Detecting Adversarial Manipulation in Intrusion Detection Systems through Explainable AI Attribution Drift

Karim Khidr | Student ID: C00315733 | MSc Cybersecurity, Privacy and Trust | Supervisor : Hisain Elshaafi | email: kakhidr@outlook.com

Ollscoil Teicneolaíochta an Oirdheiscirt
South East Technological University

Background and Research Gap

Machine Learning (ML) based Intrusion Detection Systems (IDS) improve cyber threat detection but remain vulnerable to adversarial manipulation. Most adversarial robustness research focuses on prediction errors. This study tests whether attacks can instead be detected through changes in feature-attribution explanations, even when the predicted class is preserved.

Research gap: explanation drift is rarely evaluated as a detection signal under prediction-preserving adversarial conditions.

Research Question and Objectives

Can explanation drift reliably detect adversarial manipulation in ML-IDS under prediction-preserving conditions?

- Quantify drift between clean and adversarial explanations.
- Test whether drift separates benign and manipulated samples.
- Compare behavior across Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks, Integrated Gradients (IG) and SHapley Additive exPlanations (SHAP), and two IDS datasets.
- Assess whether stronger perturbations cause greater explanation drift.

Selected references:
Goodfellow et al. (2015); Madry et al. (2018); Carlini and Wagner (2017); Lundberg and Lee (2017); Sundararajan et al. (2017).

Experimental Scope

- Datasets: CICIDS2018 (network IDS), BETH (host IDS)
- Attacks: FGSM), PGD
- Explanation methods: IG, SHAP
- Drift metrics: cosine and Euclidean distance
- Evaluation: Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) under a prediction-preserving constraint

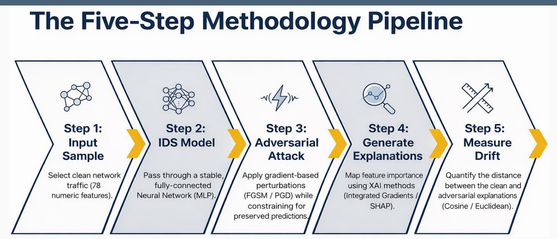
IG SHAP

Literature Insights and Contribution

Prior IDS Adversarial Studies	vs.	This Study
Focus on prediction errors		Focus on **explanation drift**
Misclassification, output degradation		Reasoning change under preserved predictions
Attack success at output level		Complementary detection signal

Contribution: explanation drift is shown to generalise across two attacks, two XAI methods, and two IDS domains.

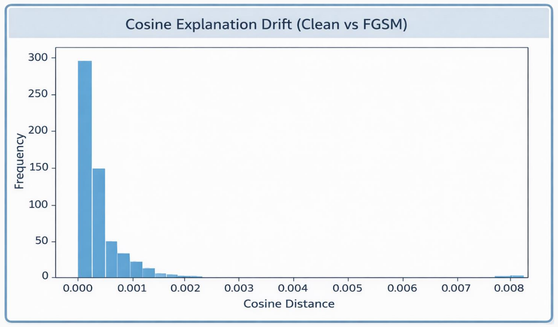
Methodology Pipeline



Controls: fixed evaluation subset, white-box threat model, prediction-preserving filtering, clean explanation stability, and epsilon sensitivity analysis.

Data and result summaries from completed dissertation experiments on **CICIDS2018** and **BETH**.

Illustrative Result: CICIDS2018 FGSM + IG



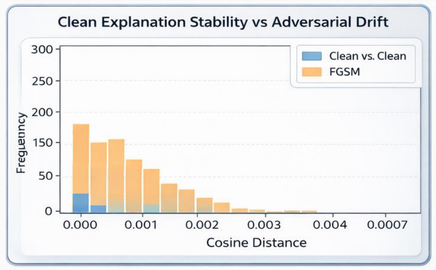
Cosine explanation drift distribution. Most adversarial samples exhibit measurable drift despite preserved predictions.

Key Findings

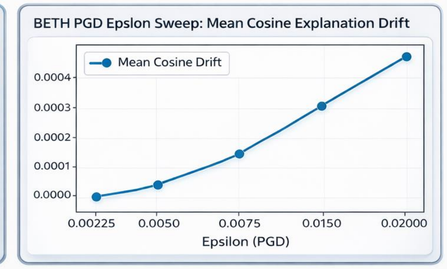
- Prediction-preserving constraints held across the controlled 500-sample evaluation subsets.
- Explanation drift was measurable for both IG and SHAP under FGSM and PGD.
- Detection remained near-perfect in the controlled setting, with ROC-AUC approaching 1.0.
- IG generally showed larger drift magnitudes on CICIDS2018, while SHAP remained highly discriminative.

Across all completed configurations, adversarial perturbations changed explanation structure without changing the predicted class label.

Validation Figures



Left: clean-vs-clean drift is negligible relative to adversarial drift.



Right: BETH PGD epsilon sweep shows monotonic drift growth with stronger perturbations.

Master Results Table

Dataset	XAI	Attack	Preserved	Mean cosine	Mean Euclidean	ROC-AUC
CICIDS2018	IG	FGSM	500/500	0.000316	0.173799	0.998
		PGD		0.000286	0.155030	1.000
	SHAP	FGSM		0.000175	0.073594	1.000
		PGD		0.000181	0.076310	1.000
BETH	IG	FGSM		0.000128	0.024850	1.000
		PGD		0.000127	0.024913	1.000
	SHAP	FGSM		0.000399	0.019551	1.000
		PGD		0.000399	0.019532	1.000

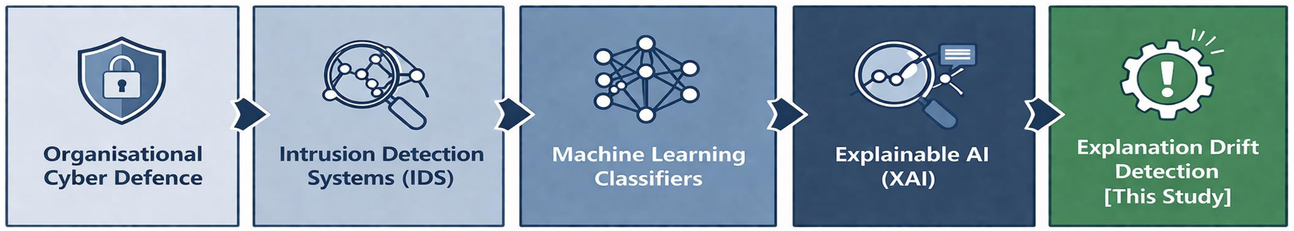
Conclusion

Adversarial perturbations consistently altered explanation structure while preserving the predicted class label. Explanation drift therefore acts as a promising complementary signal for detecting adversarial manipulation in IDS

Remaining Work

- Complete dissertation writing and integrate the full results narrative.
- Refine discussion, limitations, and future work sections.
- Prepare final screencast, viva, and industry showcase presentation.
- Prepare a conference or journal publication based on the results.

Research Context



Selected references: Goodfellow et al. (2015); Madry et al. (2018); Carlini and Wagner (2017); Lundberg and Lee (2017); Sunda rarajan et al. (2017).

Data and result summaries from completed dissertation experiments on CICIDS2018 and BETH.

PyTorch python jupyter SHAP AI → XAI

setu.ie