Project Title

Enhancing ML-DL based IDS with Robust Adversarial Defense Techniques

Group-10 Members

1. **Ahsan Md. Sajid Khan (ID 0422857)**
2. **Farhana Mim (ID 0422492)**

Ai/Ml in Cyber Security
Course Code: CPSC_5207EL_64

Spring 2024

Submitted to

Professor Sk Md Mizanur Rahman

Table of Contents

Introduction

The framework for enhancing the adversarial robustness of deep learning-based intrusion detection systems (IDS) addresses significant weaknesses in these systems. Designed to detect malicious activities in network traffic, IDS can be easily deceived by adversarial attacks that subtly modify input data, causing incorrect classifications and jeopardizing network security.

To mitigate this issue, the authors propose a straightforward framework that improves IDS robustness against adversarial attacks. This framework employs two key techniques: adversarial training and feature squeezing. Adversarial training involves using adversarial examples during training, enabling the IDS to recognize and resist deceptive patterns. Feature squeezing simplifies input data by compressing its features, aiding in the detection of adversarial manipulations by removing unnecessary variations.

Through extensive experiments, the paper demonstrates the framework's effectiveness in enhancing IDS resistance to adversarial attacks. The findings show substantial improvements in robustness, emphasizing the framework's potential to secure deep learning-based IDS against advanced threats, thus contributing to more reliable and secure network defense systems.

Background / Literature Review

Intrusion Detection Systems (IDS) play a crucial role in ensuring network security by detecting and mitigating malicious activities. Traditional IDS models, which depend on signature and anomaly detection, often fail to identify new and complex attacks. In contrast, deep learning-based IDS leverage neural networks to detect intricate patterns in network traffic, proving to be more effective. However, these advanced systems are susceptible to adversarial attacks, where small, intentional changes to input data can mislead the IDS, resulting in incorrect classifications and security breaches. This issue, initially observed in image recognition systems, was highlighted by Goodfellow et al. (2014), leading to the development of various attack and defense strategies [5].

Adversarial training, a key defense mechanism, incorporates adversarial examples during training to help models resist such manipulations. Although effective, it is resource-intensive. Another strategy, feature squeezing, introduced by Xu et al. (2017) [6], simplifies input data to

help detect adversarial perturbations. Studies by Wang et al. (2018) and Sadeghzadeh et al. (2019) examined these techniques in IDS. Building on this research, the paper proposes a framework that combines adversarial training and feature squeezing, showing through extensive experiments that this approach improves IDS robustness against adversarial attacks.

Project Objective

The primary aim of this project is to develop a robust defense mechanism for machine learning (ML) and deep learning (DL)-based intrusion detection systems (IDS) against adversarial attacks. Adversarial attacks, where attackers introduce deceptive inputs to mislead ML models, significantly threaten the effectiveness of IDS. The goal is to enhance the resilience of these systems, ensuring they can reliably detect and mitigate such attacks.

To achieve this, several key components are involved. One of the primary strategies is the use of Multi-Armed Bandits (MAB) with Thompson Sampling. MAB is a framework that dynamically selects the most effective classifiers in real-time by balancing exploration (testing different classifiers) and exploitation (using the best-performing ones). Thompson Sampling estimates the probability of success for each classifier, thereby choosing the one that maximizes the expected reward.

Another critical component is Ant Colony Optimization (ACO), inspired by the foraging behavior of ants. ACO is used to optimize the selection and combination of classifiers, solving optimization problems by finding paths that lead to optimal solutions, thus enhancing the IDS's robustness against various adversarial attacks.

Additionally, the defense mechanism involves generating adversarial samples using methods like Zeroth Order Optimization (ZOO) and Fast Gradient Sign Method (FGSM). These samples are used to test and validate the IDS, ensuring it can withstand sophisticated attacks. By training the IDS with both normal and adversarial examples, the system's resilience is significantly improved.

The project's primary goals are to analyze and understand adversarial attacks and subsequently develop a robust defense mechanism for ML and DL-based IDS. The key objectives are as follows:

1. **Enhance cybersecurity defenses against adversarial attacks.**
2. **Create an adaptable solution to combat evolving cyber threats.**
3. **Strengthen intrusion detection systems for robust protection.**

ML/DL Based IDS

An open-source platform, IDS-ML, is designed to simplify the development and testing of machine learning (ML)-based intrusion detection systems (IDS). It provides a comprehensive framework for data preprocessing, feature extraction, model training, and evaluation, facilitating the creation of effective IDS solutions. By making advanced ML techniques accessible, IDS-ML aims to enhance IDS detection capabilities and promote broader adoption of ML in cybersecurity.

