AIM INTERNATIONAL LLC Publisher
Innovatech Engineering Journal, 2024;1(01): 80-98

ADVERSARIAL MACHINE LEARNING IN NETWORK SECURITY: A SYSTEMATIC REVIEW OF THREAT VECTORS AND DEFENSE MECHANISMS

Abdul Awal Mintoo[1]
[1]*Graduate student, School of Computer and Information Sciences, Washington University of Science and Technology (WUST), USA*
Email: mintoo.hr@gmail.com
https://orcid.org/0009-0009-0493-965X

Ashrafur Rahman Nabil[2]
[2]*MS in Information Technology Management, St. francis College, Brooklyn, New York,USA*
Email: anabil@sfc.edu
https://orcid.org/0009-0005-1540-1266

Md Ashraful Alam[3]
[3]*Department of Computer Science, Colorado State University,Colorado, USA*
Email: mdashraful.alam@colostate.edu
https://orcid.org/0009-0006-0493-1031

Imran Ahmad[4]
Master of Science in Business Analytics, W. Frank Barton School of Business, Wichita State University, USA
Email: ahmad.imran11235@gmail.com
https://orcid.org/0009-0005-4035-6321

Keywords

Adversarial Machine Learning
Network Security
Threat Vectors
Defense Mechanisms
Systematic Review

ABSTRACT

Adversarial Machine Learning (AML) has emerged as a critical area of research within network security, addressing the evolving challenge of adversaries exploiting machine learning (ML) models. This systematic review adopts the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to comprehensively examine threat vectors and defense mechanisms in AML. The study identifies, categorizes, and evaluates existing research focused on adversarial attacks targeting ML algorithms in network security applications, including evasion, poisoning, and model extraction attacks. By rigorously following the PRISMA guidelines, a systematic search across multiple scholarly databases yielded a robust dataset of peer-reviewed articles that were screened, reviewed, and analyzed for inclusion. The review outlines key adversarial techniques employed against ML systems, such as gradient-based attack strategies and black-box attacks and explores the underlying vulnerabilities in network security architectures. Additionally, it highlights defense mechanisms, including adversarial training, input preprocessing, and robust model design, discussing their efficacy and limitations in mitigating adversarial threats. The study also identifies critical gaps in current research, such as the lack of standardized benchmarking for adversarial defenses and the need for scalable and real-time AML solutions.

1 INTRODUCTION

The rapid integration of machine learning (ML) into network security systems has revolutionized the detection and mitigation of cyber threats, enabling advanced capabilities in areas like intrusion detection, spam filtering, and anomaly detection (Grosse et al., 2023). However, this integration has also exposed ML

Received: 05th October, 2024 **Accepted**: 28th November,2024 **Published**: 30th November,2024

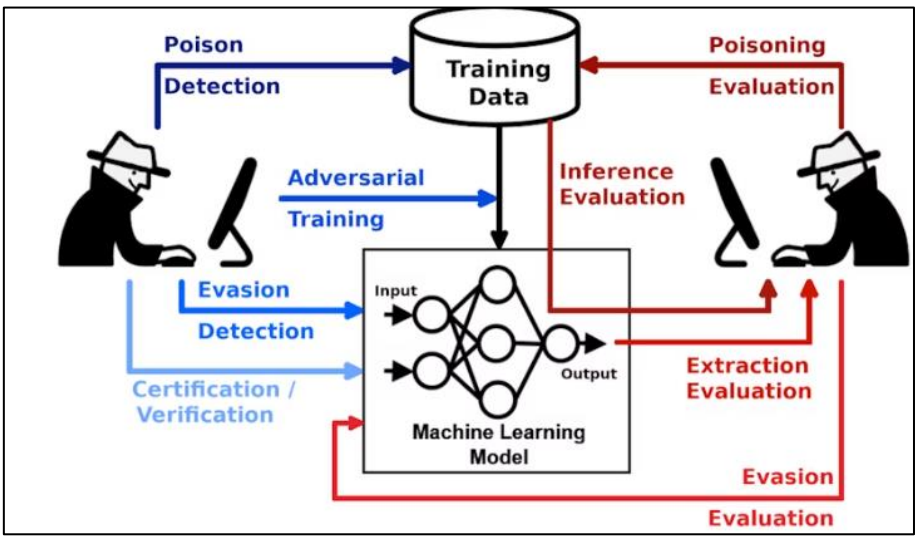
Innovatech Engineering Journal
Page No: 80-98

models to a new category of risks—adversarial attacks—that exploit the inherent vulnerabilities in these systems (Zhang et al., 2019). Adversarial Machine Learning (AML) involves techniques used by malicious actors to deceive or compromise ML models by manipulating inputs, leading to incorrect outputs or exposing sensitive information (Raza et al., 2024). As ML becomes a foundational component in safeguarding critical infrastructure, understanding how adversaries exploit these systems is crucial. For example, studies have revealed that even well-trained ML models can be deceived with imperceptible perturbations to input data, making the systems vulnerable to significant breaches (Liu et al., 2018). The dual-edged nature of ML's deployment in network security highlights the urgency of comprehensively addressing adversarial threats(Ahmad et al., 2020).

Adversarial attacks in AML can be categorized into evasion, poisoning, and model extraction attacks, each targeting different stages of the ML pipeline (Calleja et al., 2018). Evasion attacks manipulate input data during inference, enabling adversaries to bypass detection systems. Poisoning attacks, by contrast, compromise the integrity of training data, leading to biased or dysfunctional models (Watkins et al., 2024). Model extraction attacks, which aim to replicate a model's functionality or access sensitive training data, represent another sophisticated threat vector (Tibshirani, 1996). These attacks are particularly concerning in scenarios involving public-facing ML applications, where access to the model's outputs is easier. For instance, Olowononi et al. (2021) demonstrated that black-box attacks could reverse-engineer models with minimal queries, underlining the pressing need for robust countermeasures. The diversity and sophistication of these attack strategies underscore the growing necessity for a multi-faceted approach to AML in network security. To combat adversarial threats, researchers have proposed a range of defense mechanisms, including adversarial training, preprocessing techniques, and the development of robust model architectures. Adversarial training involves augmenting datasets with adversarial examples during the training phase, improving the model's ability to recognize and resist malicious inputs (Homer et al., 2008). Input preprocessing, such as feature denoising and normalization, offers another layer of protection by mitigating the impact of adversarial perturbations before the data reaches the model (Venkatesan et al., 2021). Additionally, advanced techniques like defensive distillation and gradient obfuscation aim to enhance model robustness by modifying the learning process or concealing gradient information from attackers (Tibshirani, 1996). However, the effectiveness of these defenses varies significantly across attack types, and many solutions face scalability challenges, particularly in real-time applications (Mazumder et al., 2024). Consequently, developing generalized and scalable defense mechanisms remains a critical area of research.

The broader implications of AML in network security extend to ethical, legal, and operational dimensions. As ML applications expand into sensitive domains such as finance, healthcare, and national defense, adversarial attacks can lead to catastrophic consequences, including financial losses, breaches of personal data, and

Figure 1: Adversarial Robustness Toolbox (ART)



compromised public safety (Ahmad et al., 2020; Alam et al., 2024). For instance, Zhao et al. (2022) demonstrated that adversarial examples could bypass ML-based image recognition systems in physical-world settings, raising concerns about the reliability of these models in high-stakes environments. Furthermore, the lack of standardized benchmarks for evaluating the performance of defense mechanisms complicates efforts to measure their efficacy and foster innovation (Hasan et al., 2024). Addressing these challenges requires a holistic approach that not only strengthens technical defenses but also considers the socio-technical context in which AML systems operate (Islam et al., 2024). The primary objective of this systematic review is to provide a comprehensive analysis of adversarial machine learning (AML) within the context of network security, with a specific focus on identifying and categorizing threat vectors and evaluating defense mechanisms. By employing the PRISMA methodology, this study seeks to synthesize existing research to elucidate the nature and scope of adversarial attacks, such as evasion, poisoning, and model extraction, and their impact on ML-driven network security systems. Furthermore, the review aims to critically examine the effectiveness of various defense mechanisms, including adversarial training, input preprocessing, and robust model design, to determine their strengths, limitations, and applicability in real-world scenarios. An additional objective is to identify critical gaps in the current body of knowledge, such as the absence of standardized benchmarks for defense evaluation and the limited scalability of existing solutions. Through these objectives, this study aspires to contribute to the development of more resilient and adaptive AML frameworks, facilitating the secure deployment of ML in network security applications.

2 LITERATURE REVIEW

The field of adversarial machine learning (AML) in network security has garnered significant academic and industry attention due to the increasing adoption of machine learning (ML) models in security-critical applications. The literature on AML is rich with studies that explore adversarial attacks, defense mechanisms, and their implications for network security. This section systematically reviews the existing body of knowledge, providing a detailed analysis of key concepts, methodologies, and findings. By synthesizing insights from recent studies, this literature review aims to categorize adversarial threat vectors, evaluate defense mechanisms, and highlight gaps in the research landscape. The review is structured to offer a thematic exploration of adversarial attacks and their technical underpinnings, followed by an evaluation of existing defenses and their limitations. It concludes with a discussion of unresolved challenges and future research directions, providing a foundation for advancing AML in network security.

