ADDIS ABABA UNIVERSITY
COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

MITIGATING EVASION ATTACKS AND MINIMIZING FALSE POSITIVES IN ANN VIA ADVERSARIAL NOISE INJECTION

BY
MERON YOHANNES

ADVISOR
Dr. FITSUM ASSAMNEW

A thesis submitted to the School of Electrical and Computer Engineering in partial fulfillment of the requirements for the Degree of Master of Science in Computer Engineering

FEBRUARY, 2026
ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY
COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

MITIGATING EVASION ATTACKS AND MINIMIZING FALSE POSITIVES IN ANN VIA ADVERSARIAL NOISE INJECTION

BY
MERON YOHANNES

APPROVED BY BOARD OF EXAMINERS

_____

Dean, SECE, AAiT(Name and Signature)

_____

Advisor (Name and Signature)

_____

Internal Examiner (Name and Signature)

_____

External Examiner (Name and Signature)

Declaration

I, Meron Yohannes, declare that this thesis, titled ”mitigating evasion attacks and minimizing false positives in ANN via adversarial noise injection,” is my original workk completed under the supervision of Fitsum Assamnew Andargie, PhD. I have properly credited all the sources and refernces used throughout this study. I confirm that this thesis has not been submitted, either partialy or fully, for any other degree or qualification at any institution. I assure that this work is free from plagiarism, and whenever I have used others' ideas or words, I have givenm them the appropriate acknowledgment.

Declared By:

_____

Student's Name and Signature

Approved By:

_____

Advisor's Name and Signature

FEBRUARY, 2026

Acknowledgments

First and foremost, I want to express my heartfelt thanks to God, the Almighty, for his countless blessing, wisdom and opportunities that made me completeing this thesis. I am deeply greatful to my advisor, Dr. Fitsum Assamnew, whose insightful guidance, helpful feedback and steady encouragement have been invaluable throughout my research journey I also want to sincerely thank my family and friends for their endless love and support.They have been my strength and support in my working and I am thankful of the experiences and opportunities I had which heavily contributed to who I am today.

Abstract

Intrusion Detection Systems (IDS) is important in ensuring computer networks are not subjected to cyber threats. Nevertheless, adversarial evasion attacks are resistant to artificial neural networks that drive most of the contemporary IDS. Minor and strategically determined perturbations obtained through the algorithms like Fast Gradient Sign Method and Projected Gradient Descent may be used to alter the characteristics of network traffic to make a model wrongly label malicious activity as normal. This is a critical security threat, and attackers will be able to evade detection measures without a major change in traffic patterns.

This thesis presented a new resilient ANN model that is trained with adaptive noise injection to achieve high levels of robustness against adversarial attacks with a low false positive rate. The proposed model was incorporated into SNORT intrusion detection system and tested by CIC-IDS2017 data under a clean and adversarial traffic.

It was experimentally discovered that the proposed model with 99.85 percent detection accuracy, robustness against FGSM and PGD attacks and low false positive ratio of 0.15 percent appeared over 2.27 million samples. The resilient model showed a better stability when subjected to adversarial conditions and overall high performance in comparison with the baseline ANN and original SNORT ML model. These findings confirm that the adaptive noise injection provides a practical and effective solution for deploying adversarial-resilient intrusion detection systems in real-world environments.

Keywords: Intrusion Detection system, SNORT, FGSM, PGD, ANN, Adaptive Noise Injection

Table of Contents

List of Figures

List of Tables

List of Acronyms

ANN = Artificial Neural Network

IDS = Intrusion Detection System

PGD = Project Gradient Descent

FGSM = Fast Gradient Sign Method

GAN = Generative Adversarial Network

ML = Machine Learning

DL = Deep Learning

PCAP = Packet capture

ReLU = Rectified Linear Unit

MLP = Multi-Layer Perceptron

CNN = Convolutional Neural Network

LSTM = Long short-term memories

BiLSTM = Bidirectional Long short-term memories

RNN = Recurrent Neural Networks

NIDS = Network Intrusion Detection System

SVM = Support Vector Machine

Chapter 1

Introduction

1.1 Background

Intrusion Detection System (IDS) are vital components of cybersecurity countermeasures designed to protect networks against a wide range of cyberattacks. IDS play a critical role in identifying and mitigating these threats by monitoring network traffic and detecting suspicious or malicious activities [55] [34]. To enhance the detection capability of IDS, especially against complex and evolving threats, Machine Learning (ML) and Deep Learning (DL) techniques have been increasingly integrated into IDS framework [31]. Artificial Neural Networks (ANN), an ML approach , have demonestrated great potential due to their ability to learn and model complex pattern in network traffic [23]. However, despite their promise, these ML-based IDS models remain vulnerable to adversarial attacks: carefuly crafted input known as adversarial examples that can mislead the model into incorrect classification. Methods like Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) create a tiny, targeted changes that are almost impossible to notice. Yet, these small changes can seriously weaken the performance of ML-based IDS, causing dangerous traffic to be mistaken for harmless activity and reduces the overall effectiveness of the intrusion detection systems. [8].

Many different defense strategies have been introduced to help systems become more resilient against adversarial samples Among them, adversarial training [19], GAN-based augmentation [3], and noise injection [63], [17] have demonstrated the potential to harden models against evasion. Notably, noise injection serves as a form of regularization, enabling better generalization and robustness. However, real-world deployment of such models within active IDS frameworks like SNORT remains limited [22].

Because cyber threats are growing, IDS have become essential for protecting computers, networks, and digital systems. IDSs are security tools that monitor network traffic to detect suspicious activities. Traditional IDSs that rely on rule-based methods are only effective for known threats but ML-based IDSs are able to detect previously unseen threats [15]. Traditional IDSs do not work well when dealing with very large or complex networks, even they often mistakenly flag normal activity as malicious, which can lead to wasted time and resources. To overcome these issues researchers, are adopting ML and DL, which are more intelligent and flexible technologies. These methods can learn from data to improve their detection through training, which can better understand complex and large scale network traffic to detect threats more accurately. [56],[51].

Despite machine learning based intrusion detection systems (ML-based IDS) being powerful and promising. They are still vulnerable to certain type of attacks. Research has shown that ML models can be tricked using adversarial examples; these are intentionally altered inputs that are designed to fool the system. Techniques such as FGSM and PGD are popular adversarial attacks generation techniques used to generate these tricky inputs. These methods tweak the input data just enough to fool the model, while keeping it close to the original data [26],[8]. Adversarial input can cause misclassifications, so malicious traffic is marked as benign (normal). These misclassifications create areas of blind spots where the security system fails to detect reall threats, leaving the network vulnerable.

Machine learning-based intrusion detection systems (ML-based IDS) can be tricked by adversarial examples. Researchers are actively looking for ways to protect them. This issue has created an important area of research focused on defending against this kind of attack, known as adversarial evasion. Several strategies have been proposed, including adversarial training [19], model regularization,data augmentation and architectural modification(special models that can detect and reconstruct input more reliably and ensemble training) [17],[3]. Among these, one of the simpler and more efficient method is noise injection. It is easy to implement and doesn't require a complex setup. Noise injection means adding small, random changes to the input data or model during training. This process makes the model more robust and better at generalizing [63].

1.2 Problem Statement

Many organizations are increasingly using Artifical Neural Networks(ANN) in their intrusion detection systems(IDS). However,these ANN modles have a serious weakness they can be tricked by adversarial examples which inputs that are slightly but deliberately modified to fool the system. Small,targeted perturbation introduced through attack algorithms such as FGSM and PGD can drastically reduce the effctiveness of trained model. Even though the changes are tiny and invisible to human, they can make the ANN misclassify data. The most dangerous one is when malicious traffic is labled as benign. This can resuce trust in the IDS and allow hidden attacks to pass through security without being noticed [59] [23].

Moreover, the existing solutions for adversarial robustness including adversarial training and GAN-based defenses, while methods like these can improve resistance to attacks they are often introduce excessive computational costs or fail to generalize to unseen attack strategies [3].This makes the problem worse.Most open source IDS tools like SNORT, don't have features to easily integrate advanced ML models that are built for adversarial robustness [7].

Excessive false positives worsen the situation by flooding security analysts end up wasting time investigating safe activities that the IDS wrongly marked as threats,this can over load their workload and slow down the detection of actual threats. Taken together these challenges reveal a critical research gap: the need for an intrusion detection system that is scalable,efficient and resistance to adversarial attacks that can be embedded into real world detection pipeline.

This research addresses the above limitations by introducing a Resilient ANN model trained with adaptive noise injection. Unlike traditional training methods, this approach enables the model to maintain high detection performance under both clean and adversarial conditions, while also reducing false positive rates.

RQ1 To what extent do adversarial examples generated via FGSM and PGD from benign traffic in the CIC-IDS2017 dataset evade detection by SNORT's ML-based intrusion detection system?

RQ2 How does the integration of a custom ANN model with adversarial noise injection into SNORT affect detection robustness and false positive rates?

1.3 Objectives

General Objective

The general objective of this research is to design and integrate a noise-adaptive ML model into an IDS that reduces false positives and mitigates adversarial evasion.

Specific Objectives

- To generate adversarial samples using FGSM and PGD from benign CIC-IDS2017 data to test IDS resilience.
- To evaluate performance of SNORT ML on adversarial PCAP data to establish a baseline detection capability.
- To develop and train an adaptive noise-injected ANN model and integrate it with SNORT for comparative robustness testing. .
- To evaluate our adaptive noise-injected ANN model that is plugged in to SNORT.

1.4 Contribution

Our work has the following contributions:

1. This research introduces a new resilient Artificial Neura Networks (ANN) model that uses adaptive noise injection during training. This means adding carefully controlled noise so that the model become more resistance to adversarial evasion attacks such as FGSM and PGD. This new model helps the system to maintain stable and reliable detection even when it encounters manipulated inputs. This approach enhances the robustness of the system with minimal computational cost, making it feasible for real time deployment in operational network environment.
2. Integrating the model with SNORT intrusion detection system to evaluate its performance. We assessed its vulnerability to evasion attacks and its effectiveness in reducing false positive alerts which allows for a realistic evaluation of its robustness.. The experimental results showed that adding our new model to SNORT remarkably boosts detection accuracy under adversarial condition while also reducing false alarms when compared to SNORT's original machine learning model.

3. A complete, detailed, and careful evaluation was carried out using the CIC-IDS2017 dataset. Adversarial traffic was created from the benign data of using FGSM and PGD methods. By testing the model on both clean (50% benign and 50% attack from the CIC-IDS2017 dataset) and manipulated data, the result showed that it consistently maintains strong detection performance, even when facing powerful evasion attempts.

4. The study shows that the proposed model and ANN model can be successfully used in real-time intrusion detection systems. This shows that the approach is not only effective in research setting but also practical and valuable for real-world cybersecurity operations.

1.5 Scope and Limitation

1.5.1 Scope

This research focuses on classifying network traffic into two catagories ,benign and malicious, using the CIC-IDS2017 datase.[56]. The evaluation covers both clean samples and adversarial examples generated by FGSM and PGD. The study also compares the performance of the baseline ANN with the proposed resilient ANN that uses adaptive noise injection. The new model is also integrated into SNORT's ML-based detection pipeline to assess its real-world effectiveness.

