IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.25

Adversarial Machine Learning: Attacks & Defences on Deep Neutral Network in Cyber Security

P.Rengasamy
Assistant Professor,Department of CSE(Cyber Security),
Sengunthar Engineering College (Autonomous),Tiruchengode,India
prengasamy.cse@scteng.co.in
https://orcid.org/0009-0005-0537-8373
Ajay SuryaA R,Nithish K,Selvam A,Sneha C,
Department of CSE(Cyber security)
Sengunthar Engineering College (Autonomous),Tiruchengode,India
r.ajaysurya2302@gmail.com,nithishnithish3305@gmail.com,aselvam5646@gmail.com,snehanchelladurai@gmail.com

Publication History
Manuscript Reference: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10104
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10104
Received: 30, January 2026, Revised: 13, February 2026, Accepted: 28 February 2026 Published Online: 25 March 2026
https://www.irjcs.com/volumes/Vol13/iss-03/25.CSMR26.MRCS10104.pdf
Article Citation:Rengasamy,Ajay,Nithish,Selvam,Sneha(2026),Adversarial Machine Learning on Deep Neutral Network in Cyber Security,IRJCS: International Research Journal of Computer Science, Volume 13,Issue 03 of 2026 pages 242-246 **Doi:>** https://doi.org/10.26562/irjcs.2026.v1303.25 **BibTeX Key Rengasamy@2026Adversarial**
Orcid: https://orcid.org/0009-0004-9398-7488
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2026, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1303.25 The journal's official abbreviation is IRJCS.
About the License:Copyright©2026 copyright by the authors. This article is an open access and license under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Machine learning and deep neural networks have become fundamental technologies in modern cyber security systems. These systems are widely used formal ware detection, intrusion detection, spam filtering, and threat analysis. However, recent studies show that deep learning models are vulnerable to adversarial attacks that manipulate input data to deceive the model into making incorrect predictions. Adversarial Machine Learning focuses on studying these vulnerabilities and developing techniques to defend against such attacks. Attackers can slightly modify input data in away that is almost invisible to humans but causes machine learning models to misclassify the data. These attacks pose a significant risk in security-critical applications such as network intrusion detection and malware classification. In this project, we analyze different types of adversarial attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Deep Fool attacks. We also study defense mechanisms including adversarial training, defensive distillation, and input preprocessing techniques. The objective of this work is to understand how adversarial attacks affect deep neural networks and to implement defense strategies that improve the robustness of cyber security systems.
Keywords: Adversarial Machine Learning, Cyber Security, Deep Neural Networks, Adversarial Attacks, Defense Mechanisms, Intrusion Detection

I. INTRODUCTION

Cyber security systems increasingly rely on machine learning models to detect and prevent cyber attacks. Deep neural networks are capable of learning complex patterns from large datasets, making them effective for tasks such as malware detection and network intrusion detection [1]. However, recent research has revealed that deep learning models are vulnerable to adversarial examples. Adversarial examples are specially crafted inputs designed to fool machine learning models into making incorrect predictions [2]. These inputs contain small perturbations that a real most invisible to human observers but can significantly impact the model's decision-making process. In cyber security applications, adversarial attacks can have serious consequences. For example, an attacker can manipulate malware features so that a malicious file is classified as benign by the security system [3].Similarly, attacker scan modify network traffic patterns to bypass intrusion detection systems. Adversarial Machine Learning aims to study these vulnerabilities and develop defense strategies that make machine learning systems more robust against adversarial manipulation. In this project, we investigate adversarial attacks on deep neural networks and propose defense mechanisms to mitigate these threats.

LITERATURE REVIEW

Several researchers have explored adversarial machine learning and its impact on deep learning models. Good fellow et al. introduced the Fast Gradient Sign Method (FGSM), which generates adversarial examples by adding small perturbations in the direction of the gradient of the loss function [2].This method demonstrated that even well-trained neural networks can be easily fooled by carefully crafted inputs. Kurakin et al. extended the FGSM approach and developed stronger iterative attack methods such as Projected Gradient Descent (PGD) [4]. PGD is considered one of the most powerful adversarial attack methods and is commonly used to evaluate model robustness. Paper not et al. proposed defensive distillation as a method to protect neural networks against adversarial attacks by training the model with softened probability outputs [5].

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.25

This approach aims to reduce the sensitivity of neural networks to small input perturbations. Madry et al. proposed adversarial training, which involves training neural networks using adversarial examples to improve robustness [6]. This method has shown promising results in improving the resilience of machine learning models. Despite these efforts, adversarial attacks remain a major challenge in cyber security systems. Continuous research is required to develop more effective defense strategies.

ADVERSARIAL MACHINE LEARNING

Adversarial machine learning is a field of study that focuses on the interaction between machine learning models and malicious attackers who attempt to manipulate the learning process. Adversarial attacks can occur during different stages of the machine learning pipeline, including training time and testing time [7].These attacks aim to either degrade the performance of the model or cause it to produce incorrect predictions. In cyber security systems, adversarial attacks can target malware detection models, spam filters, biometric systems, and intrusion detection systems. Attackers exploit vulnerabilities in machine learning algorithms to bypass security mechanisms. Adversarial machine learning research focuses on understanding how these attacks work and developing methods to defend against them.

TYPE OF ADVERSARIAL ATTACK

Fast Gradient Sign Method (FGSM)

FGSM is one of the earliest and most widely used adversarial attack techniques. It generates adversarial examples by adding small perturbations to the input data based on the gradient of the loss function [2].

The adversarial example is generated using the following equation: $x\_adv=x+\varepsilon sign(\nabla x J(\theta,x,y))$ Where:

* X is the original input
* E is the perturbation magnitude
* J is the loss function
* Θ represents model parameters

FGSM is computationally efficient and can quickly generate adversarial samples.

Projected Gradient Descent (PGD)

PGD is an iterative attack method that repeatedly applies small perturbations to the input while keeping the perturbation within a specified limit [4]. Compared to FGSM, PGD produces stronger adversarial examples because it performs multiple iterations of gradient updates. PGD is widely used to test the robustness of machine learning models against adversarial attacks.

Deep Fool Attack

Deep Fool is an attack algorithm that iteratively perturbs input data until it crosses the decision boundary of the classifier [8]. This method generates minimal perturbations that are sufficient to cause misclassification. Deep Fool is considered highly effective in generating adversarial examples with small modifications.

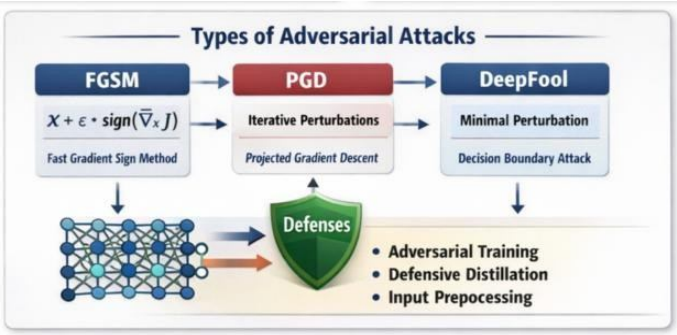

Fig.1

THREAT MODEL

In adversarial machine learning, a “threat model” defines the capabilities and goals of the attacker. Understanding the threat model is essential for designing effective defense mechanisms against adversarial attacks. There are three major components of a threat model: Attacker’s Knowledge

The attacker may have different levels of knowledge about the machine learning model:

White-box attack:

The attacker has full knowledge of the model architecture, parameters, and training data. These attacks are considered the most powerful because the attacker can compute gradients and generate effective adversarial examples [4].

Black-box attack:

In this scenario, the attacker does not have access to the internal structure of the model. Instead, the attacker queries the model and observes the outputs to create adversarial samples [5].

Gray-box attack:

The attacker has partial knowledge of the model. Attacker’s Goal. The main objectives of adversarial attacks include:

- Causing misclassification of malicious inputs
- Reducing model accuracy
- Extracting sensitive information from the model

In cyber security applications, attackers may attempt to modify malware features so that malicious files are classified as safe by detection systems [3].

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.25

Attacker's Capability

The attacker may manipulate:

- Input data
- Training data
- Model parameters

These manipulations can significantly affect the performance of deep learning models.

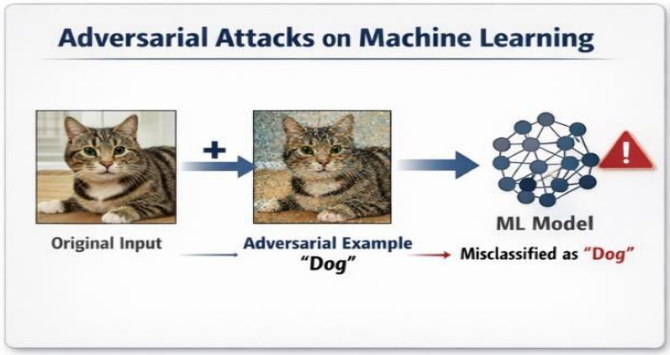

Fig.2

DEFENCE TECHNIQUE

Adversarial Training

Adversarial training is one of the most effective defense mechanisms against adversarial attacks. In this approach, adversarial examples are included during the training process to improve model robustness [6]. By exposing the model to adversarial inputs during training, the neural network learns to correctly classify both normal and adversarial data.

Defensive Distillation

Defensive distillation involves training a neural network using probability outputs generated by another network [5]. This technique reduces the sensitivity of the model to small perturbations in input data. Defensive distillation helps in smoothing the decision boundaries of neural networks.

Input Preprocessing

Input preprocessing techniques aim to remove adversarial perturbations before the data is fed into the neural network. Common preprocessing techniques include:

* Feature squeezing
* Image filtering
* Noise reduction
* Data normalization

These techniques help reduce the effectiveness of adversarial attacks.

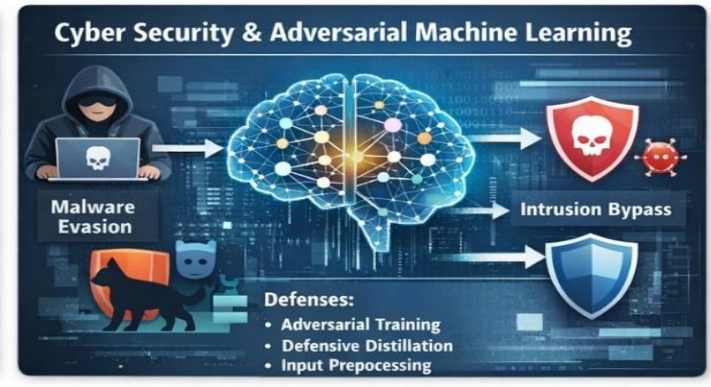

Fig.3

PROPOSED SYSTEM

The proposed system focuses on improving the robustness of deep neural networks used in cyber security applications. The system consists of the following components:

1. Data Collection
2. Data Preprocessing
3. Deep Neural Network Model
4. Adversarial Attack Generation
5. Defense Mechanism Implementation
6. Performance Evaluation

The dataset is first preprocessed to remove noise and normalize the features. A deep neural network model is then trained for cyber security classification tasks such as malware detection.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.25

Adversarial examples are generated using attack algorithms like FGSM and PGD. Defense techniques such as adversarial training are applied to improve model robustness. Finally, the performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score.

DATASET DESCRIPTION

The performance of machine learning models depends heavily on the quality of the dataset used for training and testing. In cyber security research, several publicly available datasets are used for training deep neural networks.

NSL-KDD Dataset

The NSL-KDD dataset is widely used for intrusion detection research. It contains network traffic data labeled as normal or attack.

The dataset includes several attack categories such as:

- Denial of Service (DoS)
- Probe attacks
- User to Root (U2R) attacks
- Remote to Local (R2L) attacks

This dataset is commonly used for evaluating machine learning models in network intrusion detection systems.

CICIDS Dataset

Another important dataset used in cyber security research is the CICIDS dataset. It contains realistic network traffic data that includes both benign and malicious activities.

The dataset contains multiple types of attacks including:

- Distributed Denial of Service (DDoS)
- Brute force attacks
- Botnet attacks
- Web attacks

These datasets help researchers evaluate the performance of machine learning models against cyber threats.

PERFORMANCE EVALUTION METRICS

To evaluate the effectiveness of the proposed adversarial defense system, several performance metrics are used. Accuracy Accuracy measures the percentage of correctly classified samples.

Accuracy is calculated using the formula: Accuracy=(TP+TN)/(TP+TN+FP+FN) Where: TP = True Positive TN =True Negative FP = False Positive FN=False Negative Precision Precision measure show many of the predicted positive samples are actually positive. Precision=TP/(TP+FP) Recall

Recall measures the ability of the model to correctly detect positive samples.

Recall=TP/(TP+FN) F1 Score

F1 Score is the harmonic mean of precision and recall.

F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

These metrics help evaluate the robustness of machine learning models against adversarial attacks [6].

Advantages of Adversarial Defense

The proposed adversarial defense system provides several advantages:

* Improves robustness of machine learning models
* Enhances detection accuracy of cyber threats
* Protects systems from adversarial manipulation
* Strengthens cybersecurity infrastructure

Adversarial training significantly improves the reliability of deep learning models in security- critical applications [6].

Limitations of the System

Despite the improvements in model robustness, the proposed system has certain limitations.

* Adversarial training increases computational cost.
* Some advanced attacks may still bypass defense mechanisms.
* Large datasets are required for effective training.
* Real-time detection of adversarial attacks remains challenging.

These limitations indicate that further research is required to develop more effective defense strategies.

Implementation

The implementation of the proposed system is carried out using Python and deep learning frameworks such as Tensor Flow or PyTorch.

The steps involved include:

1. Dataset collection
2. Data preprocessing
3. Model training
4. Adversarial attack generation
5. Defense mechanism implementation
6. Performance evaluation

The neural network model is trained on a cyber security dataset and tested against adversarial attacks to evaluate its robustness.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.25

Results and Discussion

The experimental results show that adversarial attacks significantly reduce the accuracy of deep neural networks. When adversarial examples are introduced, the classification accuracy of the model decreases because the input data is manipulated to fool the model. However, when defense techniques such as adversarial training are applied, the robustness of the model improves significantly. The results demonstrate that combining multiple defense techniques can help mitigate adversarial attacks and improve the security of machine learning systems.

FUTURE ENHANCEMENTS

Future research can focus on developing more advanced techniques to defend against adversarial attacks.
Some potential improvements include:
* Using *Explainable AI (XAI)* to understand model decisions.
* Implementing *reinforcement learning for adaptive defense systems*.
* Developing *real-time adversarial detection systems*.
* Applying *block chain technology for secure machine learning models*.
These improvements can significantly enhance the security and reliability of machine learning based cyber security systems.

FINAL CONCLUSION

Adversarial machine learning has become one of the most important research areas in cybersecurity. As machine learning models are increasingly used in security systems, protecting them from adversarial attacks has become a critical challenge. This project studied various adversarial attack techniques including FGSM, PGD, and Deep Fool, which can manipulate deep neural networks and cause misclassification of malicious data. Several defense mechanisms such as adversarial training, defensive distillation, and input preprocessing were implemented to improve model robustness. Experimental results demonstrate that combining multiple defense strategies significantly enhances the resilience of machine learning systems against adversarial manipulation. Future work will focus on developing more advanced and adaptive defense mechanisms to ensure the security of machine learning based cyber security systems.

REFERENCES

1. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, MIT Press, 2016.
2. Ian Goodfellow et al., “Explaining and Harnessing Adversarial Examples,” International Conference on Learning Representations, 2015.
3. Battista Biggio and Fabio Roli, “Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning,” Pattern Recognition, 2018.
4. Alexey Kurakin, Ian Goodfellow, and Samy Bengio, “Adversarial Examples in the Physical World,” 2016.
5. Nicolas Papernot et al.,“Distillation as a Defense to Adversarial Perturbations, ”IEEE Symposium on Security and Privacy, 2016.
6. Aleksander Madry et al., “Towards Deep Learning Models Resistant to Adversarial Attacks,” ICLR, 2018.
7. Huang et al., “Adversarial Machine Learning,” ACM Workshop on Artificial Intelligence and Security, 2011.
8. Moosavi-Dezfooli et al., “DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks,” CVPR, 2016.