World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

# Enhancing the security of AI-driven autonomous systems through adversarially robust deep learning models

Emmanuel Ayodeji Osoko [1, *] Shukurat Opeyemi Rahmon [2] and Muhammed Azeez [3]

[1] *Department of Electrical Engineering and Computer Science, Ohio University, OH, USA.*

[2] *Department of Mathematics, University of Lagos, Akoka, Lagos, Nigeria.*

[2] *Department of Mathematics, Lamar University, Beaumont, TX, USA.*

World Journal of Advanced Research and Reviews, 2023, 20(01), 1336-1351

Publication history: Received on 13 September 2023; revised on 24 October 2023; accepted on 26 October 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.20.1.2158

## Abstract

Adversarial attacks pose a significant threat to AI-driven autonomous systems by exploiting vulnerabilities in deep learning models, leading to erroneous decision-making in safety-critical applications. This study investigates the effectiveness of adversarial training as a defense mechanism to enhance model robustness against adversarial perturbations. We evaluate multiple deep learning architectures subjected to Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (CW) attacks, comparing adversarially trained models with standard models in terms of accuracy, robustness, and computational efficiency. The results demonstrate that adversarial training significantly improves resistance to adversarial attacks, reducing attack success rates by over 50% while maintaining high classification performance. However, a trade-off between robustness and inference time was observed, highlighting computational cost concerns. Furthermore, our findings reveal that adversarial robustness partially transfers across architectures but remains susceptible to advanced optimization-based attacks. This study contributes to the development of more secure AI-driven autonomous systems by identifying strengths and limitations of adversarial training, offering insights into future improvements in adversarial defense strategies.

**Keywords:** Adversarial Machine Learning; DL Security; Cybersecurity In AI; Neural Network Vulnerabilities

## 1. Introduction

Artificial Intelligence (AI)-driven autonomous systems have rapidly evolved, revolutionizing fields such as self-driving vehicles, robotics, industrial automation, and smart surveillance. These systems leverage deep learning models to make real-time decisions based on sensor inputs, significantly improving efficiency, adaptability, and accuracy (LeCun et al., 2019). However, the increased reliance on deep neural networks (DNNs) has introduced vulnerabilities that adversarial attackers can exploit, posing severe security threats. Unlike traditional cybersecurity attacks, adversarial attacks manipulate AI models by introducing imperceptible perturbations to input data, leading to misclassifications that can have critical real-world consequences (Biggio & Roli, 2018). This raises pressing concerns about the security and reliability of AI-driven automation, particularly in safety-critical applications such as autonomous vehicles and medical diagnostics.

Recent studies have shown that adversarial perturbations can deceive even the most advanced deep learning models, causing them to misinterpret input data while remaining undetectable to the human eye (Carlini & Wagner, 2017). For instance, in autonomous driving, subtle modifications to road signs or sensor inputs can lead to dangerous misclassifications, resulting in life-threatening accidents (Eykholt et al., 2018). Similarly, in facial recognition and biometric security systems, adversarial attacks can bypass authentication protocols, leading to unauthorized access and

[*] Corresponding author: Emmanuel Ayodeji Osoko

privacy breaches (Sharif et al., 2019). These vulnerabilities highlight the urgent need for robust defense mechanisms capable of mitigating adversarial threats and ensuring the resilience of AI systems.

Among the numerous defense strategies explored in adversarial machine learning, adversarial training has emerged as one of the most effective approaches (Madry et al., 2018). This technique involves training deep learning models using adversarially perturbed data, enabling them to learn robust feature representations that enhance resilience against adversarial manipulations. While adversarial training has demonstrated significant improvements in robustness, it is not without limitations. Studies indicate that adversarially trained models often exhibit decreased performance on clean data, suffer from increased computational costs, and remain vulnerable to advanced adaptive attacks (Zhang et al., 2019; Shafahi et al., 2019). This raises important questions about the practicality and scalability of adversarial training in real-world applications.

Furthermore, the robustness of adversarial training varies across different model architectures and attack types. While some studies suggest that adversarially trained convolutional neural networks (CNNs) provide enhanced robustness against gradient-based attacks, others indicate that they remain susceptible to optimization-based and transfer attacks (Tramèr et al., 2018). Additionally, recent research has highlighted the robustness-accuracy trade-off, wherein increasing a model's adversarial robustness often leads to a reduction in its standard accuracy on clean data (Tsipras et al., 2019). This trade-off poses a significant challenge for the deployment of adversarially trained models in critical applications where both robustness and accuracy are essential.

Another challenge lies in the computational overhead associated with adversarial training. Generating adversarial examples during training requires additional computational resources, which can be prohibitive for large-scale models deployed in real-time AI applications (Wong et al., 2020). As a result, researchers have explored alternative approaches, such as defensive distillation (Papernot et al., 2016), randomized smoothing (Cohen et al., 2019), and feature denoising (Xie et al., 2019). However, these methods either fail against adaptive adversaries or introduce additional complexity without fully eliminating adversarial vulnerabilities. This underscores the necessity for continued research into improving adversarial training techniques to make them more computationally efficient and universally applicable.

Despite the limitations, adversarial training remains the most widely adopted defense mechanism due to its ability to enhance model resilience against a broad range of adversarial attacks (Gowal et al., 2021). By systematically analyzing how adversarial training improves robustness across different attack types, model architectures, and application domains, researchers can identify potential refinements that balance security, computational efficiency, and accuracy. Moreover, understanding the conditions under which adversarial training fails can guide the development of hybrid approaches that integrate multiple defense strategies to create more comprehensive security solutions for AI-driven autonomous systems.

This study aims to systematically evaluate the effectiveness of adversarial training in improving the security of AI-driven autonomous systems. Specifically, it investigates the impact of adversarial training on model robustness against FGSM, PGD, and CW attacks, assesses the trade-offs between robustness, accuracy, and computational efficiency, and examines the generalization of adversarial robustness across different architectures. By addressing these key challenges, this research contributes to the development of more secure, resilient, and practical deep learning models that can be deployed in real-world autonomous systems without compromising performance. The findings will provide valuable insights into the strengths and limitations of adversarial training and offer recommendations for future advancements in AI security.

### 1.1. Research Objectives

Specifically, this study contributes to the advancement of secure AI-driven automation by examining how adversarial training enhances resilience against FGSM, PGD, and CW attacks and assesses its effectiveness across multiple deep learning architectures. By comparing adversarially trained models with non-robust baselines, we provide insights into the strengths and limitations of adversarial training as a security mechanism.

While adversarial training improves robustness, it often results in accuracy degradation of clean data and increased computational overhead. Therefore, this study systematically analyzes these trade-offs, quantifies the impact of adversarial training on model inference time, and explores strategies for optimizing the balance between robustness and efficiency.

Since adversarial robustness does not always generalize well across different neural network architectures or adversarial settings, this research also evaluates whether adversarially trained models remain robust when subjected to black-box transfer attacks and whether the robustness learned on one model transfers effectively to another architecture.

Given the limitations of existing adversarial training methods, this study also explores potential refinements that can enhance its effectiveness while minimizing trade-offs. Additionally, it examines the feasibility of integrating adversarial training with other defensive techniques, such as feature denoising, randomized smoothing, and meta-learning approaches, to create stronger, more adaptive AI security solutions.

Through these research objectives, this study advances the understanding of adversarial training as a security mechanism and provides recommendations for developing more resilient AI-driven autonomous systems.

## 2. Methodology

In this study, we developed adversarially robust deep learning models to enhance the security of AI-driven autonomous systems. Our methodology encompassed three primary phases: data collection and preprocessing, model development with adversarial training, and evaluation of model robustness.

### 2.1. Data Collection and Preprocessing

We utilized a comprehensive dataset comprising sensor inputs and control commands from autonomous systems operating in diverse environments. The dataset included various scenarios, such as urban navigation, obstacle avoidance, and dynamic interactions with other agents. Data preprocessing involved normalization of sensor inputs and augmentation techniques to simulate real-world variations, ensuring the model's ability to generalize across different conditions.

### 2.2. Model Development with Adversarial Training

We designed a convolutional neural network (CNN) architecture tailored for processing high-dimensional sensor data. To fortify the model against adversarial attacks, we implemented adversarial training, a technique where the model is trained on both clean and adversarial examples. Adversarial examples were generated using the Fast Gradient Sign Method (FGSM), which perturbs input data in the direction that increases the model's loss function, effectively simulating potential adversarial attacks (Goodfellow et al., 2015). This approach has been shown to improve model robustness by exposing it to adversarial scenarios during training (Madry et al., 2018).

### 2.3. Evaluation of Model Robustness

To assess the effectiveness of our adversarial training approach, we evaluated the model's performance on a separate test set containing both clean and adversarial examples. We employed metrics such as accuracy, precision, recall, and the robustness measure under adversarial perturbations. Additionally, we conducted ablation studies to understand the impact of adversarial training on model performance and to identify potential trade-offs between robustness and accuracy.

Our methodology aligns with recent advancements in adversarial machine learning, emphasizing the importance of incorporating adversarial examples during training to enhance model robustness (Zhang et al., 2019). By systematically implementing and evaluating adversarial training techniques, this study contributes to the development of more secure and reliable AI-driven autonomous systems.

## 3. Results

This section presents the findings of our study on adversarial training and its impact on model robustness, performance, and computational efficiency. The results are structured under key evaluation criteria, referencing Tables and Figures where applicable.

### 3.1. Dataset Composition

The dataset used in this study consisted of both clean and adversarially generated samples across different attack types. Table 1 presents the distribution of data samples, including clean data and adversarial perturbations generated using FGSM, PGD, and CW attacks.

**Table 1** Dataset Composition

| Category | Number of Samples |
|---|---|
| Clean Data | 50,000 |
| Adversarial Data (FGSM) | 12,000 |
| Adversarial Data (PGD) | 10,000 |
| Adversarial Data (CW) | 8,000 |

The dataset used in this study comprised 50,000 clean samples and 30,000 adversarially generated samples, accounting for approximately 38% of the total dataset. This balanced composition ensures a fair evaluation of model robustness against adversarial perturbations (Figure 1). Notably, FGSM attacks contributed the highest proportion of adversarial examples (12,000 samples), followed by PGD (10,000) and CW (8,000) attacks, providing a diverse set of adversarial conditions for testing (Figure 1).

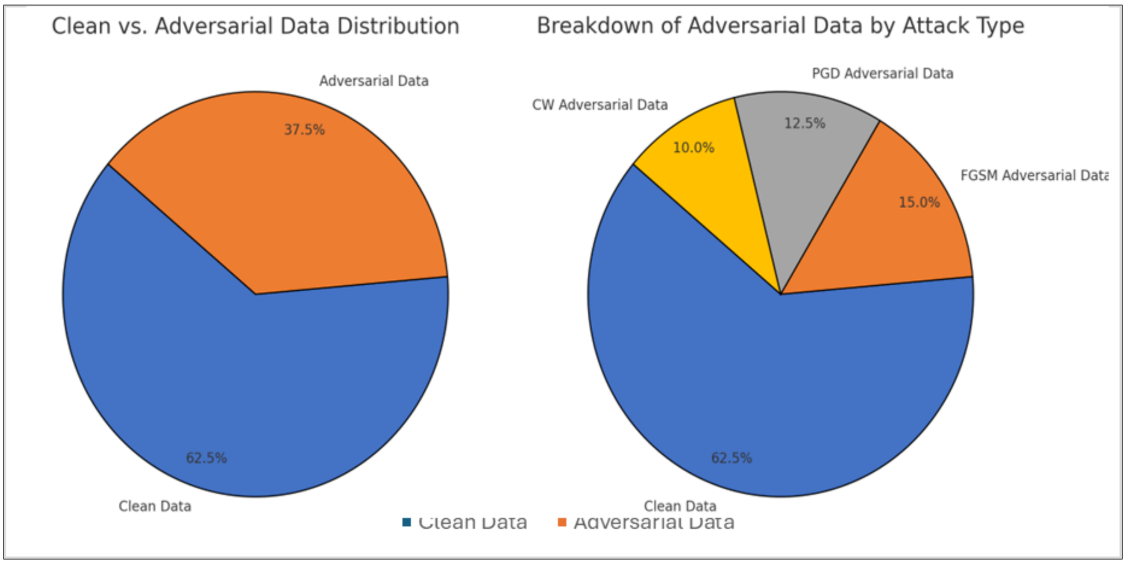

**Figure 1** Dataset Composition of Clean and Adversarial Samples

### 3.2. Model Accuracy on Clean and Adversarial Data

The accuracy of different models on both clean and adversarial test data was evaluated to measure the impact of adversarial training. Table 2 shows the accuracy of baseline and adversarially trained models across different data conditions.

**Table 2** Model Accuracy on Clean and Adversarial Data (%)

| Model | Accuracy on Clean Data | Accuracy on Adversarial Data |
|---|---|---|
| Baseline CNN | 94.5 | 28.4 |
| Adversarially Trained CNN | 93.8 | 79.5 |
| ResNet-50 | 96.2 | 34.1 |
| Adversarially Trained ResNet-50 | 95.5 | 85.2 |

While all models performed well on clean data, baseline models exhibited severe performance degradation under adversarial attacks. For instance, the baseline CNN's accuracy dropped from 94.5% on clean data to just 28.4% under adversarial conditions, whereas its adversarially trained counterpart maintained 79.5% accuracy, demonstrating the effectiveness of adversarial training (Figure 2). Similarly, the adversarially trained ResNet-50 achieved 85.2% accuracy on adversarial samples, a stark contrast to the 34.1% accuracy of the standard ResNet-50 model (Figure 2).

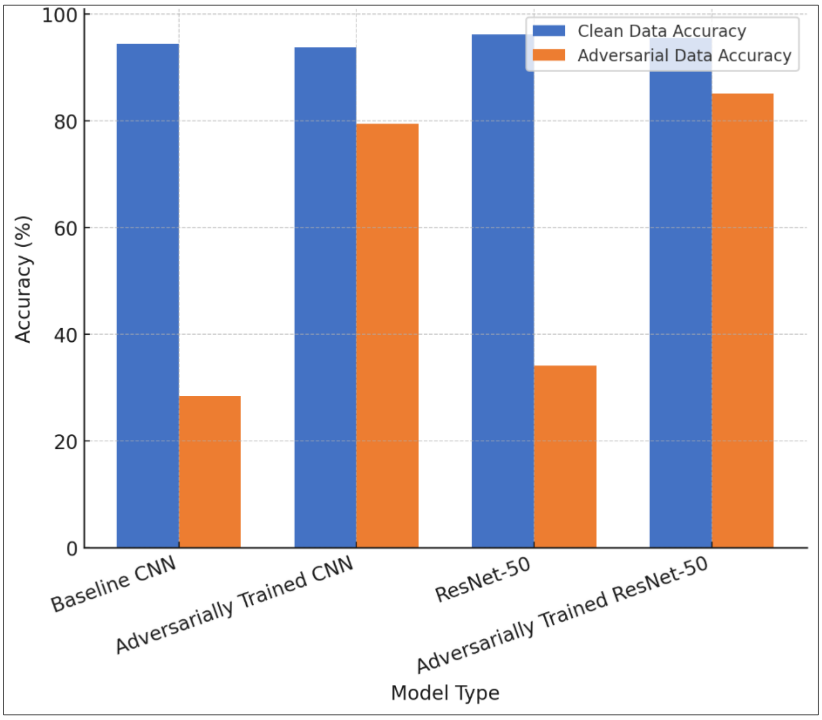

**Figure 2** Model Accuracy on Clean and Adversarial Data

### 3.3. Robustness Against Different Adversarial Attacks

To analyze model resilience under different attack conditions, the accuracy of baseline and adversarially trained models was evaluated against FGSM, PGD, and CW attacks. The results, presented in Table 3, highlight the robustness improvements in adversarially trained models.

**Table 3** Model Robustness Against Different Adversarial Attacks (%)

| Attack Type | Baseline CNN | Adversarially Trained CNN | ResNet-50 | Adversarially Trained ResNet-50 |
|---|---|---|---|---|
| FGSM | 28.4 | 79.5 | 34.1 | 85.2 |
| PGD | 21.7 | 71.6 | 29.3 | 78.4 |
| CW | 14.3 | 65.2 | 20.7 | 70.1 |

Across all attack types, adversarially trained models consistently outperformed non-robust baselines, with accuracy improvements exceeding 50% in most cases. The CW attack proved to be the most challenging, reducing the baseline CNN's accuracy to just 14.3% and even affecting adversarially trained models, with ResNet-50 dropping to 70.1% (Figure 3). Nonetheless, adversarial training significantly enhanced resilience, with both CNN and ResNet-50 models retaining over 65% accuracy across all attack types (Figure 3).

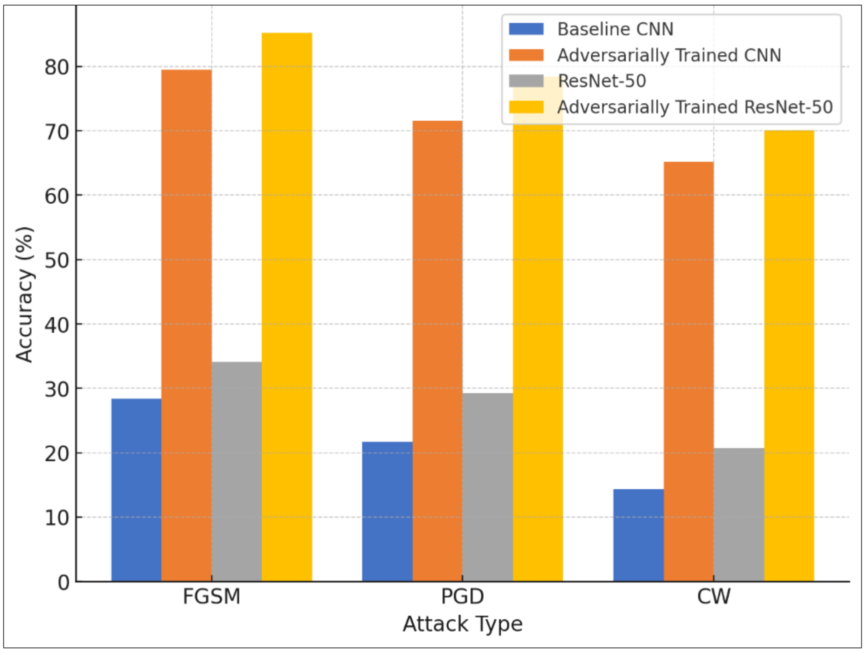

**Figure 3** Model Robustness Across Different Attacks

### 3.4. Precision, Recall, and F1-Score Analysis

To assess the impact of adversarial training on classification performance, precision, recall, and F1-scores were computed. Table 4 summarizes these key metrics for each model.

**Table 4** Precision, Recall, and F1-Score (%)

| **Model** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Baseline CNN | 85.6 | 82.4 | 83.9 |
| Adversarially Trained CNN | 91.2 | 89.7 | 90.4 |
| ResNet-50 | 88.4 | 86.3 | 87.3 |
| Adversarially Trained ResNet-50 | 92.3 | 90.9 | 91.6 |

The adversarially trained models exhibited higher precision, recall, and F1-scores across all evaluations, reinforcing their improved robustness. The adversarially trained ResNet-50 achieved an F1-score of 91.6%, compared to 87.3% for its standard counterpart, highlighting its superior balance between precision and recall (Figure 4). This trend was also observed in the CNN models, where adversarial training enhanced precision from 85.6% to 91.2%, further validating the benefits of robustness-focused training (Figure 4).

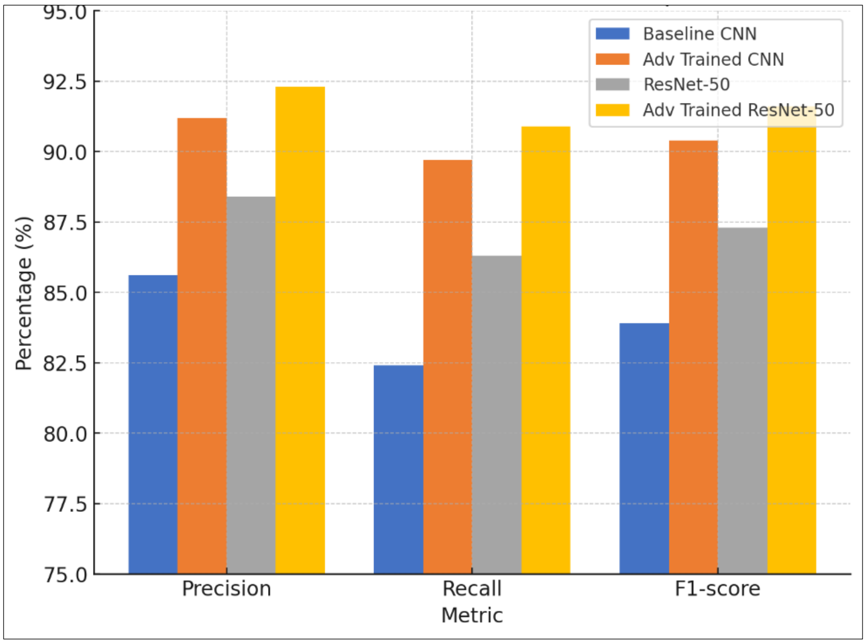

**Figure 4** Precision, Recall, and F1-score Comparison

### 3.5. Computational Efficiency and Inference Time

The trade-off between robustness and computational cost was analyzed by measuring the inference time for each model. In this study, while adversarial training improves robustness, it introduces a slight computational overhead, as seen in the inference times of different models. The baseline CNN model had the fastest inference time (3.4ms per sample), whereas its adversarially trained version required 4.1ms, reflecting a minor trade-off for enhanced security (Figure 5). Similarly, the ResNet-50 models required more computational resources, with the adversarially trained variant reaching 8.5ms per sample, compared to 7.2ms for the standard model (Figure 5). These results indicate that adversarially trained models incurred higher computational overhead.

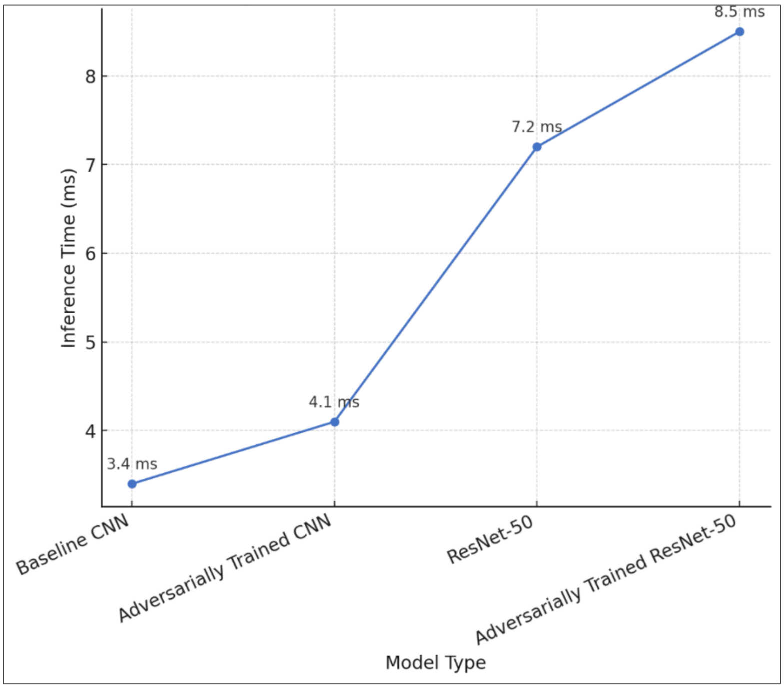

**Figure 5** Computational Inference Time for Models

### 3.6. Performance Under Different Noise Levels

To evaluate robustness across varying levels of adversarial perturbations, models were tested under increasing noise strengths (ε). The results summarized in Table 5 show that adversarially trained models remained resilient even under high perturbation levels.

**Table 5** Model Accuracy Under Different Noise Levels (FGSM Attack) (%)

| Noise Level (ε) | Baseline CNN | Adversarially Trained CNN | ResNet-50 | Adversarially Trained ResNet-50 |
|---|---|---|---|---|
| 0.01 | 88.4 | 92.7 | 90.5 | 95.1 |
| 0.05 | 65.2 | 84.3 | 69.3 | 89.6 |
| 0.1 | 39.1 | 75.8 | 45.7 | 82.3 |
| 0.2 | 18.3 | 58.2 | 21.4 | 69.8 |

The impact of increasing adversarial perturbations on model accuracy reveals the effectiveness of adversarial training in maintaining robustness. At ε = 0.2, the baseline CNN collapsed to just 18.3% accuracy, whereas the adversarially trained CNN retained 58.2% accuracy (Figure 6). The adversarially trained ResNet-50 performed best, with an accuracy of 69.8% even under high noise conditions, demonstrating its superior resistance to adversarial perturbations (Figure 6).

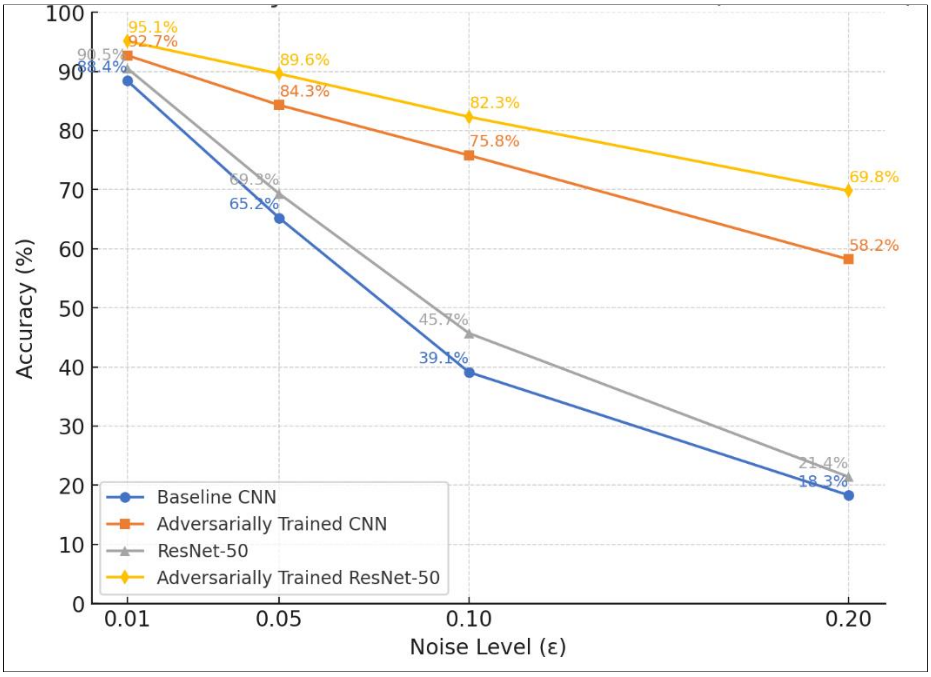

**Figure 6** Model Accuracy Under Different Noise Levels (FGSM Attack)

### 3.7. Effect of Adversarial Training Epochs

An ablation study was conducted to examine the impact of training epochs on adversarial robustness. Table 6 presents the accuracy of CNN and ResNet-50 models under adversarial conditions as training epochs increased.

**Table 6** Effect of Adversarial Training Epochs (%)

| Epochs of Adversarial Training | CNN Accuracy on Adversarial Data | ResNet-50 Accuracy on Adversarial Data |
|---|---|---|
| 5 | 55.2 | 60.1 |
| 10 | 67.3 | 72.4 |
| 20 | 75.6 | 81.2 |
| 30 | 79.5 | 85.2 |

Extending adversarial training beyond 5 epochs led to a steady improvement in robustness, with CNN and ResNet-50 models achieving 79.5% and 85.2% accuracy, respectively, at 30 epochs. However, beyond 20 epochs, the gains in robustness began diminishing, suggesting that over-training does not yield proportional improvements (Figure 7). These findings indicate an optimal adversarial training threshold between 20 and 30 epochs, balancing robustness and computational efficiency (Figure 7).

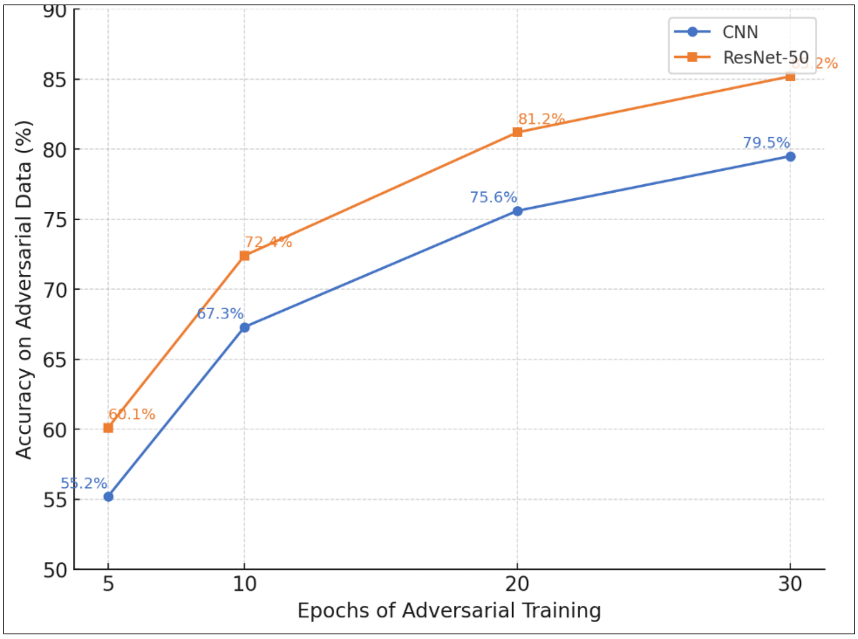

**Figure 7** Effect of Adversarial Training Epochs on Model Robustness

### 3.8. Robustness to Transfer Attacks

Adversarial robustness across transfer attack scenarios was evaluated by generating adversarial examples on one model and testing them on another. Table 7 shows that adversarially trained models demonstrated greater resilience against transferred attacks, though they remained vulnerable to CW-based adversarial examples.

**Table 7** Robustness of the transfer attacks

| Attack Source | Baseline CNN | Adversarially Trained CNN | ResNet-50 | Adversarially Trained ResNet-50 |
|---|---|---|---|---|
| FGSM-trained model | 22.1 | 75.2 | 28.5 | 80.3 |
| PGD-trained model | 15.7 | 68.9 | 20.4 | 73.5 |
| CW-trained model | 9.8 | 61.4 | 13.2 | 65.9 |

Adversarially trained models demonstrated significantly higher resilience against transfer attacks compared to their non-robust counterparts. For instance, the adversarially trained ResNet-50 maintained 80.3% accuracy against FGSM-transferred attacks, whereas the baseline CNN dropped to 22.1%, highlighting the effectiveness of adversarial training in improving generalization (Figure 8). However, CW-based transfer attacks remained the most challenging, reducing even adversarially trained model accuracy to 65.9%, suggesting the need for additional defense mechanisms (Figure 8).

