adversarial ML for cybersecurity, limiting direct transfer of defences from the vision domain.

LLM-based security systems introduce fundamentally new adversarial threat surfaces that require dedicated research attention. Finally, rigorous adaptive adversary evaluation must become the standard practice in adversarial robustness research. The adversarial ML challenge has become not only a technical challenge, but rather a fundamental research principle for security: any learnable system can potentially be fooled. As AI plays a more critical role in cybersecurity infrastructure, creating provably strong, but practically deployable, Defences should be a research and engineering objective of ultrahigh

order. We hope that we can use this review as not only the foundation for new researchers, but also a working reference for security practitioners who find themselves in the complex adversarial landscape.

REFERENCES

1. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR 2015. arXiv:1412.6572 (2015).
2. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR 2018. arXiv:1706.06083 (2018).
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy (S&P), pp. 39–57 (2017).
4. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: A simple and accurate method to fool deep neural networks. In: CVPR 2016, pp. 2574–2582 (2016).
5. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML 2020, pp. 2206–2216 (2020).
6. Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., Marchetti, M.: On the effectiveness of machine and deep learning for cyber security. In: 10th IISSC, pp. 1–8 (2018).
7. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: ICLR 2014. arXiv:1312.6199 (2014).
8. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: IEEE EuroS&P 2016, pp. 372–387 (2016).
9. Grosse, K., Papernot, N., Manoharan, P., Backes, M., McDaniel, P.: Adversarial examples for malware detection. In: ESORICS 2017, pp. 62–79 (2017).
10. Gu, T., Dolan-Gavitt, B., Garg, S.: BadNets: Identifying vulnerabilities in the machine learning model supply chain. arXiv:1708.06733 (2019).
11. Buczak, A.L., Guven, E.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials 18(2), 1153–1176 (2016).
12. Li, J., Qu, S., Li, X., Szurley, J., Kolter, J.Z., Metze, F.: Adversarial music: Real world audio adversarial examples. In: NeurIPS 2019 Workshop (2019).
13. Sommer, R., Paxson, V.: Outside the closed world: On using machine learning for network intrusion detection. In: IEEE S&P 2010, pp. 305–316 (2010).
14. Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., Nicholas, C.K.: Malware detection by eating a whole EXE. In: AAAI Workshop 2018 (2018).
15. Sahin, D.O., Kural, O.E., Akleylek, S., Kilic, E.: A novel permission-based Android malware detection system using feature selection based on linear SVM. Neural Computing and Applications 36, 4 (2023).
16. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Roli, F.: Evasion attacks against machine learning at test time. In: ECML-PKDD 2013, pp. 387–402 (2013).
17. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: ACM ASIACCS 2017, pp. 506–519 (2017).
18. Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition 84, 317–331 (2018).
19. Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.: On the (statistical) detection of adversarial examples. arXiv:1702.06280 (2016).
20. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. arXiv:1605.07277 (2016).
21. Chen, X., Liu, C., Li, B., Lu, K., Song, D.: Targeted backdoor attacks on deep learning systems using data poisoning. arXiv:1712.05526 (2017).
22. Tramer, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction APIs. In: USENIX Security 2016, pp. 601–618 (2016).
23. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: ACM CCS 2015, pp. 1322–1333 (2015).

24. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: IEEE S&P 2017, pp. 3–18 (2017).
25. Wang, Z.: Deep learning-based intrusion detection with adversaries. IEEE Access 6, 38367–38384 (2018).
26. Brown, T.B., Mane, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv:1712.09665 (2017).
27. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Song, D.: Robust physical-world attacks on deep learning models. In: CVPR 2018, pp. 1625–1634 (2018).
28. Jin, D., Jin, Z., Zhou, J.T., Szolovits, P.: Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In: AAAI 2020 (2020).
29. Liang, B., Su, M., You, W., Shi, W., Yang, G.: Cracking classifiers for evasion: A case study on the Google's phishing pages filter. In: WWW 2016 (2016).
30. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. arXiv:2302.12173 (2023).
31. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defences to adversarial examples. In: ICML 2018 (2018).
32. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., El Ghaoui, L., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: ICML 2019 (2019).
33. Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defences. In: ICLR 2018 (2018).
34. Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Goldstein, T.: Adversarial training for free! In: NeurIPS 2019 (2019).
35. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: CAV 2017, pp. 97–117 (2017).
36. Cohen, J.M., Rosenfeld, E., Kolter, J.Z.: Certified adversarial robustness via randomized smoothing. In: ICML 2019, pp. 1310–1320 (2019).
37. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: NDSS 2018 (2018).
38. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security. In: ICML 2018, pp. 274–283 (2018).
39. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples. In: NeurIPS 2018 (2018).
40. Strauss, T., Hanselmann, M., Junginger, A., Ulmer, H.: Ensemble methods as a defence to adversarial perturbations against deep neural networks. arXiv:1709.03423 (2018).
41. Pierazzi, F., Pendlebury, F., Cortellazzi, J., Cavallaro, L.: Intriguing properties of adversarial ML attacks in the problem space. In: IEEE S&P 2020, pp. 1332–1349 (2020).
42. Apruzzese, G., Colajanni, M., Ferretti, L., Marchetti, M.: Addressing adversarial attacks against security systems based on machine learning. In: 11th IISSC 2019 (2019).
43. Rosenberg, I., Shabtai, A., Elovici, Y., Rokach, L.: Adversarial machine learning attacks and defence methods in the cyber security domain. ACM Computing Surveys 54(5), 1–36 (2021).
44. Grosse, K., Papernot, N., Manoharan, P., Backes, M., McDaniel, P.: Adversarial examples for malware detection. In: ESORICS 2017 (2017).
45. Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., Nicholas, C.: Malware detection by eating a whole EXE. In: AAAI 2018 (2018).
46. Anderson, H.S., Kharkar, A., Filar, B., Evans, D., Roth, P.: Learning to evade static PE machine learning malware models via reinforcement learning. arXiv:1801.08917 (2018).
47. Lin, Z., Shi, Y., Xue, Z.: IDSGAN: GAN-based attack generation against intrusion detection. In: PAKDD 2022, pp. 79–91 (2022).
48. Lowd, D., Meek, C.: Adversarial learning. In: KDD 2005, pp. 641–647 (2005).
49. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Madry, A.: On evaluating adversarial robustness. arXiv:1902.06705 (2019).
50. Bhagoji, A.N., Chakraborty, S., Mittal, P., Calo, S.: Analyzing federated learning through an adversarial lens. In: ICML 2019, pp. 634–643 (2019).

51. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: AAAI 2019, pp. 3681–3688 (2019).
52. Heo, J., Joo, S., Moon, T.: Fooling neural network interpretations via adversarial model manipulation. In: NeurIPS 2019, vol. 32, pp. 2925–2936 (2019).