Review of Adversarial Attack Mechanisms, Detection Techniques and Defence Strategies in Critical Network Infrastructures

Francis A. Okoye., *Aghaizu Herman Chijioke., Shamsudeen Mohammed S.B

Department of Computer Engineering, Enugu State University of Science and Technology, Enugu, Nigeria

***Corresponding Author**

DOI: https://doi.org/10.51584/IJRIAS.2026.11060294

Received: 29 June 2026; Accepted: 04 July 2026; Published: 20 July 2026

ABSTRACT

Many of the critical networks are now vulnerable to complex security threats, especially those launched by adversaries against the machine learning-driven security systems used by these networks. Such attacks take advantage of weaknesses in AI systems by perturbing the model with carefully designed perturbations, which result in misclassification of malicious content as benign, compromising the system's confidentiality, integrity, and availability. The adversarial threat is unlike traditional cyberattacks; it is dynamic, adaptive and can circumvent traditional intrusion detection capabilities. This paper provides an extensive literature review on the adversarial attack methods, detection and defence techniques of critical network infrastructures. This review includes peer-reviewed publications published between 2019 and 2024 from the leading academic databases such as IEEE Xplore, SpringerLink, ScienceDirect and Google Scholar. The total number of studies analyzed were 48, covering contributions in the fields of creating adversarial attack methods, machine learning and deep learning based detection methods, and mitigation techniques. The results indicate that adversarial attacks can be divided into the following categories: evasion attacks, poisoning attacks, and exploratory attacks, where some of the more sophisticated methods, including those based on gradient, optimization, and reinforcement learning, are very effective in evading security systems. Current solutions, however, suffer from limited real-time adaptability, cross-domain generalization ability, explainability and integration across the attack lifecycle. While there are several defence mechanisms proposed, such as adversarial training, anomaly detection, and input transformation, existing defences have difficulties in being adaptable in real time, cross-domain generalizable, explainable and suitable for certain phases of the attack lifecycle. The study highlights a number of critical research challenges such as the lack of a common defence framework, inadequate real-time detection capabilities, absence of a standardized data sets and poor ability to withstand adaptive adversaries. The paper suggests the creation of multi-strategic, adaptive, and real-time adversarial threat management systems that can sustain themselves in a heterogeneous network environment.

Keywords: Adversarial Machine Learning; Cybersecurity; Critical Infrastructure; Intrusion Detection; Deep Learning

INTRODUCTION

Advanced cyber attacks are increasingly targeting critical network infrastructures like data centres, cloud platforms, and industrial control systems (Aiken and Scott, 2020). Adversarial attacks are among the most harmful, as these attacks are designed to trick the intelligent firewall and machine learning defence systems using subtle manipulation of the input features(Alhajjar *et al*., 2021). With adversarial threats, the attackers take advantage of weaknesses in AI models and have them misclassify malicious traffic as legitimate. This is a threat to the principles of computer security including confidentiality, integrity, and availability and a huge challenge for organizations that depend on automated threat detection (Alotaibi & Rassam, 2023).

Typically, adversarial attacks are created through methods like data poisoning, label flipping, and feature manipulation(Qian *et al*., 2023; Butt *et al*., 2023). The three types of poisoning are corrupted samples injected

into the training data, flipping the ground-truth labels, and bounded perturbations applied to input vectors via feature manipulation. These techniques can be used together to create a malicious threat variant that evades even the most sophisticated intrusion detection systems. These attacks are adaptive, and a multi strategic approach is needed that is more than just single method defences(He *et al*., 2023).

Researchers have suggested approaches for detection and mitigation over the years, from statistical anomaly detection, input transformation, machine learning models, to deep learning models (Apruzzese *et al*., 2019). Gradient Based Adversarial Training (GBAT), Surrogate Model (SM) and Hybrid Neural Networks (HNN) have been found to be promising. Many of the current solutions, however, are limited in their capacity to respond to new strategies used by adversaries, are not adaptive in real time, or do not combine detection and incident response (IR) (Liang *et al*., 2022). These gaps indicate the need for a holistic examination of the existing strategies and the need to create a common multi strategic approach.

The objective of this review paper is to review the techniques that are currently being used for the management of adversarial threats to the critical infrastructure of networks. We study the adversarial attacks generation and existing detection and defence techniques such as statistical methods, machine learning and deep learning techniques and identify the key research gaps (Khazane *et al*., 2024). This review is based on the analysis of the main research done in recent years and will serve as a background on how to build dynamic, real time and multi strategic security systems to effectively mitigate the adversarial threats in modern network environments.

LITERATURE REVIEW

Wallace *et al*., (2024) presents an empirical risk analysis of adversarial threats against critical infrastructure. This study examines the origins and development of these approaches, which were taken from the nuclear and atomic energy sectors. These methods were used as a basis for the risk assessment documents that control security in the water sector. The traditional risk formula of threat multiplied by vulnerability multiplied by consequence is the foundation of the current American National Standards Institute (ANSI) risk models used by the water sector. These models are based on Design Basis Threat (DBT) and Risk Analysis and Management for Critical Asset Protection (RAMCAP). This study found that the requirements listed in these methodologies are unachievable and will remain unknown in industrial systems due to the inability to define the adversary and their goals, motives, and capabilities, as well as the lack of statistically valid datasets or available intelligence of malevolent threats. The Mathematical Analysis of Defence Strategy and Countermeasures (MADSC) technique can be used to determine the probability of success of a particular threat by treating risk as a vector quantity with known parameters. The physical security of other DHS critical infrastructure sectors may be aided by the water sector's adoption of the MADSC technique as a model.