In addressing the vulnerabilities of deep learning (DL)-based IDS to adversarial attacks, another framework proposes combining adversarial training and feature squeezing to bolster IDS robustness against such threats. Adversarial attacks involve subtle modifications to input data designed to deceive the IDS, leading to incorrect classifications and compromised security. Adversarial training involves incorporating adversarial examples into the training process, while feature squeezing reduces input data complexity to detect manipulations. Extensive experimental evaluations demonstrate the framework's effectiveness in significantly improving IDS resilience to adversarial attacks.

Both contributions advance IDS by leveraging ML and DL techniques. IDS-ML offers a practical, open-source solution for developing ML-based IDS, emphasizing ease of use and accessibility. In contrast, the proposed framework focuses on enhancing the security of DL-based IDS against adversarial attacks. Together, these works highlight the potential of ML and DL to improve IDS performance and resilience, addressing both practical implementation and security challenges in cybersecurity.

Adversarial Attacks

Adversarial attacks that have been explored earlier and their impact on deep learning (DL)-based intrusion detection systems (IDS) are shown in details below:

Nature of Adversarial Attacks:

Definition and Mechanism: Adversarial attacks involve subtle, intentional modifications to input data designed to mislead DL models into incorrect classifications. These modifications, often imperceptible to humans, can cause significant errors in model predictions. Attackers craft these adversarial examples by adding carefully calculated perturbations to the original inputs, exploiting the model's sensitivity to these changes.

Impact on IDS: For DL-based IDS, adversarial attacks pose a significant threat as they can result in false negatives (failing to detect actual intrusions) or false positives (misclassifying benign activities as malicious). These attacks can undermine the reliability and effectiveness of IDS, allowing malicious activities to go undetected and potentially compromising network security.

Techniques for Adversarial Attacks:

Common Methods:

1. **Fast Gradient Sign Method (FGSM):**
    - FGSM generates adversarial examples by calculating the gradient of the loss function with respect to the input data and adjusting the input in the direction of the gradient to maximize the loss. This method adds small perturbations proportional to the sign of the gradient.
2. **Zeroth Order Optimization (ZOO):**
    - ZOO generates adversarial examples without requiring gradient information, making it effective against black-box models where internal parameters are unknown. It uses optimization techniques to approximate gradients and iteratively refine the adversarial perturbations to achieve the desired misclassification.

Workflow of ML based IDS

We utilized open-source Python code to demonstrate a general Intrusion Detection System (IDS) [1]. For this purpose, we used the CIC-IDS-2017 dataset [4]. The process involved training and evaluating the IDS using machine learning techniques. The overall workflow is depicted in the following diagram:

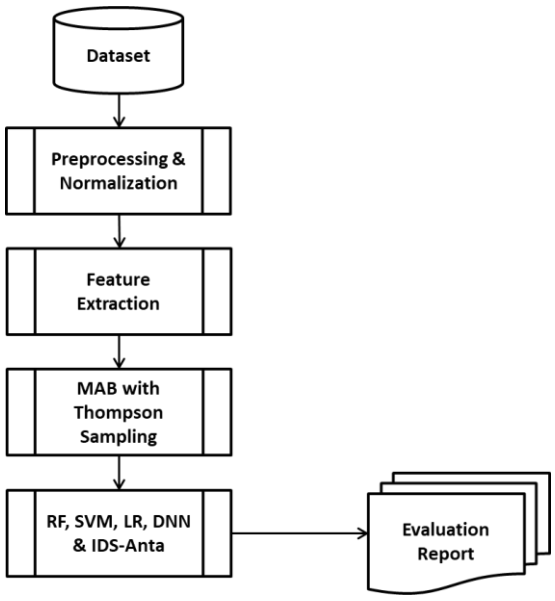


Figure 1 Machine Learning Based IDS

Datasets: The process begins with gathering the datasets.

Preprocessing & Normalization: This step involves cleaning the data and normalizing it to ensure consistency.

Feature Extraction: Relevant features are extracted from the processed data, which are crucial for training models.

MAB with Thomson Sampling: This step seems to implement a Multi-Armed Bandit (MAB) approach using Thomson Sampling, likely for model selection or optimization.

Model Training: Various machine learning models, such as Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Deep Neural Networks (DNN), and the IDS-Anta model, are trained using the features.

