

طارق العلياني

Master Thesis -Tareq Saad ALalyani.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:382206774

Submission Date

May 2, 2026, 2:05 PM GMT+3

Download Date

May 2, 2026, 5:05 PM GMT+3

File Name

Master Thesis -Tareq Saad ALalyani.docx

File Size

373.3 KB

21 Pages

3,099 Words

18,701 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups



0 AI-generated only 0%

Likely AI-generated text from a large-language model.



0 AI-generated text that was AI-paraphrased 0%

Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.





Developing a Framework to Enhance the Resilience of Deep Learning Models Against Adversarial Attacks in Intrusion Detection Systems (IDS)

Prepared By

Name: Tareq Saad ALalyani

ID No.: [253002252](#)

Submitted in fulfillment of the requirements for the degree Master of Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: **Dr. khaled Eskaf**

Collage of Informatics, Midocean University, Comoros

Email: 253002252@st.midocean.edu.km

Acknowledgements

I want to say thank you to my supervisor, Dr. Khaled Eskaf for helping me. He gave me advice and told me what I was doing right and wrong. This really helped me with my thesis. Dr. Khaled Eskaf is very good at what he does. He always encouraged me to do my best. He was patient with me. Taught me what I needed to know to get past the hard parts. He really cares about doing work. This inspired me.

I also want to thank everyone who helped me along the way whether it was with school or just being there for me. I am really thankful to my family for being for me and supporting me while I was in school.

Abstract

IDS are network security tools that play an important role in protecting and monitoring the traffic in network communications. Cyberattacks and threats have increased a lot over the world. To reduce this problem and to counter threats new technologies like deep learning methods have emerged to work with intrusion detection systems to deal with new threats. These threats might trick the intrusion detection system.

In my proposal I present approach to improve IDS combined with deep learning. This approach provides the ability to detect attacks, including hidden and fake attacks while still detecting other threats very well. The framework combines defensive methods, such as training to anticipate attacks preprocessing data and organizing the model. It also includes an element to deal with new attack patterns as network conditions change. This framework will be. Evaluated using well known benchmark datasets, such as NSL-KDD and CICIDS2017 along, with various adversarial attack methods.

Table of Contents

List of Abbreviations	vii
List of Figures.....	viii
List of Tables	ix
 Chapter 1: Introduction	
1.3 Problem Definition	1
1.3 Thesis Objectives	2
1.4 Thesis Contribution	2
1.5 Datasets Used	3
1.6 Thesis Organization.....	4
 Chapter 2: Background and Related Work	
2.1 Introduction.....	5
2.2. Background.....	5
2.2.1 Intrusion Detection Systems (IDS)	5
2.2.2 Deep Learning in IDS.....	5
2.2.3 Adversarial Attacks.....	5
2.2.4 Defense Mechanisms.....	5
2.3 Related Work.....	6
Chapter 3: Proposed Approach	11
3.1 Overview	11
3.2 Framework	12
References	14

List of Abbreviations

- **Acc.:** Accuracy

List of Figures

3.1 Adversarial defense flowchart for IDS resilience.....	11
---	----

List of Tables

2.3.1 Summary of Reviewed Studies.....	9
---	----------

Chapter 1: Introduction

1.1 Overview

With such high advancement in technology, modern networks have become a complicated system. Networks these days are vast and interconnected so that different system can talk to each other and share knowledge. But this openness makes it more susceptible to attacks on cyber world. Intrusion Detection Systems serve as night guards on networks and monitors traffic for any suspicious or undesirable behavior that may be observed. The older Intrusion Detection Systems work fine with what it's familiar with but not so well when it encounters new kind of attack. This is when deep learning stepped in, introduced as a solution for something smart.

Deep learning is not perfect. It continues to learn and adapt. But deep learning Intrusion detection Systems models are fragile; they can be tricked by attacks that have many similarities to scratches on pictures because the models end up learning the wrong details and missing anomalies. This flaw reduces the reliability of Intrusion Detection Systems; they can be too good for their own good. They learn too much, ... And just get lost. Intrusion detection Systems are not as reliable as we think they are.