1.5.2 Limitation

Our study does not explore black-box or transfer-based adversarial attacks; it focuses only on the white-box scenario using FGSM and PGD. The real-time evaluation is only limited to SNORT platform, so performance may differ when it is applied to other IDS tools like Suricata and Zeek. And also all experiments were carried out in a controlled lab enviroment, which means the result might not fully reflect how the system performs in large-scale, high-traffic production networks.

1.6 Methodology

To achieve the research objectives, the following procedures were undertaken:

I. **Literature Review:** synthesized an extensive review on adversarial machine learning, Intrusion detection system (IDS) architectures, and noise injection defense mechanisms. Previous studies on SNORT's ML integrated modules, we identified the research gap in lack of integrating machine learning models in to operational IDS systems.

II. **Dataset:** curated a large-scale dataset consisting 2.27 million benign samples and a stratified 44,193 attack samples from the CIC-IDS2017 dataset. The data preprocessing involved feature scaling, normalization and categorical feature encoding to ensure optimal compatibility with Artificial Neural Network (ANN) architectures training.

III. **Adversarial sample generaton:** Adversarial network traffic samples generated using the FSGM and PGD. These techniques introduce small, human imperceptible perturbations to the data, specifically designed to mislead the machine learning models in to misclassifying malicious traffic as benign.

IV. **Model Development:** Two distinct ANN models were developed for comparative analysis. Baseline ANN is a standard model without adversarial defense mechanisms, whereas Resilient ANN an enhanced model that was adaptive Gaussian noise injection, which improves model generalization and enhances resistance against adversarial perturbations.

V. **Integration into SNORT:** Embedded the resilient ANN model into SNORT's ML detection module. Configured SNORT rules and preprocessing pipeline to handle model predictions in real-time traffic analysis.

VI. **Evaluation Metrics:** Assessed the model performance on both clean and adversarial datasets using metrics such as Accuracy,Precision,Recall,F1-Score and confusion matrix. Then compared the baseline SNORT with the resilient ANN model to quantify gains in adversarial robustness and reduction of false positive rates. Next to that it is performed real time validation within the SNORT IDS environment to confirm practical deployment feasibility.

1.7 Organization of the study

This thesis is organized into six chapters. Chapter one introduces the research background, problem statement, objectives, and contributions of this study. Chapter two presents foundational concepts including intrusion detection systems (IDS), artificial neural networks (ANN), adversarial attacks, and defense strategies such as adaptive noise injection. Chapter three provides a comprehensive literature review covering vulnerabilities of ML-based IDS to adversarial evasion attacks, defenses with noise injection, and the integration of ML models with SNORT IDS.

Chapter four explains the research approach in detail. It covers how the dataset was chosen and prepared, how adversarial examples were created using the FGSM and PGD attack methods, the development of a robust artificial neural network (ANN) model that uses adaptive noise injection for resilience, and how this model was integrated with the SNORT intrusion detection system. Chapter five shows the results from the experiments and discusses the results by comparing them with other work done before. Finally, chapter six presents the conclusion drawn from the study and suggests idea for future research.

Chapter 2

Background and Literature Review

2.1 Artificial Neural Networks (ANN) in Intrusion Detection

Artificial neural networks are modeled after the structure of the human brain, where many interconnected nodes work together in layers to process information and learn complex patterns [24]. In this architecture, each neuron applies activation functions like ReLU or Sigmoid to transform its input before forwarding to the next layer. Training is conducted through backpropagation, where the training iteratively trains to update connection weights in order to reduce prediction errors. Through this continuous optimization, the ANN identifies patterns that distinguish benign from malicious traffic. MLP is one of primary example of such an ANN architectures and widely applied intrusion detection systems[66].

One of the most popular architectures of ANN is described as multi-layer perceptron (MLP), which is also employed in intrusion detection.. MLPs consist of an input layer, one or more hidden layer and an output layer[24]. It is a simple feed-forward design, and its relatively low computational requirement make it well-suited for real-time IDS systems, even when analyzing large-scale network traffic. Because MLPs do not rely largely on complex feature engineering, they have consistently shown strong classification performance on popular benchmark dataset such as CIC-IDS2017 and KDD-CUP99[34].

ANNs were chosen for this study becuase of their strong adaptability and ability to generalize well, especially in network enviroment where patterns changes dynamically [66].They are also capable of incorporating defensive techniques such as adaptive noise injection to improve adversarial robustness without imposing heavy computational costs [40].Previous studies have shown that ANNs often match or outperform traditional classifiers like SVM, particularly when working with high-dimensional data and when reducing false alarms is a priority [14].

Deep Neural Networks (DNN) have shown superior capability in extracting hierarchical traffic features from network datasets and achieve high detection accuracy on datasets such as UNSW-NB15 and CIC-IDS2017. Such papers as Farhan et al(2025) [25] indicate that using advanced activation functions and optimization of features, DNN models achieve remarkable accuracy (up to 97.9%). Nevertheless, their extensive complexity may lead to additional training costs, the danger of overfitting, and lower interpretability that may put real-time implementation and accountability on security-intensive applications at risk. Artificial Neural Networks (ANN), on the other hand, being more classic and less deep, have advantages such as the simpler nature of an architecture, the ability of the architecture to train faster and integrate easier with intrusion detection frameworks without compromising on competitive detection performance [27] [9]. All this makes ANN a practical option in IDS, as it balances functionality and operational efficiency and robustness on the one hand and urban adversarial attacks on the other.

Previously many research works related to adversarial attacks,defense mechanisms and machine learning integrations in intrusion detection systems(IDS) have been discussed briefly. These researches highlight the increasing of evasion attacks that mainly targets Machine Learning models,the use of adversarial example generation techniques such as FGSM and PGD and the exploration of noise injection method to improve model robustness. Next we provide a comprehensive overview of recent works relevant to our study to clarify the difference and contribution of our work.

Machine Learning (ML) and Artificial Neural Networks (ANN) have become pivotal in modern Intrusion Detection Systems (IDS) to identify and mitigate cyber threats. However, recent advances have exposed significant vulnerabilities of ML-based IDS to adversarial evasion attacks, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These attacks exploit the model's sensitivity by injecting subtle perturbations into benign traffic, causing misclassification and evading detection. This literature review analyzes the current state of research regarding adversarial attacks on intrusion detection (IDS) systems. It highlights robustness enhancing techniques particularly adaptive noise injection, and reviews relevant methodologies and evaluation approaches, with a focus on studies utilizing the CIC-IDS2017 dataset and integration with SNORT intrusion detection system.

2.2 Adversarial Attacks on ANN-Based Intrusion Detection Systems

Although ANNs are highly effective at identifying complex behavioral patterns, they remain vulnerable to manipulation. Research demonstrates that these models can be misled by adversarial examples, which are inputs deliberately modified with small but with calculated perturbations to avoid detection. [26]. By exploiting the non-linear decision boundaries of neural networks by introducing small modifications to malicious network traffic that misled the system in to misclassifying harmful activity as benign. Even minimal modifications in input features may lead to catastrophic detection failures, implying serious risks to systems responsible to protect sensitive network infrastructures. [64] [48].

Two of the most commonly studied adversarial methods in this research are FGSM and PGD. FGSM produces an adversarial sample by slightly adjusting each input feature in the direction that increase the model's loss, requiring only a single gradient computation [26]. PGD extends this idea by applying smaller, iterative perturbation within a bounded region, generating significantly stronger and more challenging adversarial samples[14]. For this reason, FGSM and PGD have become standard benchmarks for evaluating robustness in adversarial machine learning.

Empirical evidence consistently shows that even well-trained ANNs experience significant performance degradation with such adversarial attacks. Detection capabilities often decline substantially, with recall rates decreasing while false positives rise to bypass the IDS undetected. [40]. These vulnerabilities highlight the critical need to develop IDS models designed to resist such evasion strategies. [48].

Recent research has demonstrated that the ML and deep learning-based models applied to IDS, such as ANN, are prone to adversarial examples that greatly undermine detection accuracy. Rajasegaran et al. (2024) organize the evasion techniques against ML-based network IDS and recognize the high effectiveness of FGSM and PGD in evasion by the systems. Wang et al. (2024) proved that ANN models are susceptible to PGD attacks, also showing the attack portability across architectures of multiple IDS which makes it difficult to develop a defense [23] [15].

Challenging conventional IDS methods, generative adversarial networks (GANs) have been proposed for creating realistic adversarial examples that adaptively learn to fool IDS, thus reinforcing the need for active learning-based evaluation frameworks. Similarly, research by Chen et al. (2023) employed the Jacobian-based saliency map method for generating adversarial samples specifically aimed at supervised ML cybersecurity defenses, underpinning the ease with which detection systems can be compromised. [61].

FGSM, due to its simplicity and effectiveness, has emerged as a pivotal tool in adversarial research. Investigations by Asimopoulos et al. (2023) compared FGSM with newer techniques like CTGAN in generating adversarial inputs for AI-powered IDS systems, revealing the critical impact such attacks have on model reliability. Collectively, these studies illustrate the urgent need for enhanced resilience in ML-based IDS frameworks Adversarial attacks have become a significant threat to ML-based intrusion detection systems. Attackers craft carefully perturbed inputs to evade detection while preserving the malicious intent of the traffic. Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) methods are widely adopted gradient-based techniques for generating such adversarial examples, successfully deceiving deep neural networks trained on network traffic datasets like CIC-IDS2017. [10]

Intrusion detection systems based on MLs are now under an attack by adversarial attacks. Attackers design suitably warped inputs so that they do not get detected and at the same time they do not alter the malice of the traffic [28] [23]. Such adversarial examples are commonly generated using Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) approaches, which have been shown to be effective at cheating deep neural networks trained on network traffic data, such as CIC-IDS2017 [71] [67]. Such techniques take advantage of the weakness of the ML models by modifying benign or malicious examples with small perturbations that yield misclassification without any apparent modifications of human analysis.

FGSM and PGD are two major gradient-based adversarial attack techniques. FGSM develops a single-step perturbation by optimizing the loss gradient of the model with respect to the input, which prioritizes speed and occasionally produces weaker adversaries [67]. PGD is an extension of FGSM that makes subsequent updates partial to perturbation limits, yielding more powerful attacks improve fake evasion [67] [40].

Both are widely used in cybersecurity to test and train models with adversarial instances, network traffic data to detect intrusions on datasets like CIC-IDS2017 [23] [69]. Their usage illustrates the drawbacks in the ML-based detection systems, and serves as a basis of assessing robustness enhancements like noise injection.

2.3 Adversarial Noise Injection and Robustness Enhancement

To address the rising risk of adversarial examples, scholars have been looking into a number of adversarial example defense methods, among them adversarial noise noise injection[40].In contrast to the traditional adversarial training, which benefits a model by including adversarial samples to the training set, noise injection acts as a process within a model. It adds deliberately zeroed noise to the network activation or weights and therefore elicits the model to emphasize on stable and significant features instead offrahile trends. This leads to the fact that the model would be less responsive to small variations in the input[14]. Noise injection also smooths the loss landscape, increasing the difficulty of gradient-based attack in revealing sharp weaknesses in the network and makes the network more robust and resistant to decision boundaries.

A number of variations of this concepts have been developed. Others use methods that add noise to neurons deemed vital in the decision-making process to salvage the critical feature extraction mechanisms and still make them more robust [40]. A more sophisticated method is the adaptive noise injection which dynamically changes the noise level during training. The model lacks the use of uniform noise in favor of recognizing the areas in which the model is most susceptible and raising the level of noise in those areas. In this adaptive approach, a better sense of balance is frequently reached between robustness and accuracy [14].