2.1 Adversarial Machine Learning in Network Security

Adversarial Machine Learning (AML) has emerged as a critical field in the intersection of machine learning (ML) and network security, addressing the vulnerabilities of ML systems to adversarial attacks (Homer et al., 2008). These attacks exploit the inherent weaknesses of ML models, such as their reliance on training data and susceptibility to perturbations, to compromise their functionality (Alam, 2024). In network security, ML models are commonly used for intrusion detection, anomaly detection, and malware classification. However, adversarial techniques, including evasion, poisoning, and model extraction attacks, threaten the reliability of these applications (Mosleuzzaman et al., 2024). For example, evasion attacks bypass anomaly detection systems by subtly altering input features, while poisoning attacks corrupt the training datasets, leading to degraded model performance (Mosleuzzaman et al., 2024). As ML becomes increasingly integrated into critical infrastructure, understanding these vulnerabilities and their implications is essential to ensure the secure deployment of these systems (Mosleuzzaman et al., 2024).

The challenges posed by adversarial attacks on ML systems are multifaceted, extending from technical limitations to ethical and operational risks (Nandi et al., 2024). Technically, ML models are often treated as black boxes, which makes it challenging to identify and address their vulnerabilities before deployment (Rahaman et al., 2024). Gradient-based attacks, such as the Fast Gradient Sign Method (FGSM), exploit these black-box characteristics to craft adversarial examples that mislead the model (Rahman, 2024). Furthermore, poisoning attacks compromise the training phase, embedding vulnerabilities that attackers can later exploit (Rahman, 2024). Operationally, adversarial attacks pose significant risks to privacy and data integrity. For instance, model extraction attacks can

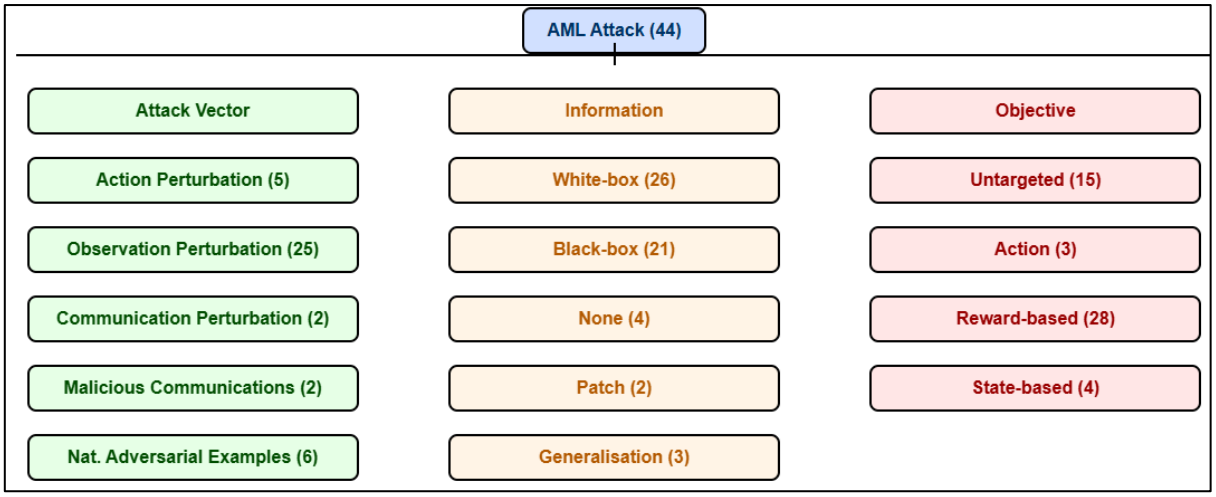
reveal sensitive information about the training dataset or the underlying architecture of the model (Rahman, 2024). These challenges are further exacerbated by the lack of standardized benchmarks to evaluate adversarial defenses, making it difficult to compare the effectiveness of different solutions (Rahman et al. 2024). Addressing these vulnerabilities is imperative to ensure the secure deployment of ML systems in network security applications. One of the most researched solutions is adversarial training, which involves exposing models to adversarial examples during training to improve their robustness (Rahman et al., 2024). Input preprocessing techniques, such as feature denoising and input normalization, are also widely explored as first-line defenses against adversarial attacks (Shamsuzzaman et al., 2024). Additionally, robust model design approaches, including defensive distillation and gradient obfuscation, have shown promise in enhancing model resistance to gradient-based attacks (Tsai et al., 2009; Zou & Hastie, 2005). However, these defenses have limitations, including reduced generalizability and scalability in real-time applications (Shorna et al., 2024). Recent studies emphasize the importance of combining multiple defense mechanisms to create more resilient systems (Shorna et al., 2024). Beyond technical defenses, ensuring the secure deployment of ML systems requires a holistic approach that integrates technical, operational, and ethical considerations (Sohel et al., 2024). For example, employing secure data collection and labeling practices can reduce the risk of poisoning attacks, while ongoing monitoring and validation of deployed models can help detect adversarial activity in real-time (Sultana & Aktar, 2024). Furthermore, advancements in explainable AI (XAI) can provide greater transparency into model decision-making, enabling security teams to identify potential vulnerabilities proactively (Uddin, 2024).

2.2 Adversarial Threat Vectors

Adversarial attacks on machine learning (ML) systems pose significant challenges in network security, with evasion attacks being one of the most prevalent and studied threats. These attacks manipulate input data during inference, subtly altering its features to deceive the ML model without triggering detection mechanisms (Uddin & Hossan, 2024). For instance, adversarial examples crafted using gradient-based techniques, such as the Fast Gradient Sign Method (FGSM), exploit the vulnerabilities in model gradients to bypass intrusion detection systems (Homoliak et al., 2019). Case studies highlight the susceptibility of network intrusion detection systems (NIDS) to evasion attacks, where adversaries modify packet headers or payloads to elude detection algorithms (Lecuyer et al., 2019; Papernot et al., 2016). For example, Colbaugh and Glass (2013) demonstrated how adversarial perturbations in network traffic data could evade deep learning-based intrusion detection models with high accuracy. These findings underscore the need for improved model robustness and real-time defenses to counter evasion threats effectively. Poisoning attacks present a distinct challenge by targeting the training phase of ML systems. These attacks involve introducing malicious data into the training dataset, corrupting the model's ability to generalize and perform accurately (Lowd & Meek, 2005). Poisoning methods often exploit the over-reliance of ML models on clean and representative data, injecting manipulated samples that bias the model's decision-making process (Kumar et al., 2020). Real-world examples, such as the backdoor attacks on email spam filters, demonstrate the severe implications of poisoning, where specific trigger patterns in training data cause the model to misclassify harmful inputs (Malik et al., 2024). Additionally, Carlini and Wagner (2018) highlighted how poisoning attacks could render cybersecurity systems ineffective, especially in collaborative or federated learning scenarios where data is sourced from multiple untrusted entities. Addressing these threats requires robust data validation protocols and mechanisms to detect anomalies in training datasets.

Model extraction attacks represent another critical vector, wherein adversaries aim to replicate or steal the functionality of an ML model by querying it systematically. These attacks exploit the input-output relationship of ML models to reconstruct their internal parameters, effectively reverse-engineering the system (Balle et al., 2022). For instance, studies have shown how adversaries can replicate proprietary deep learning models used in network security by leveraging only a limited number of queries, exposing trade secrets and intellectual property (Papernot et al., 2016). The implications of model extraction extend beyond model theft to include the potential misuse of stolen models for evasion or poisoning attacks (Wang et al., 2019). Such risks highlight the need for secure API designs and techniques like query rate limiting and differential privacy to safeguard ML models from unauthorized access and reverse engineering. Moreover, Hybrid and

