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is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources).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

List of Figures

List of Tables