IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Adversarial Attack Defense For Image Classifiers: A Lightweight Cnn Filter

[1]Tanishka Gupta, [2]Fiza Bano, [3]Shalu Tyagi, [4]Pooja Yadav
[1]Student, [2] Student, [3]Assitant Professor, [4]F Grade Scientist
[1]CSE(AI&ML),
[1]Raj Kumar Goel Institute Of Technology, Ghaziabad, India

Abstract: This paper proposes a CNN-based defense mechanism to protect machine learning classifiers from adversarial attacks. By reconstructing inputs to mitigate adversarial perturbations, the method enhances model robustness without altering the original classifier. The approach maintains high clean accuracy, computational efficiency, and seamless integration, making it an effective solution against adversarial threats. Experimental results validate its peformance across various datasets and attack scenarios.

IndexTerms – Adversarial Attacks, Convolutional Neural Networks (CNN), FGSM, PGD, Robustness, MNIST Dataset, Image Classification, Adversarial Defense, Machine Learning Security

1.INTRODUCTION

Despite their impressive performance, machine learning models are vulnerable to adversarial attacks that exploit their sensitivity to small input perturbations. (S. Qiu et al. [17]) An antagonist injects malicious data into the training dataset to change the decision boundary and can change the input features of the model. Medical imaging relies on deep learning models to identify cancer and pneumonia from X-rays and MRIs. (Z. Yang et al. [19]) How ever, small changes to these images can trick the models into making errors, which could lead to misdiagnoses or delays in treatment. This makes it essential to improve their reliability and safe ty. (H. Sun et al. [6]) Adversarial attacks significantly distort outputs, exposing the high vulnerability of GAN-based image fusion.

Scope and Purpose: Machine learning models, intense deep learning networks, are increasingly deployed in critical applications, making their vulnerability to adversarial attacks a growing concern. The two types of adversarial attacks called Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM) succeed in tricking machine learning models through the intro duction of undetectable noise in the input images. C. Wu et al. [3] demonstrate the use of SRDA and ARDA to disclose protection weaknesses in image compression algorithms based on learning frameworks as well as test defense capabilities. Such attacks reveal critical weaknesses in deep learning models particularly when these weaknesses produce hazardous misclassifications. The standard defense approaches of adversarial training together with input transformation methods fail to achieve suitable accuracy levels alongside robust protection while generating substantial processing load.

The application of convolutional neural networks (CNNs) represents a potential defense solution because these networks excel at image processing by extracting automatic hierarchical features from images. Training CNNs enables them to identify clean images from adversarial images through the acquisition of unique adversarial disturbance patterns. The main hurdle exists in developing CNN training strategies that provide universal protection across multiple attack types and ensures high accuracy and processing speed.

This paper presents a new method which uses CNNs to process images in order to eliminate adversarial noise. The proposed method strengthens defense against adversarial attacks by keeping both model accuracy and efficiency intact. The research addresses model hardening and real time detection of adversarial inputs as two separate methods according to M.-I. Nicolae et al. [9]. The method presents an interesting way to protect machine learning security systems against hostile threats without reducing operational efficiency.

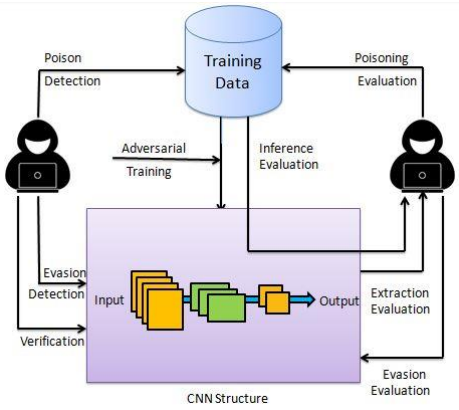

Figure 1. Adversarial Attack

2. METHODS.

2.1 Dataset: The application of CNNs to defend against adversarial attacks with (Bhatia et al. [2]) the MNIST dataset functions as the main subject of investigation. The dataset known as (Lv et al. [8]) MNIST contains 70,000 grayscale handwritten digital images divided into 60,000 training and 10,000 testing examples. The structure provides excellent conditions to evaluate different defense mechanism against adversarial attacks inside controlled systems.

The examination focuses on assessing CNN systems that attempt to generate uncorrupted images from manipulative ones while upholding correct outputs. MNIST provides exceptional conditions to evaluate adversarial defense methodologies that utilize CNNs for obtaining defense insights.

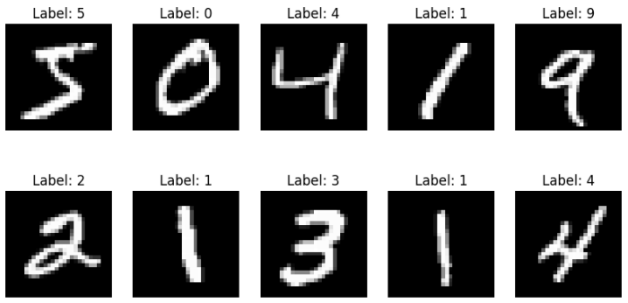

Figure 2. MNIST Dataset

2.2 Methods: The examination focuses on assessing CNN systems that attempt to generate uncorrupted images from manipulative ones while upholding correct outputs. MNIST provides exceptional conditions to evaluate adversarial defense methodologies that utilize CNNs for obtaining defense insights. Adversarial images generated using FGSM and PGD attacks are then applied to assess the model's robustness, ensuring reliable predictions despite adversarial noise. This part describes the main elements of our approach, covering the model design and training process.

The CNN model is used to predict the dataset images. It is designed for the classification of grayscale images, specifically 28×28×1, commonly found in the MNIST dataset. The model begins with an input layer that processes the image, followed by the first convolutional layer (Conv1), (Mitra et al. [11]) It uses 32 convolutional filters with dimensions 3×3, applying the ReLU activation to capture fundamental image features such as edges.

Batch Normalization (BN) is applied to stabilize and accelerate training, (Ghayoumi et al. [5]) followed by a 2×2 Max Pooling layer, which helps downsample the spatial dimensions. The network proceeds to Conv2 and

Conv3, which use 64 and 128 filters, respectively, to identify more complex features. These are followed by additional normalization and pooling layers.

(Bani-Hani et al. [1]) The output of the convolutional layers is flattened into a single-dimensional vector, which is passed to a fully connected layer with 128 units. A dropout layer with a rate of 0.5 is added to prevent overfitting. Finally, the softmax layer, with 10 output neurons, predicts the probability of each class, corresponding to the digits 0–9. This architecture effectively combines feature extraction and classification for robust image recognition tasks. (D. Liu et.al.[4]) Decoupled visual representation masking serves as an effective approach to enhance adversarial robustness in neural networks**.** (M. Terzi et.al.[12]) Image filtering techniques as a strategy to enhance the robustness of neural networks against adversarial attacks. The training progress was monitored using training and validation accuracies, along with training and validation losses.

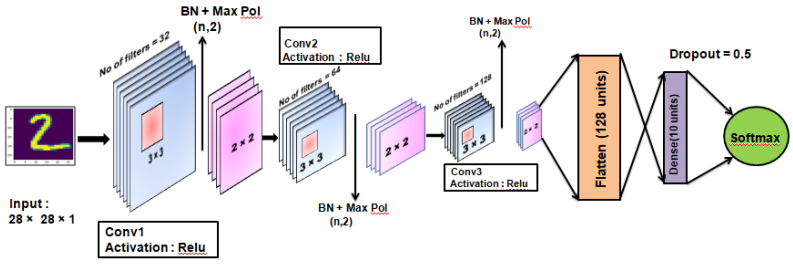

Figure 3. CNN Architecture

Table 1. summarizes the CNN architecture layers, shapes, and parameters.

Layer (Kind)	Output Shape	Parameters Count
conv2D	[batch size,26,26,32]	320
batch-normalization	[batch size,26,26,32]	120
max−pooling2D	[batch size,13,13,32]	0
Dropout	[batch size,13,13,32]	0
conv2d −1	[batch size,11,11,64]	18,496
batch−normalization −1	[batch size,11,11,64]	256
max−pooling2D−1	[batch size,5,5,64]	0
dropout − 1	[batch size,5,5,64]	0
conv2d −2	[batch size,3,3,128]	73,856
batch−normalization −2	[batch size,3,3,128]	512
max−pooling2D−2	[batch size,3,3,128]	0
dropout − 2	[batch size,3,3,128]	0
flatten	[batch size,128]	0
dense	[batch size,128]	16,512

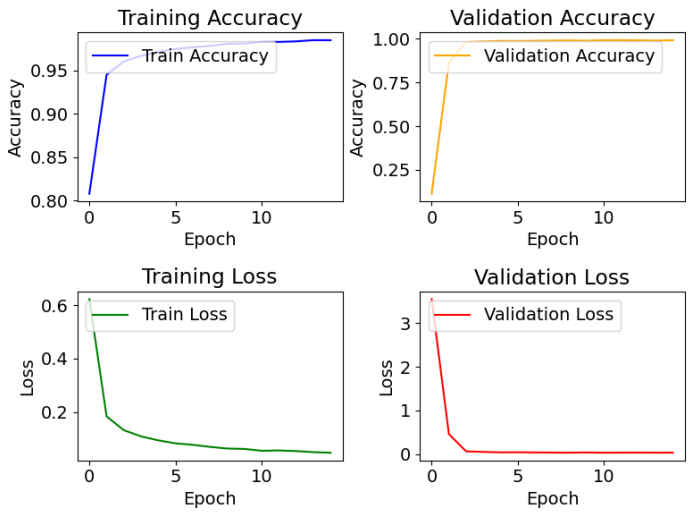

Figure 4. Training and Validation Accuracy, Loss curve

3. RESULTS: Now, we use two types of adversarial attacks, namely FGSM and PGD attacks. Demonstrating their effectiveness in deceiving image classifiers.

Qualitative Analysis: The effectiveness of adversarial attack FGSM (fast gradient sign method) was evaluated in the data set by varying the perturbation parameter (epsilon) and observing its impact on the precision and predictions of the test. The results demonstrate the impact of adversarial attacks by FGSM on the performance of the model as the perturbation parameter (epsilon) increases.

Table 2. below summarizes prediction over the 5 different epsilon values:

Epsilon	Accuracy	Loss	Test Accuracy	Prediction
0.01	0.9829	0.0449	98.74%	2
0.025	0.9747	0.0644	98.14%	2
0.05	0.9449	0.1518	95.77%	2
0.075	0.9001	0.3113	91.66%	2
0.1	0.8334	0.5540	85.63%	2

Figure 5. Prediction over 5 different epsilon values

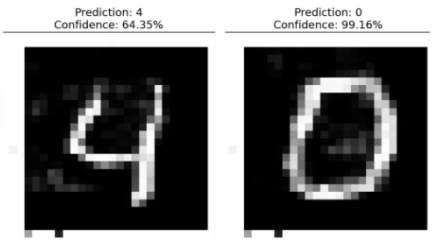

Figure 6. PGD prediction

Quantitative Analysis: The proposed CNN model demonstrates high accuracy in classifying PGD adversarial images, The defense system demonstrated excellent performance metrics with **0.9905** for both precision and recall measurements, resulting in an F1 score of **0.9904**. This performance profile highlights the effectiveness of our approach in detecting adversarial samples with high accuracy. The balanced precision and recall values suggest that the model excels not only at identifying manipulated images but also at minimizing erroneous classifications in both directions—limiting both false alarms and missed detections. The nearly perfect F1 score further validates the robustness of our detection mechanism across the evaluation dataset., making it a reliable approach for adversarial image detection.

The CNN model effectively classifies PGD adversarial images, demonstrating its robustness against perturbations. These prediction outcomes demonstrate our system's effectiveness at distinguishing legitimate images from those containing adversarial perturbations. The results confirm that the proposed lightweight CNN filter successfully identifies malicious inputs with high consistency, providing dependable protection for image classification systems under attack, making it a promising approach for secure image classification.

4. DISCUSSION

4.1 Overview: (Meiseles et al. [10]) Adversarial attacks are a major challenge to machine learning models, particularly in image classification. Methods like PGD and FGSM introduce slight perturbations to input images that cause misclassification. (Sam. et al. [15]) Using CNNs for medical image processing to enhance diagnostic accuracy and automate analysis. In fields like autonomous vehicles and medical imaging, these misclassifications can have serious consequences, highlighting the need for effective defenses.
Several defense strategies have been explored, such as (Wang et al. [18]) adversarial training, which involves training models on adversarial examples. However, this can reduce the accuracy of clean images. Input transformation methods, like image preprocessing, can reduce adversarial effects but often degrade clean image quality.

4.2 Comparison with existing model: Convolutional Neural Networks (CNNs) are effective in adversarial defense due to their ability to learn hierarchical features in images. (Sar. et al. [16]) Apply CNN models to food safety control by detecting risks such as contamination or spoilage in food images. CNNs can be used to identify and filter out adversarial noise, restoring image integrity without changing the core classifier.
CNNs have been studied for adversarial defense, with methods that focus on either classifying adversarial images or using them to clean the images. The challenge is to remove noise without distorting important features, and combining CNNs with other defense strategies can improve robustness. (J. Kim et al. [7]) Sensible adversarial learning improves robustness and is available as a Python package.
Despite their potential, CNNs require significant computational resources and may not work equally well against some types of attacks. Ensuring CNNs generalize well across various adversarial methods while maintaining accuracy remains a challenge.

5. CONCLUSION

In this study, we assessed (R. Paul et al. [14]) the robustness of a CNN model against adversarial attacks. Our experiments tested the defense mechanism against two prominent adversarial techniques: FGSM (Fast Gradient Sign Method) and PGD (Projected Gradient Descent). These attacks represent common vulnerability exploitation strategies in neural network image classifiers. Experimental results indicate a substantial decline in model accuracy under these attacks, revealing its susceptibility to adversarial perturbations. The single-step FGSM attack led to moderate accuracy degradation, while the iterative PGD attack caused severe misclassifications, significantly reducing test accuracy. These findings underscore the necessity of adversarial defense strategies, including adversarial training, input preprocessing, and robust model architectures, to enhance the resilience of CNNs against in adversarial threats.

6. REFERENCES

1. Bani-Hani, R. M., Shatnawi, A. S., & Al-Yahya, L. (2024). Vulnerability detection and classification of Ethereum smart contracts using deep learning. *Future Internet*, *16*(9), 321. https://doi.org/10.3390/fi16090321

2. Bhatia, T., Singh, P. K., Singh, K. V., Jayesh, & Rais, F. (2025). Optimized adversarial defense: Combating adversarial attacks with denoising autoencoders and ensemble learning. *AIJR Proceedings*, 150–158.

3. C. Wuet al., "On the adversarial robustness of learning-based image compression against rate-distortion attacks," arXiv [eess.IV], 2024.

4. D. Liu, T. Chen, C. Peng, N. Wang, R. Hu, and X. Gao, "Improving adversarial robustness via decoupled visual representation masking," arXiv [cs.CV], 2024.

5. Ghayoumi, M. (2023). *Generative adversarial networks in practice*. Chapman and Hall/CRC.

6. H. Sun, S. Wu, and L. Ma, "Adversarial attacks on GAN-based image fusion," Inf. Fusion, vol. 108, no. 102389, p. 102389, 2024.

7. J. Kim and X. Wang, "Robust sensible adversarial learning of deep neural networks for image classification," arXiv [cs.CR], 2022.

8. Lv, C., Gu, Y., Guo, Z., Xu, Z., Wu, Y., Zhang, F., Shi, T., Wang, Z., Yin, R., Shang, Y., Zhong, S., Wang, X., Wu, M., Liu, W., Li, T., Zhu, J., Zhang, C., Ling, Z., & Zheng, X. (2024). Towards biologically plausible computing: A comprehensive comparison. In *arXiv [cs.NE]*. http://arxiv.org/abs/2406.16062

9. M.-I. Nicolae et al., "Adversarial Robustness Toolbox v1.0.0," arXiv [cs.LG], 2018.

10. Meiseles, A., Motro, Y., Rokach, L., & Moran-Gilad, J. (2023). Vulnerability of pangolin SARS-CoV-2 lineage assignment to adversarial attack. *Artificial Intelligence in Medicine*, *146*(102722), 102722. https://doi.org/10.1016/j.artmed.2023.102722

11. Mitra, M., & Roy, S. (2023). Advancing COVID-19 diagnosis with CNNs: An empirical study of learning rates and optimization strategies. *Intelligent Control and Automation*, *14*(04), 45–78. https://doi.org/10.4236/ica.2023.144004

12. M. Terzi, M. Carletti, and G. A. Susto, "Improving robustness with image filtering," Neurocomputing, vol. 596, no. 127927, p. 127927, 2024. 12. S. G. Hodes, K. J. Blose, and T. J. Kane, "Black box phase-based adversarial attacks on image classifiers," in Automatic Target Recognition XXXIV, 2024

13. Prasad, R., & Koren, A. (2025). Safeguarding 6G: Security and Privacy for the Next Generation (1st ed.). River Publishers. https://doi.org/10.1201/9788770047951

14. R. Paul, M. Schabath, R. Gillies, L. Hall and D. Goldgof, "Mitigating Adversarial Attacks on Medical Image Understanding Systems," *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA, 2020, pp. 1517-1521, doi: 10.1109/ISBI45749.2020.9098740.

15. Samanta, S., Singh, J., Bhattacharjee, A., Kumar, S., Behera, M. (2023). Medical image processing using CNN. 2023 IEEE 3rd International Conference on Applied Electromagnetics, Signal Processing, Communication (AESPC).

16. Saranya, P., Durga, R. (2023). Food safety control using CNN model in image processing technique. 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS).

17. S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," Appl. Sci. (Basel), vol. 9, no. 5, p. 909, 2019.

18.Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., Hossain, E., & Poor, H. (2023). Adversarial attacks and defenses in machine learning-powered networks: A contemporary survey. *ArXiv*, *abs/2303.06302*. https://doi.org/10.48550/arXiv.2303.06302

19. Z. Yang, Y. Yang, Y. Xue, F. Y. Shih, J. Ady and U. Roshan, "Accurate and adversarially robust classification of medical images and ECG time-series with gradient-free trained sign activation neural networks," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 2456-2460, doi: 10.1109/BIBM49941.2020.9313442

7. ETHICAL CONSIDERATION

This research utilizes the MNIST dataset obtained through Hugging Face's publicly available data repository, which is provided under appropriate licensing terms for academic research. No additional data collection involving human participants was conducted for this study. We have adhered to Hugging Face's terms of service regarding dataset usage and citation requirements. Our adversarial attack methods were implemented solely for defensive research purposes, and we have taken care to document these approaches responsibly to promote security improvements rather than enabling potential misuse. All computational experiments were conducted with consideration for resource efficiency and environmental impact, utilizing optimized implementations where possible.