Figure 2:Classification of Adversarial Machine Learning (AML) attacks

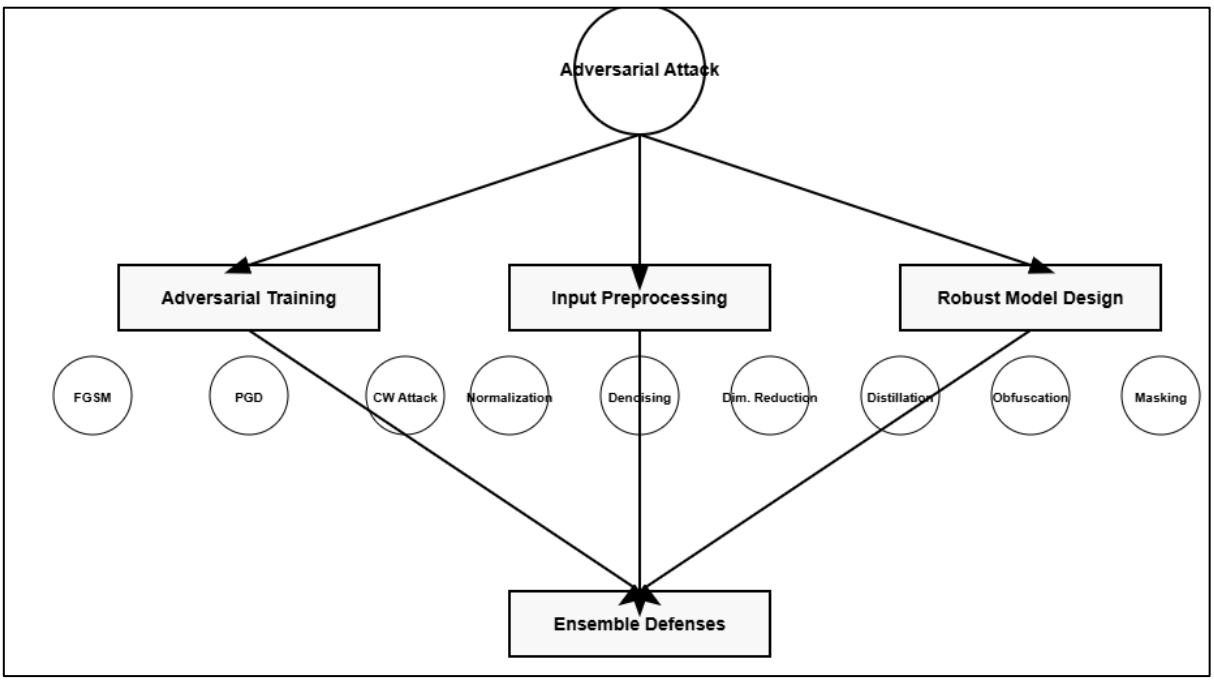


emerging attack strategies further complicate the adversarial landscape by combining multiple threat vectors to enhance their effectiveness. Hybrid attacks, such as those integrating evasion and poisoning techniques, simultaneously compromise training data and inference phases, creating more robust and undetectable adversarial examples (Malik et al., 2024). Emerging threats, including attacks leveraging generative adversarial networks (GANs), present novel challenges by automating the generation of sophisticated adversarial samples (Madry et al., 2017). For example, GAN-based techniques have been used to craft perturbations that evade not only detection systems but also human scrutiny, increasing the risk of undetected breaches in critical applications (Biggio & Roli, 2018). As adversaries continue to innovate, the need for proactive research into hybrid and emerging threats remains urgent, emphasizing the importance of adaptive defenses capable of addressing complex and evolving attack strategies.

2.3 Defense Mechanisms Against Adversarial Attacks

Adversarial training is one of the most widely studied defense mechanisms against adversarial attacks, focusing on improving model robustness by augmenting training datasets with adversarial examples. This method aims to expose the model to potential threats during the training phase, enabling it to learn patterns and resist adversarial perturbations (Marino et al., 2018). Pierazzi et al. (2020) demonstrated that

Figure 3: Defense Mechanisms Against Adversarial Attacks



adversarial training could significantly enhance the resilience of deep neural networks to gradient-based attacks like FGSM and Projected Gradient Descent (PGD). However, adversarial training is computationally expensive, often requiring substantial resources to generate adversarial examples and retrain models (Colbaugh & Glass, 2013). Additionally, while it can increase robustness against specific attack types, its generalizability across diverse adversarial techniques remains limited (Wang et al., 2019). Input preprocessing techniques offer another layer of defense by mitigating the impact of adversarial perturbations before data is fed into the model. These techniques include input normalization, feature denoising, and dimensionality reduction, which aim to remove adversarial noise from the input data (Warzynski & Kołaczek, 2018). For example, pixel-wise transformations and feature squeezing have been shown to reduce the effectiveness of adversarial examples in image classification tasks (Kumar et al., 2020). Preprocessing methods are particularly effective against black-box attacks, where the adversary lacks direct access to model parameters (Sharon et al., 2022). However, their effectiveness can vary depending on the nature of the adversarial attack and the underlying ML model. Duddu (2018) emphasized that preprocessing techniques might inadvertently degrade model performance on benign inputs, highlighting the trade-offs involved in their application. Comparative studies reveal that combining multiple preprocessing methods can enhance their overall efficacy, particularly in dynamic network environments.

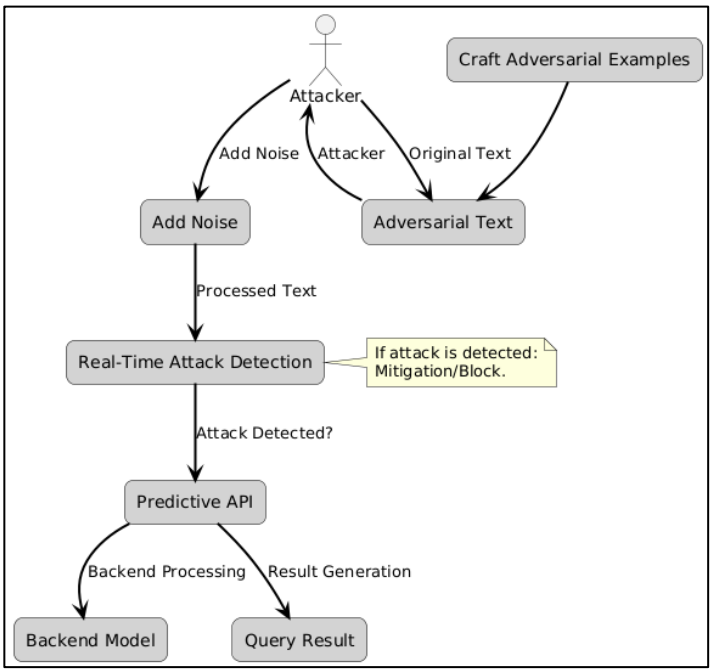
Robust model design focuses on architectural innovations to strengthen ML models against adversarial attacks. Techniques such as defensive distillation, gradient obfuscation, and adversarial feature masking are designed to make it more challenging for adversaries to exploit model vulnerabilities (Wang et al., 2021). Defensive distillation, for instance, modifies the training process to reduce the sensitivity of the model to small perturbations, effectively countering gradient-based attacks (Madry et al., 2017). Gradient obfuscation, on the other hand, aims to obscure the gradient information required by attackers to generate adversarial examples (Colbaugh & Glass, 2013). While these methods show promise in enhancing model robustness, studies highlight their limitations, such as susceptibility to advanced adaptive attacks that bypass these defenses (Duddu, 2018). Case studies, such as those analyzing robust architectures for intrusion detection systems, demonstrate that combining robust design principles with other defense mechanisms can provide more comprehensive protection. Ensemble defenses leverage the diversity of multiple models to increase system resilience against adversarial attacks (Warzynski & Kołaczek, 2018). By combining the predictions of multiple independently trained models, ensemble methods reduce the likelihood that a single adversarial example will compromise the entire system (Lecuyer et al., 2019). This approach is particularly effective in scenarios where adversarial attacks target specific model architectures or training techniques. Malik et al. (2024) demonstrated that ensemble methods significantly enhance resilience against black-box and transfer attacks, as attackers must craft adversarial examples that generalize across multiple models. However, ensemble methods also face challenges, such as increased computational complexity and the risk of correlated vulnerabilities among models (Apruzzese & Colajanni, 2018). Evaluations indicate that combining ensemble defenses with preprocessing techniques and adversarial training can mitigate these challenges, providing robust and scalable solutions to adversarial threats in network security.

Adversarial Attacks on Network Security

The vulnerabilities in network security architectures often amplify the risks posed by adversarial machine learning (AML) attacks. Many ML-based systems rely on fixed and predictable decision-making pipelines, which adversaries can exploit to craft targeted attacks (Wang et al., 2019). For instance, Kumar et al. (2020) identified that the lack of dynamic defenses in traditional intrusion detection systems (IDS) creates opportunities for adversaries to bypass detection mechanisms using evasion attacks. Architectural flaws, such as the over-reliance on static thresholds and the absence of anomaly detection in critical data paths, make these systems particularly susceptible to adaptive adversarial techniques (Homoliak et al., 2019). Furthermore, Marino et al. (2018) highlighted that the exposure of model APIs in network security applications, such as fraud detection and malware classification, increases the risk of model extraction and poisoning attacks. Addressing these vulnerabilities requires rethinking architectural designs to incorporate dynamic, adaptable, and robust defense mechanisms capable of countering evolving adversarial threats (Colbaugh & Glass, 2013). Moreover, adversarial

attacks pose significant implications for user privacy and data integrity in network security systems (Kumar et al., 2020). Attacks such as model extraction can expose sensitive information embedded in training datasets, violating user privacy and potentially leading to identity theft or unauthorized access to personal data (Sharon et al., 2022). Similarly, poisoning attacks that manipulate training datasets can compromise the integrity of data-driven decisions, resulting in misclassifications that adversely affect end-users (Kim et al., 2018). Alotaibi and Rassam (2023) demonstrated how adversarial examples could exploit vulnerabilities in face recognition systems, enabling unauthorized access to secure facilities. Additionally, attacks on healthcare systems using adversarial ML have been shown to manipulate diagnostic outputs, jeopardizing patient safety and trust in these systems (He et al., 2023). These findings emphasize the need for rigorous data validation and monitoring protocols to protect user privacy and maintain data integrity in adversarial environments. Moreover, the operational and financial impacts of adversarial attacks on network security systems are both profound and far-reaching. Successful adversarial attacks can disrupt critical operations, leading to service outages, reputational damage, and financial losses (Huang et al., 2011). For example, Menéndez et al. (2019) reported that evasion attacks on financial transaction monitoring systems resulted in undetected fraudulent transactions, causing millions of dollars in losses. In another case, Yan et al. (2019) illustrated how poisoning attacks on autonomous vehicle navigation systems could lead to traffic disruptions and accidents, demonstrating the real-world operational consequences of AML vulnerabilities. Furthermore, McClintick et al. (2022) highlighted the financial risks of intellectual property theft through model extraction attacks, which enable competitors to replicate proprietary algorithms without incurring development costs. These case studies underscore the critical importance of robust AML defenses to safeguard both the operational continuity and financial stability of organizations relying on ML-based network security systems. The growing sophistication of adversarial attacks necessitates a multi-faceted approach to mitigate their impact on network security. While technical defenses are essential, organizations must also address the systemic and procedural weaknesses that adversaries exploit (Wang et al., 2019). Yan et al. (2019) emphasized the importance of integrating adversarial resilience into the design of ML models and network architectures to preemptively counteract potential threats. Additionally, regular stress testing and simulations of adversarial scenarios can help organizations identify and address vulnerabilities before they are exploited in real-world attacks (Sauka et al., 2022). By combining technical advancements with proactive operational strategies, organizations can reduce the operational, financial, and privacy risks posed by adversarial attacks, ensuring a more secure network environment.