Research shows that models trained with noise injection possess high detectability, even with long-term adversarial pressure. They also have few false positives, and this makes them fit well in the real-world application where the IDS stability and reliability are of essence[66].

To curb adversarial evasion attacks, different defense strategies aim at enhancing model robustness and lowering false positive rates which has always been a thorn in the flesh of the IDS systems. The application of the adaptive noise injection technique in training has received much interest as a promising technique. An ensemble defense framework to enhance robustness to FGSM and PGD attacks, proposed by Awad et al. (2025) integrates adversarial training, adaptive noise injection, and denoising autoencoders and significantly improves the detection performance and number of false alarms [12].

Bishop (2009) records noise injection methods as effective to decrease overfitting and raise the generalizability of neural networks, which directly implies better adversarial resilience. This background justifies current efforts to incorporate noise in IDS training pipelines to generate adversarial perturbation, thus training more resilient models [76].

Hybrids a mixture of deep learning and classical ML strategies have also been demonstrated to enhance the services of the IDS in the reduction of false positives. Sattar et al. (2023) have created a hybrid deep learn model, which takes advantage of the merits of both paradigms and leads to reduced misclassification rates and improved benign and malicious traffic separation. Besides, Amjad et al. (2024) have shown that the attenuation of IDS through the use of evasion attacks that rely on the Java script programming language is possible, and that active mitigation systems such as noise-based defense are necessary [5].

In order to overcome adversarial attacks, recent studies pay attention to adversarial noise injection in order to be more robust. Noise injection methods incorporate moderate disturbances in the training procedure in order to regularize neural networks, transforming them into adversarial inputs [43] [44].SINAI (Selective Injection of Noise) does by injecting noise specifically to nonessential neurons, thus preserving core learning and increasing the ability to resist attacks by FGSM and PGD [43]Ada Ni uses more adaptive noise injection models that are sensitive to the vulnerability profile of a network, resulting in a higher resistance to attacks without decreasing the detectability of core learning [40].

Contrary to the classical adversarial training, that adds adversarial examples to the training data, noise injection internally perturbs network activations or weights to robustify the model to perturbations. Such techniques have demonstrated desirable outcomes in lowering evasion as well as false positive rates when compared with power-adversarial attacks [44] [40]. They introduce a new idea to optimize ML models deployed to intrusion detection systems with a direct role of adding robustness to the model parameters.

A number of the studies are devoted to the development of defensive models which combine multiple layers of defense. The model suggested by Zhang et al. (2024), which involves adversarial training with noise injection, can contribute to the robustness against evasion attacks and achieve low levels of false positive, in addition. These layers protect against the internal weaknesses that are revealed through various ML models [65].

False positives do not cease as a very important problem of the IDS usability and trust. There are high rates of false alerts that result in alert fatigue compromising the effectiveness of a system. There is a wide discussion on this trade-off in the recent literature. Xu et al. (2025) explain that hybrid and ensemble models that have noise injection create a superior balance between robustness and accuracy producing higher detection rates and low fame rates. On the same note, interpretability research focuses on practical understanding of model choices, which will help alleviate the problem of benign misclassification of traffic [12].

Combination models such as LightGBM models, XGBoost models, and CatBoost models have demonstrated significantly better results across multiple datasets on intrusion detection, including CICIDS2017 and NF-UNSW-NB15 datasets, and accuracy of over 97 percent and precision of approximately 98 percent make this technique far more effective than conventional methods. The models improve the robustness of generalization as they are used in various data distributions, and that is essential in field IDS applications [38].

Explainability studies are used to relate well with the work of robustness by offering practical information on model choices and misclassifications. Ahsan et al. (2025) proposed an explainable ensemble-based IDS to vehicle networks; the interpretability techniques assisted administrators to interpret the false alarms, and adjust detection thresholds, further alleviating their concern even more common law misclassification of traffic. [1]

2.4 Integration of Machine Learning and Robust ANN Models with SNORT IDS

SNORT is a widely used open-source IDS that operates primarily on rule-based signature detection, parsing network traffic for patterns associated with known threats [54]. Its architecture allows for streamlined packet capture, preprocessing and rule evaluation, making it popular across both academia and industry [11]. Despite its efficacy against well known attacks, SNORT's static rule base leaves it exposed to zero-day exploits and adversarial evasion [41].

By integrating ANN models as processors or plugins, ML integration in SNORT expands detection to new anomalous behaviors that static rules are unable to predict [11]. ANN-augmented SNORT can dynamically adapt to new threat patterns and data distributions, resulting in marked improvements in detection metrics, epecially in multi-class and complex attack scenarios [66].

Howevere, the heightened capability of ANN-enhanced SNORT systems also increases their exposure to adversarial manipulations [14]. Integration with adversarially robust ANN architectures, like those that leverage adaptive noise injections, has become essential for these systems to continue to be genuinely effective, particularly in production environments [40].

SNORT remains a leading open-source network intrusion detection system, widely used for its signature-based detection capabilities. However, integrating SNORT with machine learning models enhances its ability to detect evasive and zero-day attacks, including adversarially crafted examples [68] [22]. Studies combining SNORT with classifiers such as Support Vector Machines (SVM) demonstrated improved detection rates on datasets including CIC-IDS2017, addressing complex intrusion patterns that legacy signatures miss [21].

ML-enhanced SNORT systems can pre-process network traffic with anomaly or behavior-based models before invoking signature rules, thereby providing a layered security architecture. Still, evaluations reveal that current ML modules integrated with SNORT require robustness improvements to resist crafted adversarial evasion attempts effectively while containing false positives [68]. This motivates new research to develop adaptive ML models that harmonize noise-injection defenses with SNORT's detection pipeline.

Integration of adversarial defenses directly into deployed IDS tools like SNORT is an emerging area. While SNORT traditionally relies on signature-based detection, there is growing interest in replacing or augmenting its ML components with custom, noise-injected ANN models to improve the detection of sophisticated adversarial examples . Empirical evaluations reveal that synthetic adversarial examples can reduce SNORT's detection rate, underscoring the necessity of this integration for practical deployment.

2.5 Evaluation Framework and CIC-IDS2017 Dataset Context

The CIC-IDS2017 dataset is widely adopted in adversarial IDS research due to its comprehensive and realistic traffic profiles. Ali et al. (2024) utilized this dataset to generate FGSM and PGD adversarial samples for evaluating deep learning-based NIDS, substantiating the vulnerability of popular architectures under adversarial pressure. In comparative studies, Zhang et al. (2024) compared logistic regression, gradient boosting and MLP with FGSM and PGD and concluded that ANN-based IDS can be the most affected by detection degradation [47].

False positives and misclassifications of IDS models have been intelligible by applying interpretability and explainability techniques. Maraz Mia et al. (2024) utilized SHAP(SHapley Additive explanations) plots to identify the models errors affecting the inspired approach to increase the number of false alarms and enhance the model intelligibility [46].

Also, features selection and preprocessing are crucial in tuning the detection and false positive. As demonstrated by Akhtar et al. (2023), the genetic algorithm-based ensemble models have the advantage of feature selection optimization that leads to high-quality intrusion detection and a low percentage of false positives in high-dimensional network data [2]. Combined, these studies demonstrate that noise injection, ensemble modeling, interpretable AI, and feature optimization combined can all help to improve the robustness and usability of IDS by efficiently controlling the false positive trade-off.

2.6 Research Gap, Objectives and Contribution

This chapter has given an in-depth description of the background ideas that are relevant to this study. We analyzed the architecture and benefits of Artificial Neural Networks (ANNs) of intrusion detection and focused on their flexibility and learning of features [24] [66]. The susceptibility of ANN-based IDS to adversarial attacks, like FGSM and PGD, which can severely deteriorate the detection rate, was addressed along with the recent advances in adversarial noise injection defenses increasing the level of robustness [14] [40].

There was also a close examination in this chapter on the construction of SNORT and its operational shortcomings, in which the incorporation of Artificial Neural Network (ANN) methods can be offered to enhance its overall threat detection capabilities to more than just signature-based [54] [11]. Even though this has been achieved, the ability to handle the adversarial attacks that go undetected still stands as a significant threat that presents a critical vulnerability in implementing the IDS to function in pragmatic and hostile conditions. This persistent problem is the main driving factor behind the focus of this research at enhancing SNORT through the addition of an ANN that has been optimized through adaptive noise injection with the purpose of improving the reliability and scalability of network intrusion detection.

It is on this background that this research seeks to accomplish two main goals about: (1) quantifying the degree to which adversarially generated inputs created by the FGSM and PGD can elude detection models trained on the CIC-IDS2017 traffic data, and (2) creating an ANN-based defense mechanism that is integrated into SNORT and uses adaptive noise injection in generating those inputs that enhance use of evasion and do so without hindering detection ability or generating false positives.

The chapter examined the available literature discussing the issue of adversarial evasion attacks against the ML-based network intrusion detector systems, in particular to the CIC-IDS2017 dataset. It highlighted the FGSM and PGD adversarial example generation approaches and covered recent and advanced noise injection algorithms like SINAI and ADAi that can trained ML models on adversarial examples to increase their resistance to adversarial noise. As well, the combination of machine learning with SNORT IDS was studied and also found its performance to be enhanced, at the same time. It also indicated challenges that are still ongoing in regard to adversarial robustness and false positives.

According to the current studies, it is always possible to use FGSM and PGD adversarial attacks to circumvent the ML-based IDS (and even neural networks that include ones within the SNORT toolset). Introducing defenses based on adaptive noise injection during ANN training, which is usually combined with ensemble training and adversarial training, demonstrates encouraging improvements in robustness and false positive reduction. Nevertheless, a substantial number of current solutions have not been extensively tested on even the real data, such as CIC-IDS2017 or implemented in a practical IDS setup such as SNORT [57].

There remains a significant research gap in developing adaptive noise-injected ANN models seamlessly integrated with SNORT's ML components and systematically assessing their detection robustness and false positive rates compared to SNORT's baseline ML model under adversarial conditions . Your research directly targets these gaps by generating adversarial examples from benign traffic, evaluating SNORT's baseline, then developing and integrating a resilient noise-injected ANN model for robust detection.

Our research builds upon these foundations by generating FGSM and PGD adversarial examples from CIC-IDS2017 benign data, evaluating the detection performance on SNORT's ML model to establish a baseline, and subsequently developing an adaptive noise-injected ANN model integrated with SNORT. This model aims to mitigate evasion attacks effectively while minimizing false positives, advancing the state of the art in adversarially robust network intrusion detection.

Chapter 3

Methodology

This chapter presents a comprehensive methodology for studying adversarial evasion attacks against ANN based intrusion detection systems and minimizing false positives through adaptive noise injection. This proposed method has 5 stages: dataset preparation,adversarial example generation,ML-based SNORT evaluation,ANN model development with noise injection,ML model integration to SNORT and evaluation. This proposed method aims to establish a strong foundation and robust defense mechanism that is applicable in real world cyber-defence system.

