© NOV 2024 | IRE Journals | Volume 8 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1711957

Adversarial Attacks and Defense Mechanisms in AI

GEETHA ARADHYULA
Zolon Tech Inc

Abstract- ***The performance of Artificial Intelligence (AI) systems and especially those relying on deep learning has been impressive in the breadth of their application in computer vision, natural language processing, healthcare, and autonomous system domains. Nevertheless, their increased reliance on neural networks of large scale has left them vulnerable to a fatal attack--adversarial attacks. These attacks include the purposeful control of the input data by well-crafted, and typically invisible, perturbations that trigger AI models to make erroneous predictions or classifications. These counter-examples cast severe doubts on the resilience, stability, and safety of the AI-based technologies, with respect to safety-related purposes like self-driving cars, biometric authentication, and medical diagnosis. In this paper, the approach to adversarial attacks is described in detail, and they are divided into white-box, black-box, and gray-box models, depending on the knowledge of the attacker of the target system. It also discusses the various attack approaches including Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) and optimization-based approaches which exploit the sensitivities and gradients of models. Simultaneously, the paper examines the current defense mechanisms to enhance AI resilience, such as adversarial training, defensive distillation, input transformation, gradient masking, and certified robustness methods. Regardless of the great advances, the majority of defense mechanisms have a low scalability, generalization, or computing efficiency, creating a continuous arms race between the generation of adversarial attacks and defense of models. The paper ends by covering areas where the research is moving including explainable AI, robust optimization, and the incorporation of security-by-design concepts in the development of neural networks. To make AI models more trustworthy, transparent and safe to deploy in the real world, it is necessary to strengthen them against adversarial manipulation.***

I. INTRODUCTION

The performance of Artificial Intelligence (AI) systems and especially those relying on deep learning has been impressive in the breadth of their application in computer vision, natural language processing, healthcare, and autonomous system domains. Nevertheless, their increased reliance on neural networks of large scale has left them vulnerable to a fatal attack--adversarial attacks. These attacks include the purposeful control of the input data by well-crafted, and typically invisible, perturbations that trigger AI models to make erroneous predictions or classifications. These counter-examples cast severe doubts on the resilience, stability, and safety of the AI-based technologies, with respect to safety-related purposes like self-driving cars, biometric authentication, and medical diagnosis.

In this paper, the approach to adversarial attacks is described in detail, and they are divided into white-box, black-box, and gray-box models, depending on the knowledge of the attacker of the target system. It also discusses the various attack approaches including Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) and optimization-based approaches which exploit the sensitivities and gradients of models. Simultaneously, the paper examines the current defense mechanisms to enhance AI resilience, such as adversarial training, defensive distillation, input transformation, gradient masking, and certified robustness methods.

Regardless of the great advances, the majority of defense mechanisms have a low scalability, generalization, or computing efficiency, creating a continuous arms race between the generation of adversarial attacks and defense of models. The paper ends by covering areas where the research is moving including explainable AI, robust optimization, and the incorporation of security-by-design concepts in the development of neural networks. To make AI models more trustworthy, transparent and safe to deploy in the real world, it is necessary to strengthen them against adversarial manipulation.

II. BACKGROUND AND FUNDAMENTALS

To gain a better insight into the concept of adversarial attacks and their effects on Artificial Intelligence (AI) systems, one has to consider the basic concepts of machine learning and how such models arrive at decisions. Machine learning (ML) algorithms are created to be able to learn mappings between the input features and the output labels by minimizing a loss function on the training data. For example, in

supervised learning a classifier is trained to discriminate between dissimilar classes by putting up decision boundaries in high-dimensional feature spaces. Preferably, such boundaries are generalizable to unobservable data, but in general, such boundaries tend to be complicated and highly sensitive to minor distortions. Adversarial manipulation is based on this sensitivity.

2.1 Concept of Adversarial Examples

An adversarial example is a deliberately perturbed input that appears identical to a legitimate sample to the human eye but causes a machine learning model to produce an incorrect output. The perturbations are typically crafted using optimization techniques that exploit the gradients of the model's loss function. For example, adding minimal noise to an image can lead a neural network to misclassify an object, revealing a lack of robustness in the model's learned representations. These examples demonstrate that deep learning systems often rely on superficial correlations rather than truly understanding semantic features.