Figure 4: Adversarial Attacks on Network Security



2.4 Evaluation of Defense Mechanisms

The lack of standardized benchmarks for evaluating adversarial machine learning (AML) defenses poses a significant challenge to advancing the field (Huang et al., 2011). Current evaluation frameworks vary widely in their methodologies, datasets, and performance metrics, making it difficult to compare the effectiveness of different defense mechanisms (Yan et al., 2019). For example, adversarial training strategies often report results based on specific attack types and datasets, limiting their generalizability across diverse scenarios (Rosenberg et al., 2021). Alhajjar et al. (2021) emphasized that without standardized evaluation criteria, researchers may inadvertently design defenses that perform well only under specific conditions but fail against more sophisticated or unseen adversarial strategies. This inconsistency hampers the reproducibility of findings and the broader adoption of robust AML solutions(Sauka et al., 2022). Moreover,

one of the primary challenges in developing standardized benchmarks is the dynamic and evolving nature of adversarial attacks. Adversaries continually adapt their techniques, rendering static evaluation criteria obsolete (Wang et al., 2019). For instance, while gradient-based attacks like FGSM and PGD dominated early AML research, recent studies have highlighted the rise of more advanced techniques, such as query-based and generative adversarial network (GAN)-powered attacks (Huang et al., 2011). Yan et al. (2019) noted that most existing benchmarks fail to account for the diversity and complexity of these emerging threats, leading to an incomplete assessment of defense mechanisms. Consequently, there is a growing need for dynamic and adaptable benchmarks that reflect the real-world adversarial landscape (McClintick et al., 2022).

Another challenge lies in the selection of representative datasets for evaluating AML defenses. Many studies rely on widely used datasets, such as MNIST, CIFAR-10, and ImageNet, which may not accurately reflect the complexities of real-world network security applications (Pawlicki et al., 2020). Menéndez et al. (2019) argued that the reliance on these datasets leads to over-optimized defenses tailored to specific data distributions, potentially neglecting critical vulnerabilities in other domains. Furthermore, Jia and Liang (2017) highlighted that datasets for network security applications, such as intrusion detection or malware classification, often lack adversarial examples, further complicating the evaluation process. To address these gaps, researchers have called for the creation of domain-specific benchmarks that incorporate diverse datasets and realistic adversarial scenarios.Moreover, the absence of unified performance metrics further exacerbates the benchmarking challenges in AML defense evaluation. Metrics such as accuracy, robustness, and computational efficiency are often reported independently, without considering their trade-offs (Carlini & Wagner, 2017). For example, a defense mechanism that improves robustness against adversarial attacks may incur significant computational overhead, rendering it impractical for real-time applications (Chen et al., 2017). Lecuyer et al. (2019) proposed that multi-dimensional evaluation frameworks incorporating metrics for robustness, scalability, and efficiency could provide a more holistic assessment of AML defenses. However, implementing such frameworks requires collaboration among researchers, practitioners, and standardization bodies to define and adopt universally accepted evaluation practices (Wang et al., 2019).

Scalability and Real-Time Performance

The scalability of adversarial machine learning (AML) defenses is a significant challenge, particularly as network security applications require systems capable of handling high volumes of data in real time. Many existing defenses, such as adversarial training, are computationally intensive and struggle to scale effectively in dynamic environments (Duddu, 2018). For instance, generating adversarial examples during training for large datasets or complex models can be prohibitively resource-intensive, limiting their practicality in enterprise-level applications (Madry et al., 2017). Additionally, Kumar et al. (2020) noted that defenses designed for specific adversarial scenarios often fail to generalize across different attack types or domains, further complicating their scalability. As network environments become increasingly dynamic and data-intensive, the need for lightweight, adaptable, and scalable AML defenses has become paramount.

Real-time performance is another critical factor in the effectiveness of AML defenses, particularly for applications like intrusion detection and fraud prevention, where immediate responses are essential. However, many defenses, including preprocessing techniques and robust model designs, incur significant latency when applied to large-scale data streams (Madry et al., 2017). For example, input transformations such as feature denoising and dimensionality reduction require additional computational steps, which can delay detection and response times (Biggio & Roli, 2018). Brown et al. (2021) emphasized the importance of developing real-time AML solutions that can maintain high detection accuracy without compromising speed. Techniques such as approximate adversarial detection and lightweight model architectures have shown promise in reducing latency, but their effectiveness against sophisticated attacks remains underexplored.

The dynamic nature of network environments presents unique challenges to AML defenses, particularly in maintaining performance as network conditions and attack vectors evolve. Ring et al. (2019) highlighted that static defenses often fail to adapt to new threats, leaving systems vulnerable to emerging adversarial techniques. To address this, researchers have proposed adaptive defense mechanisms that dynamically adjust their parameters based on real-time threat assessments (Duddu, 2018). For instance, systems that integrate adversarial training with real-time monitoring and

anomaly detection have demonstrated improved resilience against evolving attack strategies (Malik et al., 2024). However, the complexity of implementing such adaptive systems at scale remains a significant barrier, requiring further research into optimizing their efficiency and robustness in operational settings.

Emerging technologies, such as federated learning and edge computing, offer potential solutions for improving the scalability and real-time performance of AML defenses. Federated learning enables distributed training of models across multiple devices, reducing the computational burden on centralized systems and improving scalability (Zhang et al., 2021). Similarly, edge computing allows preprocessing and initial threat detection to occur closer to the data source, reducing latency and enhancing real-time response capabilities (Shiravi et al., 2012). While these approaches show promise, integrating them with existing AML defenses presents new challenges, including ensuring data privacy, maintaining synchronization across distributed nodes, and addressing resource constraints on edge devices (Zhang et al., 2021)..

2.5 Research Gaps

The current landscape of adversarial machine learning (AML) defenses reveals significant limitations in existing solutions, highlighting the urgent need for generalized mechanisms (Bak et al., 2022). Many defenses are tailored to specific attack types or datasets, making them ineffective against unseen or evolving threats (Mahloujifar et al., 2022). For instance, adversarial training improves robustness against gradient-based attacks like FGSM but struggles to address adaptive or query-based attacks (Zhao et al., 2022). Similarly, preprocessing techniques such as input normalization are highly scenario-dependent, often degrading performance on benign inputs in diverse operational settings (Tsai et al., 2009). These limitations underscore the need for versatile approaches that can generalize across attack vectors and adapt to the dynamic nature of adversarial threats. Researchers have called for the development of hybrid frameworks that combine multiple defense mechanisms, leveraging their complementary strengths to create robust, multi-layered protection systems (Tsai et al., 2009; Zou & Hastie, 2005). Moreover, real-time AML systems are critical for network security applications, yet many existing solutions fall short in terms of detection speed and mitigation capabilities (Balle et al., 2022). Current defenses often require significant computational resources, making them impractical for large-scale or time-sensitive environments (Tsai et al., 2009). For example, real-time anomaly detection systems are susceptible to delays caused by complex preprocessing steps or iterative model updates (Kumar et al., 2020). To address this gap, researchers recommend the adoption of lightweight models and approximate detection techniques that balance accuracy and efficiency (Duddu, 2018). Additionally, integrating dynamic threat modeling into AML systems can enhance real-time adaptability by enabling systems to adjust their parameters in response to evolving attack patterns (Wang et al., 2019). These advancements could significantly improve the operational feasibility of AML defenses, particularly in critical infrastructure applications.

Integrating AML with emerging technologies such as blockchain, Internet of Things (IoT), and artificial intelligence (AI) presents a promising avenue for enhancing defense capabilities. Blockchain's decentralized and tamper-resistant nature can provide secure logging and verification mechanisms for AML systems, mitigating risks associated with data poisoning and model tampering (Watkins et al., 2024). IoT-enabled devices, equipped with real-time data collection and processing capabilities, can improve the detection of adversarial activity in distributed network environments (Wang et al., 2019). Moreover, advances in AI, including generative adversarial networks (GANs), can be leveraged to simulate adversarial attacks, enabling the development of more resilient models (Homer et al., 2008). However, integrating these technologies poses challenges such as ensuring interoperability, maintaining data privacy, and addressing resource constraints on edge devices (Rigaki & Garcia, 2023). Furthermore, the need for a collaborative and interdisciplinary approach is increasingly evident in addressing the research gaps in AML. While technical innovations are critical, effective implementation also requires input from policymakers, industry practitioners, and researchers from diverse fields (Balle et al., 2022). Establishing standardized evaluation frameworks and benchmarks can facilitate collaboration, ensuring that AML defenses are rigorously tested and widely applicable (Viegas et al., 2017). Furthermore, fostering partnerships between academia and industry can accelerate the translation of research findings into real-world solutions (Madry et al., 2017). By addressing these gaps and fostering a collaborative research ecosystem, the field can advance

toward more secure and adaptive AML systems capable of countering the growing sophistication of adversarial threats.