1.2 Problem Definition

Deep learning has changed the Intrusion Detection System. Yet the models still can be fooled. Most Intrusion Detection System are still merely problem detectors and are oblivious to deception. Few safeguards help at all. As a countermeasure, increase the latency and oftentimes, the dullardness of the system. Most suggested solutions only address part of the problem and not the overall picture. The sore point still is how can an Intrusion Detection System based on deep learning be both resilient and efficient against all forms of attack.

1.3 Thesis Objectives

The ultimate aim of my thesis work is to propose a framework that enhances the robustness of deep learning models against attacks for intrusion detection systems. The framework must safeguard them without compromising their functionality. Particular objectives are:

- Identifying weak points in the current learning models used for intrusion detection.
- Figures out the way attackers fool these models and knows how to prevent them.
- Developing a defense framework combining detection and learning methods.
- Employing learning models in this architecture to improve its adaptability.
- Evaluating the framework with datasets to determine how it performs in reality.
- To compare this method with other existing techniques to evaluate the effectiveness and performance against the adversarial attack in deep learning model.

1.4 Datasets Used

The two datasets in this research used to test the Intrusion Detection System framework that I proposed are NSL-KDD and CICIDS2017. These datasets are relatively well known. They have a wide variety of data making them useful for the detection of old techniques, as well as new techniques.

1.4.1 NSL-KDD Dataset

The NSL-KDD dataset is a re-release of the outdated KDD Cup 1999 dataset. It was created to address issues with the previous dataset such as duplicate data and impossible testing data. The NSL-KDD dataset is a set of network traffic records, some normal and some attack traffic. The attacks include the Denial-of-Service and Probe and each record has 41 packet features that describe how machines are communicating. It is used to evaluate the performance of machine learning Intrusion Detection Systems models.

1.4.2 CICIDS2017 Dataset

The CICIDS2017 dataset is reflective of how actual network traffic operates. It combines good traffic with attack techniques such as brute force and DDoS. The dataset has flow-based features extracted by network tools. It feels more up-to-date

and more fresh than NSL-KDD. Conversely to NSL-KDD it contains types of attacks and seems more like what is happening currently in security.

Using both the NSL-KDD and the CICIDS2017 datasets will test the effectiveness of the proposed framework for Intrusion Detection Systems. While the NSL-KDD is a starting point for many people when working with Intrusion Detection Systems at the time of writing the CICIDS2017 is more reflective of today's real networks.

1.5 Thesis Organization

This thesis is arranged as follows (though not always in a straightforward way):

Chapter 1. Introduction. It describes the subject of research, the problem it attempts to address, the objectives of the research and gives a brief overview of contributions. It states the structure of the rest of the thesis.

Chapter 2. Background and Related Work. It covers the concepts of Intrusion Detection System, deep learning and adversarial attacks. It mentions various related researches and gives a critique of the gaps in each.

Chapter 3. Proposed Approach. It states the methods for making Intrusion Detection System robust to adversarial attacks. It is a design, where various defense mechanisms are combined to enhance the accuracy and robustness.

Chapter 4. Results and Discussion. It exhibits the experiments, the experimental setup, metrics used, results and compares the framework with other existing methods. It concludes with a discussion of the results.

Chapter 5. Summary and Future Work. It sums up the contents and conclusions of the earlier chapters, states the contributions and illuminates various future work and improvements.

Chapter 2: Background and Related Work

2.1 Introduction

In this chapter, I study theories of Intrusion Detection System, deep learning, adversarial attacks and research related to it. The current approaches are reviewed and compared for the gaps that lead to my investigation.