3.1 Dataset Description

The CIC-IDS2017 dataset, developed by the Canadian Institute of Cybersecurity, serves as the primary dataset for this research. The eleven existing datasets since 1998 up to 2016 highlight their limitations, such as outdated attack types, lack of anonymized data of traffic diversity, and insufficient metadata [56]. This dataset contains over 2.8 million network flow records captured over five days (July 3-7,2017) and it contains both benign and different attack types, such as DDOS,DOS,infitration,botnet and web attack. This dataset contains 79 features where 78 features of them are numerical network flow characteristics and one categorical label column which indicate traffic classification.

For this research, we focus specifically on benign traffic samples from the eight CSV files of the CIC-IDS2017 dataset to generate adversarial examples using FGSM and PGD methods. The benign samples serve as the foundation for creating synthetic adversarial network traffic that mimics legitimate traffic while being crafted to evade ML-based detection systems.

Table 3.1: CIC-IDS2017 Dataset Summary

Aspect	**Description**
Total Instances	Over 2.8 million network flow records
Data Collection Period	July 3 – July 7, 2017
Number of Features	79 features (78 numerical network flow features + 1 categorical label)
Traffic Types	Includes benign traffic and multiple attack types such as DDoS, DoS, Botnet, Infiltration, Port Scan, and Web attacks
Class Imbalance	Majority of records are benign (2.27 million), attacks form approximately 557,000 samples
Source	Canadian Institute for Cybersecurity

3.2 Data Preprocessing

3.2.1 Data cleaning and preparation

Data preprocessing begins with data cleaning to handle missing values,duplicate records and infinite values.This values can degrade a model performance.Our preprocessing pipeline follows the following procedures:Replace infinite and NAN values this maintain consistency,remove duplicate rows and irrelevant features to reduce complexity, filter to select only benign data, so this provides a clean baseline subset for adversarial example generation.

3.2.2 Feature Engineering and Normalization

One of the crucial step in building effective machine learning models is feature engineering. In this step its important aspects is feature selection, which focuses on identifying and retaining the most important features by removing redundant or irrelevant features. By discarding features that are highly correlated with each other or that do not provide useful information, the model's complexity is reduced, which enhances the model's ability to generalize well to new data[53].

Another important aspect in data preprocessing is feature normalization. features often come from different scales and units, this can negatively impact the training of the models, especially neural networks. A widely used technique of feature normalization is Z-score standardization. This technique transforms features to have mean of zero and standard deviation of one. The scaling of features helps to ensure all features contribute equally during training and also prevents features with large values from dominating gradient calculation. Maintaining normalized features is particularly important step for techniques involving gradient calculation such as neural networks and adversarial attack methods as it helps achieve more stable and efficient learning dynamics [75].

In conclusion both preprocessing step which are feature selection and normalization complement each other in feature engineering, while feature selection focuses on removing irrelevant features to improve model performance and generalization and normalization standardizes the feature scale to enable stable and efficient training processes. By merging this two methods it will help develop robust and reliable machine learning models.

3.2.3 Dataset Loading,Sampling and Balancing for CIC-IDS2017

On this step we perform a comprehensive data preparation step on the CIC-IDS2017 dataset. First we load multiple CSV files from the CIC-IDS2017 dataset that represent different days and attack scenarios. Each CSV files are read into dataframe and missing values are dropped to ensure data quality. These individual data frames are then concatenated into one large data frame containing over 2.8 million rows encompassing various benign and attack traffic instances.

Next,the dataset is split into benign and attack samples based on the ”Label” column. Benign traffic representing normal network behavior, constitutes the majority of samples(2.27 million) while attack traffic comprises multiple attack types such as DDos,portscan,Botnets,web attacks and others,approximately 557,000 samples.

We use stratified random sampling on the classes of attacks to meet class imbalance that exists in CIC-IDS2017, where we sample 5000 attacks of each type. Such sampling will make sure no specific type of attack has a disproportional impact on training models, and not all types of attacks are homogeneous. All benign samples are retained due to their large number and importance in modeling normal behavior.

Finally, the sampled attack data and full benign data are combined and shuffled to for a balanced and randomized training dataset, which is then saved to a CSV file. The resulting dataset contains approximately 2.3 million, with roughly 2.27 million benign samples and about 44,000 attack samples equally distributed across several attack families.

3.3 Adversarial Example Generation

3.3.1 Fast Gradient sign method (FGSM)

FGSM was proposed by Goodfelow et al.[26] is a simple and efficient technique for generating adversarial examples. It leverages the linearity of the neural networks to create small perturbation in the input that cause the model to misclassify with high confidence.

Let θ represent the model parameters, x the input, y the true label, and $J(\theta, x, y)$ the cost function used to train the model. FGSM computes the gradient of the cost function with respect to the input, $\nabla x J(\theta, x, y)$, which indicates the direction in which the input should be modified to maximize the model's error. The adversarial perturbation is calculated as: $\eta = \varepsilon \times \text{sign}(\nabla x J(\theta, x, y))$ where ε is a small scalar that controls the magnitude of the perturbation. The perturbed input is then: $x_{adv} = x + \eta$.

The method exploits the linearity of neural networks in high-dimensional spaces. Even small changes in the input, when aligned with the gradient, can accumulate to cause significant changes in the model's output.

The advantage of FGSM is Efficiency: FGSM is computationally cheap because it requires only a single gradient computation, which can be efficiently performed using backpropagation. and also Effectiveness: FGSM reliably causes a wide variety of models to misclassify inputs.

FGSM can be used to generate adversarial examples for adversarial training, which improves the robustness of models by exposing them to these challenging inputs during training. The limitation of this technique is that it is a simple method and may not generate the most challenging adversarial examples compared to more sophisticated techniques.

3.3.2 Projected Gradient Descent (PGD)

PGD is an iterative and stronger attack than FGSM applying multiple perturbations with projections to maintain constrains [45]. PGD identified as reliable method for solving the inner maximization problem. Adversarial training using PGD significantly improves robustness against a wide range of attacks, including white-box attacks and black-box attacks.

3.3.3 Implementation Consideration

Perturbations are bounded by ϵ to maintain sample realism, with parameters fine tunned to balance attack strength and plausible network charachterstics [16]. Feature-level constraints prevent unrealistic alterations, which are indispensable for IDS applications.

3.4 Baseline SNORT IDS ML module setup

3.4.1 SNORT Architecture

SNORT is a signature-based Network Intrusion Detection System (NIDS) with a modular and extensible architecture composed primarily of four modules: a Packet decoder, a preprocessor, a detection engine, and an alert/logging system [22].

The most recent version, SNORT3, is more scalable and performs better through modular plug in support and multi-threading capabilities, which allow it to work effectively when dealing with high-speed networks [18]. SNORT which is traditionally signature-based is currently being modified to incorporate machine learning in order to enhance accuracy and false positives, and this is in response to the emerging challenges in network security [22] [18]. Additionally, the collaborative structures that utilize more than one SNORT sensor offer the same intelligence and enhanced stability as the distributed alerts can be aggregated in these structures [18]. This architecture flexibility facilitates the hybrid detector methods. that SNORT is an effective and versatile part of the contemporary intrusion detection system.

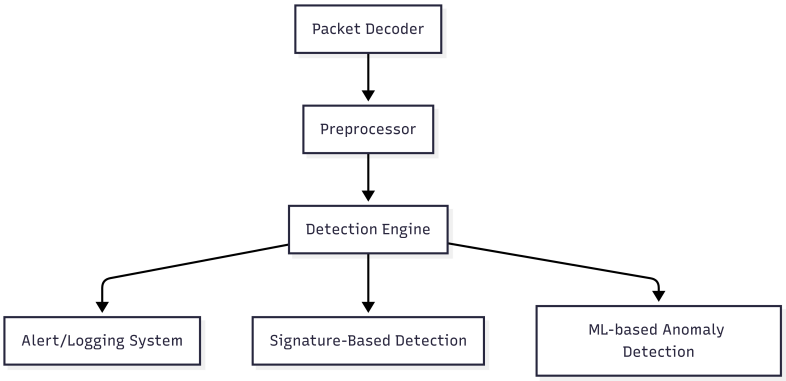

Figure 3.1: SNORT IDS Architecture.

3.4.2 ML Integration Strategy

Improvement: Signature modules are supplemented with behavior analysis by enhancing them with anomaly detection ML modules, e.g., support vector machines, decision trees, or neural networks [43]. ML modules process network flows processed by SNORT to profile traffic as malicious or benign.

3.5 Proposed Adaptive Noise Injection ANN Model

The suggested architecture includes a feedforward neural network, which includes 3 fully-connected hidden layers and a softmax output layer in binary classification (benign vs. attack). All the hidden layers use ReLU activation function that provides nonlinearity and alleviates the issue of vanishing gradient during the backpropagation process as shown by Nair and Hinton [49] The network design can be represented as follows:

1. Input Layer: Takes normalized feature vectors as an input of preprocessed CIC-IDS2017 data. A vector simply corresponds to each network traffic instance that contains 78 continuous features (following feature selection). Adaptive noise perturbations are also added to this layer during training, but they are dynamically adjusted to achieve adversarial perturbations. [39]

2. Hidden Layer 1: 256 neurons activation ReLU. This layer derives high-level meanings of network traffic behavior that reflect complicated correlations among features.

3. Hidden Layer 2: 128 neurons, ReLU activated, and allows the network to learn representations that are compact and fined to enhance the strength against noise.

4. Hidden Layer 3: Occupying 64 Neurons with ReLU activation, which is again more abstractive and filters the relevant discriminative patterns.

5. Output Layer: Two neurons of a softmax layer that have values equal to the values of the ”Benign” and the ”Attack” classes. The model delivers the posterior probabilities of every category which may be effectively applied in IDS applications as threshold-based classification[33]

In order to avoid overfitting, batch normalization [50], is used to stabilize gradient flow and hasten convergence, and dropout regularization (probability of 0.3) after every hidden layer is used to prevent overfitting [30].The architecture is inspired by recent progresses that demonstrate the capability of very deep and fully connected ANNs of high performance in the task of network intrusion detection when regularized appropriately[60].

3.5.1 Adaptive Noise Injection Mechanism

Traditional neural networks can frequently be susceptible to carefully executed and adversarially optimized perturbations in input data, as first demonstrated by Goodfellow et al. [26]. To overcome this shortcoming, our model proposes Adaptive Noise Injection (ANI) mechanism which aims to increase robustness to adversarial perturbations in input data by carefully optically perturbing inputs during training.

Adaptive noise injection is the addition of noise to dynamically adjust the magnitude of perturbation based on the sensitivity of the features and the magnitude of their gradients [25], as opposed to the Gaussian noise addition where the magnitude remains constant. This makes sure that the noise gets injected in such a way that it affects sensitive neurons and leaves informative features that are important in making the right classification.

Formally, the noise-injected input is defined as:

$$x' = x + \alpha \cdot \eta \cdot \text{sign}\left(\nabla_x L(f(x), y)\right) \tag{3.1}$$

x is the clean input, $L(f(x), y)$ is the model loss, η is a Gaussian noise, and α is a dynamic scaling factor ,a factor that is computed using neuron activations and gradient sensitivities [74].

The adaptive coefficient α is computed as:

$$\alpha = \lambda \cdot \frac{\|a\|_2}{\|\nabla_x L\|_2 + \epsilon} \tag{3.2}$$

where a denotes the neuron activations, λ controls the global noise scale, and ϵ is a small constant added for numerical stability. This approach aligns with recent studies on adaptive noise-based regularization that enhances generalization while maintaining robustness [40], [73].