Apruzzese *et al*., (2022) researched on the exposure of adversarial examples to 5G network infrastructures using wild networks. The study presented a brand-new adversarial machine learning threat model that is specifically tailored to 5G situations and is independent of the specific task that machine learning (ML) is intended to accomplish. Because of the open nature of 5G networks and the guarantees of Quality of Service (QoS), attacks against existing ML threat models do not necessitate any compromise of the target 5G system. The work also suggests a novel methodology for practical ML security evaluations using open data. It proactively assesses the threat model on six 5G-envisioned machine learning applications. The attacks have a lower entrance barrier than earlier attacks, impact both the training and inference stages, and have the potential to impair the performance of cutting-edge machine learning systems. The new myopic threat model that this study proposes, which enhances current ML threat models, is its main contribution. Lastly, the study presents a machine learning system that is resistant to attacks by design. When the actual machine learning components in SA 5G are made available for adversarial ML research, this work can serve as an inspiration for a variety of research areas, including case studies.

Lennon (2022) presents a study on the mitigation of adversarial threats in artificial intelligence infrastructure using direct action. The work states that the best way to address the threat of AML is to combine technological solutions with policies that limit public knowledge of the most effective attacks and defences. This is because adversarial learning is a constantly evolving landscape. A thorough handling of this complex problem will encourage more in-depth conversations about striking the right balance between preserving scientific curiosity

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

and the freedom of information sharing and minimising the risk of property or human damage from improper disclosure of sensitive research data. According to the study's findings, society will eventually benefit from a new era of artificially intelligent systems if the goal of secure, dependable machine learning systems is effectively achieved.

Maiorca *et al*., (2020) presents a study towards adversarial malware detection system from pdf-based attacks. The system outlines the primary attack vectors that can be used with PDF files. Second, by emphasising their machine-learning components, the study offers a thorough explanation of cutting-edge PDF detectors. Thirdly, the work demonstrates the many adversarial approaches that can be used to circumvent such systems. The study, in particular, offers a thorough taxonomy of the attacks that can be used against learning-based detectors and describes how these attacks can be used in real-world scenarios. Additionally, this study attempts to offer strong foundations for overcoming obstacles that may arise when operating in hostile contexts.The primary objective of this study is to demonstrate how adversarial machine learning has advanced our understanding of how to apply malware detection in recent years, mostly through the examination of PDF malware.

Li *et al*., (2019) presents a black box empirical study for adversarial example detectors on cloud-based system. The primary focus of this work is on the security concerns associated with cloud-based image detectors in the real world. The work specifically proposed four attacks to generate semantics-aware adversarial examples by interacting with black-box APIs alone, based on effective semantic segmentation. Additionally, this is the first attempt to conduct a comprehensive empirical study of black-box attacks against real-world cloud-based image detectors. The study shows that using image processing-based attacks can achieve a success rate of approximately 100% through thorough evaluations on five major cloud platforms: AWS, Azure, Google Cloud, Baidu Cloud, and Alibaba Cloud. Meanwhile, using semantic segmentation-based attacks has a success rate of over 90% among various detection services, such as politician and violence detection.

Demetrio *et al*., (2019) researched on the vulnerabilities of deep learning to adversarial malware binaries. In this work, explainable machine-learning techniques created to decipher the opaque judgements made by deep neural networks are used as a first step towards addressing this problem. Specifically, the study use an approach called as feature attribution, which is explainable, to determine which input features have the greatest influence on each conclusion and then modifies it to offer insightful justifications for the malware binary classification. In this instance, we discover that a recently proposed Convolutional Neural Network learns to distinguish between benign and malware samples based on the features present in the file header rather than any significant characteristic for malware detection from the data and text sections of executable files.The article concludes by emphasising how readily aware adversaries might exploit these sophisticated devices, making them far from secure. The study suggests that in the future, researchers and developers should pay more attention to this issue and become more conscious of how vulnerable these models are.

Ibitoye *et al*., (2020) surveyed on adversarial attack threats against the network security of machine learning system. The work developed a matrix to correlate the different types of adversarial attacks with a taxonomy-based classification in order to ascertain their effectiveness in causing a misclassification. It also introduced a new classification for adversarial attacks based on applications of machine learning in network security. The study also introduced a brand-new notion for machine learning in network security: an adversarial risk grid map. In order to determine whether a generalised approach to adversarial defences will be practically achievable, future research will survey on generalised defences against adversarial attacks. Deep learning is a very limited technology when it comes to encryption.

Demetrio *et al*., (2021) surveyed on the experimental evaluation of machine learning for windows malware detection. The work produced a unified framework that comprises three new attacks based on realistic, functionality-preserving changes to the Windows Portable Executable (PE) file format, in addition to including and generalising prior attacks against machine-learning models. These three attacks called Full DOS, expand, and Shifteach manipulate the DOS header, expand it, and alter the first section's content in order to inject the antagonistic payload. The study's experimental results demonstrate that these attacks perform better than current ones in both white-box and black-box scenarios, obtaining a better trade-off between evasion rate and injected payload size. They also make it possible to evade models that have proven to be resilient to earlier attacks.

