# Efficient Defense Against Adversarial Attacks

**Amruta Mankawade[1], Pavitha Nooji[2*],Aditya Kulkarni[3], Shravani Dhamne[4], Raj Dharmale[5], Jayesh Chaudhari[6]**

[1] Assistant Professor, Department of Artificial Intelligence and Data Science Engineering, Vishwakarma Institute of Technology, Pune
[2*]Assistant Professor, Department of Artificial Intelligence, Faculty of Science and Technology, Vishwakarma University, Pune, Email: pavitha.nooji@vupune.ac.in
[3,4,5,6]Student, Department of Artificial Intelligence and Data Science Engineering, Vishwakarma Institute of Technology, Pune
*Corresponding Author

Received: 15.07.2024 Revised: 13.08.2024 Accepted: 09.09.2024

**ABSTRACT**
Adversarial attacks are a significant vulnerability for deep learning models, particularly Convolutional Neural Networks (CNNs), which are widely employed in image classification and object detection. These attacks involve crafting imperceptible perturbations to input data that mislead CNNs into making incorrect predictions, posing risks in critical areas such as autonomous driving, security, and healthcare. This paper focuses on understanding the nature of adversarial attacks on CNNs, including white-box attacks, where attackers have full knowledge of the model's parameters, and black-box attacks, where attackers have limited or no access to the model's architecture. Common attack techniques such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) and many more are reviewed to illustrate how CNNs can be compromised. In response to these threats, we explore various defense mechanisms aimed at increasing CNN robustness. Adversarial training, which incorporates adversarial examples during the training process, is a prominent defense. Other approaches, like input preprocessing, gradient obfuscation, and randomization techniques, are also discussed. This work emphasizes the trade-off between the efficiency of these defenses and their ability to protect CNNs without significantly increasing computational costs.

**Keywords:** Attacks, Adversarial Training, Defenses, Neural Network

## 1. INTRODUCTION

Adversarial attacks present a growing threat to the security and dependability of deep learning models, particularly neural networks. These attacks involve making subtle changes to input data, often undetectable by humans, which lead to incorrect predictions from neural networks. Such vulnerabilities are especially concerning in critical sectors like autonomous systems, healthcare, and cybersecurity, where errors can have severe consequences. As adversarial techniques become increasingly sophisticated, they reveal new weaknesses in neural architectures, making the development of robust defense mechanisms essential.

Adversarial attacks are generally classified into two categories: targeted and untargeted. Targeted attacks manipulate the input to force the model into predicting a specific incorrect class, often with harmful or malicious intent. In contrast, untargeted attacks aim to mislead the model into making any incorrect prediction, without focusing on a particular class. Both attack types can be executed under different access conditions.

Based on the attacker's access to the model, adversarial attacks can be categorized as white-box, black-box, and gray-box attacks. In white-box attacks, attackers have complete knowledge of the model's architecture, weights, and training data, allowing for precise manipulation. Black-box attacks are carried out without any internal knowledge of the model, typically through querying it to observe outputs. Gray-box attacks fall between these two extremes, where attackers have partial knowledge of the model's internals.

Despite the existence of numerous defensemethods, such as adversarial training and preprocessing techniques, they often struggle to keep pace with new adversarial strategies and can be computationally expensive. This research focuses on a comprehensive vulnerability assessment and proposes a defense mechanism that leverages parallel computing to improve efficiency, aiming to protect neural networks while maintaining high performance in real-time applications

## 2. Objectives

This research focuses on developing a computationally efficient defense mechanism using parallelism to protect deep learning models from adversarial attacks. By leveraging parallel computing, the defense aims to enhance the model's ability to withstand various adversarial strategies without significantly increasing computational costs. In addition, the study evaluates the vulnerability of models through diverse attack techniques, such as single-pixel attacks and randomized perturbations, providing a thorough analysis of the models' security. The investigation further explores the effectiveness of advanced techniques like Denoising Autoencoders (DAE), while utilizing parallelism to optimize performance and improve defenses against adversarial threats. A comparative analysis of various attack and defense methods is conducted to identify the most effective strategies, contributing to a deeper understanding of the security landscape in deep learning models. Through this, the research aims to strengthen defenses and provide insights into optimal approaches for safeguarding neural networks.