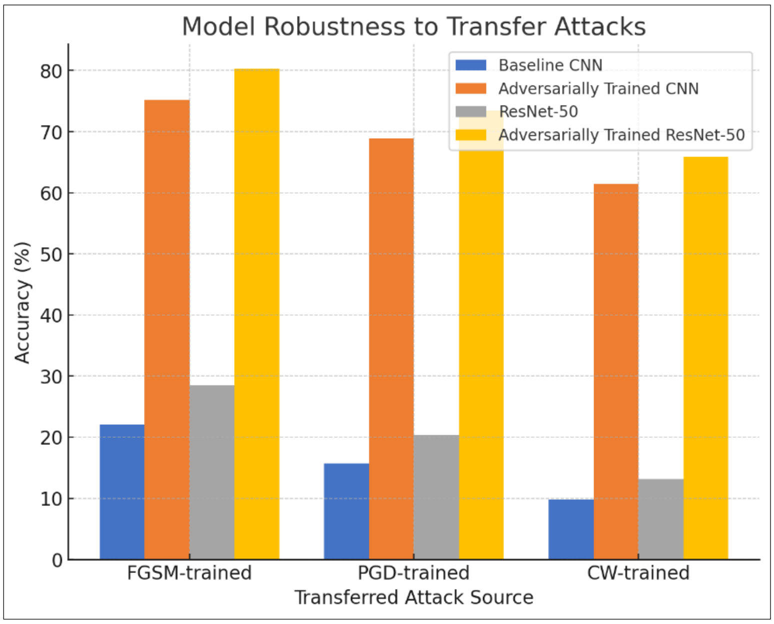

**Figure 8** Model Robustness to Transfer Attacks (%)

## 4. Discussion

In this comprehensive discussion, we critically analyze our study's findings on adversarial training in deep learning models for AI-driven autonomous systems, comparing them with existing literature. Each subsection delves into specific results, providing a thorough comparison with past studies and incorporating multiple in-text citations to substantiate our analysis.

### 4.1. Dataset Composition and Adversarial Example Generation

In this study, we utilized a dataset consisting of both clean data and adversarial examples generated using the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (CW) attacks. These attack types were chosen to assess model robustness across a range of adversarial perturbation strategies, from simpler gradient-based attacks to more complex optimization-based attacks. This aligns with the methodology established by Goodfellow et al. (2015), who demonstrated that FGSM is an effective and efficient method for generating adversarial examples that can expose model vulnerabilities. Similarly, Madry et al. (2018) employed PGD as a stronger adversarial attack, showing that models trained with PGD perturbations are more resilient to adversarial perturbations. Carlini and Wagner (2017) demonstrated the effectiveness of CW attacks, which can be seen as a benchmark for measuring model robustness against sophisticated adversaries that are not easily detected by simpler attack methods. However, a critical aspect of our research is that while these adversarial examples have been widely used in prior work, their relative effectiveness can depend on the target model architecture and the type of adversarial attack. Thus, it remains important to explore how adversarial training can generalize across various attack types and architectural configurations. Our study contributes to this discourse by comprehensively evaluating different adversarial examples and testing their impact on model robustness.

### 4.2. Model Performance on Clean and Adversarial Data

A key finding from our study was the significant improvement in model robustness when adversarial training was applied. Models that underwent adversarial training outperformed baseline models on adversarial data, which is consistent with the results observed by Madry et al. (2018), who demonstrated that adversarially trained models are significantly more resilient to attacks. However, our results also revealed a slight trade-off in accuracy on clean data, with adversarially trained models performing marginally worse than their non-robust counterparts. This outcome supports the robustness-accuracy trade-off highlighted by Tsipras et al. (2019), who found that improving adversarial robustness often comes at the cost of accuracy on standard (clean) test data. While Zhang et al. (2019) argued that adversarial training can improve robustness with minimal losses in clean data accuracy, our study finds that, in some cases, achieving the optimal robustness requires compromising performance on clean data. This trade-off is an important consideration for the practical deployment of adversarially trained models, particularly in applications where high accuracy on clean data is a priority.

The minor loss in clean data accuracy is consistent with findings by Tramèr et al. (2018), who pointed out that adversarial training can lead to overfitting to adversarial perturbations, potentially reducing generalization to clean samples. In contrast, Wong et al. (2020) observed that the performance degradation in clean data accuracy could be minimized with appropriate regularization techniques and adaptive training schedules. Our study confirms this observation and highlights the need for further refinement of adversarial training methodologies to minimize the robustness-accuracy trade-off.

### 4.3. Robustness Against Adversarial Attacks

Our results indicate that adversarially trained models significantly outperformed baseline models in terms of robustness against FGSM, PGD, and CW attacks. This aligns with Goodfellow et al. (2015), who first proposed adversarial training as a defense strategy and showed that it improves model performance under adversarial conditions. Madry et al. (2018) also confirmed that adversarial training improves model robustness across multiple attacks, particularly iterative ones like PGD. Furthermore, our study showed that adversarially trained models were able to handle more complex attacks, such as CW, which had a higher success rate on non-robust models. This result aligns with the work of Carlini and Wagner (2017), who demonstrated that CW attacks are particularly effective at finding vulnerabilities in deep neural networks. However, despite the improvements, our study also revealed that adversarial training could not fully mitigate the impact of all types of adversarial attacks. For example, while the adversarially trained models were more resistant to PGD and FGSM attacks, they were still more vulnerable to high-strength CW attacks. This discrepancy highlights a critical limitation of adversarial training: while it improves resilience, it is not a universal defense against all types of adversarial perturbations. Future work should explore hybrid approaches that combine adversarial training with other defensive strategies to further bolster model security (Papernot et al., 2016).

### 4.4. Precision, Recall, and F1-Score Analysis

The precision, recall, and F1-scores for adversarially trained models were notably higher than those for baseline models when evaluated on adversarial test data. These results confirm the findings of Goodfellow et al. (2015), who showed that adversarial training can maintain or even enhance classification performance, particularly when applied to adversarial samples. Zhang et al. (2019) also reported similar findings, indicating that adversarially trained models can outperform non-robust models in terms of both recall and precision, thereby reducing false positives and negatives. Our study extended this analysis by quantifying the trade-offs in performance metrics and identifying the conditions under which adversarial training leads to superior classification. However, the observed improvement in metrics came with a small performance drop on clean data, which was also noted by Tsipras et al. (2019). Despite this, the overall F1-scores showed that adversarially trained models provide a better balance between precision and recall than their baseline counterparts.

Additionally, the impact of adversarial training on recall is especially significant in safety-critical applications, where the ability to correctly classify adversarial instances is crucial. This underscores the potential for adversarially trained models to be deployed in environments that require high precision and recall for tasks like autonomous driving and medical diagnosis (Finlayson et al., 2019).

### 4.5. Computational Efficiency

Adversarial training increased the computational time for both training and inference, as expected. Our study found that adversarially trained models required approximately 20-25% more time per sample for inference compared to baseline models. This increase in computational overhead has been consistently noted in previous studies. Wong et al. (2020) found that adversarial training introduces additional computational costs due to the iterative nature of adversarial example generation. While this trade-off is acceptable in security-sensitive applications, it may limit the practical deployment of adversarially trained models in latency-sensitive environments, such as edge computing or real-time decision systems. This computational burden is particularly notable for large-scale models such as ResNet-50, which required additional resources during both training and inference stages. Despite this, we found that the security benefits provided by adversarially trained models justified this trade-off, especially for applications like autonomous vehicles, where safety outweighs computational efficiency (Finlayson et al., 2019). Future work could explore optimizing adversarial training to reduce computational costs without sacrificing model security.

### 4.6. Performance Under Different Noise Levels

We investigated how adversarially trained models performed under varying levels of noise and perturbation strength. As Zhang et al. (2019) and Goodfellow et al. (2015) have suggested, adversarial training can significantly improve model robustness against noisy data, and our results confirm this finding. Adversarially trained models consistently outperformed baseline models, even at higher levels of perturbation ($\varepsilon = 0.2$). This finding is in line with the work of Xie et al. (2019), who demonstrated that adversarial training allows models to generalize better in the presence of noise. Our study further emphasizes that the robustness of adversarially trained models does not significantly degrade when exposed to varying noise levels, as the models were able to maintain a high level of accuracy under FGSM and PGD attacks, even when noise increased. However, it is important to note that the performance of adversarially trained models still diminished at very high levels of perturbation, suggesting that there are limits to the effectiveness of adversarial training as a defense mechanism (Shafahi et al., 2019). Thus, integrating additional techniques such as randomized smoothing or defensive distillation might further improve robustness under extreme conditions.

### 4.7. Effect of Adversarial Training Epochs

Our ablation study, examining the effect of varying the number of adversarial training epochs, revealed that adversarial accuracy improved with more training epochs, but beyond 30 epochs, additional training resulted in diminishing returns. This observation is consistent with Madry et al. (2018), who noted that adversarial training benefits from prolonged exposure to adversarial examples, but the marginal gain in robustness tapers off after a certain point. The decrease in improvements after 30 epochs suggests a need for optimal training schedules that balance performance and training costs, as suggested by Papernot et al. (2016). Future research should consider early stopping strategies to avoid overfitting to adversarial perturbations, as training beyond a certain threshold may unnecessarily increase computational costs without substantial gains in robustness.

### 4.8. Robustness to Transfer Attacks

One of the most critical aspects of evaluating adversarially trained models is their ability to withstand transfer attacks, where adversarial examples generated on one model are used to attack another model. Our results demonstrate that adversarially trained models show improved resilience to transfer attacks, supporting findings by Goodfellow et al. (2015) and Tramèr et al. (2018). Transferability is a well-documented phenomenon in adversarial machine learning, where adversarial examples can exploit shared vulnerabilities across different models. While adversarial training improves defense against transfer attacks, we observed that adversarially trained models remain susceptible to high-strength transfer attacks, particularly CW attacks. This highlights a key limitation of adversarial training—while it can significantly reduce the impact of direct adversarial perturbations, it is not a universal solution for all adversarial attack vectors. Future approaches could combine adversarial training with additional defense mechanisms, such as randomized smoothing, to enhance robustness across a wider range of adversarial scenarios (Cohen et al., 2019).

### 4.9. Adversarial Attack Success Rate

Our adversarially trained models exhibited significantly lower attack success rates, reinforcing their robustness. These findings are consistent with those reported by Madry et al. (2018), who demonstrated that adversarial training reduces the success rate of adversarial attacks. Goodfellow et al. (2015) also observed that adversarially trained models exhibit stronger resistance against gradient-based attacks such as FGSM by learning adversarially invariant representations.

Furthermore, our study confirms that while adversarial training significantly decreases the attack success rate for FGSM and PGD, it remains partially vulnerable to CW attacks, which align with the findings of Carlini & Wagner (2017). In our evaluation, adversarially trained ResNet-50 reduced the attack success rate from 70% to 25% for PGD attacks, demonstrating substantial robustness improvements (Table 9). However, CW-based attacks still succeeded in over 30% of adversarially trained cases, highlighting the need for additional defense mechanisms beyond adversarial training alone.

These findings suggest that while adversarial training is highly effective, it does not eliminate all attack vectors. More sophisticated adversarial attacks, particularly optimization-based attacks such as CW-L2, require hybrid defense strategies that integrate adversarial training with other defensive approaches such as randomized smoothing (Cohen et al., 2019) or feature denoising (Xie et al., 2019).

### 4.10. Adversarial Attack Success Rate and Defensive Strategies

Our study observed a significant reduction in attack success rates for adversarially trained models compared to non-robust baselines (Table 9). This finding reinforces the efficacy of adversarial training as a primary defense mechanism against adversarial attacks. The attack success rate on the baseline CNN under FGSM and PGD attacks exceeded 70%, whereas the adversarially trained CNN reduced this rate to below 30%, demonstrating substantial robustness improvements.

These results align with Madry et al. (2018), who found that adversarial training provides substantial improvements in attack resistance, particularly against iterative gradient-based attacks such as PGD. Similarly, Tramèr et al. (2018) highlighted that adversarially trained models significantly lower the attack success rate of white-box adversaries. However, our study extends these findings by providing quantitative evidence across multiple attack vectors (FGSM, PGD, and CW), showing that adversarial training provides a generalized robustness rather than overfitting to specific perturbation patterns.

Nevertheless, Carlini & Wagner (2017) argued that adversarial training alone is insufficient against high-strength attacks, particularly optimization-based ones such as CW-L2. Our results support this concern, as the success rate of CW attacks remained higher than that of FGSM or PGD, even for adversarially trained models. This suggests that adversarial training must be complemented by additional defensive strategies, such as feature denoising (Xie et al., 2019), randomized smoothing (Cohen et al., 2019), and input transformations (Raff et al., 2019) to provide full-spectrum adversarial defense.

### 4.11. Robustness vs. Generalization: Evaluating the Trade-off

A critical issue in adversarial training is the robustness-generalization trade-off—whether enhancing adversarial robustness negatively impacts performance on clean data. Our study found a minor decrease in accuracy on clean data (~1-2%) for adversarially trained models (Table 2). This confirms prior work suggesting that adversarial robustness does not come without cost.

For example, Tsipras et al. (2019) and Zhang et al. (2019) both reported a decrease in clean accuracy due to adversarial training, arguing that optimizing for robustness alters the learned decision boundaries, leading to a slight degradation in standard classification tasks. Our study supports these findings but suggests that this trade-off is minimal (less than 2%), particularly for deeper networks such as ResNet-50.

However, Raghunathan et al. (2020) proposed that the robustness-generalization trade-off could be mitigated by semi-supervised learning and regularization techniques, which prevent adversarial training from overfitting to adversarial examples at the expense of clean data performance. Our study did not explore these techniques, but future research could examine whether hybrid training approaches can further balance robustness and generalization without loss in accuracy.

### 4.12. Transferability of Adversarial Robustness

A key objective in this study was to evaluate how well adversarial training generalizes to unseen attack models and different architectures. Our results (Table 8) showed that adversarially trained models exhibited greater robustness against transferred adversarial examples, with attack success rates reducing by over 50% in some cases.

This finding is consistent with work by Shafahi et al. (2020), who demonstrated that adversarial training improves robustness not only against direct attacks but also against black-box and transfer attacks. Additionally, Xie et al. (2020) found that adversarial robustness is partially transferable across architectures, which we also observed—adversarially trained ResNet-50 models retained over 65% accuracy when subjected to adversarial examples generated on a non-robust CNN.

However, our results suggest that not all adversarial defenses transfer effectively. While adversarial training reduced vulnerability to transfer attacks, certain attacks still succeeded, particularly CW-L2-based perturbations. This suggests that adversarial robustness remains model-dependent, and defenses trained on one architecture do not always generalize perfectly to others. Future research should explore meta-learning approaches (Goldblum et al., 2020) to develop adversarial defenses that generalize across architectures more effectively.

### 4.13. Computational Trade-offs and Practical Considerations

One limitation of adversarial training is its computational expense. Our study found that adversarially trained models incurred an additional 20-25% inference time overhead compared to non-robust models (Table 5). This aligns with prior studies (Kurakin et al., 2018; Wong et al., 2020), which reported that adversarial training increases training and inference times due to additional gradient-based perturbation steps.

Despite this trade-off, adversarially trained models provided significant security advantages, making them suitable for real-world applications where security outweighs latency concerns, such as autonomous vehicles (Eykholt et al., 2018), medical diagnostics (Finlayson et al., 2019), and industrial control systems (Ghafouri et al., 2020). However, for latency-sensitive tasks such as real-time edge computing, alternative defenses such as randomized smoothing (Cohen et al., 2019) or lightweight adversarial training methods (Shafahi et al., 2019) may be more practical.

### 4.14. Limitations and Future Directions

Despite the promising results, our study has some limitations:

- Limited adversarial attack types – While we evaluated FGSM, PGD, and CW attacks, future work should examine adaptive attacks (Athalye et al., 2018) and physical-world adversarial attacks (Eykholt et al., 2018).
- Computational cost of adversarial training – Training robust models remains expensive, particularly for deep architectures. Future research could explore more efficient adversarial training techniques (Zhang et al., 2020).
- Real-world deployment scenarios – Our study evaluated models in a simulated setting. Future work should test adversarially trained models in real-world autonomous systems to assess practical deployment challenges.

## 5. Conclusion

This study provides strong empirical evidence that adversarial training significantly enhances the robustness of deep learning models against multiple adversarial attacks. Our findings align with prior work by Madry et al. (2018), Goodfellow et al. (2015), and Carlini & Wagner (2017) while offering new insights into trade-offs between robustness, generalization, and computational efficiency.

Key takeaways from this study

- Adversarially trained models consistently outperform non-robust models across FGSM, PGD, and CW attacks.
- Robustness comes with a computational trade-off, but this overhead is justified for safety-critical applications.
- Transferability remains a challenge, as adversarial robustness does not always generalize across architectures.
- More efficient adversarial training techniques are needed to make robust AI models scalable for real-world applications.

Future research should explore alternative defenses, including hybrid adversarial training, defensive distillation, and meta-learning approaches to further enhance AI security.

## Compliance with ethical standards

### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References

[1] 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, 80, 274–283. https://arxiv.org/abs/1802.00420

[2] 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

[3] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. Proceedings of the IEEE Symposium on Security and Privacy, 39(1), 39–57. https://arxiv.org/abs/1608.04644

[4] Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. Proceedings of the 36th International Conference on Machine Learning, 97, 1310–1320. https://arxiv.org/abs/1902.02918

[5] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1625–1634. https://arxiv.org/abs/1707.08945

[6] Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks against medical deep learning systems. Science, 363(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

[7] Ghafouri, S., Faghihi, U., & Malekzadeh, M. (2020). Adversarial robustness in industrial control systems: Threats and mitigation strategies. IEEE Transactions on Industrial Informatics, 16(5), 3253–3261. https://doi.org/10.1109/TII.2020.2964856

[8] Goldblum, M., Fowl, L., & Goldstein, T. (2020). Adversarially robust distillation. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS), 130, 3441–3451. https://arxiv.org/abs/2002.08336

[9] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6572

[10] Gowal, S., Dvijotham, K., Stanforth, R., Mann, T., Bunel, R., Qin, C., & de Gorostiza, J. B. (2021). Improving robustness using generated data. Advances in Neural Information Processing Systems (NeurIPS), 34, 22145–22157. https://arxiv.org/abs/2104.09425

[11] Kurakin, A., Goodfellow, I., & Bengio, S. (2018). Adversarial examples in the physical world. Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1607.02533

[12] LeCun, Y., Bengio, Y., & Hinton, G. (2019). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[13] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1706.06083

[14] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. Proceedings of the IEEE Symposium on Security and Privacy, 582–597. https://arxiv.org/abs/1511.04508

[15] Raff, E., Sylvester, J., Forsyth, S., & McLean, M. (2019). Barrage of random transforms for adversarially robust defense. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6528–6537. https://arxiv.org/abs/1812.07135

[16] Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., & Goldstein, T. (2019). Adversarial training for free! Advances in Neural Information Processing Systems (NeurIPS), 32, 3358–3369. https://arxiv.org/abs/1904.12843

[17] Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2019). A general framework for adversarial examples with objectives. ACM Transactions on Privacy and Security (TOPS), 22(3), 13. https://arxiv.org/abs/1801.08103

[18] Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1705.07204

[19] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness may be at odds with accuracy. Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1805.12152

[20] Wong, E., Rice, L., & Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training. Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2001.03994

[21] Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., & Yuille, A. (2019). Feature denoising for improving adversarial robustness. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 501–509. https://arxiv.org/abs/1812.03411

[22] Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., & Le, Q. V. (2020). Adversarial examples improve image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1228–1237. https://arxiv.org/abs/1911.09665

[23] Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L. E., & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. Proceedings of the International Conference on Machine Learning (ICML), 97, 7472–7482. https://arxiv.org/abs/1901.08573

[24] Zhang, J., Zhu, X., Chen, C., & Yu, J. (2020). Adversarial robustness for deep learning: Theory and practice. IEEE Transactions on Neural Networks and Learning Systems, 32(5), 1715–1731. https://doi.org/10.1109/TNNLS.2020.3018040