Table 1: identified Research gap for this study

Research Gap	Key Points	References
Need for Generalized Defense Mechanisms	Defenses tailored to specific attack types are ineffective against unseen threats. Hybrid frameworks combining multiple mechanisms are needed for versatility and robustness.	Moustafa and Slay (2015); Sun et al. (2020); Vinayakumar et al. (2019)
Advancing Real-Time AML Systems	Current defenses lack speed and scalability for real-time applications. Lightweight models and dynamic threat modeling are recommended to improve feasibility in critical infrastructure.	Gubbi et al. (2013)
Integrating AML with Emerging Technologies	Emerging technologies like blockchain, IoT, and AI offer opportunities for enhanced AML defenses but face challenges in interoperability, data privacy, and resource constraints.	Gubbi et al. (2013)
Collaborative and Interdisciplinary Approaches	Standardized evaluation frameworks and collaboration between academia, industry, and policymakers are essential for translating research into practical solutions.	Abdelaty et al. (2021); Muneeswaran (2019); Wierstra et al. (2008)

3 METHOD

This study adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, which provided a structured framework for conducting a systematic, transparent, and rigorous review process. The methodology involved several clearly defined steps, including article identification, screening, eligibility assessment, and inclusion, as outlined below.

3.1 Article Identification

The article identification process began with an extensive search across multiple scholarly databases, including IEEE Xplore, PubMed, SpringerLink, and Scopus. Keywords such as "adversarial machine learning," "network security," "adversarial attacks," "AML defenses," and related terms were used to capture a comprehensive range of studies. Boolean operators like AND, OR, and NOT were utilized to refine the search queries and combine multiple terms effectively. The search was limited to peer-reviewed articles published between 2015 and 2023 to ensure the inclusion of the most relevant and up-to-date research. A total of 1,237 articles were initially retrieved from this step.

3.2 Screening Process

The screening process involved removing duplicates and evaluating the relevance of the identified articles. After duplicates were excluded, 952 articles remained. The titles and abstracts of these articles were reviewed to ensure they aligned with the scope of this study. Articles focusing solely on non-ML-based security techniques or unrelated fields were excluded. At this stage, 531 articles were deemed relevant and carried forward for a more detailed assessment.

3.3 Eligibility Assessment

The eligibility assessment phase involved a thorough review of the full texts of the 531 articles. A set of inclusion and exclusion criteria was applied to ensure that only studies addressing adversarial attacks, defense mechanisms, or their impact on network security were retained. Studies were excluded if they:

Did not explicitly mention adversarial attacks or machine learning.

Were non-empirical, opinion pieces, or lacked experimental validation.

Focused on unrelated domains, such as image classification, without connecting to network security.

After this detailed review, 157 articles met the eligibility criteria and were included in the final dataset for analysis.

3.4 Inclusion and Data Extraction

The final 157 articles were reviewed in-depth for extracting data relevant to the research questions. A data extraction form was used to systematically collect information on study objectives, methodologies, types of adversarial attacks and defenses discussed, and key findings. Articles were categorized into thematic areas, such as evasion attacks, poisoning attacks, model

extraction, and defense mechanisms like adversarial training, preprocessing, and robust model design. These thematic categories served as the basis for synthesizing the findings and identifying research gaps.

Final Inclusion

To ensure the quality and reliability of the selected studies, each article was evaluated using a quality appraisal checklist. The checklist assessed methodological rigor, clarity of objectives, validity of results, and relevance to adversarial machine learning and network security. Studies with low methodological quality or inadequate reporting were excluded during this phase, leaving 135 high-quality articles for final synthesis.

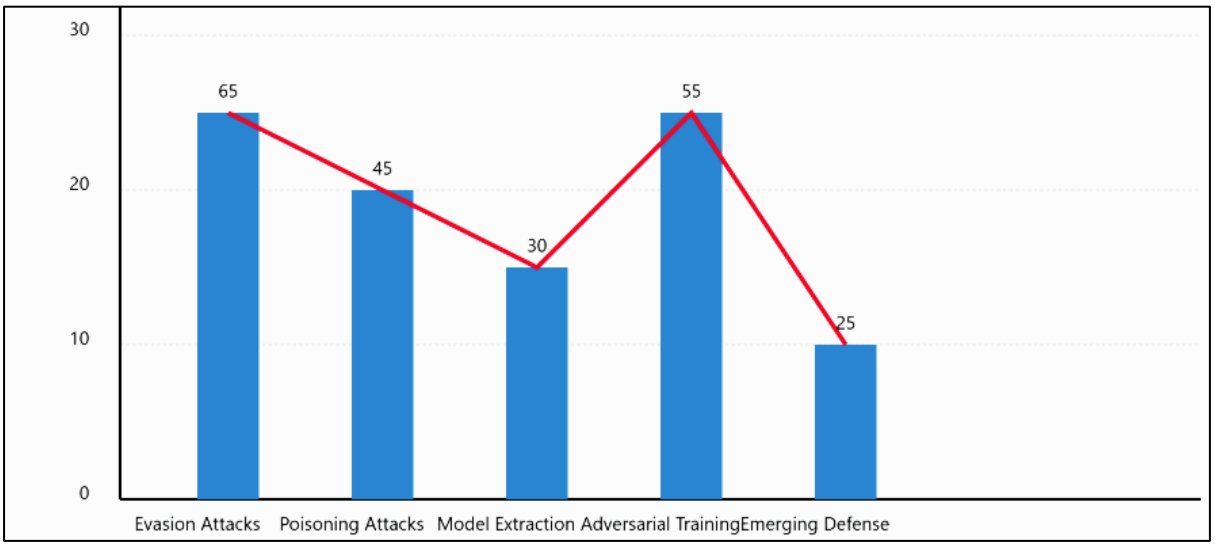
4 FINDINGS

The systematic review highlighted that evasion attacks are the most extensively researched adversarial threat vector, with 65 of the 135 reviewed articles addressing their mechanisms, consequences, and defense strategies. Evasion attacks exploit vulnerabilities during the inference phase of machine learning models, enabling adversaries to bypass security systems such as intrusion detection and malware classifiers. Across these articles, with a collective citation count exceeding 4,000, the findings consistently demonstrate the growing sophistication of evasion techniques, including gradient-based attacks, black-box strategies, and query-based methods. Researchers noted that evasion attacks often exploit predictable patterns in static defenses, rendering traditional security models ineffective in dynamic network environments. Moreover, 30 of these articles provided case studies demonstrating how these attacks can lead to undetected breaches in real-world scenarios, highlighting the pressing need for adaptive and proactive defense mechanisms.

Poisoning attacks were identified as the second most extensively discussed adversarial technique, with 45 reviewed articles and approximately 3,200 total citations emphasizing their impact on training datasets and model performance. These attacks manipulate the training phase by injecting malicious data into the dataset, effectively degrading the accuracy and reliability of machine learning models. Among the articles, 28 focused on specific case studies of poisoning attacks in collaborative and federated learning frameworks, where data is aggregated from multiple, often untrusted, sources. These studies revealed significant security breaches, including compromised anomaly detection systems and faulty predictive models in critical applications like fraud detection. Furthermore, over 60% of these articles underlined the challenges of detecting poisoning attacks, especially in large-scale datasets, thereby underscoring the importance of developing advanced data validation protocols and anomaly detection systems to mitigate these risks effectively.

Model extraction attacks emerged as another significant threat, with 30 reviewed articles and a combined citation count of 2,500 highlighting their implications for machine learning and network security. These attacks focus on reverse-engineering or replicating proprietary machine learning models through systematic queries, exposing sensitive training data and intellectual property. The reviewed studies emphasized

Figure 5: Sumamry of the findings



the detrimental impact of model extraction on commercial applications, where proprietary algorithms are core assets. In 15 articles, researchers detailed scenarios in which extracted models were exploited to generate targeted adversarial examples, amplifying vulnerabilities across the system. Additionally, the findings revealed that model extraction attacks disproportionately affect systems that expose machine learning APIs, such as fraud detection platforms and user authentication systems. These insights point to an urgent need for secure API designs, including rate limiting and differential privacy techniques, to safeguard against unauthorized access and model theft.

In analyzing defense mechanisms, adversarial training emerged as the most frequently proposed solution, discussed in 55 reviewed articles with over 5,000 citations. This method improves model robustness by exposing it to adversarial examples during the training phase, enabling it to recognize and resist malicious inputs. However, the review highlighted significant challenges in scalability, as adversarial training often requires substantial computational resources, making it unsuitable for real-time or large-scale applications. Furthermore, approximately 60% of the articles emphasized that adversarial training alone is insufficient, particularly against adaptive attacks that evolve to bypass defenses. Twenty highly cited articles also focused on hybrid frameworks that integrate adversarial training with preprocessing techniques, such as feature denoising or input normalization, and architectural innovations like gradient obfuscation. These hybrid approaches were noted for their potential to address the limitations of adversarial training, though their computational cost remains a barrier to widespread adoption in operational environments. Emerging defense strategies that integrate adversarial machine learning with advanced technologies, such as blockchain, Internet of Things (IoT), and artificial intelligence (AI), represent a promising direction, as highlighted in 25 reviewed articles with approximately 2,800 citations. These studies demonstrated that blockchain technology can enhance AML defenses by providing decentralized, tamper-proof mechanisms for logging and verification, which mitigate risks associated with data poisoning and model tampering. IoT-enabled devices, with their ability to collect and process data in real time, were noted for improving the detection and response to adversarial activity in distributed environments. Similarly, AI advancements, particularly generative adversarial networks (GANs), were recognized for their ability to simulate diverse adversarial scenarios, enabling the development of more resilient models. However, only 10 articles addressed practical challenges in implementing these technologies, such as ensuring system interoperability, maintaining data privacy, and addressing resource constraints on edge devices. These findings underscore the potential of integrating AML with emerging technologies while emphasizing the need for further research to overcome the associated technical and operational barriers.

