International Journal of Innovative Computer Science and IT Research

Volume 01, Issue 04 – Jul-Aug 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

Adversarial Machine Learning for Evaluating the Robustness of Intelligent Systems

Prof. Sophia Reynolds

Associate Professor, University of Maryland, USA

Abstract

The increasing deployment of artificial intelligence and machine learning systems in critical computing applications has created a growing need to evaluate their robustness against intentional and unintentional perturbations. Intelligent systems are often trained and evaluated under assumptions that operational data will remain similar to training distributions, but adversarial machine learning demonstrates that carefully designed input modifications can cause significant changes in model predictions. This research examines adversarial machine learning as a methodological approach for evaluating and improving the robustness of intelligent systems. The study focuses on adversarial examples, threat models, attack strategies, robustness assessment, adversarial training, defensive mechanisms, and security-oriented evaluation. A conceptual framework is developed that connects intelligent-system inputs, threat modeling, adversarial perturbation generation, model response analysis, robustness measurement, and defensive improvement. The analysis considers major categories of adversarial attacks, including gradient-based, optimization-based, query-based, and data-level attacks, while examining their relevance to image recognition, natural language processing, cybersecurity, autonomous systems, and intelligent information processing.

The research emphasizes that robustness evaluation should not be restricted to measuring standard predictive accuracy because a model may demonstrate high performance on clean data while remaining highly sensitive to carefully constructed perturbations. Robustness should instead be evaluated through multiple dimensions, including prediction stability, perturbation sensitivity, attack success, generalization, computational resilience, and recovery behavior. Adversarial training, input preprocessing, robust optimization, defensive distillation, ensemble mechanisms, uncertainty estimation, and certified robustness approaches are discussed as potential defense strategies. The study further identifies challenges related to adaptive attackers, computational cost, transferability, distribution shift, defense degradation, and evaluation reliability. The research concludes that adversarial machine learning provides an important framework for systematically testing intelligent systems under hostile or abnormal conditions. Integrating adversarial evaluation into the machine learning lifecycle can improve the reliability, security, and resilience of intelligent computing systems deployed in complex real-world environments.

Keywords: Adversarial Machine Learning, Intelligent Systems, Robustness, Adversarial Examples, Machine Learning Security, Adversarial Attacks, Robust AI, Deep Learning, Cybersecurity, Model Evaluation

International Journal of Innovative Computer Science and IT Research

Volume 01, Issue 04 – Jul-Aug 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

1. Introduction

Artificial intelligence and machine learning systems are increasingly integrated into applications involving classification, prediction, recognition, recommendation, cybersecurity, autonomous control, and intelligent information processing. These systems frequently operate in environments where inputs may be noisy, incomplete, manipulated, or intentionally modified. Conventional model evaluation generally measures performance using clean test datasets that reflect the expected data distribution. Although such evaluation is necessary, it may not adequately reveal how a model behaves when inputs are deliberately designed to exploit weaknesses in its decision process.

Adversarial machine learning provides a framework for studying these weaknesses. An adversarial example is an input that has been intentionally modified in a manner that can cause a machine learning model to produce an incorrect or undesirable prediction. In many cases, the modification may be difficult for humans to notice while remaining highly influential for the model. The discovery of such vulnerabilities has demonstrated that predictive accuracy on conventional benchmark datasets does not necessarily guarantee robustness under adversarial conditions.

Adversarial attacks can be constructed using different levels of knowledge about the target system. White-box attacks assume substantial knowledge of model architecture and parameters, while black-box attacks attempt to generate effective perturbations using limited information, often through queries or transferability. Other attacks target the training process itself by manipulating training data or labels. These different threat models demonstrate that robustness evaluation must consider not only the model architecture but also the conditions under which the intelligent system operates.

The importance of adversarial robustness extends across multiple areas of computer science and information technology. Image recognition systems may be affected by small visual perturbations, natural language systems may respond differently to carefully modified text, and cybersecurity models may be challenged by strategically altered network or system data. Autonomous systems can face additional risks because incorrect predictions may influence physical or operational decisions. Consequently, robustness evaluation should be incorporated into system development rather than treated as an optional post-deployment activity.

Defensive strategies have developed alongside adversarial attack research. Adversarial training exposes models to adversarially modified examples during training, while robust optimization attempts to improve model behavior across possible perturbation ranges. Input preprocessing, ensemble methods, uncertainty estimation, detection mechanisms, and certified robustness techniques provide additional approaches. However, defenses may introduce computational costs or become ineffective against adaptive attackers that modify their strategies according to the defense mechanism.

This research examines adversarial machine learning as an approach for evaluating the robustness of intelligent systems. The study develops a conceptual framework covering threat modeling, perturbation generation, model evaluation, robustness measurement, and defensive improvement. It analyzes major adversarial attack categories and their relevance to intelligent computing applications while identifying challenges associated with adaptive threats, transferability, computational requirements, and reliable evaluation. The research aims to provide a structured foundation for incorporating adversarial robustness assessment into intelligent-system development.

2. Objectives

Primary Objective

The primary objective of this research is to examine adversarial machine learning as a systematic approach for evaluating the robustness of intelligent systems. The study focuses on adversarial threats, robustness assessment techniques, defensive mechanisms, and methodological requirements for improving the resilience of machine learning models under abnormal or intentionally manipulated inputs.

2.1 Development of an Adversarial Robustness Framework

The first objective is to develop a conceptual framework connecting threat identification, adversarial example generation, model testing, robustness measurement, and defensive improvement. The framework emphasizes continuous evaluation of intelligent systems under both conventional and adversarial operating conditions.

2.2 Analysis of Adversarial Attack Strategies

The second objective is to examine major adversarial attack approaches, including gradient-based, optimization-based, query-based, transfer-based, and data-poisoning techniques. The objective is to understand how different attack assumptions can expose weaknesses in intelligent classification and prediction systems.

2.3 Examination of Robustness and Defense Mechanisms

The third objective is to analyze adversarial training, robust optimization, input transformation, ensemble methods, detection mechanisms, and certified robustness approaches. The objective is to identify the requirements for developing intelligent systems that maintain reliable behavior under adversarial conditions.

3. Review of Literature

3.1 Foundations of Adversarial Machine Learning

Adversarial machine learning emerged from research demonstrating that machine learning models can be sensitive to carefully designed input perturbations. Szegedy et al. showed that small changes to inputs could result in different model predictions, establishing an important foundation for adversarial-example research. Goodfellow et al. subsequently introduced the fast gradient sign method and provided an influential interpretation of adversarial vulnerability through the behavior of high-dimensional linear models.

Subsequent research demonstrated that adversarial examples could be generated using different optimization and gradient-based methods. Carlini and Wagner developed powerful optimization-based attacks, while Madry et al. formulated adversarial robustness as a robust optimization problem. These studies contributed to the understanding that robustness should be explicitly evaluated rather than inferred from standard test accuracy.

3.2 Adversarial Threat Models

Adversarial attacks can be categorized according to the attacker's knowledge, access, and objective. White-box attacks assume access to model parameters and gradients, allowing the attacker to construct

Volume 01, Issue 04 – Jul-Aug 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

highly targeted perturbations. Black-box attacks operate with more restricted information and may rely on model queries, surrogate models, or transferability.

Threat models can also differ according to the stage at which the attack occurs. Evasion attacks manipulate inputs at inference time, while poisoning attacks target training data or learning processes. Backdoor attacks introduce hidden behaviors that may be triggered by specific patterns. These distinctions are important because a defense designed for one threat model may not provide sufficient protection against another.

3.3 Robustness Evaluation and Defense

Robustness evaluation attempts to determine whether a model maintains appropriate behavior when exposed to perturbations. Adversarial accuracy, attack success rate, perturbation magnitude, and certified robustness are among the measures used to characterize model resilience. However, no single measure fully captures robustness because different attacks and threat models can reveal different vulnerabilities.

Adversarial training is one of the most extensively studied defense mechanisms. It incorporates adversarial examples into model training so that the resulting classifier becomes less sensitive to selected perturbations. Other approaches include input transformations, defensive ensembles, randomized mechanisms, robust feature learning, and formal verification. Research has also emphasized the importance of evaluating defenses against adaptive attacks rather than relying on weak or outdated attack procedures.

4. Methodology

4.1 Research Design

This study adopts a conceptual and analytical research design for examining adversarial machine learning as a framework for robustness evaluation. The methodology integrates adversarial threat modeling, perturbation generation, model testing, robustness measurement, and defense analysis. Rather than presenting fabricated experimental results, the study provides a structured methodology for evaluating intelligent systems under different adversarial conditions.

The research process begins by identifying the target intelligent system, its inputs, outputs, operational environment, and security requirements. A threat model is then defined according to attacker knowledge, capabilities, objectives, and access. Suitable adversarial scenarios are subsequently established, followed by controlled generation or simulation of adversarial conditions. Model responses are evaluated against clean and manipulated inputs to identify differences in predictive behavior.

4.2 Adversarial Robustness Evaluation Model

The proposed model represents robustness evaluation as a sequence of controlled stages:

System Definition → Threat Modeling → Perturbation Generation → Model Evaluation → Robustness Measurement → Defense Analysis → Re-Evaluation

Let a trained classifier be represented by:

fθ(x)=yf_\theta(x)=y

where xx is the input, fθf_\theta is the trained model, and yy is the predicted output.

International Journal of Innovative Computer Science and IT Research

Volume 01, Issue 04 – Jul-Aug 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

An adversarial input can be represented as:

$x' = x + \delta$

where δ represents an adversarial perturbation subject to a defined constraint.

The robustness evaluation examines whether:

$f_\theta(x') = f_\theta(x)$

remains true under the specified threat conditions. The objective is not merely to determine whether an attack can alter a prediction, but to characterize the conditions under which model behavior becomes unstable.

4.3 Key Methodological Components

The proposed methodology incorporates the following components:

- **System Profiling:** Identification of model architecture, input types, outputs, and deployment conditions.
- **Threat Modeling:** Definition of attacker capabilities, knowledge, access, and objectives.
- **Adversarial Scenario Design:** Development of controlled perturbation conditions.
- **Perturbation Generation:** Creation of adversarial inputs using appropriate attack methodologies.
- **Clean-Data Evaluation:** Measurement of baseline model performance.
- **Adversarial Evaluation:** Assessment of model behavior under manipulated inputs.
- **Robustness Measurement:** Analysis of prediction stability, attack success, and perturbation sensitivity.
- **Defense Assessment:** Evaluation of protective mechanisms.
- **Adaptive Re-Evaluation:** Testing defenses against updated or adaptive attack strategies.

5. Analysis

5.1 Architecture of Adversarial Robustness Evaluation

An effective adversarial evaluation architecture should examine intelligent systems from both normal and hostile operating perspectives. The process begins with system profiling and threat modeling. This determines what types of attacks are relevant and what level of access an adversary may have. Attack-generation mechanisms can then produce controlled perturbations within specified constraints.

Table 1: Components of an Adversarial Robustness Evaluation Framework

Component	Primary Function	Robustness Evaluation Role
System Profiling	Defines model and operational characteristics	Establishes evaluation scope
Threat Model	Defines attacker capabilities	Determines relevant attack conditions
Clean Dataset	Provides normal inputs	Establishes baseline behavior
Adversarial Generator	Creates manipulated inputs	Tests model vulnerability

Component	Primary Function	Robustness Evaluation Role
Attack Constraint	Limits perturbation conditions	Defines realistic threat boundaries
Model Evaluation	Measures system responses	Identifies prediction instability
Robustness Metrics	Quantifies resilience	Supports systematic comparison
Defense Mechanism	Reduces vulnerability	Improves resistance to attacks
Re-Evaluation Layer	Tests improved model	Determines defense effectiveness

Interpretation: Table 1 demonstrates that adversarial robustness evaluation is a complete assessment lifecycle rather than a single attack procedure. Threat modeling determines which adversarial scenarios should be considered, while clean-data evaluation establishes baseline performance.

Adversarial generation and model evaluation then reveal vulnerabilities. Robustness metrics provide evidence for comparison, while re-evaluation is essential for determining whether defensive mechanisms actually improve resilience against relevant threats.

5.2 Adversarial Attack Categories

Adversarial attacks can be distinguished according to how perturbations are generated and the amount of information available to the attacker. Gradient-based attacks use model gradients to identify directions that can modify predictions efficiently. Optimization-based approaches formulate adversarial generation as an optimization problem. Query-based attacks attempt to infer model behavior by repeatedly interacting with the target system without direct access to internal parameters.

Table 2: Major Adversarial Machine Learning Approaches for Robustness Evaluation

Attack Category	Primary Target	General Mechanism	Evaluation Purpose
Gradient-Based Attack	Model input	Uses gradient information	Tests local prediction sensitivity
Optimization-Based Attack	Model decision boundary	Optimizes perturbation objectives	Evaluates stronger targeted manipulation
Query-Based Attack	Black-box model	Uses model responses	Tests limited-access robustness
Transfer-Based Attack	Related models	Exploits transferable perturbations	Evaluates cross-model vulnerability
Data Poisoning	Training process	Manipulates training examples	Tests learning-stage resilience

Attack Category	Primary Target	General Mechanism	Evaluation Purpose
Backdoor Attack	Training/model behavior	Embeds trigger-dependent behavior	Tests hidden model vulnerabilities
Physical-World Attack	Deployed intelligent system	Uses real-world modifications	Evaluates operational robustness
Adaptive Attack	Defense mechanism	Adjusts strategy against defenses	Tests defense reliability

Interpretation: Table 2 shows that adversarial robustness cannot be assessed using a single attack technique. Different attacks expose different classes of vulnerabilities. Gradient-based attacks are useful for analyzing models with accessible gradients, while query-based and transfer-based attacks provide insight into more restricted scenarios.

Data poisoning and backdoor attacks evaluate vulnerabilities introduced during training. Adaptive attacks are particularly important because they test whether a defense remains effective when the attacker understands its mechanisms.

5.3 Robustness Evaluation Factors

Robustness should be evaluated through multiple dimensions. One important factor is prediction stability, which measures whether a model maintains consistent decisions when inputs are subjected to permitted perturbations. Another is attack success, which indicates whether an adversarial strategy can cause an undesirable prediction. Perturbation magnitude can also provide information about how much manipulation is required to influence the model.

Computational cost should also be considered. Some robustness techniques require substantial additional training or inference resources. A defense that significantly increases system latency may be unsuitable for real-time applications. Therefore, robustness evaluation should consider security, predictive performance, computational efficiency, and operational requirements simultaneously.

6. Discussion

6.1 Adversarial Evaluation as a Robustness Mechanism

Adversarial testing provides a systematic way to expose weaknesses that may remain hidden during conventional model evaluation. Standard accuracy measures demonstrate how a model performs on expected data, but they do not necessarily reveal how the model responds to intentionally manipulated inputs. Adversarial evaluation introduces controlled stress conditions that can provide additional information about decision stability.

This approach is particularly valuable for intelligent systems deployed in environments where inputs cannot be fully trusted. Cybersecurity systems, autonomous applications, intelligent monitoring, and automated decision-support platforms may encounter strategically modified information. Evaluating models under such conditions can improve understanding of their operational limitations.

Volume 01, Issue 04 – Jul-Aug 2025 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

6.2 Adversarial Training and Robust Models

Adversarial training is an important approach for improving model resilience. During training, the model is exposed to selected adversarial examples, encouraging it to learn decision boundaries that are less sensitive to those perturbations. Robust optimization provides a related perspective by attempting to optimize model performance under a defined range of possible perturbations.

However, adversarial training does not automatically provide universal robustness. A model trained against one attack family may remain vulnerable to another. Stronger defenses can also increase computational requirements or affect clean-data performance. Consequently, robustness assessment should use diverse and appropriately strong evaluation methods.

6.3 Adaptive Threats and Intelligent Defense

An important characteristic of adversarial machine learning is the continuous interaction between attacks and defenses. When a new defense mechanism becomes known, attackers can potentially adapt their strategies to bypass it. This creates an ongoing security problem in which robustness must be evaluated against adaptive threat models.

Intelligent defense mechanisms should therefore incorporate continuous monitoring, attack detection, uncertainty estimation, and periodic robustness testing. A mature security architecture should not assume that a single robustness evaluation guarantees long-term protection. Instead, adversarial assessment should become a recurring component of the machine learning lifecycle.

7. Challenges

One major challenge is the difficulty of defining realistic threat models. Adversaries may possess different levels of knowledge, computational resources, and access to the target system. An evaluation based only on unrealistic white-box assumptions may overestimate operational risk, while an evaluation using weak black-box scenarios may underestimate vulnerability. Therefore, robustness testing should consider multiple threat conditions relevant to the intended deployment environment.

Another challenge is the transferability and adaptability of adversarial attacks. An adversarial example created for one model may sometimes affect another model, creating risks even when the attacker does not have direct access to the deployed architecture. Adaptive attackers can also modify their strategies after observing defensive mechanisms. This makes it difficult to claim that a model is robust based on a limited set of attack procedures.

Computational cost and evaluation complexity also create practical difficulties. Generating strong adversarial examples may require substantial processing resources, particularly for large deep learning models. Robust training can further increase training time and infrastructure requirements. In addition, robustness results can depend on attack parameters, perturbation constraints, datasets, and evaluation procedures. Standardized and reproducible evaluation methodologies are therefore necessary for meaningful comparison.

8. Future Research

Future research should investigate automated adversarial evaluation systems capable of continuously testing deployed intelligent models. Such systems could monitor model behavior, identify emerging vulnerabilities, generate relevant adversarial scenarios, and report changes in robustness over time.

Integration with machine learning operations platforms could allow adversarial testing to become part of regular model development and deployment workflows.

Another important research direction is certified and formally verified robustness. Conventional empirical testing can evaluate only selected attack scenarios, whereas certified approaches attempt to provide mathematical guarantees within defined perturbation regions. Improving the scalability of certification methods for large neural networks remains an important challenge. Research should also explore combinations of empirical adversarial testing and formal robustness verification.

Future studies should further examine multimodal and generative intelligent systems. Modern AI systems increasingly process text, images, audio, video, and structured information simultaneously. Adversarial manipulation in one modality may influence decisions across other modalities. Robustness evaluation should therefore move beyond isolated classification models and investigate multimodal interactions, foundation models, autonomous agents, and complex AI pipelines.

9. Conclusion

Adversarial machine learning provides an important methodological foundation for evaluating the robustness of intelligent systems under intentionally manipulated or abnormal inputs. Conventional machine learning evaluation primarily measures performance on clean datasets, but such testing may not reveal vulnerabilities caused by carefully designed perturbations. This research examined adversarial machine learning through a structured framework involving system profiling, threat modeling, adversarial scenario design, perturbation generation, model evaluation, robustness measurement, defense assessment, and re-evaluation.

The analysis demonstrated that robustness should be treated as a multidimensional property involving prediction stability, attack resistance, perturbation sensitivity, computational efficiency, and operational reliability. By incorporating adversarial evaluation into the development lifecycle, organizations can obtain a more realistic understanding of how intelligent systems behave under hostile conditions. Different adversarial approaches reveal different categories of vulnerability. Gradient-based and optimization-based methods can expose weaknesses in decision boundaries, while query-based and transfer-based approaches examine robustness under restricted information conditions.

Data poisoning and backdoor attacks address vulnerabilities introduced during training, whereas physical-world and adaptive attacks evaluate the resilience of deployed systems. Defensive mechanisms such as adversarial training, robust optimization, input transformation, ensemble approaches, uncertainty estimation, and formal verification can improve resistance, but no individual mechanism should be considered universally effective. Robustness must instead be evaluated against diverse and appropriately defined threat models.

Overall, adversarial robustness should become an integral component of intelligent-system engineering rather than an isolated security exercise. Future intelligent computing environments will increasingly require models that can operate reliably despite changing data distributions, uncertain inputs, and potentially malicious manipulation. Continuous adversarial testing, adaptive defense mechanisms, formal robustness guarantees, and automated security monitoring can strengthen this capability. Research should also extend robustness evaluation to multimodal systems, foundation models, autonomous agents, and interconnected AI pipelines. By combining predictive intelligence with

systematic adversarial assessment, intelligent systems can become more secure, resilient, and dependable for real-world computer science and information technology applications.

References

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations*. https://arxiv.org/abs/1412.6572
2. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*. https://arxiv.org/abs/1312.6199
3. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy* (pp. 372–387). IEEE. https://doi.org/10.1109/EuroSP.2016.36
4. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*. https://arxiv.org/abs/1706.06083
5. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy* (pp. 39–57). IEEE. https://doi.org/10.1109/SP.2017.49
6. Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2574–2582). IEEE. https://doi.org/10.1109/CVPR.2016.282
7. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. *International Conference on Learning Representations Workshop*. https://arxiv.org/abs/1607.02533
8. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 274–283). PMLR.
9. Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*. https://arxiv.org/abs/1705.07204
10. Xu, W., Evans, D., & Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2018.23153
11. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023
12. Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems, 30*(9), 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886017
13. Wiyatno, R., Xu, A., Dia, O., & de Berker, A. (2019). Adversarial examples in modern machine learning: A review. *arXiv*. https://arxiv.org/abs/1911.05268
14. Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering, 6*(3), 346–360. https://doi.org/10.1016/j.eng.2019.12.012

15. Wilson, G., & Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology, 11*(5), Article 51. https://doi.org/10.1145/3400066
16. Han, S., Lin, C., Shen, C., Wang, Q., & Guan, X. (2023). Interpreting adversarial examples in deep learning: A review. *ACM Computing Surveys, 55*(14s), Article 328. https://doi.org/10.1145/3594869
17. Meyers, C., Löfstedt, T., & Elmroth, E. (2023). Safety-critical computer vision: An empirical survey of adversarial evasion attacks and defenses on computer vision systems. *Artificial Intelligence Review, 56*, 217–251. https://doi.org/10.1007/s10462-023-10521-4
18. Vassilev, A., Oprea, A., Fordyce, A., & Andersen, H. (2024). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (NIST AI 100-2e2023). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2023
19. Peck, J., Goossens, B., & Saeys, Y. (2024). An introduction to adversarially robust deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(4), 2071–2090. https://doi.org/10.1109/TPAMI.2023.3331087
20. Liu, J., Li, Y., Guo, Y., Liu, Y., Tang, J., & Nie, Y. (2024). Generation and countermeasures of adversarial examples on vision: A survey. *Artificial Intelligence Review, 57*, 199. https://doi.org/10.1007/s10462-024-10841-z