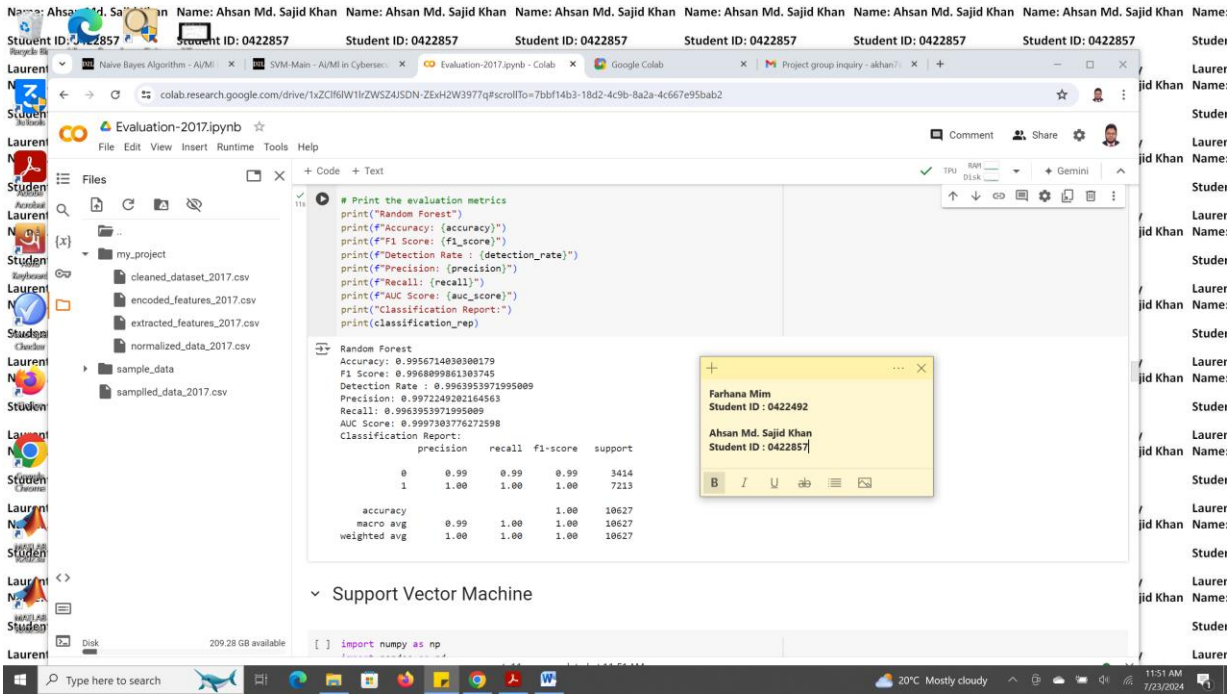
Performance Analysis: The trained models are evaluated to analyze their performance.

While executing the open-source code, we modified it for compatibility with Google Colab and obtained the following output for the CIC-IDS-2017 dataset [4] (random samples) using different classifiers:

Random Forest:

```
Accuracy: 0.9956714030300179
F1 Score: 0.9968099861303745
Detection Rate : 0.9963953971995009
Precision: 0.9972249202164563
Recall: 0.9963953971995009
AUC Score: 0.9997303776272598

Classification Report:
```