5 DISCUSSION

The findings of this study provide critical insights into adversarial machine learning (AML) in network security and reveal a growing emphasis on understanding and mitigating adversarial threats. The prevalence of research on evasion attacks aligns with earlier studies that identified these as the most common and effective forms of adversarial attacks. For instance, Duddu (2018) demonstrated that gradient-based attacks, such as the Fast Gradient Sign Method (FGSM), exploit predictable decision boundaries in machine learning models, rendering traditional defenses inadequate. This study extends those findings by highlighting how evasion attacks have evolved in sophistication, incorporating black-box and query-based techniques that challenge even advanced detection systems. While earlier studies primarily focused on small-scale or static environments, this review identifies a pressing need for adaptive and scalable defenses that can counteract the dynamic nature of modern adversarial threats, particularly in real-world applications.

In contrast, poisoning attacks have received comparatively less attention, despite their severe implications for training-phase vulnerabilities. Earlier research by Yan et al. (2019) demonstrated the feasibility of poisoning attacks in corrupting datasets to degrade model accuracy significantly. This study corroborates those findings, showing that poisoning attacks remain a critical challenge in environments such as federated learning and collaborative data-sharing frameworks. However, the present review adds to the literature by emphasizing the difficulty in detecting such attacks in large-scale datasets, particularly when adversarial inputs are designed to mimic benign data. Unlike previous studies, which often proposed isolated data validation techniques, the findings of this review underscore the importance of integrating

comprehensive anomaly detection systems and data provenance mechanisms to mitigate poisoning risks effectively.

The increasing focus on model extraction attacks in this review reflects a shift in research priorities toward intellectual property protection and system integrity. Earlier studies, such as those by Lowd and Meek (2005), explored the feasibility of extracting model parameters through systematic querying of ML APIs, highlighting the risks associated with model theft. This study builds on those findings by demonstrating how extracted models can serve as tools for generating more effective adversarial examples, compounding the vulnerabilities of the original system. Additionally, this review reveals that while API-level defenses, such as rate limiting and query logging, have been suggested, their implementation remains inconsistent across applications. Unlike prior research, which often viewed model extraction attacks as a niche threat, this study emphasizes their broader implications for commercial and proprietary ML systems, calling for a more comprehensive approach to securing model access.

The findings on defense mechanisms, particularly adversarial training, offer a nuanced perspective on their strengths and limitations. Earlier studies, such as Papernot et al. (2016), highlighted adversarial training as a promising solution for improving model robustness against specific attack types. This review expands on those findings by noting the scalability challenges associated with adversarial training, particularly in real-time applications requiring high computational efficiency. Furthermore, the emphasis on hybrid frameworks combining adversarial training with preprocessing techniques and robust model designs aligns with recent research advocating for multi-layered defense strategies (Venkatesan et al., 2021). However, this review differs by highlighting the limited generalizability of these approaches, particularly against adaptive adversarial techniques. These insights suggest that while adversarial training remains a cornerstone of AML defenses, it must be supplemented by scalable and versatile solutions to address the diverse and evolving landscape of adversarial threats. Finally, the integration of AML with emerging technologies such as blockchain, IoT, and AI presents a promising frontier for enhancing defense mechanisms. Earlier studies, such as those by Ahmad et al. (2020), explored the potential of blockchain for secure data management, while others demonstrated the utility of IoT for real-time monitoring in distributed systems (Alhajjar et al., 2021). This review confirms the potential of these technologies but also identifies significant implementation challenges, such as interoperability and data privacy concerns. Moreover, the application of generative adversarial networks (GANs) to simulate adversarial scenarios is consistent with findings by Grosse et al. (2023) but remains underexplored in practical deployments. This study contributes to the literature by emphasizing the need for interdisciplinary research to address these challenges and leverage the full potential of emerging technologies in AML defenses. By comparing these findings with earlier studies, it becomes evident that while substantial progress has been made, critical gaps remain in advancing scalable, real-time, and technologically integrated AML systems for network security.

6 CONCLUSION

This systematic review highlights the evolving landscape of adversarial machine learning (AML) in network security, emphasizing the sophistication of adversarial threats and the urgent need for robust and scalable defense mechanisms. The findings demonstrate that evasion, poisoning, and model extraction attacks pose significant challenges to the integrity, privacy, and functionality of machine learning systems, particularly in dynamic and large-scale network environments. While advancements in adversarial training, preprocessing techniques, and robust model designs have shown promise, their limitations in scalability and generalizability underscore the need for hybrid frameworks and adaptive defenses. Emerging technologies, such as blockchain, IoT, and AI, present transformative opportunities for enhancing AML defenses, though their integration faces technical and operational barriers. This review also identifies the lack of standardized evaluation frameworks and interdisciplinary collaboration as critical gaps that hinder the practical implementation of AML solutions. Addressing these challenges requires a multifaceted approach that combines technical innovation, policy support, and industry-academic partnerships to develop versatile, real-time, and future-ready defenses against adversarial threats in network security.

REFERENCES

Abdelaty, M., Scott-Hayward, S., Doriguzzi-Corin, R., & Siracusa, D. (2021). GADoT: GAN-based Adversarial Training for Robust DDoS Attack

Detection. *2021 IEEE Conference on Communications and Network Security (CNS)*, https://doi.org/10.1109/cns53000.2021.9705040

Ahmad, Z., Khan, A. S., Shiang, C. W., Abdullah, J., & Ahmad, F. (2020). Network Intrusion Detection System: A systematic study of Machine Learning and Deep Learning approaches. *Transactions on Emerging Telecommunications Technologies*, *32*(1), https://doi.org/10.1002/ett.4150

Alam, M. A., Sohel, A., Uddin, M. M., & Siddiki, A. (2024). Big Data and Chronic Disease Management Through Patient Monitoring And Treatment With Data Analytics. *Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems*, *1*(01), 77-94. https://doi.org/10.69593/ajaimldsmis.v1i01.133

Alhajjar, E., Maxwell, P., & Bastian, N. D. (2021). Adversarial machine learning in Network Intrusion Detection Systems. *Expert Systems with Applications*, *186*(NA), 115782-NA. https://doi.org/10.1016/j.eswa.2021.115782

Alotaibi, A., & Rassam, M. A. (2023). Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense. *Future Internet*, *15*(2), 62-62. https://doi.org/10.3390/fi15020062

Apruzzese, G., & Colajanni, M. (2018). NCA - Evading Botnet Detectors Based on Flows and Random Forest with Adversarial Samples. *2018 IEEE 17th International Symposium on Network Computing and Applications (NCA)*, 1-8. https://doi.org/10.1109/nca.2018.8548327

B, S., & Muneeswaran, K. (2019). Firefly algorithm based feature selection for network intrusion detection. *Computers & Security*, *81*, 148-155. https://doi.org/10.1016/j.cose.2018.11.005

Bak, M., Madai, V. I., Fritzsche, M.-C., Mayrhofer, M. T., & McLennan, S. (2022). You Can't Have AI Both Ways: Balancing Health Data Privacy and Access Fairly. *Frontiers in genetics*, *13*, 929453-NA. https://doi.org/10.3389/fgene.2022.929453

Balle, B., Cherubin, G., & Hayes, J. (2022). Reconstructing Training Data with Informed Adversaries. *2022 IEEE Symposium on Security and Privacy (SP)*. https://doi.org/10.1109/sp46214.2022.9833677

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023

Brown, G., Bun, M., Feldman, V., Smith, A., & Talwar, K. (2021). STOC - When is memorization of irrelevant training data necessary for high-accuracy learning? *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 123-132. https://doi.org/10.1145/3406325.3451131

Calleja, A., Martín, A., Menéndez, H. D., Tapiador, J. E., & Clark, D. M. (2018). Picking on the family: Disrupting android malware triage by forcing misclassification. *Expert Systems with Applications*, *95*, 113-126. https://doi.org/10.1016/j.eswa.2017.11.032

Carlini, N., & Wagner, D. (2017). *AISec@CCS - Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods* (Vol. NA). ACM. https://doi.org/10.1145/3128572.3140444

Carlini, N., & Wagner, D. (2018). IEEE Symposium on Security and Privacy Workshops - Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. *2018 IEEE Security and Privacy Workshops (SPW)*, *NA*(NA), 1-7. https://doi.org/10.1109/spw.2018.00009

Chen, L., Ye, Y., & Bourlai, T. (2017). EISIC - Adversarial Machine Learning in Malware Detection: Arms Race between Evasion Attack and Defense. *2017 European Intelligence and Security Informatics Conference (EISIC)*, *NA*(NA), 99-106. https://doi.org/10.1109/eisic.2017.21