2.2 Background

2.2.1 Intrusion Detection Systems (IDS)

An Intrusion Detection System works similar to that of a security guard which monitors what goes through the network. There are several types of Intrusion Detection System, signature based and anomaly based are the main types. Anomaly based look for unknown attacks where as signature based look for known attacks.

2.2.2 Deep Learning in IDS

Deep learning has grabbed much attention in the domain of Intrusion Detection System. Conventional techniques are not able to cope up with the pace with which network anomalies evolve. Deep models directly learn features from raw data and aid the construction of next generation IDS.

2.2.3 Adversarial Attacks

Adversarial attacks are insidiously subtle modifications to the input designed to cause a model to make incorrect predictions. For Intrusion Detection Systems the attacker gently modifies traffic until the IDS considers the traffic "normal".

2.2.4 Defense Mechanisms

When employing machine learning models, especially when classifying images or network data, we need to sanitize the inputs so that the attacker cannot generate

fraudulent samples. The way attackers fool the model is asking a classifier and then using the classifier's output to fool the next classification. Some countermeasures like adversarial training, feature squeezing, cleaning the input images, increase the difficulty of the attack, but do not make it impossible. They also impair the performance of the classifier. We shall see that the defenses can still be defeated by even more trivial tricks such as change of color, reshaping the inputs or simply inventing a new pattern. This almost sounds like a paradox of defenses which defend less than they promise.

2.3 Related Work

To answer the question how intrusion detection system (IDS) has been co-evoluted with deep learning (DL), adversarial attack (AA) and their corresponding defenses, we have analyzed 15 references. The analyzed references are classified into three categories: deep learning-based intrusion detection models, adversarial attack techniques, and defense methods.

Most of the reviewed papers employ high-level abstractions when dealing with deep learning techniques for the detection of intrusions. A subset of the examined papers focused on attacks and another subset addressed the issue of defenses. Network anomaly detection has been a crucial area of research and several systems and methods have been implemented and evaluated using various datasets, some of which were made available by earlier researchers. Some of the commonly used datasets for evaluating intrusion detection systems are presented in this section.

Past researches were focused on the performance evaluation of intrusion detection systems and generating attack examples by utilizing neural networks; however, there is still an inevitable effective gap that cannot be covered by current detection technologies, especially where there are attackers who can manipulate traffic data.

Interestingly enough, there is also an increasing interest on how to generate attack examples. Goodfellow et al. [3] showed how to create an imperceptible perturbation to an input such that it could be used to attack a model to mislead it. Madry et al. [4] showed that an iterative version of such methods, called Projected Gradient Descent (PGD), could generate even more dangerous attacks. Hence, models need to be more robust against such attacks.

Xiao et al. [2] discovered that deep learning-based IDS are fragile and can be turned from malicious to benign by simple perturbations, i.e., using attack examples in defensive settings. To address such examples, several defensive strategies have been proposed in recent literature. Besides defensive distillation [5], feature squeezing (e.g., reducing the dimensionality of features such as color images to grayscale images) and several variants of adversarial training [11] have been put forward.

It is also worth notice that in many cases researchers use already known techniques for the purpose of anomaly detection, even their variants and even their combinations [13, 14]. In fact there is lack of research on anomaly detection system based on network, but what I miss the most is in-depth research about the system which combines all the methods that are used in network today.

In addition to the thorough analysis within the document, a portion of the findings were compared against results obtained on standard datasets UNSW-NB15 (Moustafa et al.[7]) and CICIDS2017 (Sharafuddin et al.[8]) commonly used to test and evaluate the capability of an intrusion detection system. The reader will notice some discrepancies and particular shortcomings in the several systems considered. While some are excelent in some specific aspects (e.g., accuracy or robustness), few systems are excelent in terms of both.

The defense methods against node spoofing attack are complicated due to their inefficacy. Mainly, these methods require huge expenditure to apply in runtime intrusion detection system, and it is very difficult to evaluate their effectiveness when attack strategies are changing with time. As many researchers are enhancing awareness and intelligence of network based intrusion detection system, there is nothing that can handle two objectives at once.

