



# **Evaluating and Enhancing the Robustness of CNN algorithm Against Adversarial Attacks: A Case Study on MNIST**

Aththanayaka A.M.R.E.  
(Reg. No.: MS22041166)

A THESIS  
SUBMITTED TO  
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY  
IN PARTIAL FULFILMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY (CYBER SECURITY)

December 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

---

Prof Anuradha Jayakody

Approved for MSc. Research Project:

---

MSc. Programme Co-ordinator, SLIIT

Approved for MSc:

---

Head of Graduate Studies, FoC, SLIIT

# DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: 

Aththanayaka A.M.R.E

Date: 14th September 2025

# ABSTRACT

## Evaluating and Enhancing the Robustness of CNN algorithm Against Adversarial Attacks: A Case Study on MNIST

Aththanayaka A.M.R.E

MSc. in Cyber Security

**Supervisor:** Prof Anuradha Jayakody

December 2025

The Convolutional Neural Networks (CNNs) have achieved exceptional performance in computer vision tasks, particularly in image classification domains such as MNIST digit recognition. However, their susceptibility to adversarial attacks poses serious security threats that limit their deployment in real-world applications. This research examines CNNs vulnerability through systematic evaluation of five potent adversarial attacks such as FGSM, BIM, PGD, Deep Fool, and Carlini-Wagner on MNIST dataset. The baseline CNN model achieves 99.23% accuracy on clean data, However, adversarial attacks which subtly perturbed inputs designed to fool classifiers cause catastrophic performance degradation, reducing accuracy to as low as 8.91%.

To address these vulnerabilities, this study proposes CADF: a Comprehensive Cyber Attack Detection Framework which implements a multi-layered defense strategy. The framework incorporates a binary detection classifier achieving 99.56% accuracy in identifying adversarial examples, followed by a multi-class attack identifier with 93.56% accuracy in categorizing specific threat types. CADF's adaptive defense engine dynamically selects optimal countermeasures including feature squeezing, spatial smoothing, and ensemble defenses based on the identified attack characteristics. Experimental results demonstrate that CADF restores model accuracy under multi-attack scenarios while maintaining high performance on clean samples and achieving real-time processing capabilities. This integrated approach provides a scalable and efficient solution for enhancing CNN robustness without compromising computational performance, offering significant advancements in securing deep learning systems against evolving adversarial threats.

Keywords— Convolutional Neural Networks (CNN), adversarial attacks, MNIST, Cascaded Adaptive Defense Framework (CADF).

# ACKNOWLEDGEMENT

I would like to sincerely thank everyone who contributed to the successful completion of this research journey.

First and foremost, I am profoundly grateful to my supervisor, Prof. Anuradha Jayakody of the Sri Lanka Institute of Information Technology, for his exceptional guidance and mentorship throughout the 2024–2025 period. His insightful advice, unwavering support, and constructive feedback were instrumental in shaping this research. His confidence in my abilities and encouragement allowed me to explore my ideas independently while benefiting from his extensive academic expertise and professional network.

I also wish to extend my gratitude to the faculty and academic staff at SLIIT for their continuous support, particularly during coursework and evaluation stages. Their critical feedback and scholarly input greatly assisted in refining my research methodology and enhancing the overall quality of this work.

Finally, I am deeply thankful to my family for their patience, understanding, and steadfast support. Their sacrifices and encouragement, especially during the most demanding phases of this project, provided me with the motivation and determination to persevere and achieve my objectives.

This research reflects the combined efforts, guidance, and inspiration of all these individuals, and I remain genuinely appreciative of their invaluable contributions.

# TABLE OF CONTENTS

DECLARATION .....	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT .....	iv
TABLE OF CONTENTS.....	v
List of Figures .....	viii
List of Tables .....	x
Chapter 1 Introduction .....	1
1.1 Background of study .....	1
1.2 Adversarial Attacks .....	2
1.2.1 Taxonomy of Adversarial Attacks.....	3
1.2.2 Evolution of Adversarial Attack and Defense Techniques. ....	4
1.2.3 Impact of Adversarial Attacks on Model Performance and Decision-Making .....	6
1.3 Problem Statement .....	6
1.4 Research Objectives .....	8
1.5 Research Questions .....	11
1.6 Significance of the Research .....	11
Chapter 2 Literature Review .....	13
2.1 Current Limitations in Existing Research vs. Novelty / Contribution .....	24
2.1.1 Research Gaps in Current Approaches .....	24
2.1.2 Novel Contributions of This Research .....	25
Chapter 3 Methodology .....	27
3.1 Methodology Overview .....	27
3.1.1 Data Collection .....	28
3.1.2 Data Collection Process .....	28
3.1.3 Data Preprocessing .....	29
3.1.4 Development of CNN Model Architecture .....	30
3.1.5 Model Compilation and Training .....	33
3.1.6 Adversarial Attacks .....	35
3.1.1 Adversarial Dataset Generation.....	41
3.2 Comprehensive Defense Approaches Against Adversarial Attacks .....	42
3.2.1 Adversarial Training.....	42
3.2.2 Defensive Distillation .....	44
3.2.3 Randomized Smoothing .....	45
3.2.4 Feature Squeezing.....	46
3.2.5 Gradient Masking .....	47
3.2.6 Spatial Smoothing.....	49

3.2.7 Ensemble Defense with Weighted Fusion.....	49
3.3 Binary and Multi-class Classifier Training .....	50
3.4 Hierarchical Defense System Design and Integration.....	55
3.5 Evaluation Metrics .....	56
3.5.1 Precision .....	56
3.5.2 Precision & Recall .....	56
3.5.3 Macro and Weighted Averages .....	58
3.6 Research Requirements .....	58
3.6.1 Functional Requirements .....	58
3.6.2 Non-Functional Requirements.....	59
3.6.3 Software Requirements.....	59
3.6.4 Hardware Requirements .....	59
Chapter 4 Results And Discussion.....	60
4.1 Baseline Model Performance .....	61
4.2 Impact of Adversarial Attacks.....	64
4.2.1 FGSM Attack.....	64
4.2.2 PGD Attack.....	65
4.2.3 C&W Attack .....	66
4.2.4 Basic Iterative Method (BIM) Attack.....	68
4.2.5 Deep Fool Attack.....	69
4.2.6 Visualization of Adversarial Perturbations of Single-Image Attacks .....	70
4.3 Evaluation of Individual Defense Strategies .....	72
4.4 Binary Classification .....	74
4.5 Multi-class Classifier.....	78
4.6 Performance of Hierarchical Defense System .....	82
4.7 Test Cases.....	88
4.7.1 Test Case for CADF Detection Accuracy .....	88
4.7.2 Test Case for Defense Mechanism .....	89
Chapter 5 Discussion & Critical Evaluation.....	91
5.1 Interpretation of Key Findings .....	91
5.2 Robustness-Accuracy-Efficiency Trade-off Analysis.....	93
5.3 Practical Implications for Safety-Critical AI Systems .....	94
5.4 Limitations and Constraints of the Current Framework.....	95
Chapter 6 Research Timeline.....	98
Chapter 7 Conclusion and Future Work .....	99
7.1 Summary of Contributions.....	99
7.2 Key Conclusions .....	101

7.3 Recommendations for Practitioners .....	102
7.4 Future Research Directions .....	106
7.4.1 Extension to Complex and Multi-Modal Datasets .....	106
7.4.2 Real-time and Efficient Defense Mechanisms .....	107
7.4.3 Integration of Explainable AI for Enhanced Trust and Diagnostics .....	108
7.4.4 Hardware-Aware and Physically Robust Defenses .....	109
Chapter 8 Bibliography .....	110
Appendix .....	114
Appendix 1: Research Publication Progress .....	114

# List of Figures

Figure 1.1:Example of Adversarial Attack .....	2
Figure 1.2:Second Example of Adversarial Attack .....	3
Figure 1.3:Formulation of Adversarial perturbation.....	3
Figure 1.4:Data Processing and Model Deployment Pipeline .....	8
Figure 1.5: Flowchart for Adversarial Example Generation .....	8
Figure 1.6: Flowchart for Adversarial training Defense .....	9
Figure 1.7: Hierarchical Adversarial Detection Framework Workflow .....	9
Figure 1.8: Hierarchical Security Framework .....	10
Figure 3.1:MNIST DATASET .....	28
Figure 3.2:Sample Images from MNIST Dataset .....	29
Figure 3.3 Sample code from model.....	30
Figure 3.4: Building a CNN Model .....	32
Figure 3.5:Model Architecture.....	33
Figure 3.6: Implementation of Callbacks.....	34
Figure 3.7: Model training progress & callbacks logs .....	35
Figure 3.8:Implementation of FGSM Attack.....	36
Figure 3.9:Implementation of PGD Attack.....	37
Figure 3.10:Implementation of Carlini Wagner Attack.....	38
Figure 3.11:Implementation of BIM Attack .....	39
Figure 3.12:Implementation of deep-fool Attack .....	40
Figure 3.13:Implementation Of Adversarial Training.....	43
Figure 3.14: Implementation of Defensive Distillation .....	45
Figure 3.15:Implementation of Randomized Smoothing.....	46
Figure 3.16: Implementation of Feature Squeezing.....	47
Figure 3.17: Implementation of Gradient Masking .....	48
Figure 3.18: Implementation of Spatial Smoothing.....	49
Figure 3.19: Ensemble Defense with Weighted Fusion.....	50
Figure 3.20: Model Architecture of binary classifier.....	51
Figure 3.21:Model Architecture of multi class Classifier.....	53
Figure 3.22:Confusion Matrix .....	57
Figure 4.1:Model Accuracy Progress Vs Model Loss Progress .....	62
Figure 4.2:Sample Predictions with Confidence Scores.....	62
Figure 4.3:Summary of Metrics.....	63
Figure 4.4: FGSM Confusion Matrix.....	64
Figure 4.5:Confusion Matrix FGSM.....	65
Figure 4.6:Confusion Matrix PGD .....	66

Figure 4.7:Confusion Matrix - C&W.....	67
Figure 4.8:Confusion Matrix – BIM Attack .....	68
Figure 4.9: Confusion Matrix – Deep Fool Attack .....	69
Figure 4.10: VISUALIZING SINGLE IMAGE ATTACKS.....	70
Figure 4.11:Model Accuracy under different attacks and its success rates .....	71
Figure 4.12:Defenses Performance Matrix .....	72
Figure 4.13:Defenses Performance against all attacks .....	73
Figure 4.14:Defenses effectiveness against all attacks and improvement over baseline.....	74
Figure 4.15: Final Metrics Summary .....	75
Figure 4.16: Confusion Matrix of Binary Classifier.....	76
Figure 4.17: Model Accuracy Progress Vs Model Loss Progress .....	77
Figure 4.18: ROC Curve of Binary Classifier .....	77
Figure 4.19 : Precision Recall Curve of Binary Classifier.....	78
Figure 4.20 : Multi Class Metrics Summary.....	79
Figure 4.21: Confusion Matrix of Multi Class Classifier .....	80
Figure 4.22:Model Accuracy Progress Vs Model Loss Progress in Multi Class Classifier .....	81
Figure 4.23 : Overall Performance Metrics & Class-Wise F1 Scores .....	82
Figure 4.24: CADF Dashboard.....	82
Figure 4.25: Performance of Clean Image .....	83
Figure 4.26: Performance of FGSM Image .....	84
Figure 4.27: Performance of BIM Image.....	85
Figure 4.28: Performance of DeepFool Image .....	87
Figure 4.29 : Clean Image Identification .....	88
Figure 4.30: Adversarial Attack Identification .....	89
Figure 4.31: Workflow for FGSM adversarial Attack Detection & Defense .....	90
Figure 4.32:Workflow for Deep-Fool adversarial Attack Detection & Defense.....	90
Figure 6.1: Research Timeline.....	98
Figure 8.1 : Paper Publication at IJACSA Journal .....	114
Figure 8.2: Paper Publication at IEEE PuneCon 2025 .....	114

# List of Tables

Table 1: Research Gap .....	26
Table 2: Test Case 1 .....	88
Table 3: Test case for Defense Mechanism .....	89