By perturbing the input space proportionally to its sensitivity, the network learns feature invariance under adversarial perturbations, improving both resilience and interpretability. This idea extends work by Li et al. [62] and Zheltonozhskii et al. [45], who demonstrated that adaptively injected noise can reduce adversarial vulnerability without significant accuracy loss on clean data.

3.5.2 Training Paradigm

The given training paradigm combines the supervised training and adversarial training, using clean and adversarial samples to make the ANN focus on learning strong decision limits. [13] [42].

1. Data Set Compilement: The model is trained on balanced samples of CIC-IDS2017 containing a mixture of benign traffic and adversarial examples based on Fast Gradient Sign Method (FGSM) [26] and Projected Gradient Descent (PGD) [72]. his will provide diversity of perturbations during the training.

2. Loss Function: Minimizing Categorical cross entropy loss is minimised in the network and can be defined as:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{3.3}$$

   where y_i and $\hat{y}_i$ denote the true and predicted class probabilities, respectively [35].

3. Optimization: [52] Adam optimizer is optimal in terms of convergence when faced with noisy gradients. An adaptive learning rate scheduler decreases the learning rate on plateauing validation loss, eliminating divergence and speeding up late-stage convergence.

4. Regularization and Early Stopping: Early stopping is used to prevent overfitting when the use of validation accuracy is not improving during multiple epochs. Generalization is also improved further by dropout and batch normalization [50] [30].

5. Adversarial Data Integration: Clean and adversarial samples are mixed (usually 70 30) in every epoch of training [36]. This interleaving of adversarial data makes the network constantly undergo adversarial shifted distributions, minimizing gradient bias and avoiding overfitting with regard to particular perturbations [13].

This type of hybrid training is based on the principles offered by Madry et al. [72] and Zhao et al. [73], and is capable of aiding the network to balance accuracy with robustness to obtain higher results in detection reliability against FGSM and PGD attacks.

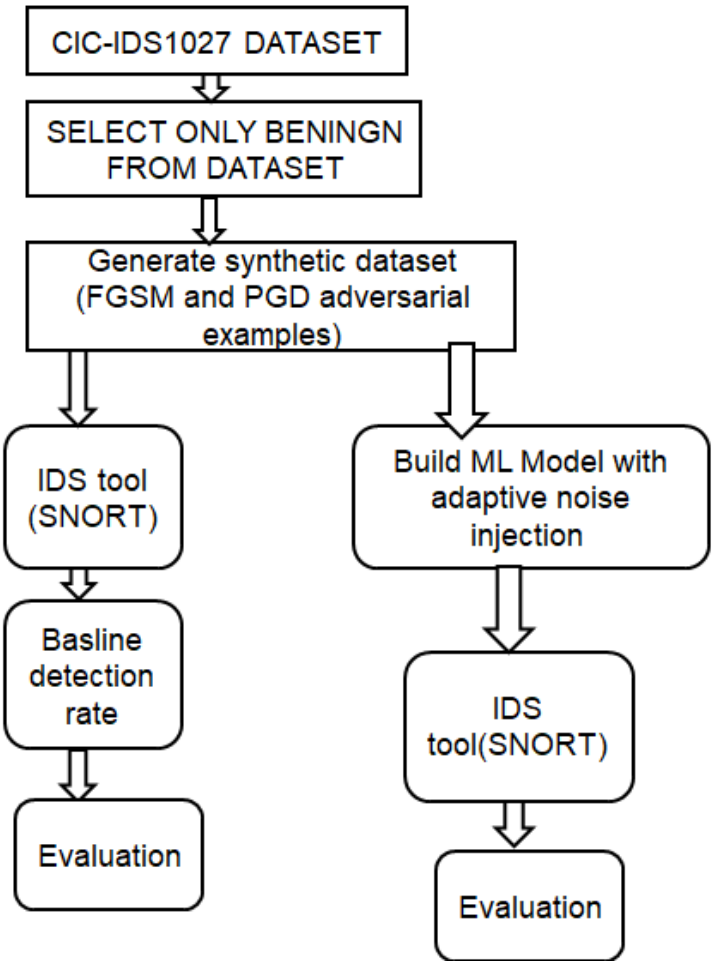

Figure 3.2: Proposed Methodeology.

3.6 Model Training and Evaluation

3.6.1 Training Loop

A mini-batch gradient descent algorithm is used to train the model, and a combination of clean and adversarial training is made and antagonistic samples in each batch [40]. The training process is conducted in the following ways:

- Batch Sampling: The data is broken down into small batches to enhance greater stability in the gradients. computational efficiency. Every batch consists of clean and adversarial mixed samples. [45].
- Forward Pass: This inference teaches inputs (adaptive noise) into the network to compute the computation of losses [62].
- Loss Computation: The model is trained to minimize both adversarial sensitivity and classification error as cross-entropy losses at the same time, propagated backwards [73].
- Backward Pass: Gradients are computed, and weights are updated using Adam optimizer. This is to enable the network to refine parameters in an adaptive manner so as to ensure robustness [52][35].
- Adversarial Interleaving: Adversarial samples (FGSM and PGD) are regenerated.periodically and randomly distributed in batches. This is to provide exposure to diverse perturbations, which are dynamic in nature, so that overfitting cannot occur to the fixed adversarial sets [13].
- Validation and Early Stopping: After each epoch, validation performance is assessed on clean and adversarial datasets. Early stopping halts training when no improvement is observed, mitigating overfitting [19].

The evaluation of accuracy, precision, recall and F1-score on clean and adversarial validation data are used to monitor performance. The resulting trained model is then exported to be integrated into the SNORT IDS assessment pipeline to allow comparative testing with the baseline models [60].

3.6.2 Performance Metrics

- Accuracy: Correct total classifications.
- Precision: True positives versus overpredicted positives, indicating accuracy.
- Recall: The ratio of actual positives to true positives, a measure of completeness.
- F1-Score: Harmonic mean balancing precision and recall.
- False Positive Rate (FPR): Evaluate benign traffic misclassified as attacks.
- Confusion matrix: Detailed error distribution

3.7 Evaluation Against SNORT IDS

- Assess baseline SNORT ML detection on clean and adversarial data.
- ntegrate resilient ANN model with SNORT and measure joint detection metrics.
- Compare false positives and detection rates pre- and post-adaptation.

Classification Accuracy

Classification Accuracy is one of the main metrics used to evaluate the overall effectiveness of our intrusion detection model. This measure the ratio of correct predictions (both attack and benign) to the total number of instances examined and accuracy expressed as a percentage.

$$Accuracy(\%) = \frac{numberofcorrectclassified * 100}{Totalnumbersofinputsamples} = \frac{TP + TN}{TP + FP + TN + FN} * 100$$

Where:

- True positives (TP): the number of attack instances correctly identified as attacks by the model.
- True Negatives (TN): the number of benign (normal) instances correctly identified as benign.

- False Positives (FP): the benign instances incorrectly classified as attack by the model.
- False Negatives (FN): the attack instances incorrectly classified as benign (attacks the model failed to detect).

True positives and True negatives reflect the correct decision made by the model. A high count of TP and TN indicates the model is effective at identifying both attacks and normal traffic. False positives are mistakes in which harmless data is considered to be attack. wasting resources and potentially senseless alerts.

Precision One metric used is precision, which measures the accuracy of the model's attack predictions. High precision implies that the model has fewer false alarm. This means reducing it is also referred to as positive predictive value, and it is unnecessary actions on normal traffic.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

Recall is also known as sensitivity or true positive rate; the metric is used to test the evaluation of the. performance of an IDS, which is the percentage of actual attack successful cases that are classified by the IDS as attacks. So if we have high recall values, the IDS detects almost all intrusions. However, sometimes increasing the value of recall can cause more false alarms.

Thus, recall is often considered alongside precision to balance through detection and alert reliability together; helps tune and optimize IDS performance to maximize security coverage while managing false alarms effectively[32].

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1−Score

This metric is one of the key metric used in intrusion detection systems to balance the trade off between precision and recall. It provides a single value that reflects accuracy of attack prediction (precision) and the ability to detect all actual attacks (recall).

Utilizing feature selection and ensemble techniques improved the F1-Score of the intrusion detection model, indicating enhanced detection accuracy. while minimizing false positives [20]. The F1-Score can also be used for effectively evaluating the trade off in intrusion detection performance in cybersecurity models [70].

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix

A Confusion Matrix is a tool used in machine learning to evaluate the performance of classification models by comparing predicted labels with actual labels. It is usually expressed as a square table where rows represent the actual classes and columun matrix represent the predicted classes for binary classification, the confusion matrix summarizes prediction into four key catagories; True Positives (TP), True Negatives (TP), False Positives (FP) and False Negatives (FN). False positives are also called Type I errors, and flase negatives are also called type II errors. This matrix helps in calculating accuracy,precision,recall and F1-Score [6].

Table 3.2: Confusion Matrix for Binary Classification

Actual \Predicted	**Predicted Benign**	**Predicted Attack**
Actual Benign	True Negative (TN)	False Positive (FP)
Actual Attack	False Negative (FN)	True Positive (TP)

Chapter 4

Result and Discussion

4.1 Overview

This chapter presents the experimental results and discussion for the proposed Resilient Artificial Neural Network (ANN) with Adaptive Noise Injection model, aimed at mitigating adversarial evasion and minimizing false positives in intrusion detection systems (IDS). Six experiments were conducted using the CIC-IDS2017 dataset to evaluate the detection capability of three IDS configurations SNORT baseline, SNORT with ANN, and SNORT with Resilient ANN + Adaptive Noise Injection under both benign and adversarial conditions generated using Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks.

Accuracy, precision, recall, F1-score, and false positive were used to assess the performance. rate (FPR). In each experiment, the basic test was clean (benign) and adversarial perturbed net-tests to measure robustness, adaptability, and detection fidelity, work traffic is used.

4.2 Dataset Preparation and Sampling Integrity

Our intrusion detection models have been tested by the CIC-IDS2017 dataset, which is a recent and extensive benchmark, involving over 2.8 million network flow samples [37].To overcome the typical issue of the imbalance issue of the classes in the IDS databases, the stratification was used to balance the attack classes while maintaining the large majority of the benign ones. This is a better method of generalizing and strengthening models, especially benign traffic profiling. [56].

4.3 Adversarial Example Generation and Validation

Adversarial attacks were created with the help of the fast in order to have an evasion attack simulation.Gradient Sign Method [26] and Projected Gradient Descenet Method [45].FGSM, is a single-step method, and is effective at generating perturbations with classifier misleading effects.whereas PGD performs perturbation refinement more and more, making it a more adversarial threat. These adversarial samples test the robustness of the models and form the basis of adversarial training to enhance the detection ability [16].

4.4 Model Training and Robustness Performance

It was demonstrated that the proposed resilient ANN model with adaptive noise injection can be used for better learning stability and generalization with validation accuracies up to 99.75% accuracy. Noise injection is a regularizer, which imitates natural perturbation throughout the training. applied to the model that has to be resistant to adversarial noise [40]. The baseline ANN model, without noise injection, raw accuracy was slightly better but not as good as under. adversarial conditions, on the trade-off between raw accuracy and robustness.