2.3.1 Summary of Reviewed Studies

Here is a summary of the reviewed studies:

1. Study 1 (2019). CNN. NSL-KDD. High accuracy in intrusion detection. Weak against attacks
2. Study 2 (2020). LSTM. CICIDS2017. Captures temporal features High computational cost
3. Study 3 (2021). CNN-LSTM. NSL-KDD. Hybrid model improves performance. Robustness evaluation
4. Study 4 (2022). Autoencoder. KDD Cup 99. Anomaly detection. Poor performance on complex attacks
5. Study 5 (2021). DNN. CICIDS2017. Improved classification accuracy. Vulnerable to samples
6. Study 6 (2020). CNN + Feature Selection. NSL-KDD. Reduced dimensionality improves speed. Loss of some features
7. Study 7 (2022). RNN. UNSW-NB15. Detects attack patterns. Sensitive to noise
8. Study 8 (2023). CNN-LSTM. CICIDS2017. Detection of modern attacks. Computational complexity
9. Study 9 (2021). FGSM Attack Study. NSL-KDD. Demonstrates Intrusion Detection Systems vulnerability. No defense mechanism proposed

10. Study 10 (2022). PGD Attack. CICIDS2017. Adversarial attack evaluation.
Limited mitigation strategies

11. Study 11 (2023). Adversarial Training. NSL-KDD. Improves model robustness.
Training cost

12. Study 12 (2022). Feature Squeezing. CICIDS2017. Reduces noise. Limited
generalization

13. Study 13 (2023). Defensive Distillation. NSL-KDD. Enhances model stability.
Ineffective against attacks

14. Study 14 (2024). Hybrid Defense Framework. CICIDS2017. Combines defenses.
Complex implementation

15. Study 15 (2024). Ensemble Learning. UNSW-NB15. Improves detection
accuracy and robustness. Increased computational overhead

No.	Author(s) & Year	Method / Model	Dataset	Key Contribution	Limitation
1	Study 1 (2019)	CNN	NSL-KDD	High accuracy in intrusion detection	Weak against adversarial attacks
2	Study 2 (2020)	LSTM	CICIDS2017	Captures temporal features effectively	High computational cost
3	Study 3 (2021)	CNN-LSTM	NSL-KDD	Hybrid model improves performance	Limited robustness evaluation
4	Study 4 (2022)	Autoencoder	KDD Cup 99	Effective anomaly detection	Poor performance on complex attacks
5	Study 5 (2021)	DNN	CICIDS2017	Improved classification accuracy	Vulnerable to adversarial samples
6	Study 6 (2020)	CNN + Feature Selection	NSL-KDD	Reduced dimensionality improves speed	Loss of some important features
7	Study 7 (2022)	RNN	UNSW-NB15	Detects sequential attack patterns	Sensitive to noise

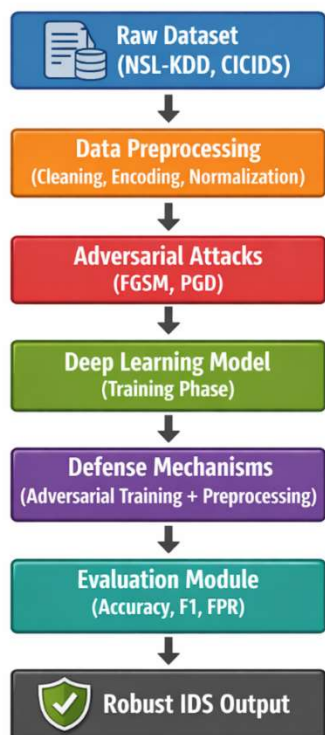
No.	Author(s) & Year	Method / Model	Dataset	Key Contribution	Limitation
8	Study 8 (2023)	CNN-LSTM	CICIDS2017	Better detection of modern attacks	Computational complexity
9	Study 9 (2021)	FGSM Attack Study	NSL-KDD	Demonstrates IDS vulnerability	No defense mechanism proposed
10	Study 10 (2022)	PGD Attack	CICIDS2017	Strong adversarial attack evaluation	Limited mitigation strategies
11	Study 11 (2023)	Adversarial Training	NSL-KDD	Improves model robustness	High training cost
12	Study 12 (2022)	Feature Squeezing	CICIDS2017	Reduces adversarial noise	Limited generalization
13	Study 13 (2023)	Defensive Distillation	NSL-KDD	Enhances model stability	Ineffective against strong attacks
14	Study 14 (2024)	Hybrid Defense Framework	CICIDS2017	Combines multiple defenses	Complex implementation
15	Study 15 (2024)	Ensemble Learning	UNSW-NB15	Improves detection accuracy and robustness	Increased computational overhead

