Adversarial Training for Network Intrusion Detection

Filippo di Gravina

Department of Computer Science
University of Bari Aldo Moro

June 27, 2025

Introduction and motivation

Network Intrusion Detection Systems have become increasingly important for network security.

Recent methods have addressed some challenges like data imbalance, but robustness against adversarial attacks remain an underexplored issue.

Related work

As a baseline for evaluating performance on clean data, we refer to the recent work by Talukder et al. [Talukder et al., 2024].

Their approach combines oversampling techniques and feature extraction to handle big and imbalanced data.

Description of the dataset

The CIC-IDS-2017 dataset [Sharafaldin et al., 2018] is a network intrusion detection dataset that contains benign and malicious labeled traffic with an imbalanced ratio.

It contains more than 2 million examples and 79 features.

In the pre-processing step, we applied some operations:

- removed rows with missing values
- dropped duplicate entries
- eliminated zero-variance features

At the end of this step, we computed the correlation matrix.

Description of the solution

Adversarially perturbed examples are inputs modified by small perturbations designed to mislead a machine learning model, while remaining nearly indistinguishable from the original data.

Adversarial training is a regularization technique in which a model is trained not only on clean inputs but also on adversarially perturbed examples, with the goal of improving robustness to these attacks.

Adversarial attacks

We used two different methods to generate adversarial examples: the Fast Gradient Sign Method [Goodfellow et al., 2015] and the Projected Gradient Descent method [Madry et al., 2019].

Fast Gradient Sign Method

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

Projected Gradient Descent method

$$x^{t+1} = \Pi_{x+S}\left(x^t + \alpha \cdot \text{sign}(\nabla_x L(\theta, x, y))\right)$$

Model selection

We compared three machine learning models:

- Logistic Regression
- Support Vector Machine
- Multi Layer Perceptron

Model selection

The MLP resulted as the best model, with the following results:

Table: Pre-attack performance

Scenario	**Accuracy**	**Precision**	**Recall**	**F1-score**
Pre-Attack	0.988	0.971	0.955	0.963

Adversarial training evaluation

We evaluated the model across three phases:

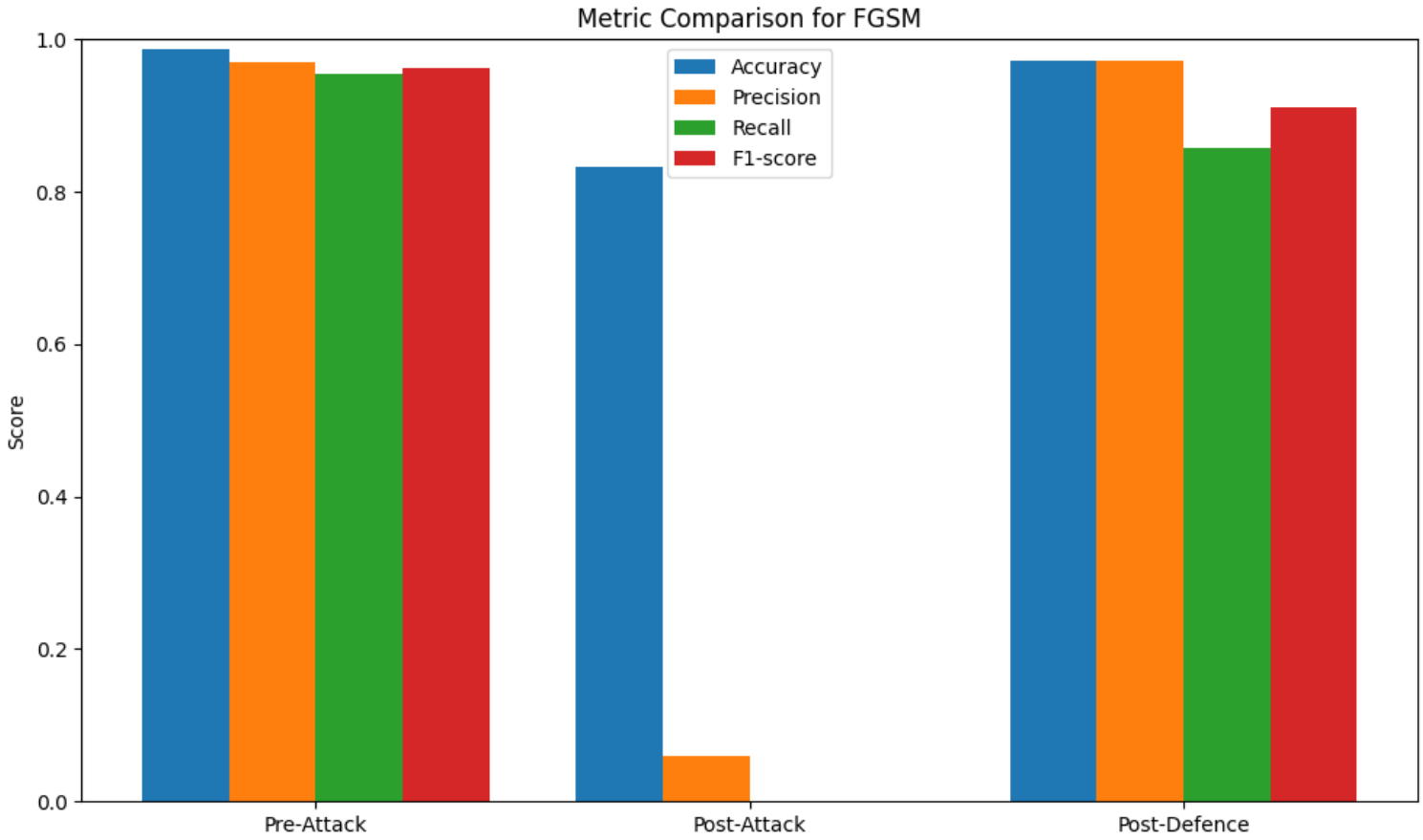
- Pre-attack: before any attack
- Post-attack: under adversarial perturbations
- Post-defense: after applying adversarial training