Ebrahimi *et al*., (2020) published a study on the use of adversarial byte-level language model for binary black-box evasion attacks against deep leaning-based static malware detectors. With no prior knowledge of the targeted anti-malware programme, the study presents a novel method for immediately learning a language model on binary executables and producing material that appears innocent. As far as is known, the suggested approach helps create the first automated attack without these constraining presumptions against DL-based anti-malware engines. Moreover, the method used does not call for costly dynamic malware analysis. To create adversarial instances, we created a novel technique called Malware Recurrent Neural Network (MalRNN), which does not require knowledge of the targeted anti-malware. Using binary executables, MalRNN directly learns a language model and produces seemingly innocent byte sequences that can simultaneously elude multiple DL-based anti-malware models.The result of the system implementation presents that the MalRNN achieved an average detection performance of 38.78%.

Rathore *et al*., (2021) researched on the detection of android system against adversarial attacks using Q-learning technique. This study constructed machine learning and deep neural network-based models for Android malware detection and examined how resilient they were to adversarial attacks. In order to achieve this, the team used reinforcement learning to develop new malware variants that the current Android malware detection algorithms will mistakenly classify as benign. Next, utilising reinforcement learning, the work offered two innovative attack tactics for white-box and grey-box scenarios, respectively: single policy attack and multiple policy attack. With a maximum of five adjustments, they were able to achieve an average fooling rate of 44.21% and 53.20% for all eight detection models, respectively, utilising a single policy attack and multiple policy attack. Using the multiple policy technique, the maximum fooled rate of 86.09% with five revisions was obtained against the decision tree-based model. Lastly, they suggested an adversarial defence tactic that, in the face of a single policy attack, triples the average fooling rate to 15.22%, boosting the detection models' resilience.

Demetrio *et al*., (2022) presents the practical attacks on machine learning on the case study of adversarial windows malware. The study offers a framework made up of two fundamental components: the optimizer that will be utilised to fine-tune the practical operations that must be described within the specified application-specific limitations. The work addressed a use case on Windows malware detection, emphasising how simple it is to instantiate our framework and launch attacks, and causing concern in the industry due to the vulnerability of deployed systems to adversarial attacks. In the future, the effort would envision the availability of tools that would assist security experts and developers in testing, debugging, applying version control, performing unit tests, and more, in addition to deploying adversarial attacks against their models. An Integrated Development Environment (IDE) similar to the one we use for standard software, where a developer has complete access to the same tools they typically use when coding, is what the authors envisage for this study.

Kong *et al*., (2021) surveyed on adversarial attacks in the age of artificial intelligence. Attack and defence are always at odds in the field of AI security. This review is based on high-quality works released since 2010 to aid researchers in entering the field of adversarial attack as soon as possible. The study served as a summary of common adversarial attacks in the domains of malware, photos, and text to assist researchers in identifying areas for their own research. The authors also introduced technologies that defend against attacks. Ultimately, the analysis pointed out a few debates and unresolved problems. The topic of security has a long history with adversarial learning. It is envisaged that subsequent researchers would be able to construct the framework of adversarial attack and defence with effectiveness, guided by this study.

Abusnaina *et al*., (2019) researched on adversarial learning attacks on graph-based Internet of Things (IoT) malware detection system. This study's primary objective is to find out how resilient these models are to adversarial learning. The study developed two methods for creating adversarial Internet of Things software: Graph Embedding and Augmentation (GEA) method and off-the-shelf methods. The work investigated eight distinct adversarial learning techniques to compel the model to misclassify using commercial adversarial learning attack tools. The goal of the GEA technique is to carefully embed a benign sample within a malicious one while maintaining the functionality and usefulness of the generated adversarial sample. Extensive tests were carried out to assess the effectiveness of the suggested approach, demonstrating that commercial adversarial attack techniques can get a 100% misclassification rate.

Anthi *et al*., (2021) researched on adversarial attacks on machine learning cybersecurity defences in industrial control systems. By creating adversarial samples and examining classification behaviours, this work investigates how adversarial learning might be applied to supervised models. Widely used supervised machine learning classifiers were trained and tested on a real power system dataset to support the experiments reported here. Furthermore, this approach takes into account realistic attacker assumptions and models. A Jacobian-based Saliency Map Attack (JSMA) was applied to the testing data in order to produce adversarial samples with a variety of combinations that alter the quantity of noise and the number of features to disturb. These samples were tested against Random Forest and J48, two of the top classifiers. When hostile samples were present, the classification performance for both models declined overall by 6 and 11 percentage points. Result of the study presents that J48 classification performance (F1-score) following adversarial training was 0.1.

Sajja and Saksham (2023) presents the use of deep learning for the detection of adversarial threats. To strengthen deep learning models against adversarial weaknesses, this calls for an integrated strategy that incorporates explainable AI, interdisciplinary collaboration, and advanced defence techniques like adversarial training and defensive distillation. In the end, it points to a bright future where AI and DL work together to overcome new obstacles. AI's deep learning potential includes the creation of reliable and understandable models as well as insights from a variety of industries, including agriculture and medical. Deep learning-enabled AI, anchored in tackling adversarial vulnerabilities, is positioned to propel industries and change the world via safety, dependability, and innovation, demonstrating the enormous potential of human brilliance unlocked by this dynamic relationship. The study's implementation outcome was not disclosed.

Stilinski and Potter (2024) researched the use of adversarial text generation in cybersecurity, exploring the potential of simulated cyber threats for evaluating Natural Language Processing (NLP)-based anomaly detection systems. Techniques for creating adversarial text involve manipulating textual material to produce minute changes that are undetectable to humans but may trick NLP-based anomaly detection systems. These methods can be used to create cyber threats that include a variety of attack scenarios and evasion tactics. These artificial threats function as difficult benchmarks to assess how resilient NLP-based anomaly detection systems are against hostile attacks. Through the use of adversarial text generation techniques, cybersecurity practitioners and researchers can produce artificial cyber threats and do more thorough assessments of NLP-based anomaly detection systems. This method assists in locating holes and weak points in current systems and provides guidance for creating stronger, more resilient cybersecurity solutions that can fend off sophisticated cyberattacks. However, the study did not report the implementation outcome of the technique adopted.

Sande-Rios et al., (2024) researched on threat analysis and adversarial model for smart grids. The smart power grid's cyber domain presents a variety of new hazards in addition to the traditional physical domain threats. In order to overcome the limitation, this work examined the smart grid's primary attack surfaces first. Next, it conducted a threat analysis from the standpoint of the adversarial model, taking into account various levels of knowledge, objectives, motivations, and capabilities. By taking into account attacks from the research literature, evaluating vulnerabilities, and dissecting Tactics, Techniques, and Procedures (TTPs) from actual attacks, the study created a model. For these situations, the model enables the definition of various players, or roles, with varying skill sets and objectives. The study did not report on the model's performance evaluation.

RESEARCH METHODOLOGY

The methodology for this review paper was a systematic literature review (SLR), which aimed to identify, analyse and synthesize previous research on adversarial threat management in critical network infrastructures. The literature search was carried out in the major digital databases, IEEE Xplore, ScienceDirect, SpringerLink and Google Scholar, from 2019 to 2024. The search terms used were: adversarial attack detection, adversarial machine learning, network intrusion detection systems, poisoning attacks, label flipping, feature manipulation, and deep learning for adversarial defence. We included peer-reviewed journal articles, conference papers, and trusted preprints that introduced or assessed the methods for attacking or defending networks that are adversarial. Studies that concentrated exclusively on non-network domains (e.g., classification of images without network security issues) or did not have an empirical evaluation were excluded. 48 relevant studies were chosen for detailed examination. This methodical approach provided comprehensive coverage and a critical evaluation of the state-of-the-art adversarial threat management techniques.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

Theory Of Adversarial Attacks

Adversarial attacks are a critical area of study in the field of machine learning and cybersecurity. These attacks involve intentionally altering input data to deceive machine learning models into making incorrect predictions. The concept is rooted in the vulnerability of these models to small, often imperceptible changes in the input data (Mercuri *et al*., 2023). This vulnerability arises because machine learning models, particularly deep neural networks, rely on complex patterns in the data that can be subtly manipulated. Understanding the theory behind adversarial attacks is essential for developing more robust and secure machine learning systems (Appruzzese *et al*., 2019).

Types and Mechanisms of Adversarial Attacks

There are several types of adversarial attacks, each with unique mechanisms and goals. Evasion attacks, for example, are designed to fool a trained model during the inference stage by adding small perturbations to input data. These perturbations are crafted to be undetectable to humans but cause the model to make incorrect predictions (Demetrio *et al*., 2022). Another type is poisoning attacks, where the attacker tampers with the training data, introducing malicious data points that corrupt the learning process and degrade the model's performance (Villegas-CH *et al*., 2024). Additionally, exploratory attacks involve probing a machine learning model to gain insights into its structure and weaknesses, which can then be exploited in subsequent attacks (Zhang *et al*., 2022). The effectiveness of these attacks is often assessed using metrics like the success rate of misclassification and the level of perturbation required.

Implications and Defence Mechanisms

The implications of adversarial attacks are profound, as they pose significant threats to the security and reliability of machine learning applications across various domains, including finance, healthcare, and autonomous systems (Demetrio *et al*., 2022). These attacks can lead to severe consequences, such as financial losses, compromised privacy, and even physical harm. Consequently, developing robust defence mechanisms is crucial (Mercuri *et al*., 2023). Approaches to defending against adversarial attacks include adversarial training, where models are trained on a mixture of normal and adversarial examples, and the use of defensive distillation, which aims to make models less sensitive to small perturbations (Alzaidy and Binsalleeh, 2024). Other strategies involve detecting and rejecting adversarial inputs before they reach the model or employing robust optimization techniques to enhance model resilience. As the field evolves, ongoing research is essential to keep pace with the advancing tactics of adversarial attackers (Zhang *et al*., 2022)

Methods of Adversarial attack tactics

Generating adversarial attacks involves various techniques (shown in Figure 1) aimed at producing diverse forms of adversarial samples from original attack dataset. These include methods such as the gradient based, optimization based and transfer based approached.

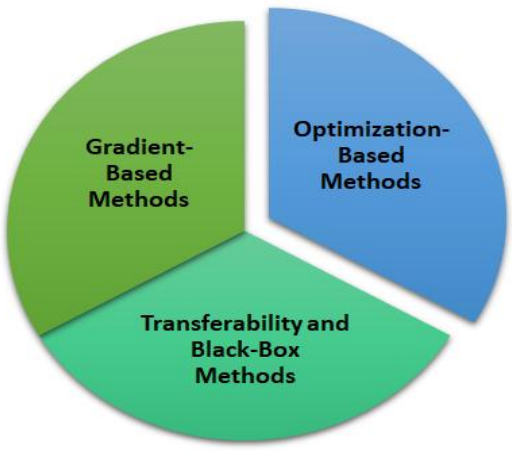

Figure 1: Methods of Adversarial Attack Generation

Gradient-Based Methods

One of the primary methods for generating adversarial samples is through gradient-based approaches, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). FGSM works by using the gradient of the loss function with respect to the input data to create perturbations (Biczyk and Wawrowski, 2023). Specifically, it adjusts the input data in the direction of the gradient to maximize the loss, thereby creating an adversarial sample that is slightly altered but causes the model to misclassify (Mercuri *et al*., 2023).

Optimization-Based Methods

Optimization-based methods, such as the Carlini & Wagner (C&W) attacks, formulate the generation of adversarial samples as an optimization problem. The goal is to find the minimal perturbation required to change the model's prediction while keeping the perturbation as imperceptible as possible. These methods involve solving a constrained optimization problem, where the objective is to minimize the norm of the perturbation subject to the constraint that the adversarial sample is misclassified (Fu *et al*., 2023).

Transferability and Black-Box Methods

Another category involves transferability and black-box methods. Transferability refers to the phenomenon where adversarial samples generated for one model can often deceive another model, even if it has a different architecture or was trained on a different dataset (Fu *et al*., 2023). This property is exploited in black-box attacks, where the attacker does not have direct access to the target model's parameters or gradients. Instead, the attacker trains a surrogate model to mimic the target model's behavior and generates adversarial samples against the surrogate (Han *et al*., 2023).

Additionally, methods such as pollution, label flipping, and feature manipulation can be employed. Pollution involves introducing malicious data points into the training set to degrade model performance (Alzaidy and Binsalleh, 2024). Label flipping changes the labels of specific data points to mislead the model during training, while feature manipulation alters key features in the input data to produce incorrect predictions. Black-box methods and these additional techniques highlight the vulnerability of models even when direct access is not possible, underscoring the need for robust security measures.

Methods of detecting Adversarial Attack

This section discussed the method of detecting adversarial attack. Some of the popular methods are the statistical approach, input transformation approach, model-based approach, machine learning and deep learning techniques respectively as shown in Figure 2.

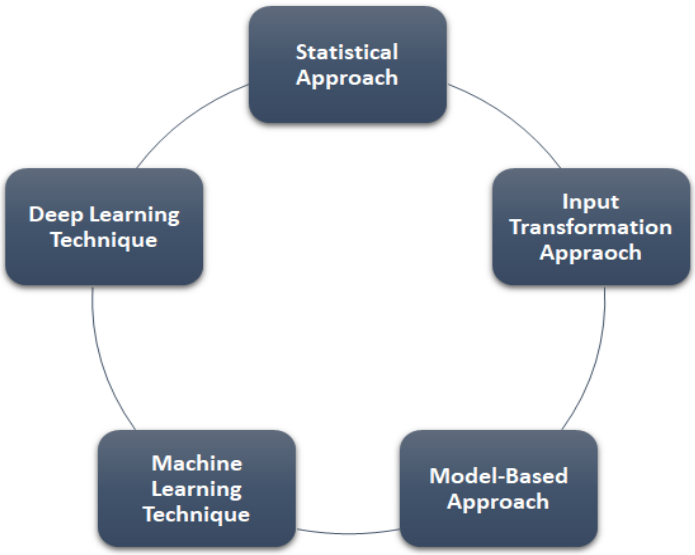

Figure 2: Methods of Adversarial Attack Detection

Statistical and Anomaly Detection Methods

One of the primary approaches for detecting adversarial attacks is statistical and anomaly detection. These methods involve analyzing the statistical properties of input data and identifying anomalies that deviate from the norm. Techniques like Principal Component Analysis (PCA) and Kernel Density Estimation (KDE) can be used to model the distribution of normal data and flag outliers as potential adversarial samples (Fu *et al*., 223).

Input Transformation and Feature Squeezing

Input transformation and feature squeezing are another set of techniques used to detect adversarial attacks. Input transformation involves applying various preprocessing steps to the input data, such as image transformations or data augmentation techniques, and then observing the model's predictions (Liang *et al*., 2022). If the predictions change significantly after these transformations, it may indicate the presence of adversarial perturbations. Feature squeezing reduces the degrees of freedom available to the adversary by coalescing similar input features, thus making it harder for adversarial perturbations to be effective (Fu *et al*., 2023)

Model-Based Approaches

Model-based approaches involve training secondary models or enhancing existing models to detect adversarial attacks. One common technique is adversarial training, where a model is trained on a mixture of regular and adversarial examples to improve its robustness and ability to recognize adversarial inputs (Liang *et al*., 2022). Another approach is to use an auxiliary network specifically designed to detect adversarial samples. This auxiliary network can be trained using a combination of normal and adversarial data to learn features indicative of adversarial manipulation (Han *et al*., 2023).

Machine Learning in Adversarial Attack Detection

Machine learning techniques play a crucial role in the detection of adversarial attacks. At its core, machine learning involves training models to recognize patterns and make predictions based on data. When applied to adversarial attack detection, these models can be trained to distinguish between benign and adversarial inputs (Alzaidy and Binsalleeh, 2024). The key advantage of using machine learning for detection is its ability to learn from data, adapting to new types of attacks and improving detection accuracy over time (Liang *et al*., 2022).

Deep Learning Techniques for Enhanced Detection

Deep learning, a subset of machine learning, offers advanced techniques for detecting adversarial attacks by leveraging neural networks with multiple layers (deep neural networks) (Han *et al*., 2023). Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), excel at extracting complex features from high-dimensional data, making them well-suited for identifying adversarial perturbations (Alzaidy and Binsalleeh, 2024). By training deep learning models on both clean and adversarially perturbed data, these models can develop a robust understanding of the characteristics of adversarial attacks. Techniques such as adversarial training, where the model is exposed to adversarial examples during training, and defensive distillation, which aims to smooth the model's decision boundaries, further enhance the detection capabilities of deep learning models (Alzaidy and Binsalleeh, 2024).

DISCUSSIONS

The results of the literature review reveal a rapidly changing field of adversarial machine learning, especially in critical network infrastructures like industrial control systems, 5G networks, cloud platforms, and IoT systems. One common theme in the studies is adversarial attacks are not abstract concepts but real, modifiable, and more and more automated attacks that can thwart contemporary machine learning security systems.

One of the key trends found is the evolution of static attack models to adaptive and intelligent adversaries. Generally, previous works have been mostly on gradient-based or optimization-based attack methods including FGSM, PGD, and C&W methods. More recent papers (Rathore *et al*., 2021; Ebrahimi *et al*., 2020), however,

show that attackers today are using reinforcement learning and generative models to dynamically produce adversarial samples that are adjusted to the nature of the system responses. This is a considerable decrease in the effectiveness of the traditional fixed rule defence structures and makes it clear why it is important to have learning security systems.

Yet another significant note is the ever-expanding attack surface created by distributed systems today. Cloud computing services, 5G networks and IoT environments are intrinsically dependent on interconnected services and data sharing, making them more susceptible to black-box and transfer-based attacks. The research by Li *et al*. (2019); Apruzzese *et al*. (2022) reveals that attackers can exploit exposed APIs and common model interfaces without accessing the internal system, making security enforcement more complex and decentralized.

The literature also shows that defence mechanisms based on machine learning and deep learning are promising but are not yet sufficiently well developed. Methods like adversarial training, defensive distillation, feature squeezing, and anomaly detection make for better resilience, but they have limited efficacy against adaptive or multi-vector attacks. This indicates that the existing defence strategies focus mainly on reacting to attacks rather than proactively stopping them, and respond to known attack patterns, instead of trying to predict new ones.

Moreover, there exists an evident gap between accuracy during detection in controlled environments and during real world deployment. It is observed that many models achieve high accuracy on the lab data sets, but are not robust in real-time conditions as the data sets cause bias, concept drift and lack representation of adversarial instances. This further highlights the disconnect between research and practice in cyber security.

Black-box nature of deep learning-based security models is a common concern in the literature. Currently, CNNs, RNNs and hybrid models are the best models for detecting the objects, but their decision-making process is still not transparent. This hampers the levels of trust and adoption in critical industries like healthcare, smart grid, and industrial control systems, where explainable decision-making is crucial for compliance and incident response.

A major consideration is the lack of uniform adversarial defence mechanisms. Current solutions focus more on detection, mitigation and response separately. So, there is no comprehensive system that can detect and classify, mitigate and recover from the adversarial attacks. This fragmentation hampers efficiency of operations and response time in the event of a cyber incident.

Lastly, there is a need to have common sets of data and benchmarking systems. Model evaluation is not consistent, especially without realistic and diverse adversarial datasets, and this makes it hard to compare new solutions. This challenge is even more pronounced in more recent areas of application, like 5G and IoT, where adversarial behavior is not fully understood and is not extensively covered in existing datasets.

CONCLUSION

This review has systematically discussed adversarial threats in CNIs, their generation mechanisms, detection techniques and mitigation strategies. The study has shown that the adversarial machine learning techniques have grown from using only perturbations to produce adversarial examples to highly complex, adaptive, and intelligent techniques that can circumvent modern security systems using AI. The literature shows that, critical infrastructures like IoT networks, cloud platforms, 5G systems, and industrial control environments are more and more relying on machine learning based security mechanisms, which makes them more vulnerable across the literature. Poisoning attacks, feature manipulation and transfer-based adversarial strategies have proved to be effective against traditional intrusion detection systems.

While various defence methods, from statistical anomaly detection to deep learning-based adversarial training, have been suggested, none of them are completely effective against the changing tactics used by attackers. Current solutions are still domain-specific, not adaptable in real-time, and are not strong enough to counter adaptive attackers who continually evolve their strategies. Moreover, gaps that persist across the review include the lack of datasets, the lack of explainability in models, the absence of a common defence architectures, and

the lack of integration between detection and response systems. All these constraints contribute to the challenges in real-world deployment of reliable and scalable adversarial defence systems.

To facilitate the summary, the adversarial threats can be summarized as a continuing and dynamic challenge that cannot be met with simple detection methods. Integrated, intelligent and adaptive solutions are needed that can perform well in a variety of dynamic and changing network environments.

Future Research Directions

The following are some areas of suggested future research based on the gaps in the literature:

1. Development of Real-Time Adaptive Defence Systems: Future research should be directed toward the development of a Real-Time Adaptive Defence System that can not only learn from new attacks but adapt to an evolving attack pattern. This should include embedding streaming data analytics and online learning systems to help mitigate threats in time.

2. Multi-Strategic Hybrid Security Frameworks: Unified frameworks integrating the statistical methods, machine learning and deep learning techniques are required. Combining technologies can increase robustness: strengths of several technologies combined are better than the weaknesses of a single technology.

3. Cross-Domain Generalizable Models: Current models are built for a particular domain and are not able to be generalized to different domains. In the future, transfer learning and domain adaptation methods have to be investigated to create models that can be applied across the IoT, cloud, 5G, and industrial systems.

4. Adversarially Robust and Explainable AI Systems: Enhancing Model Interpretability is critical for certain applications. The future holds promise for further research on techniques for adversarial detection that are explainable, allowing for transparency in the decision-making process while still retaining a high level of robustness against attacks.

5. Standardized Benchmark Datasets for Adversarial Security: Developing realistic, large-scale, and continually growing datasets is essential. Such datasets need to cover the current attack vectors such as multi-stage, reinforcement learning, and adaptive attacks.

6. Integration of Detection, Mitigation, and Response Mechanisms: Future systems should not be based on detection alone but incorporate full lifecycle security management including automatic response, mitigation and recovery mechanisms in a single architecture.

7. Adversarial Attack Forecasting and Proactive Defence: Future systems should be able to predict adversarial behaviours with historical and behavioural patterns in order to be proactive in defence rather than reactive in detection.

Successful adversarial threat management needs to be more than just a series of defensive measures where it has to be intelligent, adaptive and integrated, and constantly evolve to match the innovations of the adversary.

REFERENCES

1. Abusnaina A., Khormali A., Alasmary H., Park J., Anwar A., & Mohaisen A., (2019) Adversarial Learning Attacks on Graph-based IoT Malware Detection Systems. CNS-1809000, NRF grant 2016K1A1A2912757, NVIDIA GPU Grant Program (2018 and 2019), and a Cyber Florida Seed Grant.
2. Aiken, J., & Scott, P. (2020). Critical network infrastructures and data management. Journal of Information Security Studies, 12(3), 45–59.
3. Alhajjar, E., Maxwell, P., & Bastian, N. D. (2021). Adversarial machine learning in network intrusion detection systems. Expert Systems with Applications, 186, 115782. https://doi.org/10.1016/j.eswa.2021.115782

4. 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. https://doi.org/10.3390/fi15020062
5. Alzaidy, S., & Binsalleeh, H. (2024). Adversarial attacks with defense mechanisms on convolutional neural networks and recurrent neural networks for malware classification. Applied Sciences, 14(4), 1673. https://doi.org/10.3390/app14041673
6. Anthi E., Williams L., Rhode M., Burnap P., & Wedgbury A., (2021) Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems. Journal of Information Security and Applications 58 (2021) 102717https://doi.org/10.1016/j.jisa.2020.102717
7. Apruzzese *et al*., (2022) Wild Networks: Exposure of 5G Network Infrastructures to Adversarial Examples. IEEE Transactions On Network And Service Management. DOI: 10.1109/TNSM.2022.3188930
8. Apruzzese G., Colajanni M., Ferretti L., & Marchetti M., (2019) Addressing Adversarial Attacks Against Security Systems Based on Machine Learning. 2019 11th International Conference on Cyber Conflict: 2019 © NATO CCD COE Publications, Tallinn
9. Biczyk, P., & Wawrowski, P. (2023). Gradient-based adversarial methods in deep learning security. Journal of Cybersecurity Research, 5(1), 23–34.
10. Butt, U. J., Hussien, O., Hasanaj, K., Shaalan, K., Hassan, B., & al-Khateeb, H. (2023). Predicting the Impact of Data Poisoning Attacks in Blockchain-Enabled Supply Chain Networks. *Algorithms*, *16*(12), 549. https://doi.org/10.3390/a16120549
11. Demetrio L., Biggio B., & Roli F., (2022) Practical Attacks on Machine Learning: A Case Study on Adversarial Windows Malware. IEEE Security & Privacy Published by the IEEE Computer SocietyarXiv:2207.05548v1 [cs.CR] 12 Jul 2022
12. Demetrio L., Biggio B., Lagorio G., Roli F., & Armando A., (2019) Explaining Vulnerabilities of Deep Learning to Adversarial Malware Binaries. arXiv:1901.03583v2 [cs.CR] 24 Jan 2019
13. Demetrio L., Scott E., BiggioB., LagorioG., ArmandoA., & Roli F., (2021) Adversarial EXEmples: A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection. In Proceedings of ACM TO EDIT conference. ACM, New York, NY, USA, 31 pages. https://doi.org/TOEDIT
14. Ebrahimi M., Zhang N., Hu J., Raza M., & Chen H., (2020) Binary Black-box Evasion Attacks Against Deep Learning-based Static Malware Detectors with Adversarial Byte-Level Language Model. Artificial Intelligence Lab, The University of ArizonaarXiv:2012.07994v1 [cs.CR] 14 Dec 2020
15. Fu, Y., Xia, E., Huang, D., & Jing, Y. (2023). Adversarial attack defense method for continuous-variable quantum key distribution systems. Applied Sciences, 13(17), 9928. https://doi.org/10.3390/app13179928
16. Han C., Qin R., Wang L., Cui W., Li D.,&Yan B., (2023) Adversarial Example Detection and Restoration Defensive Framework for Signal Intelligent Recognition Networks. Appl. Sci. 2023, 13, 11880. https://doi.org/10.3390/app132111880
17. He, X., Zhou, Y., & Lin, F. (2023). Game theory approaches in cybersecurity. IEEE Transactions on Information Forensics and Security, 18(4), 1123–1135. https://doi.org/10.1109/TIFS.2023.3245678
18. Ibitoye O., Abou-Khamis R., Matraway A., & Shafiq O., (2020) The Threat of Adversarial Attacks Against Machine Learning in Network Security: A Survey. arXiv:1911.02621v2 [cs.CR] 4 Oct 2020
19. Khazane H.,Ridouani M.,Salahdine F., &Kaabouch N., (2024) A Holistic Review of Machine Learning Adversarial Attacks in IoT Networks. Future Internet 2024, 16, 32. https://doi.org/10.3390/fi16010032
20. Kong Z., Xue J., Wang Y., Huang L., Niu Z., & Li F., (2021) A Survey on Adversarial Attack in the Age of Artificial Intelligence. Hindawi Wireless Communications and Mobile Computing Volume 2021, Article ID 4907754, 22 pages https://doi.org/10.1155/2021/4907754
21. Lai, A. C. T., Ke, P. F., & Ho, A. (2019). Ransomware deception through adversarial attacks. arXiv preprint arXiv:1902.01221. https://doi.org/10.48550/arXiv.1902.01221
22. Lennon, J. (2022). Mitigation of adversarial threats in artificial intelligence infrastructure using direct action. AI & Society, 37(4), 1123–1135. https://doi.org/10.1007/s00146-021-01234-5
23. Li X., Ji S., Han M., Ji J., Ren Z., Liu Y., & Wu C., (2019) Adversarial Examples Versus Cloud-based Detectors: A Black-box Empirical Study. arXiv:1901.01223v4 [cs.CV] 14 Sep 2019
24. Liang H., He E., Zhao Y., Jia Z., & Li H., (2022) Adversarial Attack and Defense: A Survey. Electronics 2022, 11, 1283. https://doi.org/10.3390/electronics11081283

25. MaiorcaD., BiggioB., & GiacintoG.,(2019) Towards Adversarial Malware Detection: Lessons Learned from PDF-based Attacks. ACM Comput. Surv. 1, 1, Article 1 (January 2019), 35 pages. https://doi.org/10.1145/3332184
26. Mercuri V., Saletta M.,&Ferretti C., (2023) Evolutionary Approaches for Adversarial Attacks on Neural Source Code Classifiers. Algorithms 2023, 16, 478. https://doi.org/10.3390/a16100478
27. Qian, C., Zhang, M., Nie, Y., Lu, S., & Cao, H. (2023). A Survey of Bit-Flip Attacks on Deep Neural Network and Corresponding Defense Methods. *Electronics*, *12*(4), 853. https://doi.org/10.3390/electronics12040853
28. Rathore H., Sahay S., & Nikam P., (2021) Robust Android Malware Detection System against Adversarial Attacks using Q-Learning. arXiv:2101.12031v1 [cs.CR] 27 Jan 2021
29. Sajja, S., & Saksham, R. (2023). Deep learning for adversarial threat detection. AI Review, 56(3), 789–802. https://doi.org/10.1007/s10462-022-10123-4
30. Sande-Rios, J., Canal-Sanchez J., Manzano-Hernandez C., & Pastrana S., (2024) Threat analysis and adversarial model for Smart Grids. arXiv:2406.11716v1 [cs.CR] 17 Jun 2024 https://doi.org/10.1109/TSG.2024.3145679
31. Stilinski, M., & Potter, R. (2024). Adversarial text generation in cybersecurity for NLP anomaly detection. Computers & Security, 130, 103456. https://doi.org/10.1016/j.cose.2024.103456
32. Villegas-Ch, W., Jaramillo-Alcázar, A., & Luján-Mora, S. (2024). Evaluating robustness of deep learning models against adversarial attacks. Big Data and Cognitive Computing, 8(1), 8. https://doi.org/10.3390/bdcc8010008
33. Wallace, J., Foster L., Schartau K., Lewin D., Keyes C., (2024) Empirical Risk Analysis Methodology for Adversarial Threats against Critical Infrastructure. ASCE 04023036-1 J. Infrastruct. Syst., 2024, 30(1): 04023036. https://doi.org/10.1111/risa.13876