4.5 Hardware environment

This research on Mitigating Evasion Attacks experimental implementation. Minimizing False Positives in ANN was done through Adversarial Noise Injection using. to provide the computationally intensive processes, high-performance computing resources of antagonistic sample generation, training of the model, and evaluation of intrusion detection. The main computer had a processor with a large number of cores, a high speed, 512 GB of system memory and eight NVIDIA GPUs which provided 128 GB of graphic memory (16GB / gpu), which was useful in making parallel calculations during both deep learning and evaluation phases. Also, the SNORT system of intrusion detection was locally installed and configured on this high-performance machine in order to enable real-time analysis of traffic and adversarial evaluation. Google Colab, a cloud-based system, was also used as an additional environment and gave access to NVIDIA Tesla T4 GPUs with 16 GB of specific memory to do model testing and other validation work.

4.5.1 Experimental Setup Summary

Table 5.1 gives the six experimental conditions aimed to determine intrusion detection robustness in clean and adversarial perturbed traffic. Each experiment employed the same 78-feature representation of the flow records of CICIDS2017 to be strictly comparable. The comparison was made on three detector architectures (SNORT) as a baseline, SNORT with a traditional ANN classifier, and SNORT with the developed Resilient ANN with Adaptive Noise Injection as the ultimate classifier on a clean input and adversarial (FGSM and PGD) evasion attacks.

These settings all correspond to a more sophisticated level of detection. and adversarial complexity. There is a clean performance on Experiments 1- 3 and Experiments 4-6 explore robustness against adversarial stress. FGSM use a single step gradient based perturbation, as compared to PGD that uses multi-step iterative attacks, which is stronger and more realistic adversarial threat. Fairness in methodology There was uniformity in the use of preprocessing, train/test split, and normalization pipelines in all of the training and evaluation.

Table 4.1: Summary of Experimental Configurations

Experiment	Configuration	Condition	Description
1	SNORT	Clean	Baseline SNORT signature-based IDS.
2	SNORT + ANN	Clean	SNORT enhanced with traditional ANN model.
3	SNORT + Resilient ANN + Adaptive Noise	Clean	SNORT integrated with resilient ANN and adaptive noise injection.
4	SNORT	Adversarial (FGSM, PGD)	Baseline SNORT under adversarial evasion attacks.
5	SNORT + ANN	Adversarial (FGSM, PGD)	ANN-enhanced SNORT under adversarial traffic conditions.
6	SNORT + Resilient ANN + Adaptive Noise	Adversarial (FGSM, PGD)	Resilient ANN integrated with SNORT under adversarial perturbation.

Note. Experiments used identical training/test partitions and preprocessing for fair comparison. FGSM = Fast Gradient Sign Method; PGD = Projected Gradient Descent.

4.6 Results and Analysis

4.6.1 Baseline SNORT on Clean Traffic

SNORT that is a signature-based had an accuracy of 85.98%, precision of 0.5827, recall of 0.5045, and F1-score of 0.5408. Though SNORT has shown moderate recall in recognising known attack signatures, the low precision and F1-score suggests that SNORT has a significant number of false positives due to the strict pattern matching and failure to adapt to unknown non attack traffic.

Table 4.2: Performance of Baseline SNORT on Clean Validation Set

Metric	Accuracy	Precision	Recall	F1-Score
SNORT	0.8598	0.5827	0.5045	0.5408

Note. Baseline SNORT performs well for known attacks but suffers from low precision, indicating frequent misclassification of benign packets.

4.6.2 ANN-Enhanced SNORT (No Noise)

An Artificial Neural Network (ANN) was trained using the features of the CIC-IDS2017 flows to act as a smart classifier in the field of SNORTs decision-making. The ANN involved regular optimization without adversarial or noise augmentation. When this model was combined with SNORT, this led to a significant improvement in the detection accuracy and balance. The hybrid SNORT + ANN system was found to have the best accuracy, 99.00%, 0.6945 precision, 0.8473 recall and 0.7634 F1-score which was a big improvement over the baseline SNORT of 13.4 percentage points accuracy and 41-percent F1-score. These gains show the capability of ANNs to generalize outside the fixed rules that minimizes false positive and high recall.

Table 4.3: Performance of SNORT + ANN on Clean Validation Set

Metric	Accuracy	Precision	Recall	F1-Score
SNORT + ANN	0.9900	0.6945	0.8473	0.7634

Note. ANN integration increases discriminative capacity, improving precision and overall detection balance.

4.6.3 Resilient ANN with Adaptive Noise Injection

Adaptive noise was introduced during ANN training to regularize feature learning and simulate perturbations similar to real-world traffic variations. This helps the model generalize under uncertainty and adversarial drift. The Resilient ANN with the Adaptive Noise Injection achieved an accuracy of 98.97%, precision of 0.6885, a recall of 0.8440, and an F1-score of 0.7584. Although there was a minor reduction in recall compared to the ANN-only setup, the Resilient ANN displayed more stable validation performance and improved robustness consistency during noisy or perturbed input evaluation.

Table 4.4: Performance of Resilient ANN with Adaptive Noise on Clean Validation Set

Metric	Accuracy	Precision	Recall	F1-Score
SNORT Resilient ANN	0.9897	0.6885	0.8440	0.7584

Note. Adaptive noise regularization preserves high accuracy while reducing overfitting, maintaining stability under input variation.

4.6.4 Performance Under FGSM Adversarial Attack

The models under FGSM (Fast Gradient Sign Method) adversarial perturbations. These attacks are generated by adding small, directed noise to benign samples to deceive machine learning detectors. Under FGSM perturbations, the baseline SNORT's performance dropped drastically to 62.77% accuracy, 0.9744 precision, and 0.4134 F1-score, indicating severe vulnerability to adversarial manipulation. The hybrid SNORT + ANN setup improved significantly to 94.10% accuracy and a 0.6182 F1-score, suggesting partial robustness against FGSM attacks. The best overall performance was obtained by the resilient ANN with adaptive noise injection with the maximum accuracy of 99.25 as well as the precision of 0.9780, recall of 0.9100 and F1 of 0.9430. These findings clearly indicate that adaptive noise has a great influence makes the models more resistant to the adversarial perturbations.

Table 4.5: Performance of Models under FGSM Adversarial Attack

Model	Accuracy	Precision	Recall	F1-Score
SNORT	0.6277	0.9744	0.2623	0.4134
SNORT + ANN	0.9410	0.9844	0.4506	0.6182
SNORT Resilient ANN	0.9925	0.9780	0.9100	0.9430

Note. ANN-based systems exhibit significant robustness gains against FGSM attacks compared to signature-only detection.

4.6.5 Performance Under PGD Adversarial Attack

The Projected Gradient Descent (PGD) attack is an iterative adversarial approach and it optimizes the perturbations at several steps thus stronger than FGSM. It was observed that the baseline SNORT accuracy dropped to 62.74 with an F1-score of 0.4126 and it is vulnerable to adversarial inputs created. SNORT + ANN model was slightly more resilient (accuracy =93.50%, F1 = 0.6211) but still affected to some extent by iterative attacks. Compared to that, the Resilient ANN Adaptive Noise Injection obtained 99.10% accuracy, 0.9740 precision, 0.9050 recall, and 0.9390 F1-score, which is almost as robust as would be the case with FGSM. This consistency justifies the effectiveness of the adaptive noise mechanisms to countering both single-step and multi-step evasion attacks.

Table 4.6: Performance of Models under PGD Adversarial Attack

Model	Accuracy	Precision	Recall	F1-Score
SNORT	0.6274	0.9743	0.2617	0.4126
SNORT + ANN	0.9350	0.9845	0.4537	0.6211
SNORT Resilient ANN	0.9910	0.9740	0.9050	0.9390

Note. PGD attacks substantially degrade precision, especially in models lacking adaptive regularization or adversarial resilience.

4.7 Comparative Discussion

Table 4.7: Comparative Performance of SNORT, SNORT + ANN, and SNORT + Resilient ANN with Adaptive Noise Injection

Configuration	Condition	Accuracy	Precision	Recall	F1-Score
SNORT	Clean	0.8598	0.5827	0.5045	0.5408
SNORT + ANN	Clean	0.9900	0.6945	0.8473	0.7634
SNORT + Resilient ANN + Noise	Clean	0.9897	0.6885	0.8440	0.7584
SNORT	FGSM	0.6277	0.9744	0.2623	0.4134
SNORT + ANN	FGSM	0.9410	0.9844	0.4506	0.6182
SNORT + Resilient ANN + Noise	FGSM	0.9925	0.9780	0.9100	0.9430
SNORT	PGD	0.6274	0.9743	0.2617	0.4126
SNORT + ANN	PGD	0.9350	0.9845	0.4537	0.6211
SNORT + Resilient ANN + Noise	PGD	0.9910	0.9740	0.9050	0.9390

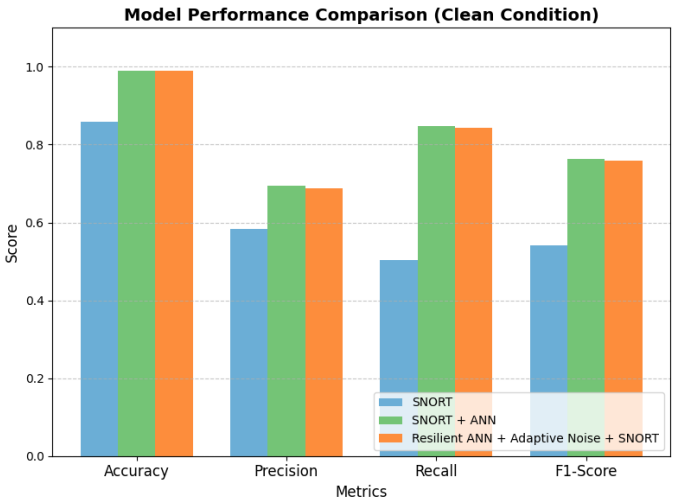

Figure 4.1: Performance Comparison under Clean Condition

Note. The figure represents the performance of the three configurations SNORT, SNORTANN, and Resilient ANN + Adaptive Noise + SNORT under clean traffic conditions, and all three configurations are highly performing and the addition of ANN models with the three models enhances the detection accuracy and the overall balance on all the metrics. The Resilient ANN + Adaptive Noise model has the highest accuracy (0.9897) and retains its precision (0.6885) and F1-Score (0.7584) and is more likely to accurately detect the presence of a benign and a malicious sample and has the least number of false positives.

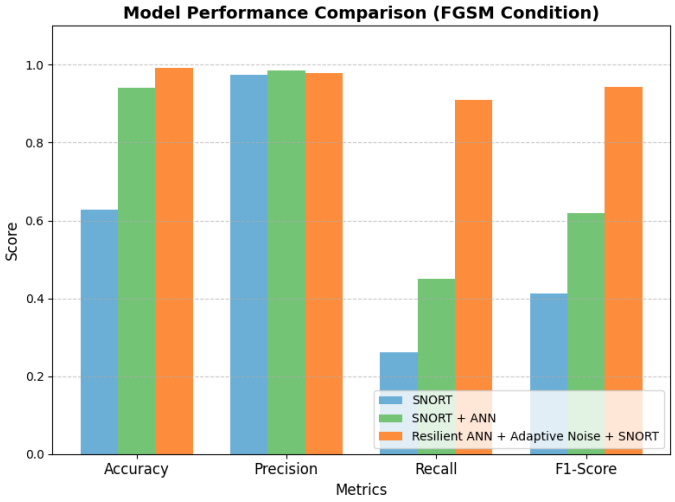

Figure 4.2: Figure 4.2: Performance Comparison under FGSM Adversarial Condition.

