WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Adversarial Attacks on Machine Learning Models and their Defenses

[1] R. Selvalakshmi, [2] Dr R. Sudhamani, [3] Dr N. Ani Brown Mary

[1] II - M.Sc Computer Science, Reg. No: 24081206503912005, Sarah Tucker College, Tirunelveli – 7
[2] Assistant Professor, Department of Computer Science, Sarah Tucker College, Tirunelveli - 7
[3] Assistant Professor, Department of Computer Science, Sarah Tucker College, Tirunelveli – 7
[1] selvalakshmi0209@gmail.com, [2] sudhamani.r2003@gmail.com, [3] anibrownvimal@gmail.com

ABSTRACT

Machine Learning (ML) models are now widely used in areas such as image classification, fraud detection, and security systems. However, these models can be tricked by adversarial attacks, where the input data is slightly changed to make the model give wrong predictions. This project, titled "Adversarial Attacks on Machine Learning (ML) Models and Their Defenses," aims to show how easily ML models can be fooled and how they can be protected. The system is developed using Python and the Flask web framework to provide an easy-to-use interface. Users can upload images or input data and observe how the ML model responds normally and when attacks are applied. Popular attack methods such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are used to create modified inputs that change the model's output. Defense methods like adversarial training and input preprocessing are added to improve the model's strength. The interface shows real-time comparisons of original and attacked inputs, along with details such as accuracy and confidence scores. Users can switch between attack and defense modes to clearly understand the impact. Overall, this project stresses the importance of building secure ML systems and provides a simple platform to study adversarial attacks and defenses.

Keywords: Machine Learning, Adversarial attack, FGSM, PGD, Adversarial training, Preprocessing.

1. INTRODUCTION

Adversarial attacks have become a major concern in modern technology because they can intentionally fool machine learning systems using very small and invisible changes to input data. These tiny modifications, known as adversarial perturbations, force a model to make wrong predictions though the data looks normal to the human eye. Such attacks can create serious risks in applications like security systems, authentication, surveillance, banking, and other areas where accurate predictions are essential.

In recent years, many researchers have studied different types of adversarial attacks and how they affect the performance of ML models. Attack techniques such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are commonly used to generate manipulated data that confuses the model. These attacks expose weaknesses in ML systems and show that models can be misled easily. To address this issue, various defense strategies such as data preprocessing, input denoising and adversarial training are used to make the systems more robust and reliable.

Machine Learning

Machine Learning (ML) plays a key role in this project by learning patterns from data and making accurate predictions. Models such as Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) are trained on clean data and then tested with adversarial samples to observe changes in their accuracy. ML is also used in the defense stage, where models are retrained with cleaned or adversarial data to improve robustness. This helps in understanding both the weaknesses and strengths of ML systems and supports the development of more secure and reliable applications.

2. LITERATURE SURVEY

Machine Learning (ML) models are widely used in critical applications but are vulnerable to adversarial attacks. Small, imperceptible changes can mislead even highly accurate models, highlighting the need for robust defenses. Researchers have therefore studied attack mechanisms, model weaknesses, and effective mitigation strategies. The following studies summarize key contributions in this area.

Chakraborty et al. (2021) [1] Chakraborty and colleagues presented a detailed survey on adversarial attacks and defense strategies, analyzing gradient-based methods such as FGSM, BIM, and PGD across multiple ML models. Their findings showed that adversarial samples can easily bypass traditional classifiers, exposing serious safety issues. The study emphasized the limitations of existing defenses and highlighted the need for more reliable and robust adversarial countermeasures. Tian et al. (2022) [2] Tian et al. investigated the vulnerability of CNN-based systems under adversarial perturbations, demonstrating significant performance degradation across various tasks. Their experiments revealed that adversarial inputs can bypass detection systems with minimal modifications. The study concluded that adversarial trained models offer improved robustness compared to standard ML models. Hao and Tao (2022) [3] Hao and Tao examined the behavior of deep learning models under different adversarial attack strengths, showing that iterative methods cause greater instability than single-step attacks. Their evaluation highlighted the susceptibility of widely used ML models to even small perturbations. They recommended improved robustness testing and stronger resilience strategies for practical deployment.

Elsisi et al. (2024) [4] Elsisi and co-authors proposed adversarial-aware defense mechanisms that combine preprocessing with enhanced model training to reduce attack influence. Their results showed improved stability and accuracy when models were retrained with purified or adversarial samples. The study demonstrated that hybrid defense strategies are more effective than isolated techniques. Siniosoglou et al. (2021) [5] Siniosoglou and colleagues introduced a unified anomaly detection and classification framework capable of identifying adversarial disruptions in ML systems. Their model efficiently detected unusual behavior and maintained high accuracy in complex environments. The work emphasized the importance of integrating anomaly detection for securing ML-based applications. Berghout et al. (2023) [6] Berghout and team developed a robust data-engineering pipeline designed to mitigate the effects of adversarial manipulations. Their approach focused on filtering and normalizing input data before model inference, significantly improving system resilience. The study reinforced the need for secure data handling to protect ML models from adversarial risks.

Panda and Das (2021) [7] proposed a smart grid architecture that integrates control, optimization, and data analytics for future power networks. Their work emphasized the role of ML-based decision systems in managing renewable energy sources. The study also highlighted that vulnerabilities in data-driven models can negatively affect system reliability if not properly secured. Hao and Tao (2022) [8] analyzed adversarial attacks on deep learning models used in smart grid applications. Their experiments showed that even small adversarial perturbations can significantly degrade model performance. The authors concluded that stronger robustness evaluation and defense strategies are essential for real-world deployment. Elsisi et al. (2024) [9] introduced an IoT-enabled smart grid framework that incorporates adversarial-aware deep learning defenses. By combining preprocessing techniques with retraining, their approach improved system stability under attack conditions. The results demonstrated enhanced accuracy and resilience against adversarial inputs. Siniosoglou et al. (2021) [10] presented a unified deep learning-based anomaly detection and classification framework for smart grid environments. Their model effectively identified unusual and adversarial patterns in data streams. The study highlighted the importance of anomaly detection as an additional security layer for ML-based systems.

3. METHODOLOGY

The proposed system is designed to analyze the robustness of a Convolutional Neural Network (CNN) against adversarial attacks and to improve its stability using suitable defense techniques. The overall workflow follows a structured process starting from dataset preparation to final performance evaluation. The system architecture is divided into five main functional components, which together enable effective adversarial attack detection and mitigation. These components are listed below:

1. Dataset Loading and Preprocessing Module
2. CNN Training and Feature Learning Module
3. Adversarial Attack Generation Module
4. Defense Mechanisms and Robust Training Module
5. Prediction and Performance Evaluation Module

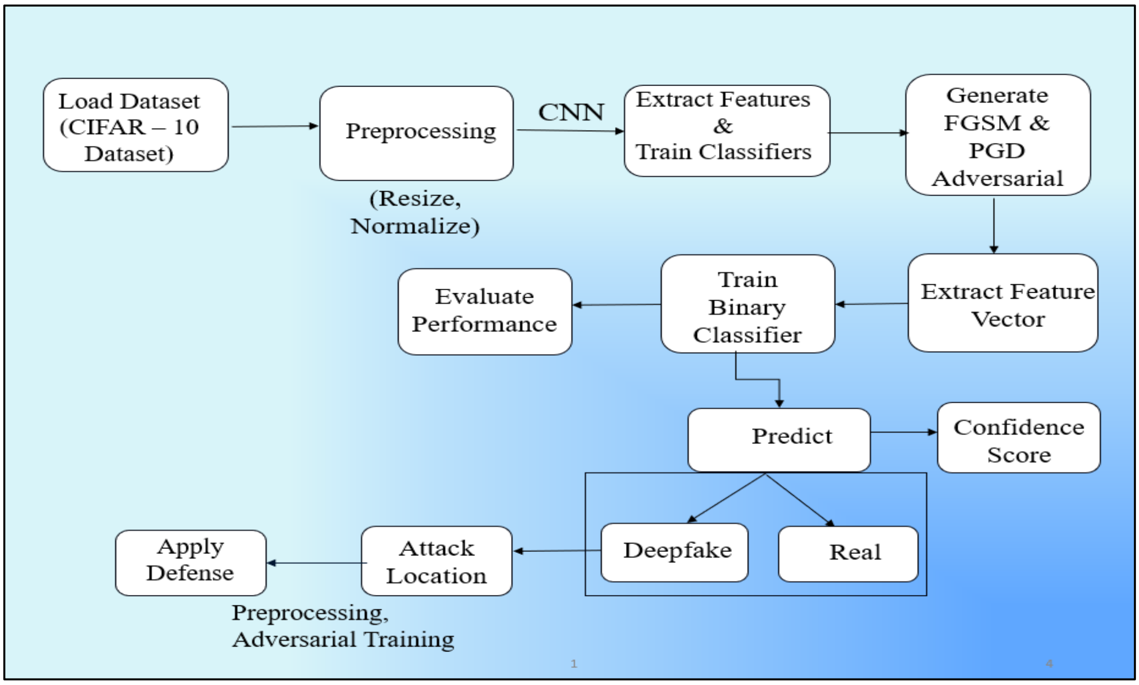

Figure 3.1: System Architecture of the Adversarial Attack on Machine Learning model