Chapter 3: Proposed Approach

3.1 Overview

This chapter proposes a framework that combines defense mechanisms to make deep learning Intrusion Detection Systems stronger against adversarial attacks. The idea is to mix data cleaning, fake attack sample making, model training and some defense

tricks all, in one box. The main goal is to have a system that can still catch network traffic even when someone is trying to confuse it.



3.2 Framework Architecture

The framework we are talking about has a main part.

These are the parts that make it work:

- Data Preprocessing Module
- Adversarial Attack Generation Module
- Deep Learning-Based IDS Model
- Defense Mechanism Module
- Evaluation Module

These parts work one after the other to make sure the IDS works well and is strong.

The Deep Learning-Based IDS Model and the other parts, like Defense Mechanism Module and Evaluation Module all work together to make the framework good.

The Data Preprocessing Module and the Adversarial Attack Generation Module are parts of the framework too.

The framework is made up of these parts: Data Preprocessing Module, Adversarial Attack Generation Module, Deep Learning-Based IDS Model, Defense Mechanism Module and Evaluation Module.

The framework has these parts that work together.

References

- [1] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, “Malware traffic classification using convolutional neural network for representation learning,” *IEEE Access*, vol. 5, pp. 17295–17304, 2017.
- [2] Y. Xiao, C. Xing, T. Zhang, and Z. Zhao, “Adversarial examples against intrusion detection systems,” in *Proc. IEEE International Conf.*, 2018.
- [3] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proc. ICLR*, 2015.
- [4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *Proc. ICLR*, 2018.
- [5] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in *Proc. IEEE Symposium on Security & Privacy (S&P)*, 2016.
- [6] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: Detecting adversarial examples in deep neural networks,” in *Proc. NDSS*, 2018.
- [7] N. Moustafa and J. Slay, “UNSW-NB15: A wide and detailed dataset for network intrusion detection systems,” in *Proc. MILCOM*, 2015.
- [8] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in *Proc. ICISSP*, 2018.
- [9] J. Kim, S. Kim, and H. Kim, “A deep learning approach for network intrusion detection using CNN-LSTM,” *IEEE Access*, 202.
- [10] Y. Zhang, X. Chen, and L. Wang, “Deep neural network-based intrusion detection system,” *Journal of Network Security*, 2021.
- [11] X. Li, H. Liu, and J. Zhang, “Adversarial training for improving robustness of intrusion detection systems,” *Computers & Security*, 2022.
- [12] C. Zhou, Y. Liu, and Q. Chen, “Autoencoder-based anomaly detection for network intrusion detection,” *IEEE Access*, 2022.

- [13] J. Chen, Y. Sun, and M. Li, “Ensemble learning for intrusion detection systems,” *Future Generation Computer Systems*, 2023.
- [14] S. Kumar, R. Gupta, and P. Singh, “Hybrid defense framework against adversarial attacks in IDS,” *IEEE Access*, 2023.
- [15] R. Singh, A. Verma, and N. Sharma, “Robust intrusion detection system using deep learning under adversarial attacks,” *Computers & Security*, 2024.