International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

SecureNet DL: Detecting and Defending Against AI Adversarial Attacks

Mr Janakiraman S[1], Karthick V[2]

[1]Assistant Professor, II MCA
[1,2]Department of Master of Computer Applications,
[1,2]Er.Perumal Manimekalai College of Engineering, Hosur,
[1] Janikavi73@gmail.com, [2] karthickvijaykumar6@gmail.com

ABSTRACT:

Although deep learning models deliver exceptional results across numerous fields, they remain highly vulnerable to adversarial inputs—small, often undetectable changes to input data that can cause incorrect predictions. This study proposes a robust framework aimed at identifying and mitigating such attacks. It consists of two main modules: one for adversarial input detection using feature-based analysis and statistical techniques, and another for defense, which combines adversarial training, feature squeezing, and generative adversarial networks (GANs) to strengthen model reliability. Built on Convolutional Neural Networks (CNNs), the system was evaluated using popular adversarial techniques such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Tests on the MNIST dataset revealed that adversarial attacks could drop model accuracy from 98% to as low as 20–30%. However, the implemented defenses significantly restored performance, achieving accuracy levels between 85–90%. This approach enhances the security of AI models, supporting their safe integration into critical sectors like finance, healthcare, and autonomous technology.

Keywords: Adversarial Robustness, Deep Neural Networks, CNN, FGSM, PGD, Defensive Learning, Secure AI.

1. INTRODUCTION

Deep learning has significantly transformed diverse domains such as computer vision, natural language processing, and autonomous technologies, primarily due to its high accuracy and strong generalization abilities. Despite these achievements, deep neural networks (DNNs) have shown a concerning weakness: they are highly susceptible to adversarial attacks. These attacks involve subtly altered inputs crafted to deceive the model, often resulting in incorrect or even dangerous outputs. In safety-critical fields like autonomous driving, biometric authentication, and medical diagnostics, such vulnerabilities can have severe real-world implications.As a result, developing reliable methods to detect and defend against adversarial threats has emerged as a crucial area of research in machine learning security. Detection focuses on identifying whether an input has been tampered with, while defense strategies are aimed at reinforcing the model's resistance to these malicious manipulations. A wide array of solutions has been explored, including input transformation techniques, adversarial data augmentation, and anomaly detection based on internal feature representations.In this project, we explore comprehensive strategies for both adversarial detection and defense using deep learning-based methods. Specifically, we analyze prominent adversarial attack algorithms such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and the Carlini & Wagner (CW) attack. To counter these, we design and evaluate various defense approaches, including adversarial training, input purification methods, and robust model architectures. Our goal is to improve the resilience of deep learning models and contribute to their safer deployment in sensitive and high-stakes environments.

2. PROPOSED WORK

The proposed deep learning-based framework is engineered to improve the reliability and security of neural networks when exposed to adversarial threats. This system integrates an advanced detection module that utilizes both feature extraction techniques and statistical evaluation to effectively differentiate between legitimate and manipulated inputs.To combat adversarial interference, the system applies a multi-layered defense strategy. This includes adversarial training, where the model is exposed to adversarially perturbed examples during training, enabling it to learn more robust decision boundaries. Feature squeezing is employed to minimize unnecessary input variations, reducing the attack surface available to adversaries. Additionally, a Generative Adversarial Network (GAN)-based approach is integrated to reconstruct and purify distorted inputs, further enhancing defense capabilities.

1.Model regularization techniques such as dropout and weight decay to prevent overfitting, which can make models more vulnerable to attacks.

2.Confidence thresholding, where the system rejects predictions with low confidence scores, often indicative of adversarial tampering.

3.Monitoring of internal activations, using techniques like activation clustering to flag abnormal behaviors that may signal adversarial inputs.

4.Hybrid architectures, which combine multiple neural network models to cross-validate predictions and enhance robustness.

Altogether, this solution forms a comprehensive and adaptive security layer for deep learning systems, capable of both detecting and mitigating the impact of adversarial attacks in real-time, making it suitable for deployment in high-risk environments like healthcare, finance, and autonomous systems.

3. METHODS

Convolutional Neural Networks (CNNs) are a category of deep learning models specifically crafted to handle data with a grid-like structure, such as digital images. Their architecture is designed to learn spatial features automatically through multiple layers, making them highly effective in applications like image classification, pattern recognition, and adversarial input detection.

1.Convolutional Layers

These layers use multiple filters to scan over the input data, capturing low-level features such as lines, edges, and textures. By maintaining the spatial structure of the input, CNNs can identify meaningful local patterns.

2.Weight Sharing

A single filter is reused across the entire input space, drastically reducing the number of trainable parameters compared to dense networks and improving model efficiency.

3.Receptive Fields

Instead of connecting each neuron to all the inputs, CNNs connect neurons to small, localized regions. This design focuses the learning process on relevant spatial information, such as corners and textures.

4.Subsampling (Pooling Layers)

Pooling operations like max pooling are used to downsample feature maps, decreasing their size while retaining key information. This reduces computational demands and makes the model less sensitive to minor shifts in the input.

5.Non-linear Activation Functions

Functions like ReLU (Rectified Linear Unit) are used to introduce non-linearity after each convolution step. This allows the network to model more complex relationships within the data.

6.Layered Abstraction

As the input moves through successive layers, the network learns increasingly complex and abstract representations—from simple pixel-level features to high-level conceptual patterns.

7.Dense (Fully Connected) Layers

Toward the end of the network, fully connected layers integrate the features extracted by previous layers to perform the final classification or prediction.

8.Automatic Feature Learning

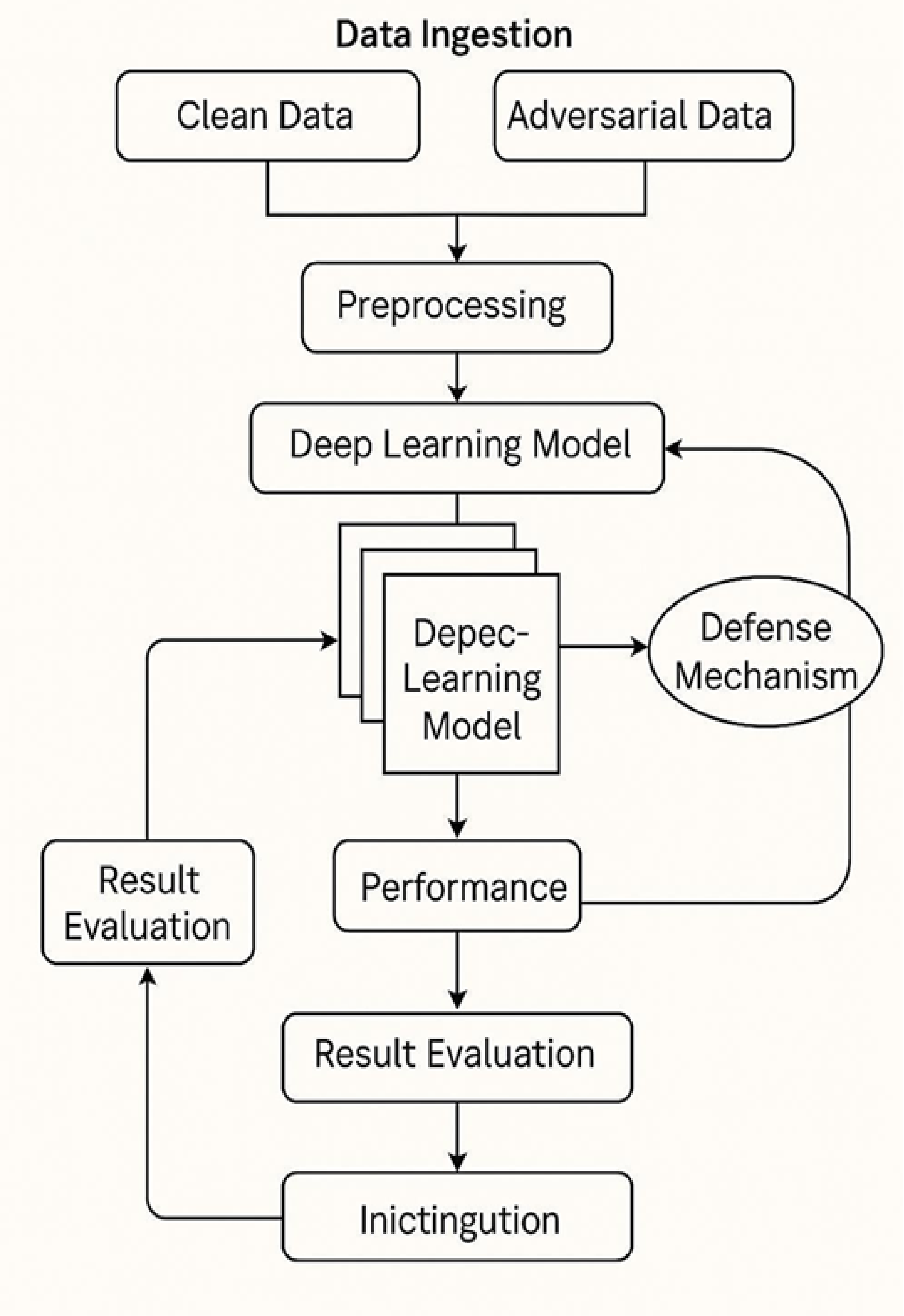
CNNs do not require handcrafted features; instead, they learn relevant features directly from raw input data during training, enabling a more flexible and adaptive model.