The workflow of the proposed adversarial attack and defense system is shown in Figure 3.1. The process starts with dataset loading and preprocessing of input images, followed by CNN model training. Adversarial examples are then generated using FGSM and PGD attacks to evaluate model vulnerability. Defense techniques such as adversarial training, denoising, quantization, and smoothing are applied to improve robustness. Finally, the system evaluates performance using accuracy and confidence-based metrics.

3.1 Dataset Loading and Preprocessing

In this stage, the CIFAR-10 dataset is loaded and prepared for experimentation. The dataset consists of 60,000 RGB images belonging to 10 different object classes. Out of these, 50,000 images are used for training and 10,000 images are reserved for testing. To ensure uniformity and stable learning, preprocessing operations such as resizing and normalization are applied. These steps help reduce variations in pixel values and improve convergence during model training.

3.2 CNN Training and Feature Learning

After preprocessing, the images are used to train a Convolutional Neural Network. The CNN learns hierarchical features such as edges, textures, and object shapes through multiple convolutional and pooling layers. Initially, the model is trained only on clean images to establish a baseline classification performance. This trained model serves as a reference point for analyzing the impact of adversarial attacks.

3.3 Adversarial Attack Generation

To evaluate the vulnerability of the trained CNN, adversarial examples are generated using the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These techniques add small and carefully crafted perturbations to the original images that are not easily noticeable by humans but can cause incorrect predictions. This step highlights how standard CNN models can be misled under adversarial conditions.

3.4 Defense Mechanisms and Robust Training

To improve robustness, several defense strategies are applied to the model. Adversarial training is performed by retraining the CNN using a combination of clean and adversarial images, allowing the model to learn more resilient feature representations. In addition, preprocessing-based defenses such as denoising, quantization, and smoothing are applied to reduce the effect of adversarial noise. These techniques help suppress minor perturbations and improve prediction consistency.

3.5 Prediction and Performance Evaluation

Finally, the defended CNN model is evaluated using both clean and adversarial test images. The system outputs predicted class labels along with confidence scores, indicating the reliability of each prediction. Performance is assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, and confidence score. These metrics provide a comprehensive understanding of the model's effectiveness before and after applying defense strategies.

4. EXPERIMENTAL RESULTS

The proposed adversarial attack detection system was evaluated using a Kaggle image dataset prepared for adversarial robustness studies. The dataset contains labeled images representing genuine and manipulated samples suitable for deep learning-based classification. For experimental evaluation, the dataset was divided into training and testing phases to assess the performance of the CNN model. Total 50,000 images were used for training, while 10,000 images were reserved for testing, ensuring sufficient data for learning and evaluation. The dataset selection provides balanced class representation and appropriate visual complexity, making it suitable for analyzing adversarial vulnerability and defense effectiveness.

Dataset Description

Table 4.1: Dataset Description of CIFAR-10

Dataset	**CIFAR-10**
Total Images	60,000
Training Set	50,000
Testing Set	10,000
Image Size	32×32
Image Type	RGB

The above Table 4.1 illustrates the characteristics of the CIFAR-10 dataset used in this study. The balanced distribution of images across training and testing sets enables the CNN model to learn meaningful visual patterns effectively and ensures unbiased performance evaluation during classification and adversarial robustness analysis.

The Convolutional Neural Network (CNN) was first trained on images that had undergone preprocessing steps such as normalization and resizing, which helped improve the training process and model convergence. To test the model's vulnerability, adversarial examples were created using the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These adversarial examples involve small, carefully designed changes that are almost impossible to detect but can mislead the model into making incorrect predictions, thereby highlighting the weaknesses of standard CNN models.

To improve the model's robustness, several defense techniques were implemented, such as adversarial training, denoising, quantization, and smoothing. Adversarial training involved retraining the CNN with both normal and adversarial examples to increase its resilience. Denoising was used to reduce the effects of adversarial noise, while quantization and smoothing were applied to minimize the impact of minor changes and improve the consistency of predictions

Performance Evaluation Metrics

The performance of the proposed CNN-based adversarial detection system was evaluated using Accuracy, Precision, Recall, F1-Score, and Confidence Score. These metrics provide a reliable assessment of the model's classification capability under clean, adversarial, and defended conditions. The following terms are used to compute the evaluation metrics:

- **True Positive (TP):** Number of adversarial images correctly classified as adversarial
- **True Negative (TN):** Number of clean images correctly classified as clean
- **False Positive (FP):** Clean images incorrectly classified as adversarial
- **False Negative (FN):** Adversarial images incorrectly classified as clean

Accuracy

Accuracy measures the overall correctness of the model by comparing the number of correctly classified samples to the total number of samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Precision

Precision indicates how many of the samples predicted as adversarial are actually adversarial. It reflects the reliability of positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall

Recall measures the ability of the model to correctly identify all adversarial samples present in the dataset.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score

The F1-Score provides a balanced evaluation by combining Precision and Recall, making it suitable for adversarial classification tasks.

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4}$$

Table 4.2**:** Classification Performance Metrics of the Proposed CNN Model

Metric	**Value (%)**
Accuracy	91.22
Precision	84.1
Recall	83.6
F1-Score	83.8

The results shown in Table 4.2 indicate that the proposed CNN-based system achieves high accuracy in detecting adversarial images. Balanced precision and recall values demonstrate the model’s effectiveness in correctly identifying manipulated inputs while minimizing misclassifications. The F1-Score further highlights the system’s stability and reliability in handling adversarial attacks.

Performance Evaluation Using Accuracy and Confidence

The impact of various defense techniques on the CNN model was evaluated by comparing the accuracy and confidence scores before and after applying these defenses.

Table 4.3: Performance Evaluation using Accuracy and Confidence

Condition	**Accuracy (%)**	**Confidence Score (%)**
Before Defense	88.06	60.21
After Adversarial Training	89.42	61.50
After Denoising	90.18	61.73
After Quantization	90.75	61.01
After Smoothing	91.22	62.45

The variation in model performance under different defense conditions is further illustrated using an accuracy graph, as shown in Figure 4.1. The accuracy graph illustrates the CNN model’s performance across multiple scenarios, clearly demonstrating its effectiveness in detecting adversarial attacks in real time. This visual representation highlights the improved robustness and adaptability of the system after applying defense techniques, reinforcing its suitability for practical and secure machine learning applications.

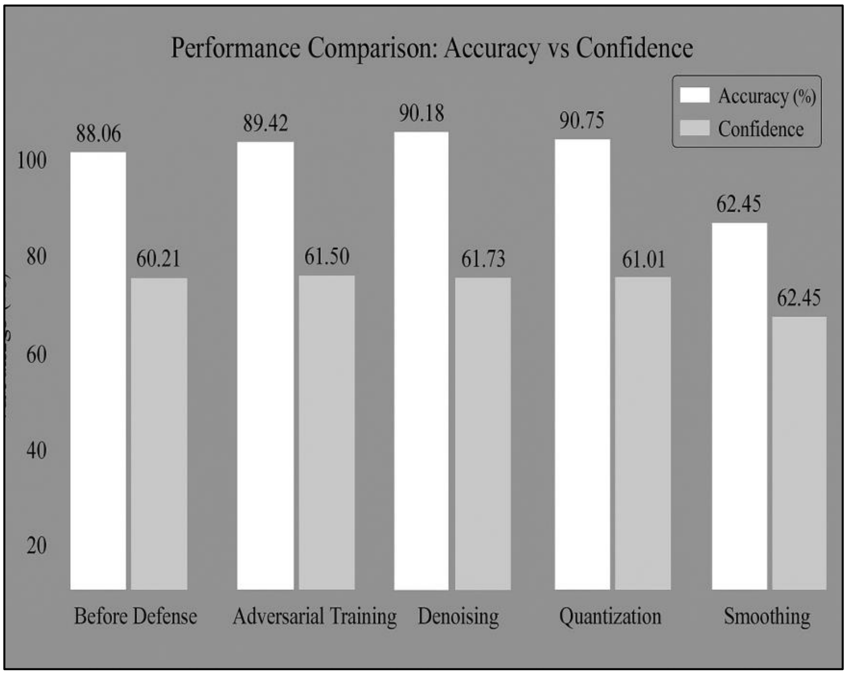

Figure 4.1: Performance Comparison – Accuracy vs Confidence

The effectiveness of each defense strategy was determined by measuring the improvements in accuracy and confidence scores relative to the model without any defenses.

Table 4.4: Improvement in Model Performance After Defenses

Defense Method	**Change in Accuracy (ΔA)**	**Change in Confidence (ΔC)**
Adversarial Training	+1.36	+1.29
Denoising	+2.12	+1.52
Quantization	+2.69	+0.80
Smoothing	+3.16	+2.24

In addition to numerical evaluations, a qualitative analysis was conducted by visually inspecting the classification outcomes under both normal and adversarial conditions. The model without any defenses showed significant misclassification when exposed to adversarial examples. After applying the proposed defense techniques, the model demonstrated better prediction consistency and higher confidence levels, successfully classifying most adversarial examples.

Overall, the experimental results show that although CNN-based image classification models are naturally vulnerable to adversarial perturbations, the use of appropriate defense strategies can significantly improve both accuracy and confidence.

The results confirm the effectiveness of the proposed framework in identifying and mitigating adversarial attacks, making it a suitable approach for secure and practical machine learning applications.

5. CONCLUSION

Machine learning models can be easily affected by adversarial attacks, where very small and barely noticeable changes in input data can cause wrong predictions. The use of Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks in a simple Flask-based system clearly shows how these attacks reduce model reliability. Applying defense methods such as adversarial training and preprocessing helps improve accuracy and prediction confidence. The proposed system connects theory with practical testing and provides an easy-to-use platform for understanding adversarial attacks and defenses, supporting the development of safer and more reliable machine learning applications.

References

[1] M. Massaoudi, H. Abu-Rub, S. S. Refaat, I. Chihi, and F. S. Oueslati, ‘‘Deep learning in smart grid technology: A review of recent advancements and future prospects,’’ IEEE Access, vol. 9, pp. 54558–54578, 2021.

[2] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, ‘‘A survey on adversarial attacks and defences,’’ CAAI Trans. Intell. Technol., vol. 6, no. 1, pp. 25–45, Mar. 2021.

[3] S. M. A. A. Abir, A. Anwar, J. Choi, and A. S. M. Kayes, ‘‘IoT-enabled smart energy grid: Applications and challenges,’’ IEEE Access, vol. 9, pp. 50961–50981, 2021.

[4] L. Phiri, ‘‘A framework for cyber security risk modeling and mitigation in smart grid communication and control systems,’’ Ph.D. dissertation, Dept. Eng., Univ. Zambia, Lusaka, Zambia, 2023.

[5] J. Tian, B. Wang, J. Li, and Z. Wang, ‘‘Adversarial attacks and defense for CNN based power quality recognition in smart grid,’’ IEEE Trans. Netw. Sci. Eng., vol. 9, no. 2, pp. 807–819, Mar. 2022.

[6] A. E. L. Rivas and T. Abrão, ‘‘Faults in smart grid systems: Monitoring, detection and classification,’’ Electr. Power Syst. Res., vol. 189, Dec. 2020, Art. no. 106602.

[7] D. K. Panda and S. Das, ‘‘Smart grid architecture model for control, opti mization and data analytics of future power networks with more renewable energy,’’ J. Cleaner Prod., vol. 301, Jun. 2021, Art. no. 126877.

[8] J. Hao and Y. Tao, ‘‘Adversarial attacks on deep learning models in smart grids,’’ Energy Rep., vol. 8, pp. 123–129, May 2022.

[9] M. Elsisi, C.-L. Su, and M. N. Ali, ‘‘Design of reliable IoT systems with deep learning to support resilient demand side management in smart grids against adversarial attacks,’’ IEEE Trans. Ind. Appl., vol. 60, no. 2, pp. 2095–2106, Mar. 2024.

[10] I. Siniosoglou, P. Radoglou-Grammatikis, G. Efstathopoulos, P. Fouliras, and P. Sarigiannidis, ‘‘A unified deep learning anomaly detection and classification approach for smart grid environments,’’ IEEE Trans. Netw. Service Manage., vol. 18, no. 2, pp. 1137–1151, Jun. 2021.