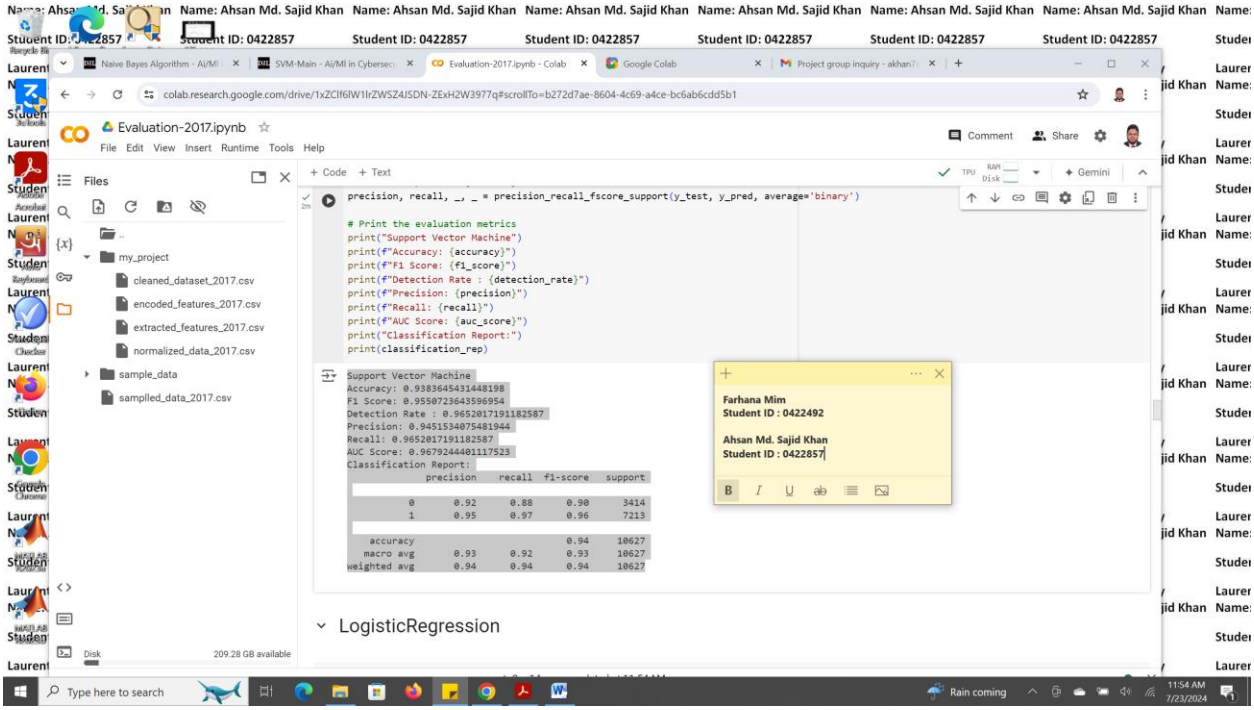
	precision	recall	f1-score	support
0	0.99	0.99	0.99	3414
1	1.00	1.00	1.00	7213
accuracy			1.00	10627
macro avg	0.99	1.00	1.00	10627
weighted avg	1.00	1.00	1.00	10627



SVM

```
Support Vector Machine
Accuracy: 0.9383645431448198
F1 Score: 0.9550723643596954
Detection Rate : 0.9652017191182587
Precision: 0.9451534075481944
Recall: 0.9652017191182587
AUC Score: 0.9679244401117523
Classification Report:
```

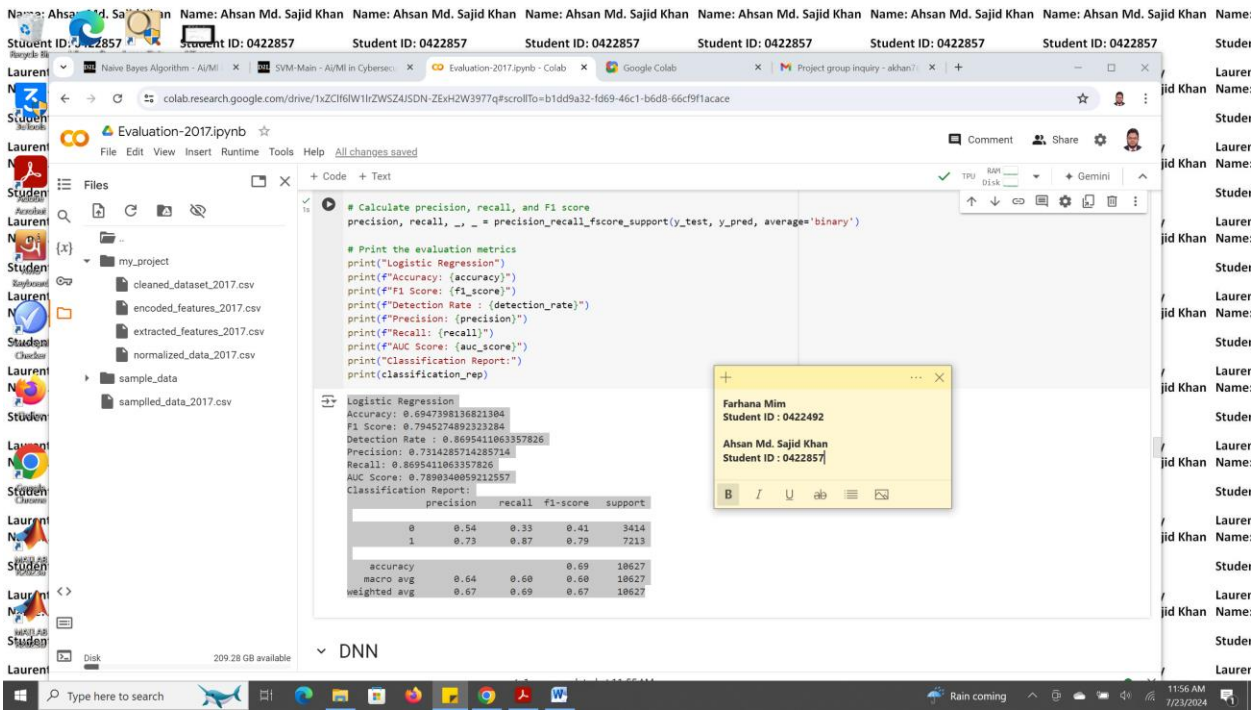
	precision	recall	f1-score	support
0	0.92	0.88	0.90	3414
1	0.95	0.97	0.96	7213
accuracy			0.94	10627
macro avg	0.93	0.92	0.93	10627
weighted avg	0.94	0.94	0.94	10627