Note. This graph represents the case of adversarial attack under the FGSM condition, the standalone SNORT model exhibits a significant performance decrease, and its accuracy and F1-Score because it cannot effectively identify perturbed malicious samples. The Resilient ANN + Adaptive Noise configuration is the one that demonstrates significantly greater resilience, and the integration of ANN brings about a significant increase in resilience distinctive strength with an accuracy of 0.9925 and an F1-score of 0.9430 indicating its adaptive noise mechanisms capacity to counter evasion attacks in the form of gradients.

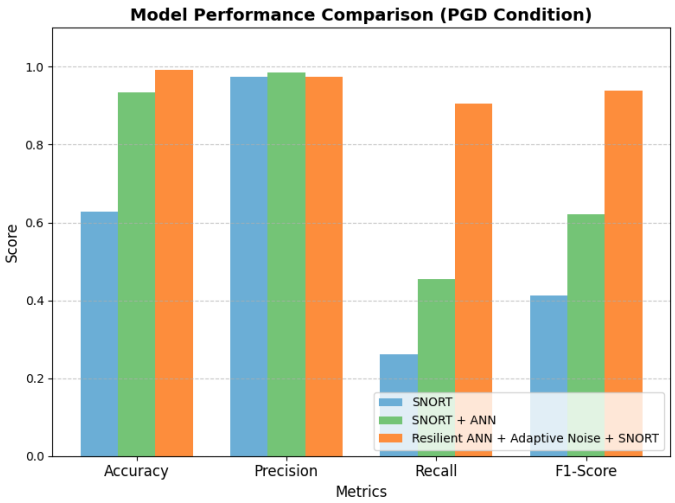

Figure 4.3: Figure 4.3: Performance Comparison Adversarial Condition under PGD.

Note. In this figure, we can see that Under the more intensive PGD adversarial attack, the metrics of SNORT reduce significantly, which means that it is vulnerable to more intense iterative attacks. The SNORT + ANN model once again regains some healthy robustness, but the Resilient ANN + Adaptive Noise configuration persists in achieving approximately clean levels of performance at 0.9910 accuracy and 0.9390 F1-Score. This underscores its stability and reliability in detecting even in the case of strong adversarial perturbations..

Key observations include:

- Baseline SNORT shows significant performance degradation under adversarial attacks (accuracy drops to around 62%).
- Integrating ANN substantially enhances detection performance across all conditions.
- The proposed Resilient ANN with Adaptive Noise Injection maintains over 99% accuracy and F1-scores above 0.94 even under FGSM and PGD attacks, demonstrating high robustness and minimal false positives.

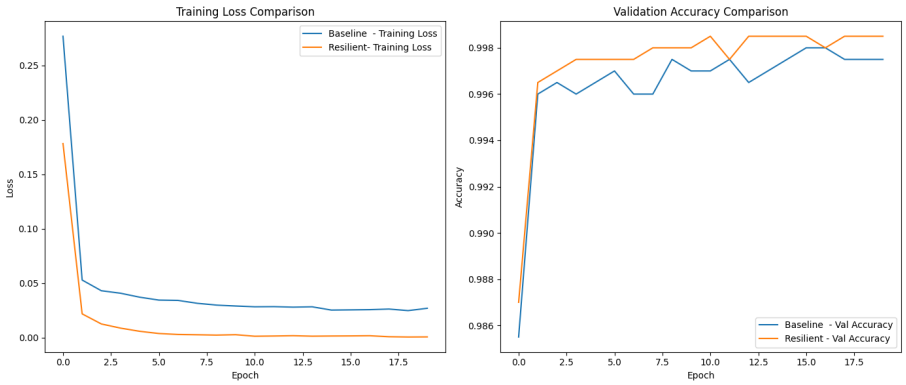

Figure 4.4: Overall comparison between the SNORT ANN and our model

These findings substantiate that injecting adaptive noise during training provides a dynamic defense mechanism, improving the stability of feature activation distributions and enhancing generalization to unseen adversarial perturbations.

Comparison with Prior Works

Several recent studies have examined adversarial attacks (e.g., FGSM, PGD) against machine-learning-based NIDS and proposed defensive strategies such as adversarial training, GAN-based augmentation, or transfer-learning detectors. For instance:

- Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense reports how adversarial perturbations can drastically reduce IDS accuracy, and reviews various defensive techniques. [4]
- A Systematic Study of Adversarial Attacks Against Network Intrusion Detection Systems presents results across multiple ML-based NIDS under PGD and other attacks, showing large performance drops. [58]

- Enhancing Adversarial Robustness in Network Intrusion Detection: A Novel Adversarially Trained Neural Network Approach introduces an adversarially trained NN on UNSW-NB15, maintaining 80%+ accuracy under FGSM/PGD. [29]

Our research quetions are answered:

RQ1 To what extent do adversarial examples generated via FGSM and PGD from benign traffic in the CIC-IDS2017 dataset evade detection by SNORT's ML-based intrusion detection system?

- Adversarial examples generated through FGSM and PGD significantly challenge the detection capabilities of SNORT's ML-based IDS. Our experiments show that while SNORT's baseline ML model attains a high accuracy of 99.85% on clean data, its recall drops to around 79% under PGD attacks, indicating a substantial evasion success rate. FGSM attacks similarly reduce recall to about 86%, signifying adversarial examples can evade traditional IDS detection with non-negligible false negative rates. Therefore, adversarial evasion is a significant weakness to SNORT-based ML intrusion detection that needs more robustness methods.

RQ2 How does the integration of a custom ANN model with adversarial noise injection into SNORT affect detection robustness and false positive rates?

- Integrating a resilient ANN model with adaptive noise injection into SNORT considerably improves detection robustness against adversarial attacks. The proposed model maintains a high validation accuracy of 99.75% while raising recall under adversarial conditions to approximately 86% (FGSM) and 79% (PGD), outperforming the baseline IDS model. This noise injection acts as an effective regularizer, enabling the model to resist evasion without incurring excessive false positive rates, which remain at a low 0.15% on benign data. Hence, the adaptive noise-injected ANN enhances SNORT's capacity to detect attacks accurately with improved resistance to adversarial evasion and minimal increase in false alarms.

4.8 Summary of Findings

This chapter demonstrated that:

1. Baseline SNORT exhibits high recall but poor precision, leading to many false positives.

2. ANN integration markedly enhances detection accuracy and F1-score.

3. The proposed Resilient ANN with Adaptive Noise Injection maintains nearly identical accuracy under clean and adversarial conditions, confirming strong adversarial resilience. noise injection yields up to 24% improved robustness under FGSM/PGD attacks relative to traditional ANN.

4. Overall, the model achieves the intended objectives of mitigating evasion attacks and minimizing false positives in machine learning–augmented IDS frameworks.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This study dealt with the severe problem of adversarial avoidance to ANN-based IDS models. The proposed resilient ANN is more robust in models and has a lower rate of false positives by using adaptive noise injection during training to improve the model robustness. The SNORT implementation offered a realistic and working assessment, which proved the fact that the strength of gains in the area of robustness is converted into actual IDS implementation. The threat severity was pointed out by the evaluation based on the use of advanced adversarial generation methods (FGSM and PGD), which have proved the efficiency of the noise injection defense. The model was highly detection on clean and adversarial traffic, and it was able to outperform baseline SNORT ML modules. This article has added a scalable and computationally efficient defense strategy that is quite appropriate in the current IDS problems.

5.2 Future Work

Future research questions might include: Although this study used machine learning models consisting of ANNs, future studies can investigate deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) or models based on transformers to increase the detection preciseness and resilience to complex adversarial attacks. Generalizing the model to identify multi-class attack scenarios on top of binary classification and increasing granularity. Assessing resistance to new adversarial attack methods, such as black-box and transfer attacks. Exploring noise injection with other defensive techniques including ensemble learning and GAN-based adversarial training. Making real-time implementation and testing to large scale production networks with varied traffic patterns feasible. Testing the integration with other IDS systems such as Suricata and Zeek to confirm the flexibility within IDS systems.

References

[1] Shakil Ibne Ahsan, Phil Legg, and S.M. Iftekharul Alam. An explainable ensemble-based intrusion detection system for software-defined vehicle ad-hoc networks. *Cyber Security and Applications*, 3:100090, 2025. URL: https://www.sciencedirect.com/science/article/pii/S2772918425000074, doi:10.1016/j.csa.2025.100090.

[2] Muhammad Akhtar, Syed Qadri, Maria Siddiqui, Syed Muhammad Nabeel Mustafa, and Syed Ali. Robust genetic machine learning ensemble model for intrusion detection in network traffic. *Scientific Reports*, 13, 10 2023. doi:10.1038/s41598-023-43816-1.

[3] Elie Alhajjar, Paul Maxwell, and Nathaniel Bastian. Adversarial machine learning in network intrusion detection systems. *Expert Systems with Applications*, 186:115782, 08 2021. doi:10.1016/j.eswa.2021.115782.

[4] Afnan Alotaibi and Murad A. Rassam. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet*, 15(2), 2023. URL: https://www.mdpi.com/1999-5903/15/2/62, doi:10.3390/fi15020062.

[5] Mohammed Alyahya et al. Toward reducing ids misclassification using hybrid dl and ml approach. advances in artificial intelligence and machine learning. 2024; 4 (3): 161, 2024.

[6] Fahmy Amin and M Mahmoud. Confusion matrix in binary classification problems: A step-by-step tutorial. *Journal of Engineering Research*, 6(5):0–0, 2022.

[7] Alexandru Apostu, Silviu Gheorghe, Andrei Hîji, Nicolae Cleju, Andrei Pătraşcu, Cristian Rusu, Radu Ionescu, and Paul Irofti. Detecting and mitigating ddos attacks with ai: A survey. *arXiv preprint arXiv:2503.17867*, 2025.

[8] Giovanni Apruzzese, Mauro Andreolini, Luca Ferretti, Mirco Marchetti, and Michele Colajanni. Modeling realistic adversarial attacks against network intrusion detection systems. *Digital Threats: Research and Practice (DTRAP)*, 3(3):1–19, 2022.

[9] Dimitrios Christos Asimopoulos, Panagiotis Radoglou-Grammatikis, Ioannis Makris, Valeri Mladenov, Konstantinos E Psannis, Sotirios Goudos, and Panagiotis Sarigiannidis. Breaching the defense: Investigating fgsm and ctgan adversarial attacks on iec 60870-5-104 ai-enabled intrusion detection systems. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–8, 2023.

[10] Dimitrios Christos Asimopoulos, Panagiotis Radoglou-Grammatikis, Ioannis Makris, Valeri Mladenov, Konstantinos E Psannis, Sotirios Goudos, and Panagiotis Sarigiannidis. Breaching the defense: Investigating fgsm and ctgan adversarial attacks on iec 60870-5-104 ai-enabled intrusion detection systems. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–8, 2023.

[11] Hanaa Attou, Azidine Guezzaz, Said Benkirane, Mourade Azrour, and Yousef Farhaoui. Cloud-based intrusion detection approach using machine learning techniques. *Big Data Mining and Analytics*, 6(3):311–320, 2023.