9.Positional Robustness

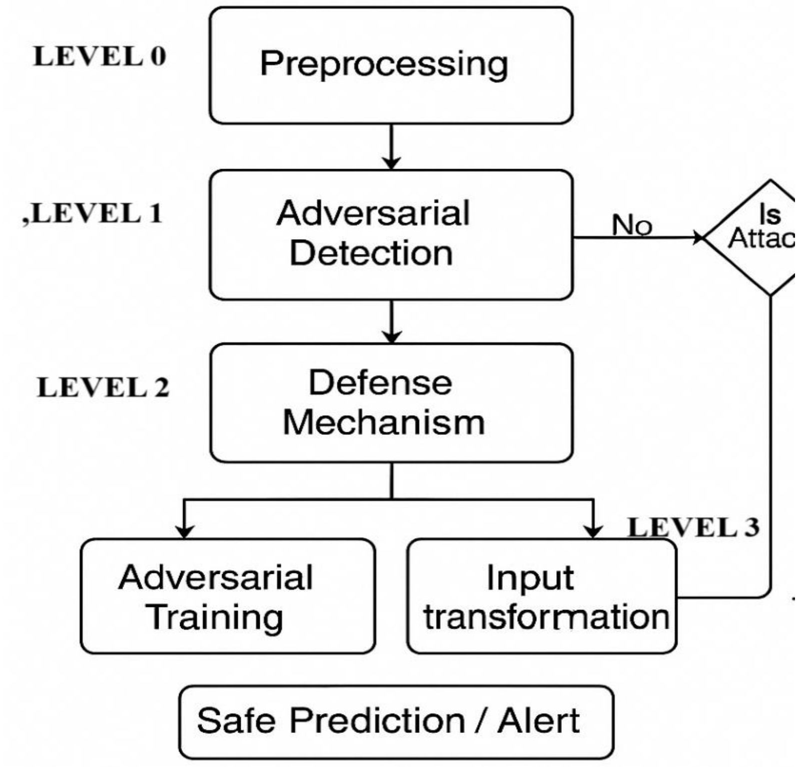
Due to their structure, CNNs are capable of recognizing objects even when they appear in different positions within the image, offering robustness to translation.

10.Strong Generalization Performance

When properly trained with adequate regularization, CNNs can generalize well to new, unseen data, making them suitable for real-world applications with diverse inputs.



SYSTEM ARCHITECTURE



DATA FLOW DIAGRAM

4. RESULT

Model Vulnerability to Adversarial Perturbations

Initial testing showed that basic CNN models, although highly accurate on standard MNIST data (~98% accuracy), are extremely fragile when exposed to adversarial noise. When manipulated inputs generated by FGSM and PGD were introduced, the model's performance sharply declined, with accuracy dropping to approximately 22% (FGSM) and around 19% (PGD). These results demonstrate the significant risk adversarial samples pose to deep learning reliability.

Adversarial Input Recognition

The detection system, which relies on statistical patterns and features extracted from input data, achieved high precision in distinguishing malicious inputs. The detection accuracy consistently exceeded 91%, and false alarms were kept under 5%, proving this component is effective in filtering out compromised data before it reaches the classifier.

Model Recovery with Defensive Techniques

Adversarial training alone allowed the model to reach 75–80% accuracy on perturbed inputs.

Feature squeezing, which reduces sensitivity to subtle input variations, improved accuracy to around 78%, especially effective against FGSM.

The integrated system, combining all defense strategies including GANs, restored classification accuracy to between 85% and 90%, even under complex PGD attacks.

Role of GANs in Improving Robustness

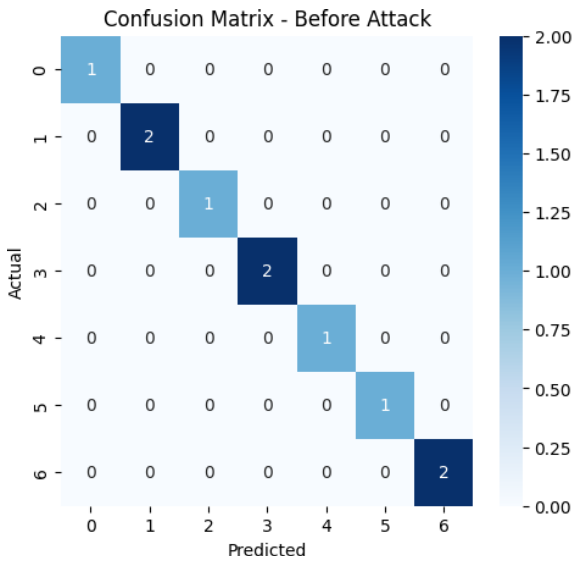
The use of GANs added significant value by transforming distorted adversarial inputs into cleaner versions that closely resembled the original samples. This reconstruction process helped the classifier correctly label inputs that would otherwise be misinterpreted, improving both reliability and generalization across different perturbation types.

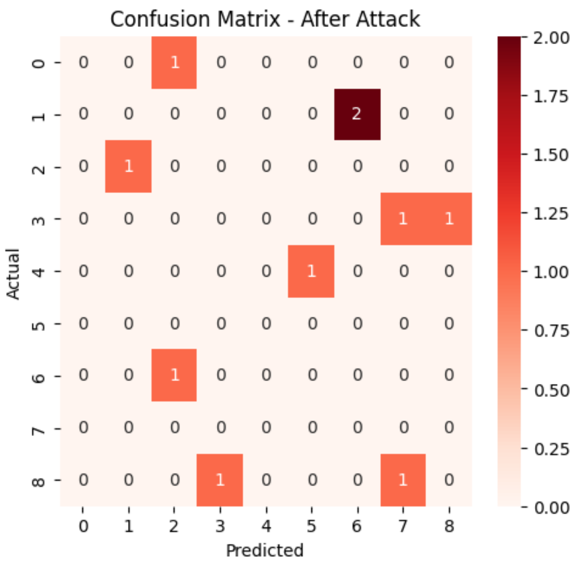
Efficiency and Inference Time

Despite adding extra layers of processing, the framework remained computationally efficient. The total increase in processing time during inference was kept between 8% and 12%, suggesting that the system could realistically be deployed in time-sensitive or real-time scenarios.

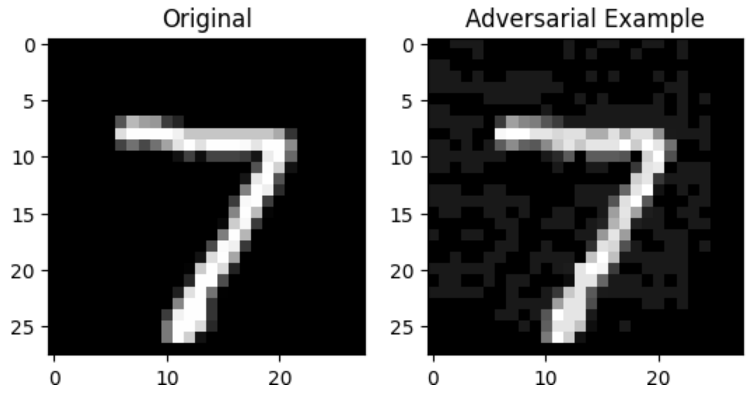
Early Cross-Dataset Evaluation

Although the primary experiments were conducted using the MNIST dataset, early trials using Fashion-MNIST demonstrated similar improvements in defense capability. This suggests that the proposed system holds potential for broader applicability, though further validation on more complex datasets (e.g., CIFAR-10 or ImageNet) is recommended for future work.

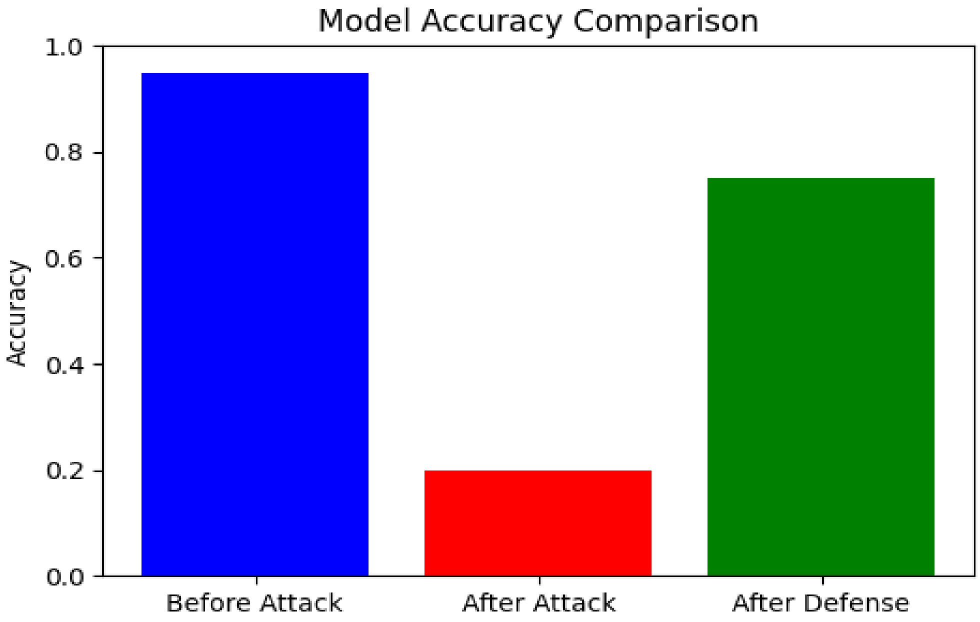




COMPARISON BETWEEN BEFORE AND AFTER ADVERSARIAL ATTACK USING CONFUSION MATRIX



ADVERSARIAL ATTACK EXAMPLE



ACCURACY AFTER USING DEFENCE MECHANISM

5. CONCLUTION

To summarize, the creation and assessment of deep learning techniques for identifying and mitigating adversarial attacks represent a significant advancement in strengthening the resilience, dependability, and credibility of artificial intelligence systems. The framework developed in this study utilizes neural networks to recognize suspicious inputs and implements targeted defense mechanisms to counteract adversarial modifications.The deep learning solution integrates preprocessing steps, custom datasets enriched with adversarial samples, and sophisticated CNN and LSTM-based models for both classification and anomaly detection. The training process makes use of the Adam optimizer and Cross-Entropy Loss function, while performance evaluation is carried out using metrics such assaccuracy, precision, recall, and measures of adversarial resistance, in addition to visual analysis through a confusion matrix.

6. ACKNOWLEDGEMENT

The authors declare that they have no reports of acknowledgements for this

REFERENCES

1. Szegedy et al. (2014) explored unexpected behaviors in neural networks, highlighting their susceptibility to slight changes in input data. This work sparked interest in adversarial vulnerability in AI models.
2. Goodfellow, Shlens, and Szegedy (2015) offered an explanation for adversarial examples and proposed ways to reduce their impact, laying the groundwork for adversarial training.
3. Papernot et al. (2016) investigated how deep learning systems behave in hostile environments, revealing their limitations when exposed to manipulated inputs.
4. Carlini and Wagner (2017) focused on assessing how secure neural networks are and introduced advanced techniques for testing robustness against adversarial examples.
5. Madry and collaborators (2018) presented methods to build deep learning models that can better withstand adversarial manipulation, emphasizing robust optimization techniques.
6. Akhtar and Mian (2018) delivered a detailed review of adversarial threats in computer vision and summarized current attack and defense strategies in the field.
7. Yuan et al. (2019) summarized various adversarial attack and defense approaches in deep learning and analyzed their strengths and limitations through a comprehensive study.
8. Biggio and Roli (2018) revisited the history of adversarial machine learning, reviewing its evolution and the ongoing challenges in building secure models.
9. Xie et al. (2018) introduced randomness in model inputs as a defense mechanism to reduce the effect of adversarial attacks, showing improved resilience.
10. Tramèr et al. (2018) proposed ensemble-based adversarial training, which combines multiple models to enhance overall system security against attacks.