E-ISSN: 1116 - 3240 P-ISSN: 3026-9334
MPRI-JAIMR | INTERNATIONAL JOURNAL OF AFRICAN INNOVATION & MULTIDSCIPLINARY RES.
VOL. 10 NO. 2 – OCTOBER, 2025

EVALUATING THE ROBUSTNESS OF CONVOLUTIONAL NEURAL NETWORKS AGAINST ADVERSARIAL IMAGE PERTURBATIONS IN REAL-WORLD SCENARIOS

ISHOLA A.K.; OLOFIN,B.B.; ISHOLA,P.E.; UMOREN,U.M.; & ADEOLA, P.
Department of Computing, Anchor University Lagos, Nigeria
Corresponding Author: aishola@pgs.aul.edu.ng
DOI: https://doi.org/10.70382/mejaimr.v10i2.077

Abstract

Convolutional Neural Networks (CNNs) have achieved remarkable success in computer vision tasks but remain highly vulnerable to adversarial image perturbations, especially in real-world environments. This study evaluates the robustness of leading CNN models against a range of adversarial attacks—both digital and physical—under realistic conditions such as lighting variations, occlusions, and sensor noise. Experimental results reveal that even state-of-the-art models experience significant performance degradation when exposed to combined adversarial and environmental challenges. The findings underscore the urgent need for more resilient training strategies and comprehensive evaluation frameworks that account for real-world complexities. By integrating adversarial training, defensive distillation, and input transformation techniques, this research contributes to the development of more reliable and secure CNN-based systems for practical applications.

Keywords:
Convolutional Neural Networks (CNNs), Adversarial Attacks, Image Perturbations, Model Robustness, Real-World Scenarios, Adversarial Training, Evaluation Metrics, Defense Mechanisms

Introduction

Convolutional Neural Networks (CNNs) have achieved huge success in the field of computer vision and image processing. However, their implementation in real-world scenarios has demonstrated significant weakness to adversarial image

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

perturbations, irreducible alterations to input images that can lead to noise, and misclassification.

Related works have highlighted the efficiency of adversarial training in enhancing the usefulness of CNNs against adversarial attacks. Eleftheriadis, et al, 2024 proposed a formidable adversarial training techniques that incorporates a representative ensemble of adversarial thereby improving the resilience of deep neural networks (DNNs) to various perturbation challenges and case studies. Furthermore, Smagulova, et al, 2024, evaluated the robustness and transferability of adversarial attacks on different image classification neural networks, demonstrating that convolution-based transformer models can facilitate superior features and performance against newly generated and transferable attacks compared to conventional CNN architectures.

The recent studies by Barr, 2025, introduced APE-GAN, an adversarial training technique that proved CNN resilience against advanced adversarial attacks while optimising performance for dynamic systems with limited computational resources and synergies. This methods pointed the essence of developing defence mechanisms that not only enhance robustness but also ensure that efficiency of CNNs is beneficial in industrial practices.

Despite the rigorous studies, difficulties remain in evaluating the robustness of CNNs against adversarial perturbations in dynamic and defiant real-world environments in Silva, et al, 2020. The conventional evaluation methods fail to account for the complexities introduced by environmental factors, residual noise, logical distortions, and physical-world transformations. There is a requirement for detailed evaluation frameworks that present real-world conditions to dutifully and significantly assess and enhance the features and viability of CNNs against adversarial attacks.

Related Works

The recent related studies have focused on evaluating the hugeness of CNNs against adversarial attacks. Eleftheriadis, et al, 2024 proposed a general adversarial training methodology that uses multiple attack strategies and distance metrics to assess robustness effectively.

Smagulova, et al, 2024 evaluated the robustness and transferability of adversarial attacks on different image processing neural networks and their findings indicate that convolution-based transformer models shows superior strength against newly generated and transferable attacks compared to conventional CNN architectures.

Wang, et al, 2025, introduced the Robust Feature Coverage Attack (RFCoA), a perturbation-based method to improve the robustness and transferability of physical-world adversarial examples. Their approach involves extracting robust features and

overlaying them onto non-perturbed images, enhancing the stealthiness and effectiveness of adversarial attacks in a controlled condition.