Adversarial training evaluation: results

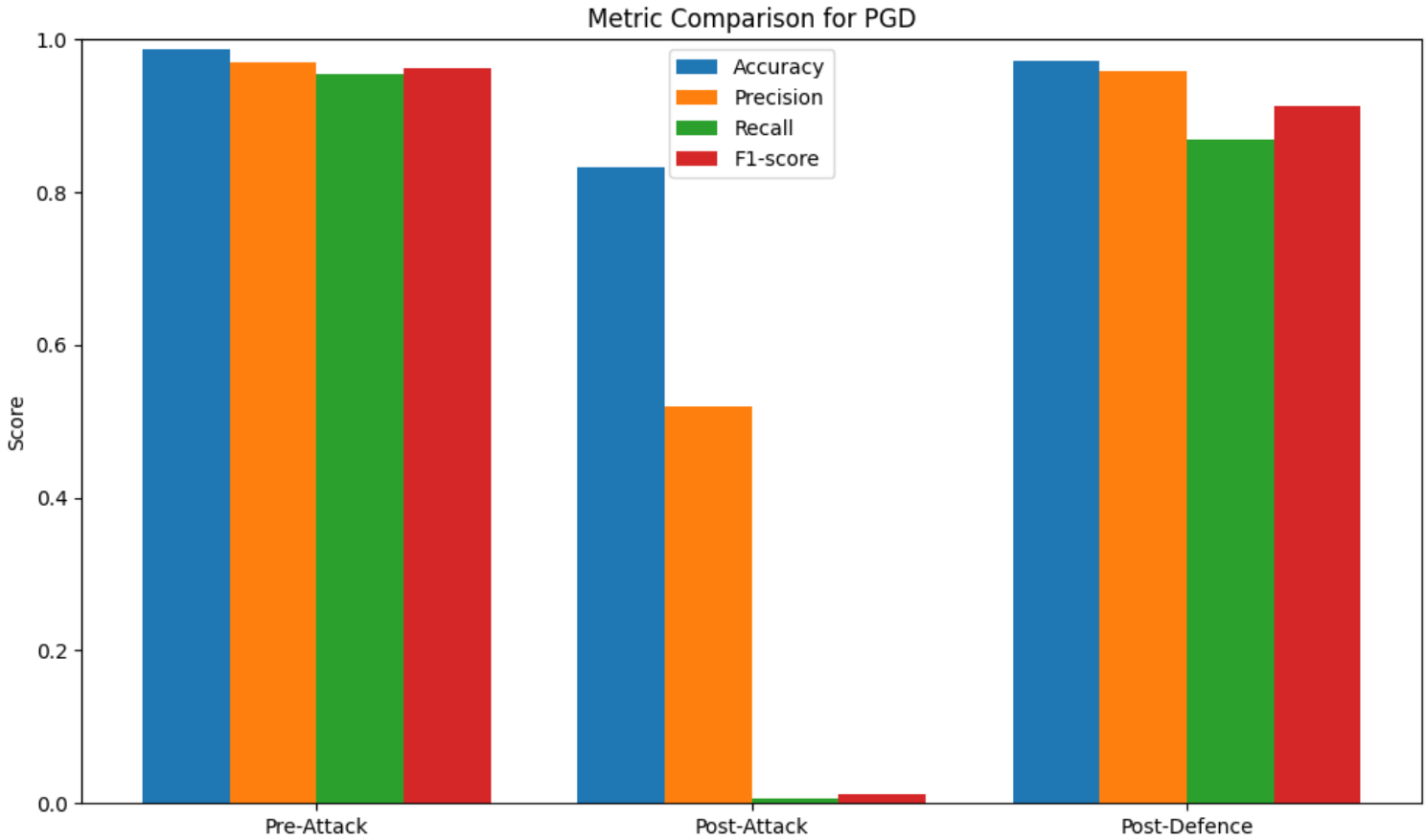
Table: Adversarial performance

Scenario	**Accuracy**	**Precision**	**Recall**	**F1-score**
Pre-Attack	0.988	0.971	0.955	0.963
Post FGSM	0.832	0.059	0.000	0.000
Post-Defense (FGSM)	0.972	0.972	0.858	0.911
Post PGD	0.832	0.520	0.006	0.012
Post-Defense (PGD)	0.972	0.958	0.869	0.912

FGSM attack plot



PGD attack plot



Adversarial training evaluation: comments

- On the clean data, the model demonstrated strong performance across all metrics
- When subjected to adversarial attacks, the model's performance dropped significantly. The inherent class imbalance in the dataset causes the model to overly rely on the majority class, making it more vulnerable to adversarial attacks that exploit this bias
- Fine-tuning on adversarial attacks proved to be an effective countermeasure

Conclusions and limitations

- Our MLP model performs very well on clean data
- However, it remains vulnerable to adversarial attacks such as FGSM and PGD despite the implemented regularized techniques, such as dropout and batch normalization
- Adversarial training effectively restores robustness, confirming the value of this defense strategy

Future work

- The simplicity and the gradient exposure of the MLP architecture make it a suitable baseline for further robustness evaluations
- Future work could explore the application of additional and more complex adversarial attacks

References I


Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015).
Explaining and harnessing adversarial examples.


Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019).
Towards deep learning models resistant to adversarial attacks.


Sharafaldin, I., Lashkari, A. H., and Ghorbani, A. A. (2018).
Toward generating a new intrusion detection dataset and intrusion traffic characterization.
In *International Conference on Information Systems Security and Privacy*.

References II


Talukder, M. A., Islam, M. M., Uddin, M. A., Hasan, K. F., Sharmin, S., Alyami, S. A., and Moni, M. A. (2024).
Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction.