Grenze International Journal of Engineering and Technology, January Issue

A Survey on Adversarial Attacks and Defense Mechanisms in Artificial Intelligence

Akshat Shree Ram Pandey[1], Sanjit Kamath Udyavar[2], Abhinav Varma[3], Mohammed Jiyad Herial[4] and Padmavathi S[5]
[1-4]Department of Computer Science & Engineering (Cyber Security), Dayananda Sagar College of Engineering, Bangalore, India
Email: akshatshreerampandey@gmail.com, sanjitkamathu@gmail.com, abhinavvarma03@gmail.com, mjiyad119@gmail.com
[5]Assistant Professor, Department of Computer Science & Engineering (Cyber Security), Dayananda Sagar College of Engineering, Bangalore, India
Email: padmavathi-cscyber@dayanandasagar.edu

***Abstract*— Adversarial attacks use the limitations of Artificial Intelligence models to generate mistakes and catastrophic failures of the system. They are a significant threat to Artificial Intelligence systems. Our objective in this work is to provide a systematic study of the different types of attacks we can encounter, their effects on AI systems, and the methods that are currently in place to combat them. We also discuss the nature of these attacks and how they impact the security of the AI model. Finally, we discuss the developments that need to be undertaken to make AI systems safer against these types of adversarial attacks.**

***Index Terms*— Adversarial attacks, Artificial Intelligence, AI security, attack mitigation, AI Robustness, systematic study.**

I. INTRODUCTION

Artificial intelligence can be defined as the ability of a computational system to perform those tasks restoration of injured human faculties, namely, learning, reasoning, problem-solving, perception, and decision-making. This Computer Science discipline focuses on developing and studying methods and software that enable a machine to, first, perceive its environment and then use learning and intelligence to act so as to maximize the chances of attaining a goal set before it.

AI encompasses a range of instruments and techniques that enable application in industries such as banking, healthcare, security, and transportation. Nevertheless, there are still a lot of significant problems with existing AI systems, including the enduring bias, privacy concerns, job displacement, and opaque decision-making.

However, AI systems are not infallible. An emerging problem for the AI sector is the increase of adversarial attacks, adversarial attackers intentionally manipulate input data to force models to make incorrect predictions or release sensitive information. These kinds of attacks may be used in domains which have a very low tolerance of failure such as self driving vehicles or medical systems. Thus, it is critical to understand and defend against adversarial attacks as we increase our reliance on AI systems by the day for various applications across the globe.

Grenze ID: 01.GIJET.12.1.201

II. LITERATURE SURVEY

Goodfellow et al. (2015), originates FGSM which perhaps is the most seminal adversarial attack ever introduced into adversarial machine learning, a gradient-based white box attack that can produce adversarial examples at lightning speed. Further, the method is largely non-robust and subject to defenses such as adversarial training and iterative attacks. From here came Madry et al. (2018), who proposed the PGD attack and adversarial training, setting the stage for many present-day defense frameworks. Nevertheless, the approach is expensive and difficult to extend to perturbations it has never seen.

Carlini and Wagner (2017) suggest a powerful optimization-based adversary (the CW attack) that accomplishes near-perfect success while keeping perturbations to the minimum. This method, although highly effective, is time-consuming and less practical for real-time or large-scale applications. Looking toward the physical world in its concern with adversarial robustness, Looking toward the physical world in its concern with adversarial robustness, Kurakin et al. (2016) show how adversarial examples are transferable to the real environment, thus substantiating the feasibility of physical attacks in real life. The drawback to this, however, is that the method usually requires constrained conditions for reproducibility and has very little generalization in changing environments.

Papernot et al. (2016) focus on the black-box threat landscape by developing transferability-based attacks that allow adversarial inputs to be generated without direct model access. These methods rely heavily on the similarity between surrogate and target models, with success rates diminishing as models diverge. The foundational work of Szegedy et al. (2014) identifies the inherent vulnerability of neural networks to small perturbations, ultimately laying the conceptual basis for adversarial ML research, though their work lacks practical implementation details and extensive empirical evaluation.

Further exploring transferability, Tramer et al. (2018) investigate the space of adversarial examples that can move between models, emphasizing the viability of black-box attacks, though their study focuses primarily on image classification and provides limited insight on broader domains. Biggio and Roli (2018) provide a longitudinal analysis of adversarial machine learning, highlighting major trends, evolving challenges, and key directions for future research, but their work remains largely theoretical and offers limited practical mitigation strategies.

Ren et al. (2020) present a comprehensive taxonomy of adversarial attacks and defenses for deep learning systems, serving as broad overviews of the field, but their work is mostly conceptual, with no experimental validation and benchmarking. Domain-specific applications are addressed by Modas et al. (2020), who explore the adversarial vulnerabilities of sensors used in autonomous vehicles, especially LiDAR and camera systems, justifying the need for domain-aware adversarial analysis, although the work is mostly theoretical and lacks real-world experimental validation.

From the perspective of defenses, Feng et al. (2020) defend using randomized smoothing and regularization to certify provable robustness, but it is costly to compute and unable to scale to large datasets, while Tao et al. (2023) espouse a theoretical framework toward a unification for the adaptation of step-size in gradient-based attacks, allowing for better understanding and enhancement of such attacks, though empirical results are only for smaller-scale model architectures. In their turn, Kreuk et al. (2018) move the threat model into discrete input spaces, showing that adversarial examples can be used to manipulate binary malware classifiers while preserving their original functionality. The experiments, however, involve only one detection model and therefore transition the question of generalization to other security domains. A detailed survey of these primary contributions and the limitations imposed upon them is presented for reference ease in Table.

TABLE I. SUMMARY OF RELEVANT PAPERS IN ADVERSARIAL ATTACKS AGAINST ML SYSTEMS

Year	Author(S)	Paper Name	Contribution	Limitations
2015	Goodfellow et al.	Explaining and Harnessing Adversarial Examples (FGSM)	Introduces the Fast Gradient Sign Method, A foundational gradient-based white-box attack for generating adversarial examples efficiently.	Limited robustness; vulnerable to defences such as adversarial training and iterative attacks.
2018	Madry et al.	Towards Deep Learning Models Resistant to Adversarial Attacks (PGD)	Proposes the Projected Gradient Descent attack and adversarial training, forming the basis of many defense frameworks.	Computationally intensive; generalization to unseen perturbations remains a challenge.
2017	Carlini & Wagner	Towards Evaluating the Robustness of Neural Networks (C&W Attack)	Presents an optimization-based adversarial attack achieving high success rates while minimizing perturbations.	Time-consuming and less efficient in real-time or large-scale scenarios.

2016	Kurakin et al.	Adversarial Examples in the Physical World	Demonstrates that adversarial examples remain effective in real-world physical settings, supporting feasibility of physical attacks.	Requires controlled conditions for reproducibility; limited generalization to dynamic environments.
2016	Papernot et al.	Practical Black-Box Attacks Using Adversarial Examples	Develops transferability-based black-box attacks, enabling adversarial input generation without model access.	Success rates depend heavily on similarity between surrogate and target models.
2014	Szegedy et al.	Intriguing Properties of Neural Networks	Identifies the vulnerability of neural networks to small perturbations, establishing the conceptual foundation for adversarial research.	Lacks practical implementation details and empirical robustness evaluations.
2018	Tramer et al.	The Space of Transferable Adversarial Examples	Explores the transferability of adversarial examples between models, contributing to the study of black-box attack viability.	Focused mainly on image classification; limited cross-domain analysis.
2018	Biggio & Roli	Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning	Provides a longitudinal analysis of adversarial machine learning, highlighting trends, challenges, and future directions.	Theoretical overview; limited emphasis on practical mitigation techniques.
2020	Ren et al.	Adversarial Attacks and Defenses in Deep Learning	Offers a comprehensive taxonomy of adversarial attack and defense strategies in deep learning systems.	Lacks experimental validation and benchmarking.
2020	Modas et al.	Towards Robust Sensing for Autonomous Vehicles	Investigates adversarial vulnerabilities in AV sensors (e.g., LiDAR, camera), supporting domain-specific impact analysis.	Emphasis on theoretical modelling; real-world experimental validation is limited.
2020	Feng et al.	Regularized Training and Tight Certification for Randomized Smoothed Classifier	Proposes a certified defense using randomized smoothing and regularization, contributing to provable robustness techniques.	High computational cost; limited scalability for large datasets.
2023	Tao et al.	Adapting Step-size: A Unified Perspective to Analyse and Improve Gradient-based Attacks	Provides a unified theoretical framework for step-size adaptation in gradient-based adversarial attacks.	Limited empirical analysis on large-scale or diverse model architectures.
2018	Kreuk et al.	Adversarial Examples on Discrete Sequences for Beating Whole-Binary Malware Detection	Demonstrates that adversarial attacks can manipulate discrete input sequences in malware classifiers while preserving functionality.	Evaluated on a single detection model; generalization to other security systems untested.

III. Helpful Hints

A. Adversarial Attack Techniques

Adversarial attacks that threaten artificial intelligence models can be categorized into four main groups, based on their methodology and the knowledge assumed about the target system.

Gradient-Based Attacks

Gradient-based attacks lie among some well-established methods employed to fool a learning model. These methods are granted direct access to the gradients involved in the model. With methods like the Fast Gradient Sign Method (FGSM), input data can be modified in minute, calculated ways, with the changes deliberately made in the direction that would maximally increase the model's loss.

This process facilitates the generation of adversarial examples with relatively little effort. Another well-established approach is that of Projected Gradient Descent (PGD), which generalizes the notion with a couple of iterations of minor modifications, allowing the adversarial example to grow in its effectiveness as it remains close to the original data.

Through iterative refinement, PGD generates inputs that consistently challenge model robustness and highlight weaknesses in the decision boundaries of the model.

Carlini & Wagner Attacks

Carlini and Wagner (C&W) attacks are a type of adversarial method that operate from an optimization perspective. Unlike attacks that simply aim to induce any form of misclassification, C&W methods attempt to alter an input in the smallest way possible to coerce a model into misclassifying it. The attack is therefore

modeled mathematically as an optimization problem, balancing the conflicting objectives of misclassification and minimal distortion.
DeepFool is a similar technique that identifies the closest point on a model's classification boundary and perturbs the given data just enough to cross that boundary. Both methods create adversarial examples that are highly effective and, from a human perspective, almost impossible to detect. This presents serious challenges for any detection approach.

Black Box Attacks
Such attacks constitute the category of black box attacks, since our prospective attacker has no way to enter the internal mechanics or architecture of the target model. The attacker only interacts with the system as an outsider, presenting inputs and analysing what outputs are given to try to identify some pattern in the system's behaviour. Attackers may do this by querying the model many times, iteratively improving their attack strategy, or they may train a different model to reproduce the target model's output. Using the output of these surrogate models, attackers can then start developing adversarial examples that would fool the protected systems, irrespective of their secrecy or complexity. These significance strikes on the point that security through obscurity does not work. Such are vulnerable machine learning systems even intractable to public eyes.

B. Impact of Adversarial Attacks

With the continued evolution of Artificial Intelligence (AI) in applied domains, the threats of adversarial attacks are real across multiple domains.

Healthcare
In areas where AI or machine learning is used for classification of clinical images, slight changes to an input can induce missed diagnoses or incorrect treatment strategies that could adversely affect the patient. Adversarial agents can undermine the information integrity in electronic health records (EHR), and unmonitored or unverified EHRs may destroy trust in diagnoses made by AI systems.

Cyber security
Adversary actions are used to change exploitation or intrusion detection capabilities of detection algorithms, leaving sensitive data compromised or systems at risk.

Finance
Slight changes to input data easily mislead AI or ML systems, particularly fraud detection algorithms, algorithmic trading models, and credit scoring models, allowing for financial crimes.

C. How Do Adversarial Attacks Work

Adversarial attacks target machine learning models by exploiting weaknesses through carefully manipulated or malicious inputs. Generally, the attack lifecycle has three stages:

- The attacker probes the AI model, often via reverse engineering, to identify vulnerabilities.
- They craft adversarial examples—inputs specifically designed to exploit discovered weaknesses.
- These malicious samples are deployed, prompting the model to err in its predictions or classifications.

Examples include editing images of stop signs to mislead self-driving vehicles, subtly altering spam messages to evade detection, and tweaking malware files to evade antivirus software. Such attacks can operate both digitally and physically, bringing to light unpredictable risks in diverse applications.

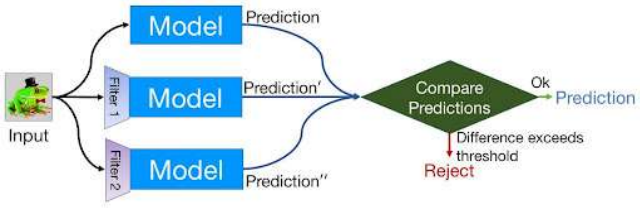

Figure 1. Feature squeezing assesses differences in model outputs between original and compressed inputs to spot adversarial manipulations.

D. Defensive Techniques

Adversarial Training
One of the primary strategies to improve model robustness against adversarial attacks is to incorporate adversarial examples directly into the training process. By augmenting the training dataset with crafted perturbed

inputs designed to mislead the model, the learning algorithm becomes better equipped to identify and correctly classify such manipulations. This approach helps the model develop resilience against specific, known attack types. However, adversarial training demands significant computational resources and may not generalize effectively to novel or unforeseen adversarial examples beyond those seen during training.

Gradient Masking

Gradient masking aims to protect a model by making its gradient information—used by many attack algorithms—difficult or impossible to exploit. This is often achieved through methods such as introducing randomness, obfuscation, or smoothing in the gradient computation process. Although this can temporarily reduce vulnerability by hindering attackers' ability to compute useful perturbations, gradient masking does not provide a fundamental solution. Sophisticated adversaries can often circumvent such defenses, exploiting alternative pathways or bypassing obfuscated gradients altogether, rendering gradient masking an incomplete security measure.

Feature Squeezing

Feature squeezing defends against adversarial inputs by simplifying the input data, thereby limiting the potential for subtle perturbations to influence the model's decision. Techniques may include reducing image resolution, bit-depth compression, or spatial smoothing, which remove fine-grained details that adversaries typically manipulate. By interpreting the difference between the results produced by original input provided to models against those on squeezed versions, discrepancies indicative of adversarial presence can be detected. Although effective for certain attack types, aggressive feature squeezing can lead to loss of legitimate detail and degrade performance on non-adversarial inputs.

Ensemble Methods

Ensemble defenses leverage the combined strength of multiple models or classifiers to improve overall robustness. By aggregating predictions from diverse architectures or training regimes, the system dilutes the impact of adversarial inputs targeting any single model. This redundancy often reduces the chance of successful attacks, especially those exploiting specific model vulnerabilities. However, ensembles increase computational costs and may still be vulnerable to attacks that transfer seamlessly between the models in the ensemble.

Defensive Distillation

Defensive distillation is a training technique where a "student" model learns from the softened outputs (probability distributions) of a "teacher" model rather than from hard labels. This smooths the decision boundaries and reduces the model's sensitivity to small input changes, making it more challenging for adversarial perturbations to shift classifications. While defensive distillation can offer some resistance against simple attacks, more advanced adversaries have found ways to circumvent these defenses.

Detection Mechanisms

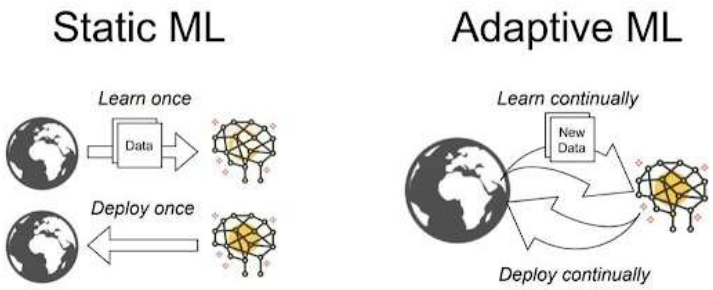

Figure 2. Illustration of continuous learning whereby AI models update iteratively with new adversarial examples

Defenses that depend on detection employ specialized detectors that process inputs and produce potential adversarial inputs before they are processed through the primary classification pipeline. The detectors used when deploying detection-based defenses will analyze statistical properties, distributions of features, or patterns of behavior in the inputs to flag atypical inputs or outliers.

Although these mechanisms add an additional level of security, the effectiveness of detection-based defenses is contingent on detection sensitivity and attacker maturity since quality adversarial examples can closely resemble real data and therefore pass detection.

Overall, while there is no technically-proven perfect defense, adversarial training is the most academically proven and effective means of improving model robustness. In addition, the effectiveness of adversarial training can be bolstered when combined with defenses like feature squeezing and detection mechanisms. This underlines the value of an adaptive, multi-layer, and defense-in-depth security approach.

E. Possible Solutions

Hybrid Defense Mechanisms

One of the popular trends implies the incorporation of multiple defensive techniques, e.g., adversarial training, anomalous detection systems, ensemble modeling, and ensemble learning. With these various techniques, the system acquires numerous strata of security, minimizing vulnerabilities of any one. Hybrid defenses have the advantage of combining both exposure through training to adverse stimuli, the ability to detect suspicious or strange workings of numbers, and the robustness generated by aggregating forecasts of several varied models to alleviate the influence of manipulations by opponents.

Explainable AI

Explainable artificial intelligence (XAI) is a collection of procedures and strategies that enable machine learning algorithms to create output and results that are understandable and dependable for human users. Deep learning is commonly mentioned in relation to explainable artificial intelligence, which is a crucial component of the transparency, accuracy, and fairness of a machine learning paradigm. When it comes to using artificial intelligence, organizations that aspire to build trust from the public can profit from XAI. They are shown to be able to have a better understanding of the behavior of an AI model and identify potential AI related issues with the assistance of XAI.

Robust Model Architectures

Designing models that have innate structural durability is another effective strategy for defending against adversarial attacks. This includes building architecture that houses redundant pathways that increase the number and complexity of decision boundaries. Applying regularization methods such as dropout or weight constraints is also effective.

Continuous Learning

Continuous learning is a process in which a model learns from newer data streams without being re-trained. It is also known as continuous machine learning.

Continuous learning models, in contrast to traditional approaches, which include training models on a static dataset, deploying them, and then updating themselves at fixed intervals, update their parameters in an iterative manner in order to reflect new variations in the data.

Through the process of learning from the most recent iteration and continuing to update its knowledge as new data becomes available, the model is able to better itself. Therefore, a model is able to be more adaptable against novel forms of adversarial attacks if it is able to analyze and learn from current trends.

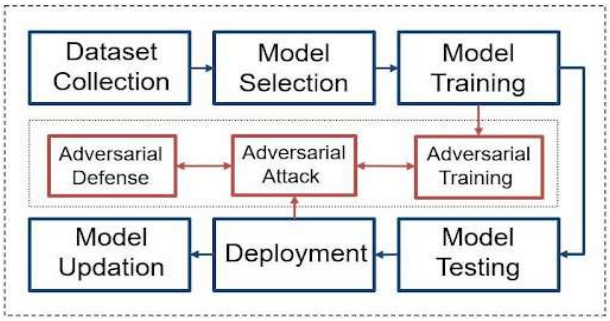

Figure 3. Adversarial Lifecycle

Standardized Benchmarks

Setting standard baselines is the key to promoting the progress in the field understanding of adversarial robustness. This includes the creation of datasets whose design is targeted at evaluation of defenses in a variety of adversarial scenarios, and creation of standardized metrics that objectively measures how a model can resist attacks. Common standards enable fair comparison, lead to advancements and roadmap jointly towards predictable, safe AI systems that can be deployed in the real-world environment.

F. Inference

The fact that adversarial attack methods are increasingly getting better highlights the importance of AI systems having defense strategies that are both flexible and strong. The inverse relationship between accuracy and defense integrity is a significant problem with the currently present defense strategies. Typically, fortifying a model against adversarial perturbations has a negative impact on its accuracy, resulting in poorer results when working with clean, unaltered data.

Usually, once AI models step outside lab conditions, they encounter new hindrances like random noise, shifting data patterns, and inputs that don't match what they were trained on. That makes holding accuracy constant a real challenge. The only way forward is to build defenses that can bend without breaking and adapt when required to.

IV. Conclusion

Adversarial attacks serve as a growing cautionary flag to the security, safety, resiliency, and integrity of AI-driven systems. While there are many different forms of attack and defense, the fact remains that many of the current methods of defense are so hardly robust to adapt to how quickly the adversaries can change their approaches. As a result, adversarial machine learning needs to continue searching for new and more robust methods of finding robust solutions that can be robust, scalable, and transferable for many types of problems. Therefore, research going forward in adversarial machine learning needs to focus on building and developing defense strategies that are capable of responding to newer adversarial attacks while monitoring the commonplace behavior of the input to respond in real time without losing any accuracy on the benign input. If that could be accomplished in the future, we can begin to alleviate some of the distrust and uncertainty, society has placed on AI, especially in the areas of health care, autonomous vehicles, and financial services.

Acknowledgment

We express our deepest gratitude to Dayananda Sagar College of Engineering for providing an environment conducive to research and learning. Special thanks to our Head of the Department Dr. Mohammed Tajuddin, and other professors for their invaluable guidance and support.

References

[1] F. O. Catak and S. Y. Yayilgan, "Deep neural network-based malicious network activity detection under adversarial machine learning attacks," 2020.
[2] S. M. Kazmi, N. Aafaq, M. A. Khan, and A. Saleem, "From Pixel to Peril: Investigating Adversarial Attacks on Aerial Imagery," 2023.
[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," ICLR, 2015.
[4] A. Madry et al., "Towards deep learning models resistant to adversarial attacks," ICLR, 2018.
[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," IEEE S&P, 2017.
[6] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool neural networks," CVPR, 2016.
[7] F. Tramèr et al., "The space of transferable adversarial examples," arXiv, 2017.
[8] Y. Liu et al., "Delving into transferable adversarial examples and black-box attacks," ICLR, 2017.
[9] K. Eykholt et al., "Robust physical world attacks on deep learning visual classification," CVPR, 2018.
[10] T. B. Brown et al., "Adversarial patch," arXiv, 2017.
[11] X. Yuan et al., "Adversarial examples: attacks and defenses for deep learning," IEEE TNNLS, 2019.
[12] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: a survey," IEEE Access, 2018.
[13] A. Athalye et al., "Synthesizing robust adversarial examples," ICML, 2018.
[14] N. Papernot et al., "The limitations of deep learning in adversarial settings," IEEE EuroS &P, 2016.
[15] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: detecting adversarial examples in neural networks," NDSS, 2018.
[16] E. Tabassi et al., "A taxonomy and terminology of adversarial machine learning," NIST IR 8269, 2019.
[17] J. H. Metzen et al., "Detecting adversarial perturbations in deep neural networks," arXiv, 2017.
[18] J. Chen et al., "Adversarial text generation via gradient-based optimization," ICLR, 2020.
[19] K. Ren et al., "Adversarial attacks and defenses in deep learning," Engineering, 2020.
[20] GeekForGeeks, "Explainable artificial intelligence," 2025.
[21] H. Chen, "Robustness enhancement of neural networks," 2025.
[22] iMerit, "A complete introduction to continual learning," 2025.
[23] Nature, "Hybrid defense mechanisms for adversarial robustness," 2025.