Logistic Regression

```
Logistic Regression

Accuracy: 0.6947398136821304

F1 Score: 0.7945274892323284

Detection Rate : 0.8695411063357826

Precision: 0.7314285714285714

Recall: 0.8695411063357826

AUC Score: 0.7890340059212557

Classification Report:
```

	precision	recall	f1-score	support
0	0.54	0.33	0.41	3414
1	0.73	0.87	0.79	7213
accuracy			0.69	10627
macro avg	0.64	0.60	0.60	10627
weighted avg	0.67	0.69	0.67	10627



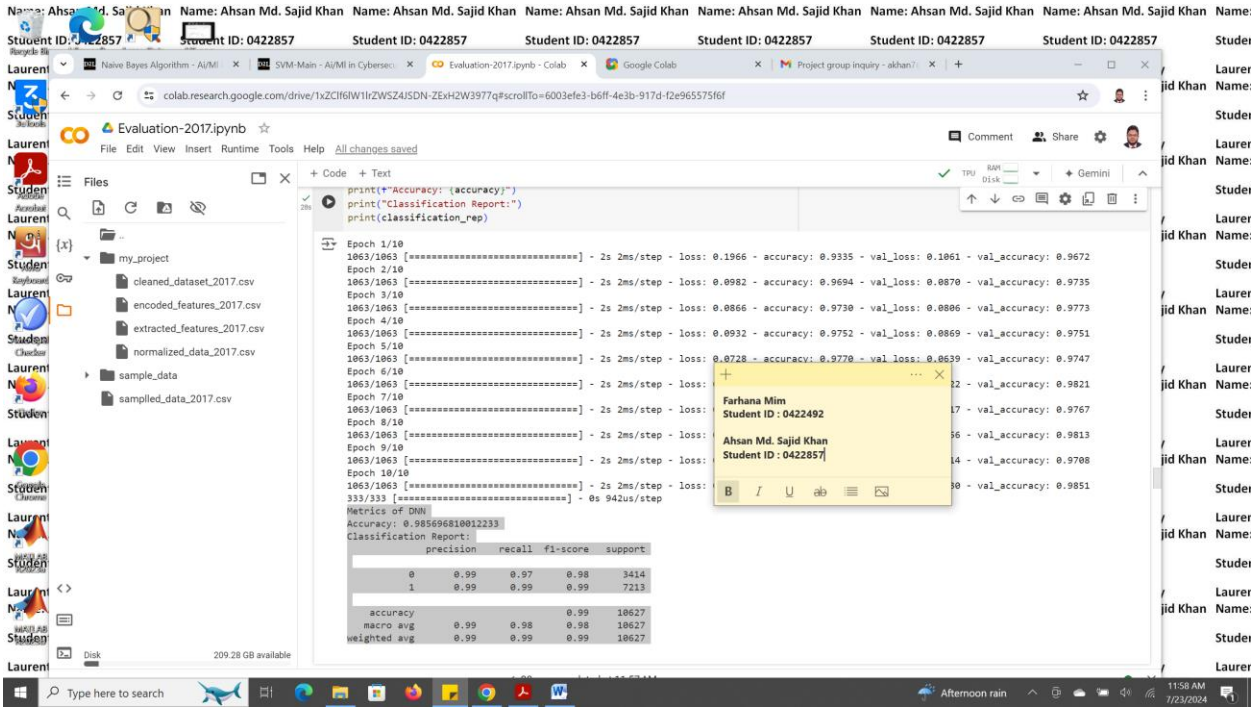
DNN

Metrics of DNN

Accuracy: 0.985696810012233

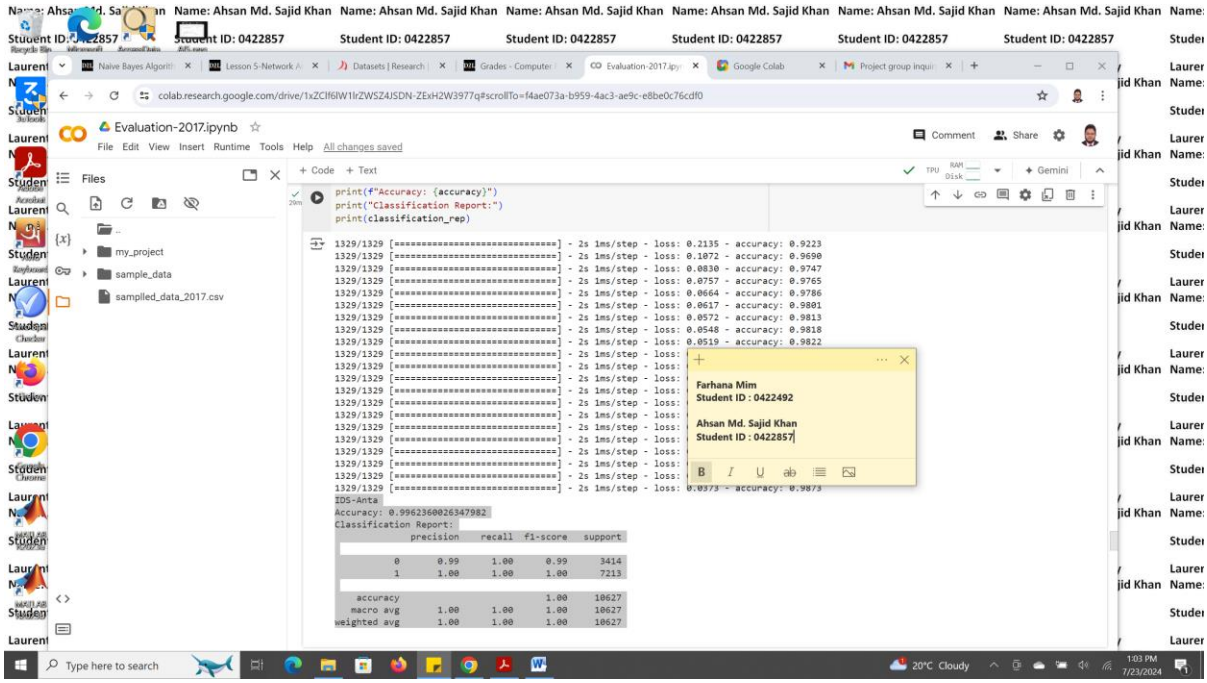
Classification Report:

	precision	recall	f1-score	support
0	0.99	0.97	0.98	3414
1	0.99	0.99	0.99	7213
accuracy			0.99	10627
macro avg	0.99	0.98	0.98	10627
weighted avg	0.99	0.99	0.99	10627



IDS Anta

```
IDS-Anta

Accuracy: 0.9962360026347982

Classification Report:

              precision    recall  f1-score   support

           0       0.99      1.00      0.99      3414

           1       1.00      1.00      1.00      7213

    accuracy                           1.00     10627

   macro avg       1.00      1.00      1.00     10627

weighted avg       1.00      1.00      1.00     10627
```



Workflow of Adversarial Attack Generation and Detection

In this process, we used the extracted features dataset to apply adversarial attacks and then followed the procedures below for classification and detection:

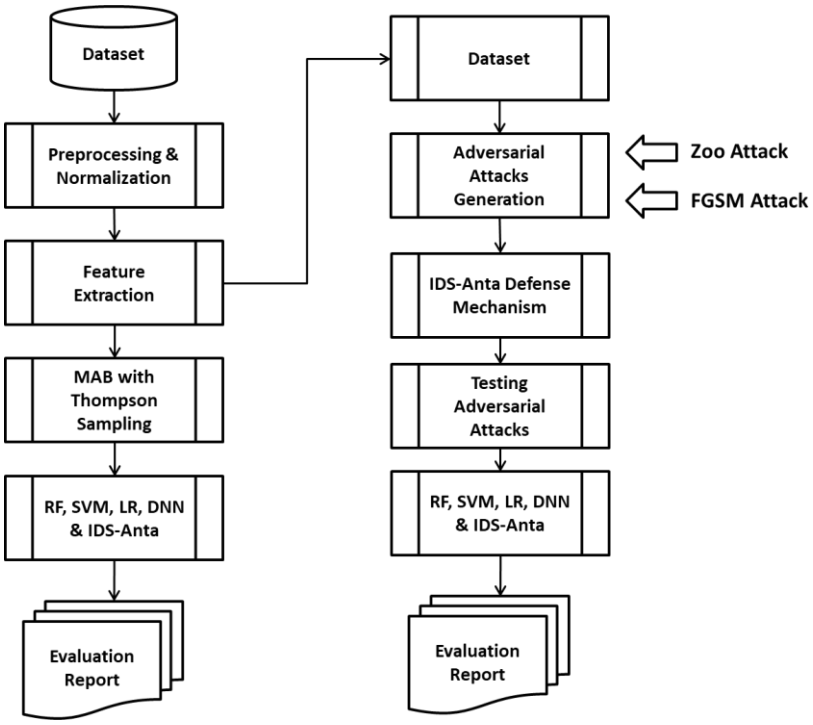


Figure 2: Adversarial Attack Generation and Detection

Adversarial Attack Generations: This step involves generating adversarial attacks based on the 2017 dataset using methods like ZOO attack and FGSM attack.

IDS-Anta: Defense Mechanism: The IDS-Anta's defense mechanisms are tested against these adversarial attacks.

Testing Adversarial Attacks: This step involves testing the IDS-Anta against the generated adversarial attacks to evaluate its robustness.

Performance Analysis: Similar to the ML based IDS, the performance of this, including its resilience to adversarial attacks, is analyzed.

The IDS-Anta open-source software repository helps researchers develop strategies to defend against adversarial attacks on machine learning and deep learning-based intrusion detection systems (IDS). It addresses several key questions: the overall process of designing ML- and DL-based IDS, the importance of preprocessing and feature extraction, using Multi-Armed Bandit

(MAB) with Thomson Sampling to select effective classifiers and improve detection rates, generating adversarial samples with ZOO and FGSM attacks, and enhancing IDS performance against these attacks with combined techniques. The repository includes Python code implementations for ML and DL-based IDS, MAB, and generating adversarial attacks using ZOO and FGSM [1].

After generating the zoo adversarial attacks and following the Figure-2 process, we have the following outputs. Due to resource constraints, we minimized the number of iterations in a few cases.

```
Test Accuracy: 0.9945422038204573

Classification Report:

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      3414

           1       1.00      1.00      1.00      7213

    accuracy                           0.99     10627

   macro avg       0.99      0.99      0.99     10627

weighted avg       0.99      0.99      0.99     10627


Test Accuracy (Adversarial): 0.48414416109908726

Classification Report (Adversarial):

               precision    recall  f1-score   support

           0       0.38      0.97      0.55      3414

           1       0.95      0.25      0.40      7213

    accuracy                           0.48     10627

   macro avg       0.66      0.61      0.47     10627

weighted avg       0.77      0.48      0.45     10627
```

Non-Adversarial Metrics:

Precision: 0.9945468513565168

Recall: 0.9945422038204573

F1-score: 0.9945438832199173

Detection Rate: [0.99267721 0.99542493]

AUC Score: 0.9940510693484418

Adversarial Metrics:

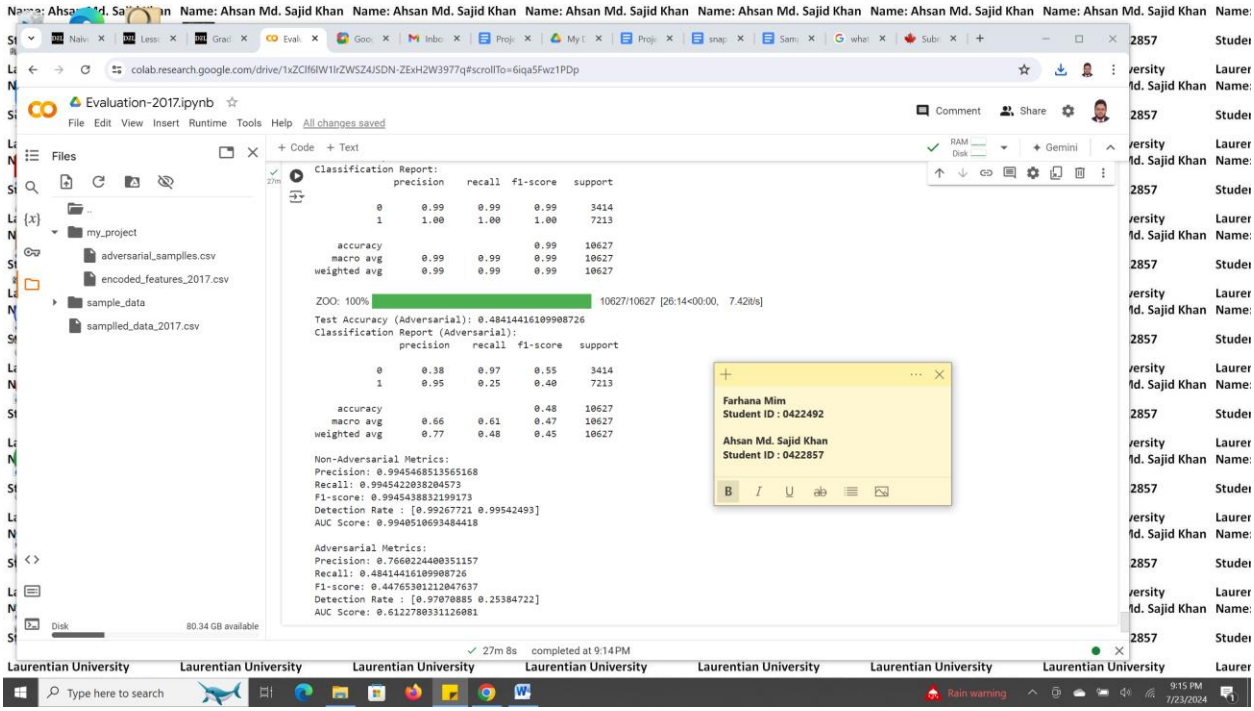
Precision: 0.7660224400351157

Recall: 0.48414416109908726

F1-score: 0.44765301212047637

Detection Rate: [0.97070885 0.25384722]

AUC Score: 0.6122780331126081



Defense Mechanisms

Proposed Framework: A combined approach has been proposed of adversarial training and feature squeezing to enhance the robustness of DL-based IDS against adversarial attacks.

Adversarial Training: This technique involves incorporating adversarial examples into the training dataset, allowing the model to learn and recognize these deceptive patterns. By exposing the model to various adversarial examples during training, it becomes better equipped to resist manipulative inputs during real-world deployment.

Feature Squeezing: Feature squeezing reduces the complexity of input data by compressing its features, aiding in detecting adversarial manipulations. Simplifying the input data limits the adversary's ability to introduce effective perturbations, making it easier for the IDS to identify and mitigate adversarial attacks.

Evaluation Results

For the studied paper [1] the proposed framework (IDS-Anta) effectiveness is validated through extensive experiments, showing significant improvements in IDS robustness. Tests using various adversarial attack methods demonstrate that the proposed framework can effectively mitigate the impact of adversarial attacks, thereby enhancing the security and reliability of DL-based IDS.

In summary, the research underscores the critical threat posed by adversarial attacks to DL-based IDS and presents a robust defense framework combining adversarial training and feature squeezing. This approach enhances the IDS's ability to detect and resist adversarial manipulations, ensuring more reliable network security.

Recommendations

The key recommendations emphasize a holistic approach to improving the adversarial robustness of DL-based IDS. Firstly, incorporating adversarial examples into the training dataset and continuously updating the training process is essential. Secondly, simplifying input data through feature squeezing and combining it with other defensive strategies creates a robust, multi-layered defense mechanism. Adopting hybrid defense strategies that adapt dynamically to new threats further strengthens IDS robustness. Enhancing model interpretability with explainable AI and greater transparency helps security analysts understand model decisions and identify vulnerabilities. Continuous monitoring and evaluation using standardized benchmarks ensure the IDS remains effective. Ongoing research and collaboration among academia, industry, and government agencies are crucial for developing advanced solutions. Additionally, training users and stakeholders on adversarial attacks and establishing best practices for secure IDS deployment are recommended. Lastly, ensuring the scalability and optimizing the performance of defense mechanisms for real-world deployment is vital.

References

[1] Barik, K., & Misra, S. (2024). IDS-Anta: An open-source code with a defence mechanism to detect adversarial attacks for intrusion detection system. *Software Impacts*, 100664. https://doi.org/10.1016/j.simpa.2024.100664

[2] A simple framework to enhance the adversarial robustness of deep learning-based intrusion detection system. (n.d.). https://arxiv.org/html/2312.03245v1

[3] Yang, L., & Shami, A. (2022). IDS-ML: An open source code for Intrusion Detection System development using Machine Learning. *Software Impacts*, *14*, 100446. https://doi.org/10.1016/j.simpa.2022.100446

[4] *IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB*. (n.d.). https://www.unb.ca/cic/datasets/ids-2017.html

[5] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014, June 10). *Generative adversarial networks*. arXiv.org. https://arxiv.org/abs/1406.2661

[6] Xu, W., Evans, D., Qi, Y., University of Virginia, & evadeML.org. (n.d.-b). Feature Squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed Systems Security Symposium (NDSS) 2018* [Conference-proceeding]. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-4_Xu_paper.pdf