## 3. Scope and Methodology

This study focuses on evaluating the robustness of deep learning models against adversarial attacks using the MNIST dataset, which consists of hand-written digits. The methodology is divided into several key steps, covering dataset preparation, model implementation, adversarial attack generation, defense mechanisms, and a comparative analysis of model performance under attack.

**Dataset Selection:** The MNIST dataset is chosen for this study as it is a widely-used benchmark for image classification tasks, particularly for evaluating the performance of Convolutional Neural Networks (CNNs). The dataset consists of 60,000 training images and 10,000 testing images, each representing a grayscale hand-written digit from 0 to 9. Images are normalized to have pixel values between 0 and 1 for consistency across models and attacks.

**Model Implementation:** A standard **Convolutional Neural Network (CNN)** is implemented for image classification on the MNIST dataset. The CNN architecture includes convolutional layers followed by pooling layers and fully connected layers. Activation functions such as ReLU are applied between layers, and softmax is used at the output layer for class predictions. The model is trained on the MNIST dataset using stochastic gradient descent (SGD) or a similar optimization algorithm, with cross-entropy as the loss function.

**Adversarial Attack Implementation:** Several adversarial attack methods are employed to test the vulnerability of the CNN model:

**Fast Gradient Sign Method (FGSM):** This attack computes the gradients of the loss function with respect to the input and creates perturbations by adding a scaled version of the sign of the gradients to the input. FGSM is fast and effective but generates relatively larger perturbations.

**Iterative Fast Gradient Sign Method (I-FGSM):** An iterative version of FGSM that applies smaller perturbations over multiple steps. The input is iteratively updated using the gradient sign, making the attack more effective than single-step FGSM.

**Momentum Iterative Fast Gradient Sign Method (MI-FGSM):** An advanced version of I-FGSM, this method incorporates momentum during the gradient update process to stabilize the attack and escape from local minima. The momentum term helps improve the attack’s performance and increases the likelihood of success against models trained with basic defenses like adversarial training.

**Projected Gradient Descent (PGD):** PGD is an iterative attack that refines adversarial examples by iteratively updating the input with smaller perturbations, while constraining them within a predefined boundary. This attack is more powerful than FGSM.

**Carlini-Wagner (CW) Attack:** CW is a sophisticated attack that minimizes the perturbation size while maintaining the attack’s effectiveness, making it harder to detect. It is often used as a benchmark for testing model defenses.

**Transferability Attack:** This attack exploits the phenomenon where adversarial examples generated for one model can successfully fool another model, even with a different architecture. We generate adversarial examples on a surrogate model and test their effectiveness on the target CNN, examining the cross-model transferability of adversarial attacks.

**Simulated Power Attack (SPA):** SPA simulates power-based physical attacks by introducing perturbations that mimic power fluctuations or hardware faults. These perturbations test a model's robustness in real-world conditions where power instability may affect performance, particularly in energy-sensitive or resource-constrained systems.

**DeepFool Attack:** DeepFool is an iterative method that generates minimal perturbations to misclassify inputs by progressively linearizing the model's decision boundaries. It's effective in creating subtle adversarial examples with minimal changes.

**Defense Mechanisms Against Adversarial Attacks:** To improve the model's robustness against adversarial attacks, the following defense techniques are implemented:

**Adversarial Training:** The CNN model is trained not only with clean examples but also with adversarial examples generated by the FGSM and PGD methods. This process helps the model generalize better against adversarial perturbations and become more robust.

**Defensive Distillation:** A distilled model is trained using softened logits from the original CNN. By lowering the temperature during the training process, the model becomes less sensitive to small changes in input, improving its resilience to attacks.

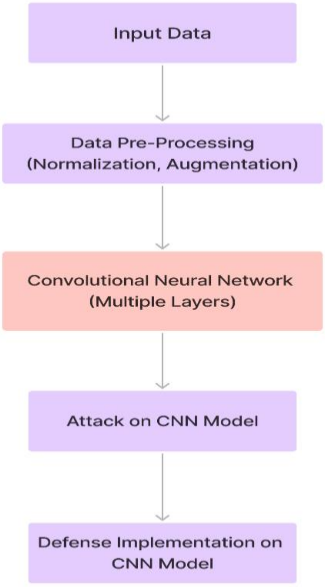
**Gradient Masking:** This method modifies the CNN architecture to reduce the effectiveness of gradient-based attacks. By obscuring gradient information, it becomes more challenging for attackers to compute precise perturbations, limiting the success of attacks like FGSM and PGD.

**Comparative Study of Models:** The study conducts a comparative evaluation of the CNN model's performance under different adversarial attack scenarios and defense mechanisms. The following metrics are used for comparison:

**Accuracy under attack:** The classification accuracy of the model is evaluated when subjected to adversarial examples generated by FGSM, PGD, and CW attacks.

**Perturbation size:** The magnitude of perturbations required for each attack is recorded to assess the effectiveness of both the attacks and defenses.

**Defense performance:** The effectiveness of adversarial training, defensive distillation, and gradient masking is evaluated based on the model's ability to maintain accuracy and withstand different attacks. This comparative study helps in identifying the most robust defense strategy for CNNs on the MNIST dataset.


Input Data
Data Pre-Processing (Normalization, Augmentation)
Convolutional Neural Network (Multiple Layers)
Attack on CNN Model
Defense Implementation on CNN Model

## 4. LITERATURE REVIEW

The paper explores white-box and black-box attacks, like Fast Gradient Sign Method (FGSM), on Support Vector Machine (SVM) and Convolutional Neural Network (CNN). It examines defenses such as adversarial training, gradient hiding, defensive distillation, and feature squeezing. A novel approach using Generative Adversarial Networks (GANs) is proposed to reduce the impact of adversarial attacks on both white-box and black-box models [1].

This paper categorizes adversarial attacks into targeted and non-targeted types, along with corresponding defense strategies. It implements Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Basic Iterative Method (BIM). A randomization-based defense method achieves a score of 92.4% among 107 submissions, while a High-level Guided Denoiser (HGD) removes adversarial noise by training a DNN-based denoiser on 20,000 ImageNet images (20 per class)[2].

This paper investigates how adversarial perturbations can weaken trained policies in deep reinforcement learning. Huang et al. utilize the Fast Gradient Sign Method (FGSM) to construct a surrogate loss across various algorithms,Liang et al. present word-level perturbations to mislead deep neural network-based text classifiers, using the TextBugger tool to show how adversarial text can cause misclassifications. A key limitation of this research is its shallow exploration of the causation behind adversarial samples, noting the difficulties in understanding how high-dimensional data geometry and insufficient training data contribute to vulnerability[3].

The paper employs MEFDroid, a multi-model ensemble framework that integrates predictors and hybrid deep learning techniques for enhanced imbalanced Android malware detection.The malware detection performance is improved using ESAES, EDAES, and EDAFS algorithms, surpassing classical machine learning and traditional sampling methods. However, it notes that many heuristic defenses struggle against adaptive adversaries in white-box settings, raising doubts about their effectiveness[4].

The paper employs comprehensive feature preprocessing using the Yeo-Johnson power transformation for standardization, utilizing nearly two dozen metrics to interpret linear model weights effectively. It introduces the Fast-tack algorithm for scalable attacks, enabling efficient computation of perturbations ranked by predicted impact, while a linear model is trained to enforce budget constraints for subtle perturbations.To deepen the understanding of adversarial attacks on graph neural networks, the paper formulates null hypotheses, analyzes the Cora-ML dataset, and references relevant works to enhance model robustness [5].

The paper proposes an efficient defense mechanism against Fast Gradient Sign (FGS) adversarial attacks on deep learning models. It employs a Denoising Autoencoder (DAE) trained on both clean and adversarially perturbed.However, the defense shows less robustness in black-box threat models due to gradient mismatches between the adversary and target model.The evaluation focuses solely on FGS attacks, overlooking performance against more complex adversarial methods. Tested on MNIST and Fashion-MNIST datasets, the defense demonstrates robust accuracy retention across varying perturbation magnitudes, outperforming baseline methods and 2.4x faster computational speed [6].

This survey highlights the growing concerns over the vulnerability of deep neural networks (DNNs) to adversarial attacks, which threaten the large-scale deployment of deep learning models.The paper reviews the literature on adversarial attacks, primarily focusing on evasion attacks, and proposes a systematic analysis framework inspired by the lifecycle of Advanced Persistent Threats (APT). This framework provides a structured approach to understanding both attacks and defenses, allowing for the combination of multiple defensive strategies at different stages[7].

This survey provides a comprehensive overview of adversarial attacks in artificial intelligence (AI), emphasizing the urgent need for research in this area as AI applications expand. The paper explains the significance, concepts, types, and dangers of adversarial attacks. It reviews key attack algorithms and defense strategies across image, text, and malicious code domains, helping researchers quickly identify relevant study areas[8].

This study investigates the vulnerability of artificial intelligence (AI) applications in oncology, particularly focusing on the susceptibility of convolutional neural networks (CNNs) to white- and black-box adversarial attacks during weakly-supervised classification tasks. The research reveals that vision transformers (ViTs) match CNN performance at baseline but exhibit significantly greater robustness against adversarial attacks. The findings support the notion that ViTs are more reliable learners in computational pathology, suggesting that AI models in this field should favor ViTs over CNN-based classifiers to enhance protection against adversarial perturbations [9].

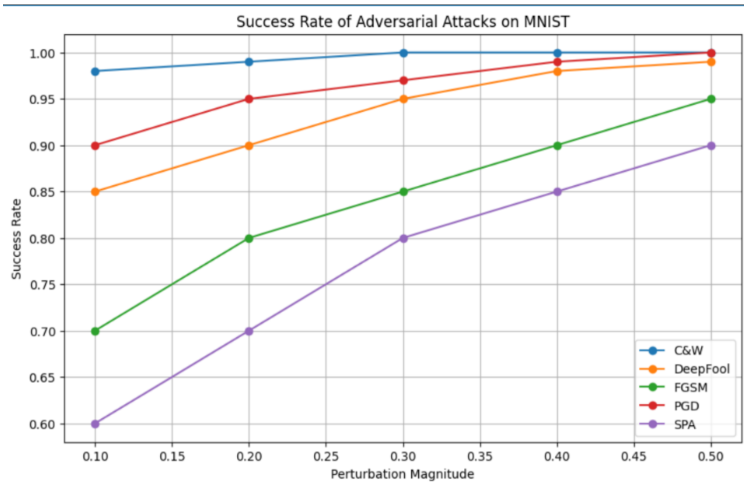
This research focuses on adversarial machine learning, specifically examining how adversarial attacks impact the accuracy of machine learning-based Intrusion Detection Systems (IDSs). Utilizing a Generative Adversarial Network (GAN), the study generated synthetic intrusion traffic to test two types of adversarial attacks: poisoning and evasion. Experiments were conducted on Decision Tree and Logistic Regression models using the CICIDS2017 dataset. The results revealed that evasion attacks significantly decreased the testing accuracy of both network intrusion detection models, with the Decision Tree model being more adversely affected than Logistic Regression[10].

## 5. RESULT AND DISCUSSION

The results demonstrate a significant impact of adversarial attacks on the accuracy of the model, followed by considerable improvements after applying defense mechanisms. For FGSM, IFGSM, and MIFGSM attacks, accuracy dropped drastically to 11%, 12.2%, and 12.03%, respectively. However, after employing defensive distillation as a defense mechanism, the model's accuracy was restored to 91% for all three attacks.With Projected Gradient Descent (PGD) attacks, using 0.1 epsilon over 12 iterations resulted in an accuracy of 18%, which further dropped to 12% with 0.3 epsilon. After defense, with adversarial training, accuracy improved to 98% for 0.1 epsilon and 97% for 0.3 epsilon.In the Carlini-Wagner attack, accuracy after the attack was 96%, but after applying adversarial training, it increased to 98%. For the transferability attack, accuracy fell to 16%, but adversarial training with a surrogate model improved it to 97%. Lastly, for the DeepFool attack, the model’s accuracy dropped drastically to 1.84%, and after defense, it was further reduced to85%, indicating some challenges in mitigating this particular attack effectively.Overall, defense techniques like defensive distillation and adversarial training demonstrated strong resilience against most adversarial attacks, successfully restoring or even enhancing model performance in most cases.Simulated Power Analysis (SPA), the mean and standard deviation of loss metrics reveal the impact of adversarial attacks on model performance. Initially, the model exhibited a stable mean loss of approximately 0.055 with a low standard deviation, indicating robust performance. The mean loss increased significantly, reflecting a degradation in accuracy, while the standard deviation rose, indicating greater variability in predictions. After employing defense strategies, such as defensive distillation, the mean loss decreased to around 0.071, and the standard deviation reduced, highlighting improved stability and reliability in the model's performance.

| Attacks | Accuracy After Attack | Accuracy After Defense | Defense Technique |
|---|---|---|---|
| **FGSM** | 11.86% | 91.91% | Defense Distillation |
| **IFGSM** | 12.2% | 91.81% | Defense Distillation |
| **MIFGSM** | 12.03% | 91.48% | Defense Distillation |
| **Carlini Wagner** | 96% | 98% | Adversarial Training |
| **Transferability** | 16% | 97% | Adversarial Training with Surrogate Model |

| | | | |
|---|---|---|---|
| **Deep-fool** | 1.84% | 85% | Adversarial Training |
| **PGD** | 12.03% | 97.64% | Adversarial Training |



## 6. Findings

This research provided key insights into the effectiveness of different defense strategies against various adversarial attacks in deep learning. Defensive distillation proved to be particularly effective against attacks such as FGSM, IFGSM, and MIFGSM, while adversarial training demonstrated resilience against more complex attacks, including Projected Gradient Descent (PGD) and Carlini-Wagner. Additionally, the study delved into the characteristics of black-box, white-box, and gray-box attacks, each posing distinct challenges based on the attacker's level of knowledge about the model. This exploration not only deepened my understanding of adversarial attack and defense mechanisms but also contributed to advancing deep learning by identifying strategies to improve model robustness against emerging adversarial threats.

## 7. Limitations and Research Gaps

One limitation of this research is the need for more powerful computational resources, such as GPUs, to handle the high complexity and faster processing of adversarial attacks and defenses. This restricted the ability to perform more extensive experimentation, especially with larger datasets and deeper models. Additionally, while certain defenses, such as defensive distillation and adversarial training, were effective in specific attack scenarios, their generalizability across different datasets remains unexplored, indicating a research gap. The relatively high accuracy (96%) after the Carlini-Wagner attack suggests that the model may not fully detect or respond to this type of attack, highlighting a need for further refinement in defenses against such sophisticated methods. Future research should focus on enhancing model understanding of more subtle attacks like Carlini-Wagner, as well as broadening testing to include diverse datasets and architectures to ensure robust, scalable defense mechanisms across various scenarios.

## 8. CONCLUSION

The research highlights the critical role of implementing tailored defense mechanisms to protect deep learning models from a range of adversarial attacks. Techniques such as defensive distillation and adversarial training have shown effectiveness against attacks like FGSM, IFGSM, MIFGSM, PGD, and Carlini-Wagner, though there remains a need for further enhancement, especially in defending against complex attacks like Carlini-Wagner. The findings also emphasize the necessity for more computational power, particularly GPUs, and for testing defenses on a wider variety of datasets toensure generalizability. Additionally, the model's vulnerability to transferability and Deepfool attacks signals areas that require deeper exploration and refinement. This research contributes to improving the resilience of deep learning models, providing a foundation for future studies to strengthen defenses against evolving adversarial attacks.

## REFERENCES

[1] Anirban Chakraborty Manaar Alam, Vishal Dey, Anupam Chattopadhyay and Debdeep Mukhopadhya. A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology (2020).

[2] Mesut Ozdag. Adversarial Attacks and Defenses Against Deep Neural Networks: A Survey Complex Adaptive Systems Conference with Theme: Cyber Physical Systems and Deep Learning, CAS (2018).

[3] Kui Ren a b, Tianhang Zheng c. Adversarial Attacks and Defenses in Deep Learning (2019).

[4] Shreya Goyal, Mitesh Khapra. A Survey of Adversarial Defences and Robustness in NLP(2023).

[5] Daniel Zügner, Oliver Borchert, Amir Akbarnejad, Stephan Günnermann Adversarial Attacks on Graph Neural Networks: Perturbations and Their Patterns (2023).

[6] Rajeev Sahay, Rehana Mahfuz, Aly El Gamal. A Computationally Efficient Method for Defending Adversarial Deep Learning Attacks (2019).

[7] Shuai Zhou, Chi Liu. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity (2022).

[8] Zixiao Kong, Jing feng Xue, Yong Wang.A Survey on Adversarial Attack in the Age of Artificial Intelligence (2021).

[9] Narmin Ghaffari Laleh1, Daniel Truhn, Gregory Patrick Veldhuizen. Adversarial attacks and adversarial robustness in computational pathology (2022).

[10] Ebtihaj AlshahraniID, Daniyal Alghazzawi, Reem Alotaibi, Osama Rabie1 Adversarial attacks against supervised machine learning based network intrusion detection system (2022).