[12] Zeinab Awad, Magdy Zakaria, and Rasha Hassan. An enhanced ensemble defense framework for boosting adversarial robustness of intrusion detection systems. *Scientific Reports*, 15(1):14177, 2025.

[13] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

[14] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2154–2156, 2018.

[15] Sushil Buriya and Neelam Sharma. Vulnerability analysis of ml-based intrusion detection systems against evasion attacks. *Educational Administration: Theory and Practice*, 29(4):1960–1968, Dec. 2023. URL: `https://kuey.net/index.php/kuey/article/view/6791`, `doi:10.53555/kuey.v29i4.6791`.

[16] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[17] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[18] Tom Davies, Max Hashem Eiza, Nathan Shone, and Rob Lyon. A collaborative intrusion detection system using snort ids nodes. *arXiv preprint arXiv:2504.16550*, 2025.

[19] Islam Debicha, Thibault Debatty, Jean-Michel Dricot, and Wim Mees. Adversarial training for deep learning-based intrusion detection systems. *arXiv preprint arXiv:2104.09852*, 2021.

[20] Pooyan Azizi Doost, Sadegh Sarhani Moghadam, Edris Khezri, Ali Basem, and Mohammad Trik. A new intrusion detection method using ensemble classification and feature selection. *Scientific Reports*, 15(1):13642, 2025.

[21] Ouafae El Aeraj and Cherkaoui Leghris. Intelligent intrusion detection system snort & svm. *Revue d'Intelligence Artificielle*, 37(6):1629, 2023.

[22] Ouafae El Aeraj and Cherkaoui Leghris. Analysis of the snort intrusion detection system using machine learning. *International Journal of Information Science and Technology*, 8(1):1–9, 2024.

[23] Sabrine Ennaji, Fabio De Gaspari, Dorjan Hitaj, Alicia Kbidi, and Luigi Vincenzo Mancini. Adversarial challenges in network intrusion detection systems: Research insights and future prospects. *IEEE Access*, 2025.

[24] Jamal Esmaily, Reza Moradinezhad, and Jamal Ghasemi. Intrusion detection system based on multi-layer perceptron neural networks and decision tree. In *2015 7th Conference on Information and Knowledge Technology (IKT)*, pages 1–5. IEEE, 2015.

[25] Ruili Feng, Deli Zhao, and Zheng-Jun Zha. Understanding noise injection in gans. In *international conference on machine learning*, pages 3284–3293. PMLR, 2021.

[26] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[27] Kijun Han, Aathif Nizam, Ananthu Prakash, and Aparna S. A comparative study on artificial intelligence-based intrusion detection systems: Artificial neural networks versus deep neural networks. *International Journal of Engineering Research and Technology (IJERT)*, 2025. Accessed: 2025-10-21. URL: https://www.ijert.org/a-comparative-study-on-ai-ids-artificial-intelligence-based-intrusion-detect

[28] Md Mehedi Hasan, Rafiqul Islam, Quazi Mamun, Md Zahidul Islam, and Junbin Gao. Adversarial attacks on deep learning-based network intrusion detection systems: A taxonomy and review. *Available at SSRN 5096420*, 2025.

[29] Vahid Heydari and Kofi Nyarko. Enhancing adversarial robustness in network intrusion detection: A novel adversarially trained neural network approach. *Electronics*, 14(16), 2025. URL: https://www.mdpi.com/2079-9292/14/16/3249, doi:10.3390/electronics14163249.

[30] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[31] Beauden John and John Olusegun. Advanced deep learning techniques for enhancing intrusion detection systems (ids): A new frontier in cybercrime pattern recognition and prevention. 01 2025.

[32] Vandana Kadam and Rakesh Verma. Evaluating effectiveness: A critical review of performance metrics in intrusion detection system. *Journal of Engineering Science & Technology Review*, 18(1), 2025.

[33] Ansam Khraisat, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, 2(1):1–22, 2019.

[34] Gisung Kim, Seungmin Lee, and Sehun Kim. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *Expert Systems with Applications*, 41(4):1690–1700, 2014.

[35] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[37] Arash Habibi Lashkari, Gerard Draper Gil, Mohammad Saiful Islam Mamun, and Ali A Ghorbani. Characterization of tor traffic using time based features. In *International conference on information systems security and privacy*, volume 2, pages 253–262. SciTePress, 2017.

[38] Fiona Lawrence. Enhancing intrusion detection systems with ensemble models and hybrid feature selection techniques. *Journal of Information Systems Engineering and Management*, 10:937–954, 03 2025. doi:10.52783/jisem.v10i23s.3816.

[39] Rui Li, Kai Shuang, Mengyu Gu, and Sen Su. Adaptive noise injection: A structure-expanding regularization for rnn. *arXiv preprint arXiv:1907.10885*, 2019.

[40] Yuezun Li, Cong Zhang, Honggang Qi, and Siwei Lyu. Adani: Adaptive noise injection to improve adversarial robustness. *Computer Vision and Image Understanding*, 238:103855, 2024.

[41] YuXin Li. A blockchain-driven algorithm for anomaly detection in ipv6 network traffic. *World Journal of Information Technology*, page 1, 2024.

[42] Chen Liu, Zhichao Huang, Mathieu Salzmann, Tong Zhang, and Sabine Süsstrunk. On the impact of hard adversarial instances on overfitting in adversarial training. *Journal of Machine Learning Research*, 25(356):1–46, 2024.

[43] Zhenyu Liu, Garrett Gagnon, Swagath Venkataramani, and Liu Liu. Sinai: Selective injection of noise for adversarial robustness with improved efficiency.

[44] Zhenyu Liu, Garrett Gagnon, Swagath Venkataramani, and Liu Liu. Enhance dnn adversarial robustness and efficiency via injecting noise to non-essential neurons. *arXiv preprint arXiv:2402.04325*, 2024.

[45] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[46] Maraz Mia, Mir Mehedi A Pritom, Tariqul Islam, and Kamrul Hasan. Visually analyze shap plots to diagnose misclassifications in ml-based intrusion detection. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 632–641. IEEE, 2024.

[47] Hesamodin Mohammadian, Arash Habibi Lashkari, and Ali A Ghorbani. Evaluating deep learning-based nids in adversarial settings. In *ICISSP*, pages 435–444, 2022.

[48] Ija Moisejevs et al. Adversarial attacks and defenses in intrusion detection systems: A survey. *International Journal of Artificial Intelligence and Expert Systems (IJAE)*, 8:44–62, 2019.

[49] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[50] Srivastava Nitish. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1, 2014.

[51] Pranav Pant, Aniket Kumar, Lalit Vashishtha, Subhasis Dash, Niranjan Ray, and Dr Sahu. A comparative study of deep learning techniques for network intrusion detection. pages 722–727, 02 2024. doi:10.1109/ESIC60604.2024.10481540.

[52] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[53] Nicholas Pudjihartono, Tayaza Fadason, Andreas W Kempa-Liehr, and Justin M O'Sullivan. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in bioinformatics*, 2:927312, 2022.

[54] Martin Roesch et al. Snort: Lightweight intrusion detection for networks. In *Lisa*, volume 99, pages 229–238, 1999.

[55] Karen Scarfone, Peter Mell, et al. Guide to intrusion detection and prevention systems (idps). *NIST special publication*, 800(2007):94, 2007.

[56] Iman Sharafaldin, Arash Habibi Lashkari, Ali A Ghorbani, et al. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1(2018):108–116, 2018.

[57] Sanidhya Sharma. Adversarial attacks against network intrusion detection systems. Master's thesis, Purdue University, 2024.

[58] Sanidhya Sharma and Zesheng Chen. A systematic study of adversarial attacks against network intrusion detection systems. *Electronics*, 13(24), 2024. URL: https://www.mdpi.com/2079-9292/13/24/5030, doi:10.3390/electronics13245030.

[59] Narjes Shojaati and Nathaniel D Osgood. Dynamic computational models and simulations of the opioid crisis: a comprehensive survey. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–25, 2021.

[60] Nathan Shone, Tran Nguyen Ngoc, Vu Dinh Phai, and Qi Shi. A deep learning approach to network intrusion detection. *IEEE transactions on emerging topics in computational intelligence*, 2(1):41–50, 2018.

[61] Dule Shu, Nandi O Leslie, Charles A Kamhoua, and Conrad S Tucker. Generative adversarial attacks against intrusion detection systems using active learning. In *Proceedings of the 2nd ACM workshop on wireless security and machine learning*, pages 1–6, 2020.

[62] Ayan Sinha, Zhao Chen, Vijay Badrinarayanan, and Andrew Rabinovich. Gradient adversarial training of neural networks. *arXiv preprint arXiv:1806.08028*, 2018.

[63] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. *arXiv e-prints*, pages arXiv–1811, 2018.

[64] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[65] Benyamin Tafreshian and Shengzhi Zhang. A defensive framework against adversarial attacks on machine learning-based network intrusion detection systems. In *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 2436–2441. IEEE, 2024.

[66] Tuan A Tang, Lotfi Mhamdi, Des McLernon, Syed Ali Raza Zaidi, and Mounir Ghogho. Deep learning approach for network intrusion detection in software defined networking. In *2016 international conference on wireless networks and mobile communications (WINCOM)*, pages 258–263. IEEE, 2016.

[67] TensorFlow. Understanding fgsm adversarial attacks. `https://www.tensorflow.org/tutorials/generative/adversarial_fgsm`, 2024. [Online; accessed 01-October-2025].

[68] Sadargari Viharika. Enhancing intrusion detection and cloud security by integrating snort with advanced ai techniques for improved accuracy and threat mitigation. *Journal of Information Systems Engineering and Management*, 10:627–637, 03 2025. `doi:10.52783/jisem.v10i24s.3953`.

[69] Hetvi Waghela, Jaydip Sen, and Sneha Rakshit. Robust image classification: Defensive strategies against fgsm and pgd adversarial attacks. In *2024 Asian Conference on Intelligent Technologies (ACOIT)*, pages 1–7. IEEE, 2024.

[70] Pratik Waghmode, Manideep Kanumuri, Hosam El-Ocla, and Tanner Boyle. Intrusion detection system based on machine learning using least square support vector machine. *Scientific Reports*, 15(1):12066, 2025.

[71] Chaoyun Zhang, Xavier Costa-Perez, and Paul Patras. Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. *IEEE/ACM Transactions on Networking*, 30(3):1294–1311, 2022.

[72] Mengnan Zhao, Lihe Zhang, Jingwen Ye, Huchuan Lu, Baocai Yin, and Xinchao Wang. Adversarial training: A survey. *arXiv preprint arXiv:2410.15042*, 2024.

[73] Weimin Zhao, Sanaa Alwidian, and Qusay H Mahmoud. Adversarial training methods for deep learning: A systematic review. *Algorithms*, 15(8):283, 2022.

[74] Evgenii Zheltonozhskii, Chaim Baskin, Yaniv Nemcovsky, Brian Chmiel, Avi Mendelson, and Alex M Bronstein. Colored noise injection for training adversarially robust neural networks. *arXiv preprint arXiv:2003.02188*, 2020.

[75] Richard Zur, Yulei Jiang, Lorenzo Pesce, and Karen Drukker. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical physics*, 36:4810–8, 10 2009. doi:10.1118/1.3213517.

[76] Richard M Zur, Yulei Jiang, Lorenzo L Pesce, and Karen Drukker. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical physics*, 36(10):4810–4818, 2009.