Hu, et al, 2024, proposed the Feature Pattern Consistency

Constraint (FPCC) method to improve adversarial tactics by aligning latent feature patterns with correct image classifications, their approach strengthens the model's tendencies to maintain correct predictions even in the presence of adversarial attacks. This method offers a clear direction for improving CNN resilience without depending on adversarial training.

The simulated attacks are often under controlled, digital conditions and then verify and validated the cases in narrow physical setups (fixed lighting, angles, backgrounds) of imaged based adversarial perturbation. Consequently, measured success rates often undermine real-world effectiveness or miss failure modes that appear under natural variability (lighting, motion blur, occlusion) and similarly environmental conditions and processes in Kurakin, et al, 2018.

Fawole, O., & Rawat, D. ,2025 shows that defenses are often validated against the attacks they were tuned or provision for, thereby producing a subtle illusion of robustness. When attackers adapt (different loss, expectation over transformation, or black-box methods), many defenses or stronghold fail. The evaluations techniques that do not include adaptive adversaries overstate protection.

The practicality and dimensionality of mounting real-world attacks such as printing high-resolution patches, physically placing adversarial object and initiating and maintaining stealth and the computational and proportionally latency cost of defenses on adversarial training and robust preprocessing are rarely quantified, limiting assessments of real operational feasibility in mission critical systems as shown in Eykholt, et al 2018.

In order to address the limitations observed in the related works, the following methodology is proposed.

Dataset Selection

The research work focus on the resilience of Convolutional Neural Networks (CNNs) against adversarial image perturbations in controlled scenerios. The benchmark datasets that closely demonstrated real-world conditions are used such as CIFAR-10 for evaluating image classification models and ImageNet dataset providing a comprehensive evaluation platform for CNNs and Medical Imaging Datasets to evaluate the performance of CNNs.

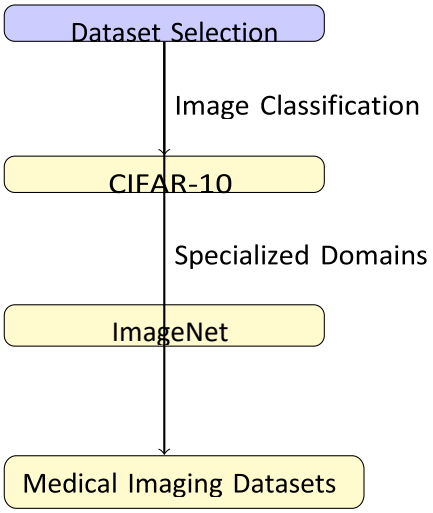
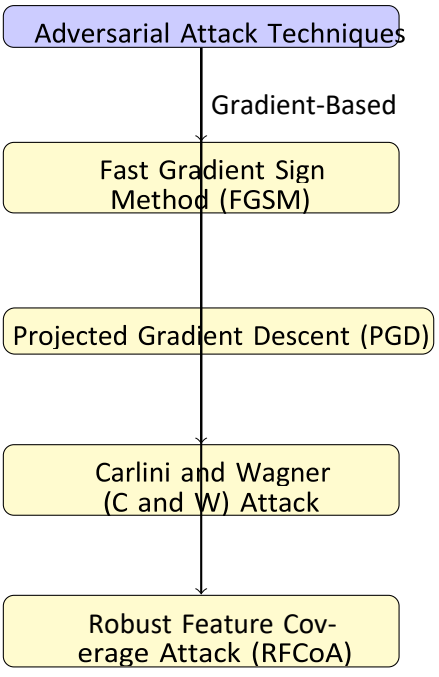


Figure 1. Dataset Selection Process for Evaluating CNN Robustness

Adversarial Attack Techniques

The research uses adversarial attack methods to perturb the input images and evaluate the CNNs' robustness, first,: Fast Gradient Sign Method (FGSM) is a white-box attack that perturbs the image in the direction of the gradient of the loss function in Lad, et al 2024; Second, Projected Gradient Descent (PGD) is an iterative method that applies small perturbations to the image, projecting them back onto the feasible set in Gupta, et al, 2018; Third, Carlini and Wagner (C and W) Attack which is a powerful attack that formulates the perturbation as an optimisation problem in Carlini, N., & Wagner, D. , 2017. Fourth, Robust Feature Coverage Attack (RFCoA) is a novel perturbation-based method that improves the transferability and suitability of physical-world adversarial examples by covering features of other classes onto the predictive features in clean images in Wang, et al, 2025.



Optimization-Based and Perturbations

Figure 2. Adversarial Attack Techniques Used to Evaluate CNN Robustness

Formulation Adversarial Attack Techniques.

Given a machine learning model $f : X \rightarrow Y$, where X represents the input space and Y represents the output space (e.g., class labels), an adversarial attack seeks to perturb the input $\mathbf{x} \in X$ such that the model's prediction is altered, typically leading to misclassification.

An adversarial attack can be formulated as:

$\mathbf{x}^* = \mathbf{x} + \delta(\mathbf{x})$ (1)

where:

- $\mathbf{x} \in X$ is the original input sample,
- $\mathbf{x}^* \in X$ is the adversarially perturbed input,

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

- $\delta(\mathbf{x})$ is the perturbation function applied to the input, which is typically designed to be small but enough to change the model's output.

The perturbation $\delta(\mathbf{x})$ is constrained based on the attack type, which may involve different norms (e.g., L_p norms) to limit the magnitude of the perturbation. For example:

$$\|\delta(\mathbf{x})\|_p \leq \epsilon \qquad (2)$$

where p is typically 2 (Euclidean norm) or infinity (maximum perturbation per feature), and ϵ is a small constant controlling the magnitude of the perturbation.
The objective of an adversarial attack is to maximize the perturbation's effectiveness, leading to a misclassification:

$$\hat{y} = f(\mathbf{x}^*) \neq f(\mathbf{x}) \qquad (3)$$

where $\hat{y}$ is the incorrect output, and $f(\mathbf{x})$ is the correct class label for the original input.

Evaluation Metrics

The Adversarial Robustness Evaluation Benchmark (AREB) which is a benchmark is employed to supports a holistic evaluation of adversarial defence methods, considering multiple attack strategies and distance metrics in Croce, F.,2020.

Formulation of Evaluation Metrics

Given a machine learning model $f : X \rightarrow Y$, where X represents the input space and Y represents the output space (e.g., class labels), the evaluation metrics are used to assess the performance of the model, especially its robustness in the presence of adversarial perturbations. These metrics help quantify how well the model is performing on both clean and adversarially perturbed inputs.
Let $\mathbf{x} \in X$ be an input sample and $\hat{y} = f(\mathbf{x})$ be the model's predicted label for $\mathbf{x}$. We define the following evaluation metrics:
Classification Accuracy The classification accuracy Acc is the proportion of correctly classified samples to the total number of samples:

$$Acc = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(f(\mathbf{x}_i) = y_i\right) \qquad (4)$$

where:

- N is the total number of samples,
- $\mathbf{1}(A)$ is the indicator function that returns 1 if the condition A is true and 0 otherwise,
- $\mathbf{x}_i$ is the i-th input sample, and y_i is its true label.

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

Attack Success Rate (ASR)

The attack success rate *ASR* measures how often adversarial attacks successfully mislead the model:

$$ASR = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}\left(f(\mathbf{x}_i^*) \neq y_i\right) \quad (5)$$

where:

- *M* is the number of adversarial samples, • $\mathbf{x}_i^*$ is the adversarial perturbation of $\mathbf{x}_i$,
- $f(\mathbf{x}^*_i)$ is the model's predicted output on the adversarial sample.

Confidence Score.

The confidence score *C* is a measure of the certainty of the model's prediction, typically computed as the maximum probability output by the softmax function:

$$C(\mathbf{x}) = \max(\sigma(f(\mathbf{x}))) \quad (6)$$

where σ represents the softmax function and max returns the highest probability value. A higher confidence score indicates higher certainty in the model's prediction.

Distance Metrics The L_p-distance metrics measure the magnitude of the perturbation applied to the input:

$$\|\delta(\mathbf{x})\|_p = \left(\sum_{i=1}^{d}|x_i - x_i^*|^p\right)^{1/p} \quad (7)$$

where:

- $\delta(\mathbf{x}) = \mathbf{x}^* - \mathbf{x}$ is the perturbation,
- $\mathbf{x}^*$ is the adversarial input,
- *p* is the norm (typically 2 for Euclidean or infinity for max norm), and *d* is the dimensionality of the input space.

Adversarial Robustness Evaluation Benchmark (AREB).

The Adversarial Robustness Evaluation Benchmark *AREB* is a formidable metric used to evaluate the model's resilience against a variety of adversarial attacks across multiple datasets and perturbations. It is computed as:

$$AREB = \frac{1}{T}\sum_{t=1}^{T}Acc(t) \quad (8)$$

where:

- *T* is the number of test attacks or scenarios,

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

- $Acc(t)$ is the accuracy of the model on the t-the attack.

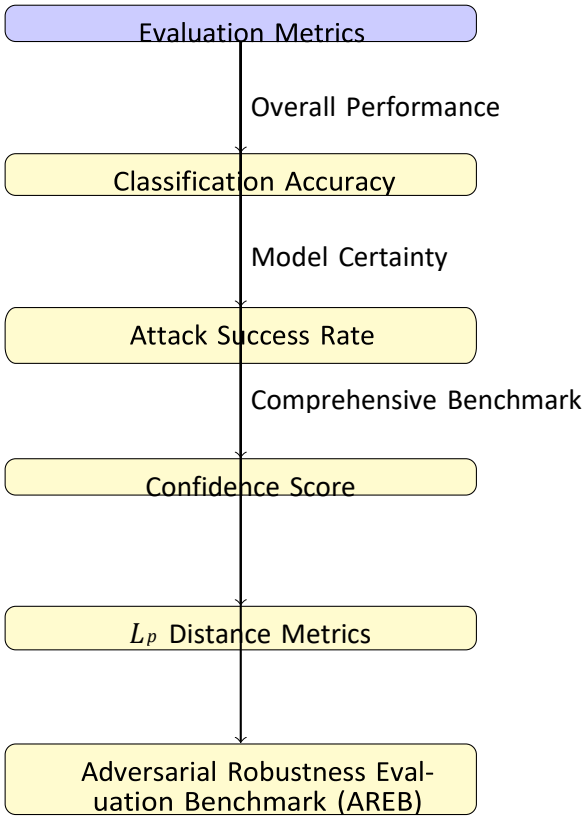


Figure 3. Evaluation Metrics for Assessing CNN Robustness

Defence Mechanisms

The defence mechanism of the CNNs is engineered using Adversarial training which incorporating adversarial use cases into the training process to improve model resilience.

3.4.1. Formulation of Defence Mechanisms. Given a machine learning model $f : X \rightarrow Y$, where X represents the input space and Y represents the output space, defence mechanisms are designed to reduce the effectiveness of adversarial attacks, i.e., to ensure that the model's predictions remain correct even when adversarial perturbations are applied to the applicable inputs.

Let the original input be $\mathbf{x} \in X$, and let $\mathbf{x}^* \in X$ represent an adversarially perturbed input. A defence mechanism D transforms the adversarial input $\mathbf{x}^*$ back to a more robust input, denoted $\mathbf{x}^D$, where:

$$\mathbf{x}^D = D(\mathbf{x}^*) \qquad (9)$$

The transformed input $\mathbf{x}^D$ is then passed to the model:

$$\hat{y} = f(\mathbf{x}^D) \qquad (10)$$

where $\hat{y}$ is the predicted output after the defence mechanism is applied.

Defence mechanisms typically aim to minimize the adversarial impact by applying one or more of the following strategies:

- Adversarial Training: The model is trained with adversarial examples generated from the training data, which strengthens its robustness. Let $\mathbf{x}_{adv}$ be an adversarial sample generated from the input $\mathbf{x}$, the model then learns:

$$L(f(\mathbf{x}_{adv}), y) \quad for all \mathbf{x}_{adv} \in X \qquad (11)$$

where L is a loss function (e.g., cross-entropy) and y is the correct class label.

- Defensive Distillation: The model is trained to output probability distributions instead of class labels. The transformed probability distribution reduces the sensitivity to small perturbations:

$$f_{distill}(\mathbf{x}^*) = softmax(Wf(\mathbf{x}^*))$$

where W is a weight matrix for the softmax function, and $f_{distill}(\mathbf{x}^*)$ is the output after distillation.

- Gradient Masking: A mechanism to hide the gradients used by adversaries to craft adversarial examples. The gradients $\nabla_x f(\mathbf{x})$ are masked to reduce their usefulness in attack generation:

$$\mathbf{x}^* = \mathbf{x} - \epsilon \nabla_x f(\mathbf{x}) \quad with \nabla_x f(\mathbf{x}) being masked \qquad (12)$$

- Input Transformation: Input transformations such as random noise addition or JPEG compression are applied to inputs before passing them to the model:

$$\mathbf{x}^T = T(\mathbf{x}^*) \qquad (13)$$

where T is a transformation function (e.g., random cropping, scaling).

The goal of any defence mechanism is to ensure that after applying the defence D, the model's prediction remains correct even when adversarial perturbations are present: $f(\mathbf{x}^D) = f(\mathbf{x})$ $for all adversarial inputs \mathbf{x}^*$. (14)

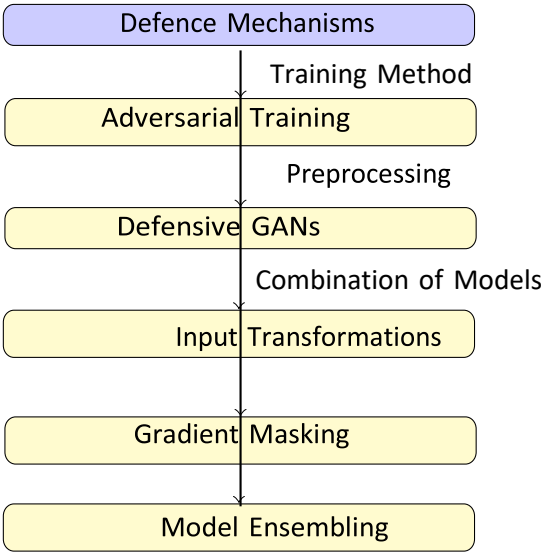


Figure 4. Defence Mechanisms to Protect CNNs from Adversarial Attacks

Results and Discussion

The research is implemented using popular deep learning frameworks- TensorFlow and PyTorch. The evaluations is performed in a scenerio where the attacker has predefined levels of cross-functional knowledge about the model. The model is distorted by using environmental factors such as lighting variations, occlusions, and sensor noise.

The Figure 5 shows that optimal stopping point for this model will be around epoch 5-6, where validation performance levels off before over-fitting becomes more severe.
The Figure 6 shows the accuracy of a machine learning model over 10 training epochs, with two lines tracking different metrics: The blue line represents Training Accuracy, which starts around 0.49 (49The orange line represents Validation Accuracy, which starts higher at about 0.56

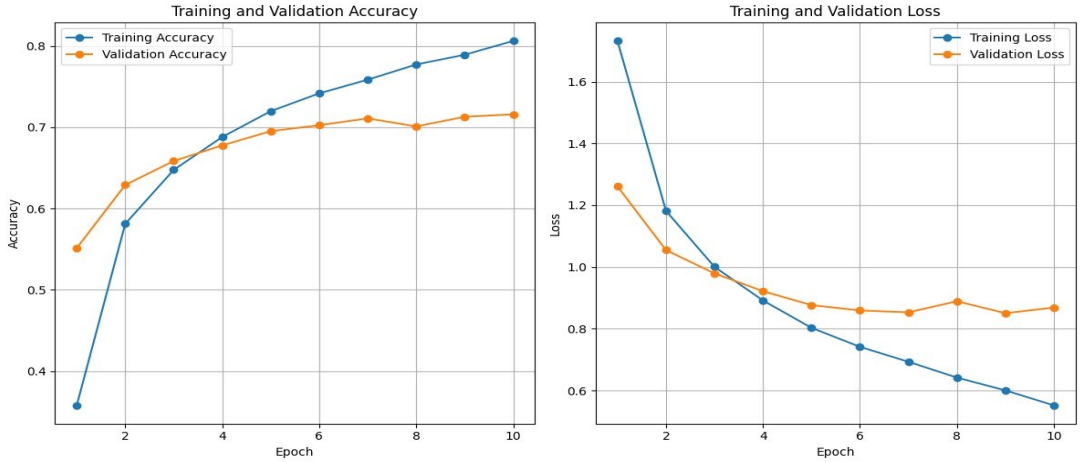


Figure 5. Accuracy and Loss Measurement

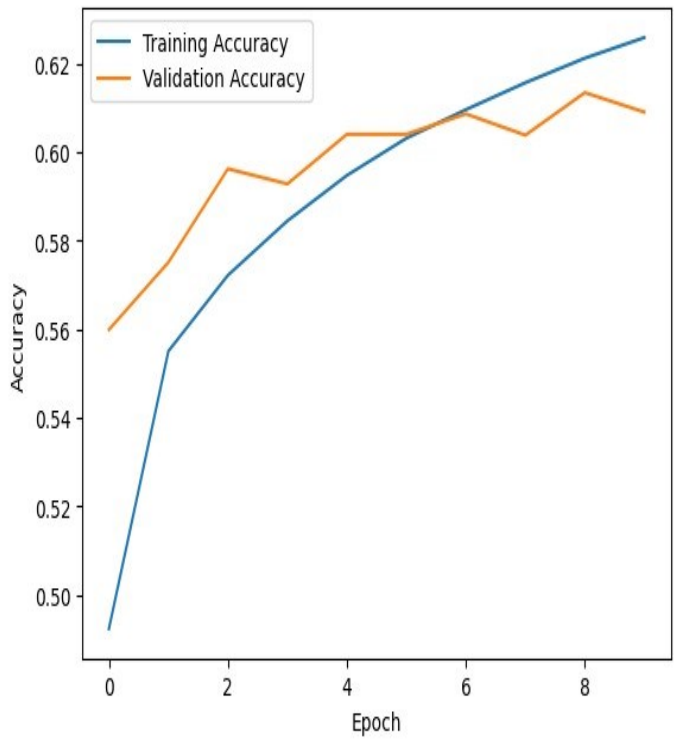


Figure 6. Training and Validation Accuracy
(56The relationship between these lines suggests potential over fitting beginning around epoch 8, as the training accuracy continues to improve while validation accuracy starts to decline. The convergence of both lines around epoch 6 followed by their divergence indicates that epoch 6-7 might be an optimal stopping point for this model to achieve the best generalisation performance and reliability
The results are further explained in Figure 7 below:

EDITOR-IN-CHIEF
Prof. A.I. Mteiye [MALAWI]

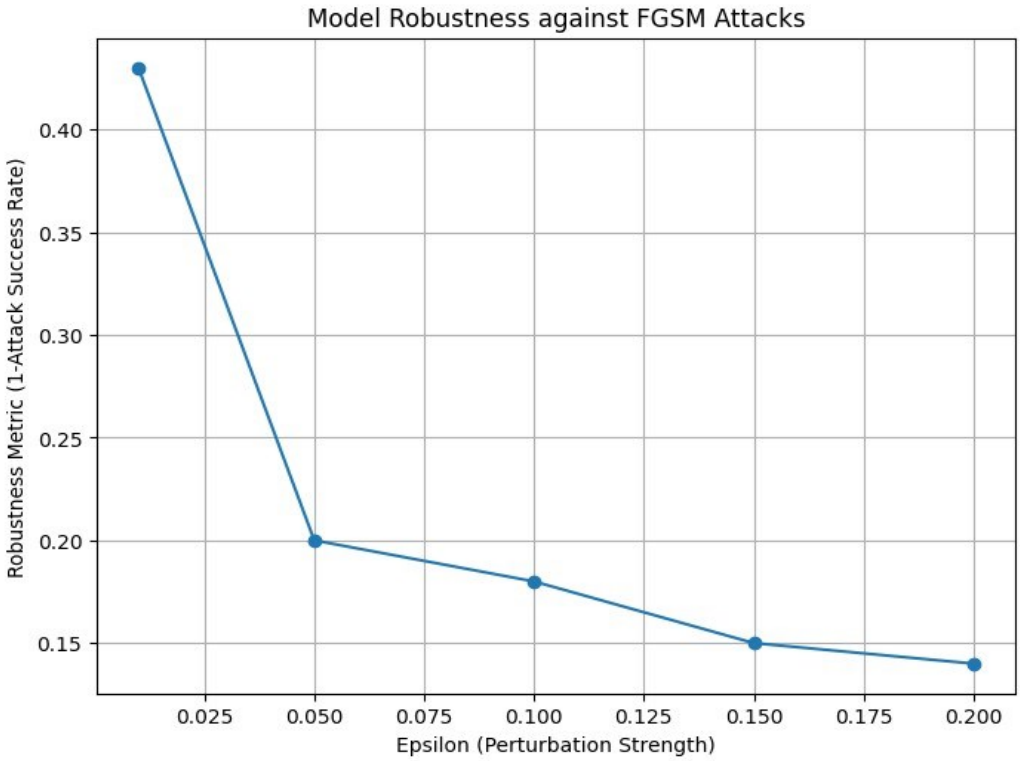


Figure 7. Model Robustness and Perturbation

Conclusion

This work investigates the robustness of Convolutional Neural Networks (CNNs) against both digital and physical adversarial image perturbations under realistic and simulated environmental conditions, including lighting variations, occlusions, and sensor noise. The experimental findings demonstrate that even state-of-the-art CNN models was impeded with a significant performance degradation when simultaneously exposed to adversarial manipulations and real-world distortions or perturbation. The use of TensorFlow and PyTorch to implement the model, the work revealed that the models exhibited early indication of overfitting beyond epoch 6, as training accuracy continued to improve while validation accuracy declined intermittently . The optimal stopping point, therefore, lies around epochs 6–7, where the model achieves the best balance between performance and generalization at best.

These results reinforce the vulnerability of CNNs in uncontrolled environments and highlight the importance of incorporating robustness-enhancing and perturbation-degradation strategies such as adversarial training, defensive distillation, and input transformation into the learning pipeline and workflow adjustment. Again, the work emphasizes the need for comprehensive evaluation frameworks that understand and real-world complexities rather than idealized testing conditions. Realistically, this work

contributes and model the development of more resilient, secure, and practically deployable CNN-based systems capable of maintaining reliability and performance even under adversarial and environmental challenges under uncertain conditions and nuances.

References

Barr, M. (2025). A Robust Neural Network against Adversarial Attacks. Engineering, Technology & Applied Science Research, 15(2), 20609-20615.

Carlini, N., & Wagner, D. (2017, November). Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security* (pp. 3-14).

Chaudhury, S., & Yamasaki, T. (2024, April). Adversarial Robustness of Convolutional Models Learned in the Frequency Domain. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7455-7459). IEEE.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., ... & Hein, M. (2020). Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*.

Eleftheriadis, C., Symeonidis, A., & Katsaros, P. (2024). Adversarial robustness improvement for deep neural networks. Machine Vision and Applications, 35(3), 35.

Fawole, O., & Rawat, D. (2025). Recent Advances in Vision Transformer Robustness Against Adversarial Attacks in Traffic Sign Detection and Recognition: A Survey. *ACM Computing Surveys*, *57*(10), 1-33.

Gupta, H., Jin, K. H., Nguyen, H. Q., McCann, M. T., & Unser, M. (2018). CNN-based projected gradient descent for consistent CT image reconstruction. *IEEE transactions on medical imaging*, *37*(6), 1440-1453.

Hu, J., Ye, J., Feng, Z., Yang, J., Liu, S., Yu, X., ... & Song, M. (2024). Improving adversarial robustness via feature pattern consistency constraint. arXiv preprint arXiv:2406.08829.

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial intelligence safety and security* (pp. 99-112). Chapman and Hall/CRC.

Lad, A., Bhale, R., & Belgamwar, S. (2024, April). Fast gradient sign method (FGSM) variants in white box settings: A comparative study. In *2024 International Conference on Inventive Computation Technologies (ICICT)* (pp. 382-386). IEEE.

Pietrosanti, G. M., Rossolini, G., Biondi, A., & Buttazzo, G. (2025). Benchmarking the spatial robustness of DNNs via natural and adversarial localized corruptions. arXiv preprint arXiv:2504.01632.

Silva, S. H., & Najafirad, P. (2020). Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*.

Smagulova, K., Bacha, L., Fouda, M. E., Kanj, R., & Eltawil, A. (2024). Robustness and transferability of adversarial attacks on different image classification neural networks. Electronics, 13(3), 592.

Wang, Y., Chou, Y., Zhou, Z., Zhang, H., Wan, W., Hu, S., & Li, M. (2025, April). Breaking barriers in physical-world adversarial examples: Improving robustness and transferability via robust feature. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 8, pp. 8069-8077).