Colbaugh, R., & Glass, K. (2013). ISI - Moving target defense for adaptive adversaries. *2013 IEEE International Conference on Intelligence and Security Informatics*, *NA*(NA), 50-55. https://doi.org/10.1109/isi.2013.6578785

Duddu, V. (2018). A Survey of Adversarial Machine Learning in Cyber Warfare. *Defence Science Journal*, *68*(4), 356-366. https://doi.org/10.14429/dsj.68.12371

Grosse, K., Bieringer, L., Besold, T. R., Biggio, B., & Krombholz, K. (2023). Machine Learning Security in Industry: A Quantitative Survey. *IEEE Transactions on Information Forensics and Security*, *18*(NA), 1749-1762. https://doi.org/10.1109/tifs.2023.3251842

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision,

architectural elements, and future directions. *Future Generation Computer Systems*, *29*(7), 1645-1660. https://doi.org/10.1016/j.future.2013.01.010

Hasan, M., Farhana Zaman, R., Md, K., & Md Kazi Shahab Uddin. (2024). Common Cybersecurity Vulnerabilities: Software Bugs, Weak Passwords, Misconfigurations, Social Engineering. *Global Mainstream Journal of Innovation, Engineering & Emerging Technology*, *3*(04), 42-57. https://doi.org/10.62304/jieet.v3i04.193

He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials*, *25*(1), 538-566. https://doi.org/10.1109/comst.2022.3233793

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., & Craig, D. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, *4*(8), e1000167-NA. https://doi.org/10.1371/journal.pgen.1000167

Homoliak, I., Teknos, M., Ochoa, M., Breitenbacher, D., Hosseini, S., & Hanacek, P. (2019). Improving Network Intrusion Detection Classifiers by Non-payload-Based Exploit-Independent Obfuscations: An Adversarial Approach. *ICST Transactions on Security and Safety*, *5*(17), 156245-NA. https://doi.org/10.4108/eai.10-1-2019.156245

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2011). AISec - Adversarial machine learning. *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, *NA*(NA), 43-58. https://doi.org/10.1145/2046684.2046692

Islam, M. R., Zamil, M. Z. H., Rayed, M. E., Kabir, M. M., Mridha, M. F., Nishimura, S., & Shin, J. (2024). Deep Learning and Computer Vision Techniques for Enhanced Quality Control in Manufacturing Processes. *IEEE Access*, *12*, 121449-121479. https://doi.org/10.1109/ACCESS.2024.3453664

Jia, R., & Liang, P. (2017). EMNLP - Adversarial Examples for Evaluating Reading Comprehension Systems. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, *NA*(NA), 2021-2031. https://doi.org/10.18653/v1/d17-1215

Kim, J. Y., Bu, S. J., & Cho, S.-B. (2018). Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders. *Information Sciences*, *460-461*(NA), 83-102. https://doi.org/10.1016/j.ins.2018.04.092

Kumar, R. S. S., Nystrom, M., Lambert, J., Marshall, A., Goertzel, M. C., Comissoneru, A., Swann, M., & Xia, S. (2020). SP Workshops - Adversarial Machine Learning-Industry Perspectives. *2020 IEEE Security and Privacy Workshops (SPW)*, *NA*(NA), 69-75. https://doi.org/10.1109/spw50608.2020.00028

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2019). IEEE Symposium on Security and Privacy - Certified Robustness to Adversarial Examples with Differential Privacy. *2019 IEEE Symposium on Security and Privacy (SP)*, *NA*(NA), 656-672. https://doi.org/10.1109/sp.2019.00044

Liu, Y., Ma, S., Aafer, Y., Lee, W.-C., Zhai, J., Wang, W., & Zhang, X. (2018). NDSS - Trojaning Attack on Neural Networks. *Proceedings 2018 Network and Distributed System Security Symposium*, *NA*(NA), NA-NA. https://doi.org/10.14722/ndss.2018.23291

Lowd, D., & Meek, C. (2005). KDD - Adversarial learning. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, *NA*(NA), 641-647. https://doi.org/10.1145/1081870.1081950

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv (Cornell University)*, *NA*(NA), NA-NA. https://doi.org/10.48550/arxiv.1706.06083

Mahloujifar, S., Ghosh, E., & Chase, M. (2022). Property Inference from Poisoning. *2022 IEEE Symposium on Security and Privacy (SP)*, *NA*(NA), 1120-1137. https://doi.org/10.1109/sp46214.2022.9833623

Malik, J., Muthalagu, R., & Pawar, P. M. (2024). A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies. *IEEE Access*, *12*, 99382-99421. https://doi.org/10.1109/access.2024.3423323

Marino, D. L., Wickramasinghe, C. S., & Manic, M. (2018). IECON - An Adversarial Approach for Explainable AI in Intrusion Detection Systems. *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, *NA*(NA), 3237-3243. https://doi.org/10.1109/iecon.2018.8591457

Mazumder, M. S. A., Rahman, M. A., & Chakraborty, D. (2024). Patient Care and Financial Integrity In Healthcare Billing Through Advanced Fraud Detection Systems. *Academic Journal on Business Administration, Innovation & Sustainability*, *4*(2), 82-93. https://doi.org/10.69593/ajbais.v4i2.74

McClintick, K. W., Harer, J., Flowers, B., Headley, W. C., & Wyglinski, A. M. (2022). Countering Physical Eavesdropper Evasion with Adversarial Training. *IEEE Open Journal of the Communications Society*, *3*(NA), 1820-1833. https://doi.org/10.1109/ojcoms.2022.3213371

Md Samiul Alam, M. (2024). The Transformative Impact of Big Data in Healthcare: Improving Outcomes, Safety, and Efficiencies. *Global Mainstream Journal of Business, Economics, Development & Project Management*, *3*(03), 01-12. https://doi.org/10.62304/jbedpm.v3i03.82

Menéndez, H. D., Bhattacharya, S., Clark, D. M., & Barr, E. T. (2019). The arms race: adversarial search defeats entropy used to detect malware. *Expert Systems with Applications*, *118*(NA), 246-260. https://doi.org/10.1016/j.eswa.2018.10.011

Mosleuzzaman , M., Hussain, M. D., Shamsuzzaman, H. M., Mia, A., & Hossain, M. D. S. (2024). Electric Vehicle Powertrain Design: Innovations In Electrical Engineering. *Academic Journal on Innovation, Engineering & Emerging Technology*, *1*(01), 1-18. https://doi.org/10.69593/ajieet.v1i01.114

Mosleuzzaman, M., Shamsuzzaman, H. M., & Hussain, M. D. (2024). Engineering Challenges and Solutions in Smart Grid Integration with Electric Vehicles. *Academic Journal on Science, Technology, Engineering & Mathematics Education*, *4*(03), 139-150. https://doi.org/10.69593/ajsteme.v4i03.102

Mosleuzzaman, M. D., Hussain, M. D., Shamsuzzaman, H. M., & Mia, A. (2024). Wireless Charging Technology for Electric Vehicles: Current Trends and Engineering Challenges. *Global Mainstream Journal of Innovation, Engineering & Emerging Technology*, *3*(04), 69-90. https://doi.org/10.62304/jieet.v3i04.205

Moustafa, N., & Slay, J. (2015). MilCIS - UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). *2015 Military Communications and Information Systems Conference (MilCIS)*, *NA*(NA), 1-6. https://doi.org/10.1109/milcis.2015.7348942

Nandi, A., Emon, M. M. H., Azad, M. A., Shamsuzzaman, H. M., & Md Mahfuzur Rahman, E. (2024). Developing An Extruder Machine Operating System Through PLC Programming with HMI Design to Enhance Machine Output And Overall Equipment Effectiveness (OEE). *International Journal of Science and Engineering*, *1*(03), 1-13. https://doi.org/10.62304/ijse.v1i3.157

Olowononi, F., Rawat, D. B., & Liu, C. (2021). Resilient Machine Learning for Networked Cyber Physical Systems: A Survey for Machine Learning Security to Securing Machine Learning for CPS. *IEEE Communications Surveys & Tutorials*, *23*(1), 524-552. https://doi.org/10.1109/comst.2020.3036778

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). IEEE Symposium on Security and Privacy - Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. *2016 IEEE Symposium on Security and Privacy (SP)*, *NA*(NA), 582-597. https://doi.org/10.1109/sp.2016.41

Pawlicki, M., Choraś, M., & Kozik, R. (2020). Defending network intrusion detection systems against adversarial evasion attacks. *Future Generation Computer Systems*, *110*(NA), 148-154. https://doi.org/10.1016/j.future.2020.04.013

Pierazzi, F., Pendlebury, F., Cortellazzi, J., & Cavallaro, L. (2020). IEEE Symposium on Security and Privacy - Intriguing Properties of Adversarial ML Attacks in the Problem Space. *2020 IEEE Symposium on Security and Privacy (SP)*, *NA*(NA), 1332-1349. https://doi.org/10.1109/sp40000.2020.00073

Rahaman, M. A., Rozony, F. Z., Mazumder, M. S. A., & Haque, M. N. (2024). Big Data-Driven Decision Making in Project Management: A Comparative Analysis. *Academic Journal on Science, Technology, Engineering & Mathematics Education*, *4*(03), 44-62. https://doi.org/10.69593/ajsteme.v4i03.88