2.2 Taxonomy of Attacks

Adversarial attacks can be broadly classified into three categories based on the attacker's level of knowledge about the target model:

1. White-box attacks: The attacker has full access to the model's architecture, parameters, and gradients. Common methods include the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD).

2. Black-box attacks: The attacker has no internal knowledge of the model and can only observe input-output behavior. Attacks are generated using techniques such as transferability or surrogate model training.

3. Gray-box attacks: The attacker possesses partial information, such as the model architecture but not the exact parameters, combining elements of both white-box and black-box approaches.

2.3 Threat Models and Assumptions

A threat model defines the attacker's goals, capabilities, and knowledge about the system. Common goals include misclassification (causing the model to produce incorrect labels), targeted attacks (forcing the model to output a specific wrong class), and untargeted attacks (causing any misclassification). The assumptions about the attacker's access—whether they can modify training data, query the model, or inject malicious inputs—determine the strength and feasibility of the attack scenario.

2.4 Metrics for Evaluating Attack Success and Model Robustness

Evaluating adversarial robustness requires quantitative measures that capture both attack effectiveness and defense performance. Common attack success metrics include *attack success rate (ASR)*, which measures the proportion of successful misclassifications, and *perturbation magnitude*, often expressed in terms of L_0, L_2, or L_∞ norms. Meanwhile, robustness metrics evaluate how well a model maintains accuracy under adversarial conditions. These include *robust accuracy*, *certified robustness bounds*, and *adversarial confidence scores*. Together, these metrics help assess the trade-off between model accuracy, resilience, and computational cost.

III. TYPES OF ADVERSARIAL ATTACKS

Adversarial attacks can manifest at various stages of the machine learning pipeline, from training to deployment. Depending on the attacker's objective and the point of intervention, these attacks can be broadly categorized into evasion attacks, poisoning attacks, model inversion and extraction attacks, and physical world attacks. Each category exploits different vulnerabilities within AI systems, posing unique challenges to their robustness and security.

3.1 Evasion Attacks

Evasion attacks occur during the inference phase, where an adversary manipulates input data to deceive a trained model without altering its internal parameters. The attacker generates carefully crafted perturbations that are often imperceptible to humans but significantly affect the model's prediction.

Two of the most well-known examples are:

1. Fast Gradient Sign Method (FGSM): Introduced by Goodfellow et al. (2015),

© NOV 2024 | IRE Journals | Volume 8 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1711957

FGSM adds a small perturbation to the input data in the direction of the gradient of the loss function. The adversarial example is computed as
$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$,
where ϵ controls the perturbation size. Despite its simplicity, FGSM can effectively fool deep networks.

2. Projected Gradient Descent (PGD): A more advanced iterative version of FGSM, PGD repeatedly applies gradient-based perturbations and projects the modified input back onto the permissible data space. PGD is considered one of the strongest first-order attacks and serves as a benchmark for evaluating model robustness.

Evasion attacks highlight the fragility of decision boundaries and underscore the need for robust model design and defensive training strategies.

3.2 Poisoning Attacks

Poisoning attacks occur during the training phase, where an adversary corrupts the dataset used to train the model. By injecting malicious or mislabeled data, the attacker can manipulate the model's behavior once it is deployed. Poisoning attacks are particularly dangerous because they compromise the model before deployment and are difficult to detect.

Common examples include:

1. Label Flipping: The attacker changes the labels of certain training samples to distort the decision boundary. For example, altering labels in a spam detection dataset may cause the model to misclassify legitimate messages as spam or vice versa.

2. Backdoor Attacks: The attacker implants hidden triggers in the training data so that when a specific pattern appears in an input (e.g., a small patch or color), the model produces an incorrect output. These attacks are stealthy and have serious implications for real-world systems.

3.3 Model Inversion and Extraction Attacks

Model inversion and extraction attacks target the confidentiality of AI models rather than their accuracy.

1. In a model inversion attack, the adversary exploits access to the model's outputs (e.g., confidence scores or gradients) to reconstruct sensitive features of the training data. This can lead to privacy breaches, especially in applications involving biometric or medical data.

2. Model extraction attacks aim to steal the intellectual property of AI models. The attacker queries the target model and uses the responses to train a surrogate model that replicates its functionality. Such attacks can undermine the commercial value of proprietary AI systems and expose them to further exploitation.

3.4 Physical World Attacks

While many adversarial attacks operate in digital environments, recent research has demonstrated that adversarial examples can also succeed in the physical world. These attacks involve real-world manipulations that persist under varying lighting, angles, and distances.

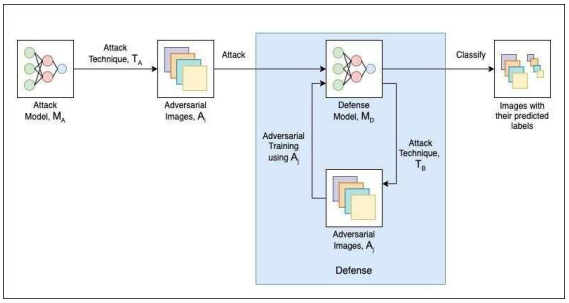
Examples include:

1. Adversarial Patches and Stickers: By applying specially designed stickers or patterns to objects—such as traffic signs—attackers can cause computer vision systems in autonomous vehicles to misclassify them.

2. 3D Adversarial Objects: Physical objects can be 3D-printed with adversarial textures that confuse perception models used in robotics or surveillance.

Such attacks expose the practical risks of adversarial vulnerabilities in safety-critical environments and emphasize the need for defenses that remain effective beyond digital simulations.

© NOV 2024 | IRE Journals | Volume 8 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1711957



IV. DEFENSE MECHANISMS

The growing threat of adversarial attacks has prompted extensive research into methods that can enhance the robustness and reliability of AI models. Defense mechanisms aim to either prevent the generation of effective adversarial examples or mitigate their impact on model predictions. These strategies can be broadly classified into six major categories: adversarial training, gradient masking and obfuscation, defensive distillation, input preprocessing, detection-based defenses, and certified or provable defenses. Each approach has its own strengths, limitations, and applicability across different domains.

4.1 Adversarial Training

Adversarial training is one of the most widely used and effective defense techniques. It involves augmenting the training dataset with adversarial examples so that the model learns to correctly classify both clean and perturbed inputs. The process can be formalized as a min–max optimization problem, where the model minimizes the worst-case loss under adversarial perturbations.

For instance, models trained using Projected Gradient Descent (PGD) adversarial samples have demonstrated enhanced robustness against gradient-based attacks. However, adversarial training significantly increases computational costs and may reduce generalization performance on clean data. Despite these trade-offs, it remains one of the strongest empirical defenses available.

4.2 Gradient Masking and Obfuscation

Gradient masking and obfuscation are techniques designed to hide or distort the gradient information that attackers rely on to craft adversarial examples. This can be achieved by introducing non-differentiable components, randomization, or gradient clipping within the model.

While these methods can temporarily impede certain white-box attacks, they often provide a false sense of security. Adaptive attackers can bypass gradient obfuscation by approximating gradients through numerical methods or transfer-based attacks. As a result, this defense category is generally considered insufficient on its own and should be combined with more robust methods.

4.3 Defensive Distillation

Defensive distillation is a training technique that leverages the concept of knowledge distillation to improve model robustness. Initially proposed by Papernot et al. (2016), it trains a secondary “student” model using the softened outputs (probability distributions) of a pre-trained “teacher” model. This process smooths the decision boundaries and reduces the model’s sensitivity to small input perturbations.

Although defensive distillation can improve resistance to some gradient-based attacks, later studies showed that it can still be circumvented by stronger or adaptive adversarial methods. Nevertheless, it remains a foundational concept in the development of modern robustness techniques.

4.4 Input Preprocessing

Input preprocessing techniques focus on transforming or sanitizing input data before feeding it into the model, thereby reducing the impact of adversarial perturbations. Common approaches include:

- Denoising filters to remove high-frequency noise components;
- Image compression (e.g., JPEG or bit-depth reduction) to eliminate imperceptible perturbations;
- Feature squeezing, which reduces the input space by combining similar features, making it harder for small perturbations to cause large output changes.

These methods are simple and computationally efficient but may degrade model accuracy on clean data. Additionally, adaptive attackers can often

© NOV 2024 | IRE Journals | Volume 8 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1711957

design perturbations that survive preprocessing transformations.

4.5 Detection-Based Defenses

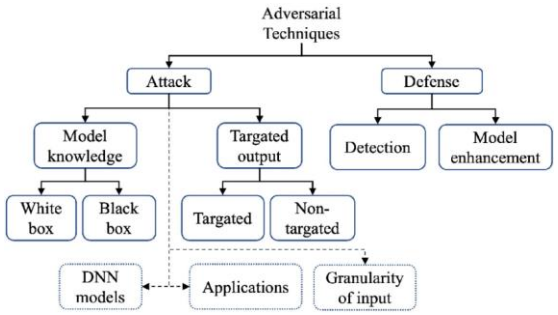
Detection-based defenses aim to identify whether an input sample is adversarial before classification. These methods typically train a separate detector or monitor abnormal behaviors in internal model activations, confidence scores, or input statistics. Techniques include statistical anomaly detection, secondary classifier models, and confidence thresholding.

While these approaches can effectively detect certain types of adversarial examples, they are often limited by transferability and may fail against novel or unseen attack strategies. Thus, they are generally used in combination with other defensive measures to provide layered protection.

4.6 Certified and Provable Defenses

Certified defenses represent a promising direction in adversarial robustness research. Unlike empirical defenses that rely on heuristic resistance, certified methods provide formal mathematical guarantees that a model's predictions will remain unchanged within a specified perturbation bound.

Approaches such as randomized smoothing, interval bound propagation (IBP), and Lipschitz regularization establish theoretical limits on how much an input can be perturbed without altering the output. Although these methods can offer provable robustness, they often involve substantial computational overhead and are typically limited to small-scale models.



V. EVALUATION AND BENCHMARKING

The evaluation of adversarial attacks and defense mechanisms is a crucial component in understanding the robustness, reliability, and practical viability of machine learning models. Robust benchmarking enables researchers to compare defense strategies objectively and to identify vulnerabilities across different architectures and datasets. This section presents the major datasets, performance metrics, and frameworks commonly used to assess adversarial robustness.

5.1 Datasets and Models Used

A variety of benchmark datasets are employed in adversarial research to evaluate both attack potency and defense resilience. These datasets differ in complexity, dimensionality, and domain, offering a diverse testing ground for adversarial analysis.

1. MNIST: A dataset of handwritten digits (0–9) that provides a simple and widely used benchmark for testing fundamental adversarial techniques. Due to its low resolution and grayscale nature, it is often used in introductory robustness studies.

2. CIFAR-10 and CIFAR-100: These datasets consist of small color images across 10 or 100 classes, respectively. They are commonly used to assess deep convolutional neural networks (CNNs) under more realistic and complex visual conditions.

3. ImageNet: A large-scale dataset with over a million high-resolution images across 1,000 categories. It serves as the standard benchmark for evaluating the scalability and transferability of adversarial attacks and defenses in state-of-the-art architectures such as ResNet, VGG, and Inception.

Other specialized datasets—such as SVHN, Fashion-MNIST, and GTSRB (German Traffic Sign Recognition Benchmark)—are also employed in domain-specific adversarial research, especially for physical-world and autonomous driving scenarios.

5.2 Robustness Metrics

Evaluating the effectiveness of adversarial attacks and defenses requires well-defined quantitative metrics. The most common robustness metrics include:

1. Accuracy Under Attack: Measures the classification accuracy of a model when subjected to adversarial perturbations. A lower accuracy indicates higher vulnerability.

2. Perturbation Norms: Quantify the magnitude of adversarial modifications. The most frequently used norms are:

    a. L_0 norm: Counts the number of pixels changed.

    b. L_2 norm: Measures the Euclidean distance between clean and perturbed samples.

    c. L_∞ norm: Represents the maximum change to any individual feature or pixel.

3. Attack Success Rate (ASR): The percentage of adversarial examples that successfully fool the model.

4. Robust Accuracy / Certified Robustness: Measures the fraction of examples that remain correctly classified within a predefined perturbation radius.

5. Computational Cost: Evaluates the time or resources required to perform an attack or defense, a crucial factor for real-time applications.

Together, these metrics enable a comprehensive assessment of the trade-offs between robustness, model accuracy, and computational efficiency.

5.3 Common Frameworks and Libraries

To facilitate standardized evaluation and reproducibility, several open-source libraries have been developed to implement adversarial attacks and defenses across multiple machine learning platforms.

The most widely used include:

1. CleverHans: A Python library developed by Google Brain for benchmarking adversarial robustness. It provides implementations of various attacks (FGSM, PGD, DeepFool) and defenses for TensorFlow-based models.

2. Foolbox: A flexible framework compatible with multiple deep learning libraries (PyTorch, TensorFlow, JAX). It supports gradient-based and decision-based attacks and provides tools for evaluating model robustness under standardized conditions.

3. Adversarial Robustness Toolbox (ART): Developed by IBM, ART offers a comprehensive suite of attack, defense, and verification algorithms. It is designed for industrial-scale evaluation and supports integration with TensorFlow, PyTorch, Keras, and MXNet.

These frameworks have become essential tools for researchers to benchmark new algorithms, compare robustness across datasets, and reproduce experimental results in a consistent and transparent manner.

VI. CHALLENGES AND LIMITATIONS

Despite significant progress in understanding and mitigating adversarial vulnerabilities, several challenges and limitations persist in developing truly robust Artificial Intelligence (AI) systems. Current defense mechanisms often struggle to achieve a balance between robustness, computational efficiency, and generalization, while attackers continue to design increasingly sophisticated techniques to bypass them. This section discusses the primary challenges that hinder practical adversarial defense deployment.

6.1 Trade-off Between Robustness and Model Accuracy

One of the most critical issues in adversarial defense research is the trade-off between model robustness and accuracy on clean data. Models trained for higher robustness—particularly through adversarial training—tend to sacrifice performance on unperturbed inputs. This phenomenon arises because robustness-oriented optimization prioritizes stability under perturbations, which can restrict the model's ability to capture fine-grained distinctions in the data.

Striking a balance between these two competing objectives remains an open problem, as overemphasis on robustness may lead to underfitting, while

focusing solely on accuracy increases vulnerability to adversarial manipulation.

6.2 Computational Cost of Defenses

Many state-of-the-art defense mechanisms, such as adversarial training or certified robustness methods, are computationally expensive. They often require generating adversarial examples during every training iteration, which significantly increases training time and memory consumption. For large-scale datasets like ImageNet or models with millions of parameters, these approaches become impractical for real-world applications. Additionally, certified defenses demand rigorous mathematical verification, further increasing resource requirements. This computational burden limits the scalability and adoption of adversarial defenses in industry settings.

6.3 Adaptive and Transferable Attacks

A major challenge in adversarial robustness research is the emergence of adaptive and transferable attacks. Adaptive attackers can modify their strategies to specifically target and bypass existing defenses by exploiting their weaknesses. Similarly, transferable attacks—where adversarial examples generated for one model successfully deceive another—demonstrate that robustness cannot be guaranteed even for models that have never encountered those specific perturbations.

This adaptability of adversaries exposes the fragility of many defense methods, which often perform well only under specific, predefined attack conditions but fail against more general or unseen threat models.

6.4 Lack of Standardized Benchmarks

The absence of standardized evaluation benchmarks poses a significant limitation in adversarial machine learning research. Differences in datasets, attack configurations, and evaluation criteria make it difficult to fairly compare the performance of various defense methods. Moreover, many studies rely on a limited set of attack types or specific parameter settings, leading to inconsistent and sometimes misleading conclusions about robustness. Efforts such as RobustBench and standardized libraries like ART and Foolbox have begun to address this issue, but a universally accepted framework for evaluating adversarial robustness across models and domains is still lacking.

VII. CONCLUSION

The increasing integration of Artificial Intelligence (AI) and Machine Learning (ML) systems into critical domains such as healthcare, autonomous transportation, finance, and security underscores the importance of ensuring their robustness and trustworthiness. This paper has examined the growing threat of adversarial attacks—malicious manipulations that exploit the vulnerabilities of deep learning models—and the corresponding defense mechanisms designed to mitigate these risks.

A comprehensive overview was presented, covering various categories of adversarial attacks including evasion, poisoning, model inversion and extraction, and physical-world attacks. Each of these attack types targets different stages of the AI pipeline, revealing that even highly accurate models can be susceptible to small, carefully crafted perturbations. To counter these threats, a range of defense mechanisms have been developed, such as adversarial training, gradient masking, defensive distillation, input preprocessing, and certified defenses, each contributing uniquely to enhancing model resilience.

However, despite substantial research progress, numerous challenges remain unresolved. Current defense techniques often face trade-offs between robustness and accuracy, high computational costs, and limited generalization against adaptive or transferable attacks. Furthermore, the lack of standardized benchmarks continues to hinder consistent evaluation and comparison of robustness across models and datasets.

Looking forward, the development of trustworthy and secure AI systems requires a holistic approach that combines theoretical robustness guarantees, efficient training algorithms, and practical deployment strategies. Future research should emphasize explainable and interpretable AI, certifiable robustness, and security-by-design principles that integrate robustness into every stage of model development. Collaboration between academia, industry, and policymakers will be essential to establish global standards and ethical frameworks for secure AI deployment.

REFERENCES

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[2] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

[3] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *Proceedings of the 2017 IEEE Symposium on Security and Privacy*.

[4] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

[5] Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

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

[7] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. *2016 IEEE Symposium on Security and Privacy (SP)*.

[8] Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). DeepFool: a simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

[10] Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv preprint arXiv:2003.01690*.

[11] Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., & Hossain, E. (2023). Adversarial attacks and defenses in machine learning-powered networks: A contemporary survey. *arXiv preprint arXiv:2303.06302*.

[12] Wu, B., Zhu, Z., Liu, L., Q. Liu, He, Z., & Lyu, S. (2023). Attacks in adversarial machine learning: A systematic survey from the life-cycle perspective. *arXiv preprint arXiv:2302.09457*.

[13] NIST (Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M.). (2025). *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (NIST AI 100-2e2025). National Institute of Standards and Technology.

[14] Carlini, N. (2019). A complete list of all adversarial example papers. Retrieved from https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

[15] Ibitoye, O., Abou-Khamis, R., elShehaby, M., Matrawy, A., & Shafiq, M. O. (2025). The threat of adversarial attacks against machine learning in network security: A survey. *Journal of Electronics and Electrical Engineering*.

[16] Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access, 6*, 14410–14430.

[17] Zhou, Y., Kantarcioglu, M., & Xi, B. (2019). A survey of game theoretic approach for adversarial machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(e1259).

[18] Dusdu, V. (2018). A survey of adversarial machine learning in cyber warfare. *Defence Science Journal, 68*, 356.

[19] Bhagoji, A. N., Cullina, D., & Mittal, P. (2018). Dimensionality's curse: defenses fail for high-dimensional spaces. *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

[20] Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*.

[21] Salman, H., Yang, G., Newsam, S., Basu, S., & Natarajan, K. (2020). Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1805.11774*.

[22] Raghunathan, A., Steinhardt, J., & Liang, P. (2018). Certified defenses against adversarial examples. *Proceedings of the 6th International*

Conference on Learning Representations (ICLR).

[23] Lecuyer, M., Atlidakis, V., Gehr, T., Miracle-Sole, C., & Vechev, M. (2019). Certified robustness to adversarial examples with differential privacy. *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*.

[24] Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: improving robustness to adversarial examples. *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

[25] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2020). Adversarial training is not robust enough — toward better defenses. *ArXiv Preprint*.