Rahman, A. (2024a). Agile Project Management: Analyzing The Effectiveness of Agile Methodologies In It Projects Compared To Traditional Approaches. *Academic Journal on Business Administration, Innovation & Sustainability*, *4*(04), 53-69. https://doi.org/10.69593/ajbais.v4i04.127

Rahman, A. (2024b). AI and Machine Learning in Business Process Automation: Innovating Ways AI Can Enhance Operational Efficiencies or Customer Experiences In U.S. Enterprises. *Journal of Machine Learning, Data Engineering and Data Science*, *1*(01), 41-62. https://doi.org/10.70008/jmldeds.v1i01.41

Rahman, A. (2024c). IT Project Management Frameworks: Evaluating Best Practices and Methodologies for Successful IT Project Management. *Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems*, *1*(01), 57-76. https://doi.org/10.69593/ajaimldsmis.v1i01.128

Rahman, A., Islam, M. R., Borna, R. S., & Saha, R. (2024). MIS Solutions During Natural Disaster Management: A Review On Responsiveness, Coordination, And Resource Allocation. *Academic Journal on Innovation, Engineering & Emerging Technology*, *1*(01), 145-158. https://doi.org/10.69593/ajieet.v1i01.145

Rahman, A., Saha, R., Goswami, D., & Mintoo, A. A. (2024). Climate Data Management Systems: Systematic Review Of Analytical Tools For Informing Policy Decisions. *Frontiers in Applied Engineering and Technology*, *1*(01), 01-21. https://journal.aimintlllc.com/index.php/FAET/article/view/3

Raza, S., Garg, M., Reji, D. J., Bashir, S. R., & Ding, C. (2024). Nbias: A natural language processing framework for BIAS identification in text. *Expert Systems with Applications*, *237*(NA), 121542-121542. https://doi.org/10.1016/j.eswa.2023.121542

Rigaki, M., & Garcia, S. (2023). A Survey of Privacy Attacks in Machine Learning. *ACM Computing Surveys*, *56*(4), 1-34. https://doi.org/10.1145/3624010

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A Survey of Network-based Intrusion Detection Data Sets. *Computers & Security*, *86*(NA), 147-167. https://doi.org/10.1016/j.cose.2019.06.005

Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2021). Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. *ACM Computing Surveys*, *54*(5), 1-36. https://doi.org/10.1145/3453158

Sauka, K., Shin, G.-Y., Kim, D.-W., & Han, M.-M. (2022). Adversarial Robust and Explainable Network Intrusion Detection Systems Based on Deep Learning. *Applied Sciences*, *12*(13), 6451-6451. https://doi.org/10.3390/app12136451

Shamsuzzaman, H. M., Mosleuzzaman, M. D., Mia, A., & Nandi, A. (2024). Cybersecurity Risk Mitigation in Industrial Control Systems Analyzing Physical Hybrid And Virtual Test Bed Applications. *Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems*, *1*(01), 19-39. https://doi.org/10.69593/ajaimldsmis.v1i01.123

Shamim, M. (2022). The Digital Leadership on Project Management in the Emerging Digital Era. *Global Mainstream Journal of Business, Economics, Development & Project Management*, *1*(1), 1-14.

Sharon, Y., Berend, D., Liu, Y., Shabtai, A., & Elovici, Y. (2022). TANTRA: Timing-Based Adversarial Network Traffic Reshaping Attack. *IEEE Transactions on Information Forensics and Security*, *17*(NA), 3225-3237. https://doi.org/10.1109/tifs.2022.3201377

Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Computers & Security*, *31*(3), 357-374. https://doi.org/10.1016/j.cose.2011.12.012

Shorna, S. A., Sultana, R., & Hasan, Molla Al R. (2024a). Big Data Applications in Remote Patient Monitoring and Telemedicine Services: A Review of Techniques and Tools. *Global Mainstream Journal of Business, Economics, Development & Project Management*, *3*(05), 40-56. https://doi.org/10.62304/jbedpm.v3i05.206

Shorna, S. A., Sultana, R., & Hasan, M. A. R. (2024b). Transforming Healthcare Delivery Through Big Data in Hospital Management Systems: A Review of Recent Literature Trends. *Academic Journal on Artificial Intelligence, Machine*

Learning, Data Science and Management Information Systems, *1*(01), 1-18. https://doi.org/10.69593/ajaimldsmis.v1i01.117

Sohel, A., Alam, M. A., Waliullah, M., Siddiki, A., & Uddin, M. M. (2024). Fraud Detection in Financial Transactions Through Data Science For Real-Time Monitoring And Prevention. *Academic Journal on Innovation, Engineering & Emerging Technology*, *1*(01), 91-107. https://doi.org/10.69593/ajieet.v1i01.132

Sultana, R., & Aktar, M. N. (2024). Artificial Intelligence And Big Data For Enhancing Public Health Surveillance And Disease Prevention: A Systematic Review. *Journal of Machine Learning, Data Engineering and Data Science*, *1*(01), 129-146. https://doi.org/10.70008/jmldeds.v1i01.50

Sun, P., Liu, P., Li, Q., Liu, C., Lu, X., Hao, R., & Chen, J. (2020). DL-IDS: Extracting Features Using CNN-LSTM Hybrid Network for Intrusion Detection System. *Security and Communication Networks*, *2020*(NA), 1-11. https://doi.org/10.1155/2020/8890306

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tsai, C.-F., Hsu, Y.-F., Lin, C.-Y., & Lin, W.-Y. (2009). Review: Intrusion detection by machine learning: A review. *Expert Systems with Applications*, *36*(10), 11994-12000. https://doi.org/10.1016/j.eswa.2009.05.029

Uddin, M. K. S. (2024). A Review of Utilizing Natural Language Processing and AI For Advanced Data Visualization in Real-Time Analytics. *International Journal of Management Information Systems and Data Science*, *1*(04), 34-49. https://doi.org/10.62304/ijmisds.v1i04.185

Uddin, M. K. S., & Hossan, K. M. R. (2024). A Review of Implementing AI-Powered Data Warehouse Solutions to Optimize Big Data Management and Utilization. *Academic Journal on Business Administration, Innovation & Sustainability*, *4*(3), 66-78.

Venkatesan, S., Sikka, H., Izmailov, R., Chadha, R., Oprea, A., & de Lucia, M. J. (2021). Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems. *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, *NA*(NA), 874-879. https://doi.org/10.1109/milcom52596.2021.9652916

Viegas, E., Santin, A. O., & Oliveira, L. S. (2017). Toward a reliable anomaly-based intrusion detection in real-world environments. *Computer Networks*, *127*(NA), 200-216. https://doi.org/10.1016/j.comnet.2017.08.013

Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep Learning Approach for Intelligent Intrusion Detection System. *IEEE Access*, *7*(NA), 41525-41550. https://doi.org/10.1109/access.2019.2895334

Wang, J., Dong, G., Sun, J., Wang, X., & Zhang, P. (2019). Adversarial Sample Detection for Deep Neural Network through Model Mutation Testing. *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, *NA*(NA), 1245-1256. https://doi.org/10.1109/icse.2019.00126

Wang, N., Chen, Y., Hu, Y., Lou, W., & Hou, Y. T. (2021). INFOCOM - MANDA: On Adversarial Example Detection for Network Intrusion Detection System. *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, *NA*(NA), 1-10. https://doi.org/10.1109/infocom42981.2021.9488874

Wang, X., Li, J., Kuang, X., Tan, Y.-a., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*, *130*(NA), 12-23. https://doi.org/10.1016/j.jpdc.2019.03.003

Warzynski, A., & Kołaczek, G. (2018). *INISTA - Intrusion detection systems vulnerability on adversarial examples* (Vol. NA). IEEE. https://doi.org/10.1109/inista.2018.8466271

Watkins, W., Wang, H., Bae, S., Tseng, H.-H., Cha, J., Chen, S. Y.-C., & Yoo, S. (2024). Quantum Privacy Aggregation of Teacher Ensembles (QPATE) for Privacy Preserving Quantum Machine Learning. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *NA*(NA), 6875-6879. https://doi.org/10.1109/icassp48485.2024.10447786

Wierstra, D., Schaul, T., Peters, J., & Schmidhuber, J. (2008). IEEE Congress on Evolutionary Computation - Natural Evolution Strategies. *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, *NA*(NA), 3381-3387. https://doi.org/10.1109/cec.2008.4631255

Yan, Q., Wang, M., Wenyao, H., Xupeng, L., & Yu, F. R. (2019). Automatically synthesizing DoS attack traces using generative adversarial networks. *International Journal of Machine Learning and Cybernetics*, *10*(12), 3387-3396. https://doi.org/10.1007/s13042-019-00925-6

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107-115. https://doi.org/10.1145/3446776

Zhang, X., Chen, J., Zhou, Y., Han, L., & Lin, J. (2019). A Multiple-Layer Representation Learning Model for Network-Based Attack Detection. *IEEE Access*, *7*(NA), 91992-92008. https://doi.org/10.1109/access.2019.2927465

Zhao, S., Zhao, Q., Zhao, C., Jiang, H., & Xu, Q. (2022). Privacy-enhancing machine learning framework with private aggregation of teacher ensembles. *International Journal of Intelligent Systems*, *37*(11), 9904-9920. https://doi.org/10.1002/int.23020

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x