Enhancing Trustworthiness of Deep Learning-Based IDS

A Framework Combining Uncertainty Quantification and XAI

Majd Shalak

Telecom SudParis/Projet SuperviZ
Workshop on AI for Cybersecurity

Tuesday 16th September, 2025

Network Security: Towards Trustworthy AI-Driven IDS

Critical IDS Requirements

- **High accuracy**
- **Quantifiable trust**
- **Interpretability**
- **Adaptive learning**

DL: Promise & Pitfalls

+ **Superior performance**
- **No uncertainty metrics**
- **Black-box nature**
- **Adversarial vulnerability**

Research Challenge

Dilemma: Need DL accuracy + transparency for mission-critical security

Solution: Trustworthy DL

- **Uncertainty Quantification:** Conformal Prediction, MC-Dropout, BNNs
- **XAI Frameworks:** Interpretability for DL decisions

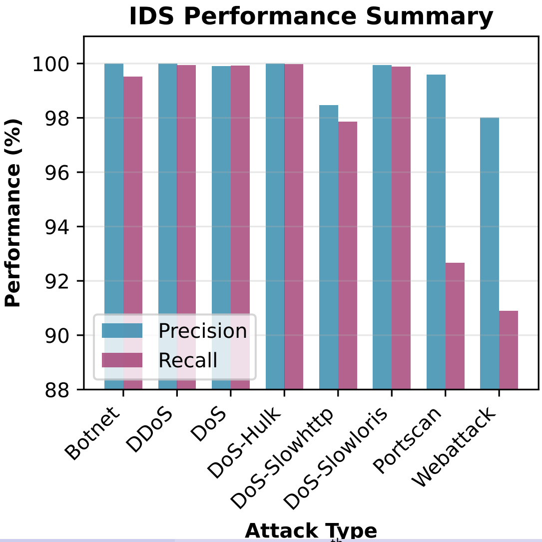
MLP Binary Classification for IDS: Results Overview

Methodology

- Model: Multi-Layer Perceptron (MLP) developed for binary classification
- Dataset: CIC-IDS2017 in NetFlow format.
- Highly imbalanced dataset (minority attack class)
- Each MLP is trained on a specific attack type

Results

- Excellent performance
- Precision: 98.0% - 99.9%
- Recall: 90.9% - 99.9%



Uncertainty Quantification with Monte-Carlo Dropout

Addressing the Critical Need for Confidence Estimation in IDS

Why Uncertainty Matters

- NNs provide **point predictions**.
- NNs can be overconfident even when wrong
- It allows to know *when not to trust* predictions

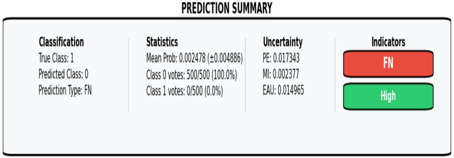
Monte-Carlo Dropout Solution

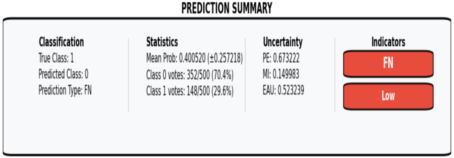
- Distinguish between:
  - Aleatoric: inherent data noise (irreducible)
  - Epistemic: model knowledge gaps (reducible)
- Transforms existing deterministic MLP into probabilistic model
- Minimal computational cost.

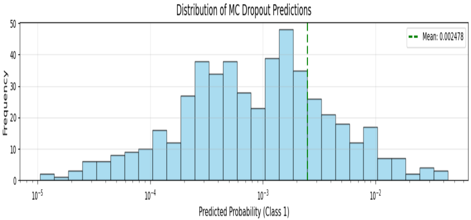
MCD Implementation Process

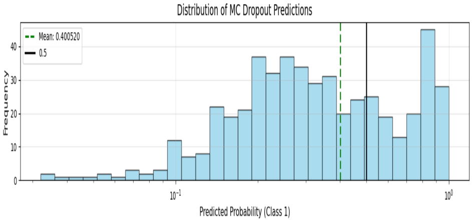
1. Enable dropout during inference
2. Perform T stochastic forward passes
3. Generate predictions $\{\hat{y}_t\}_{t=1}^{T}$
4. Compute mean prediction: $\bar{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$
5. Quantify uncertainty via:
   - **Variance**: $\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2$
   - **Entropy**: $H = -\sum_c p(c)\log p(c)$

Uncertainty with Monte-Carlo Dropout

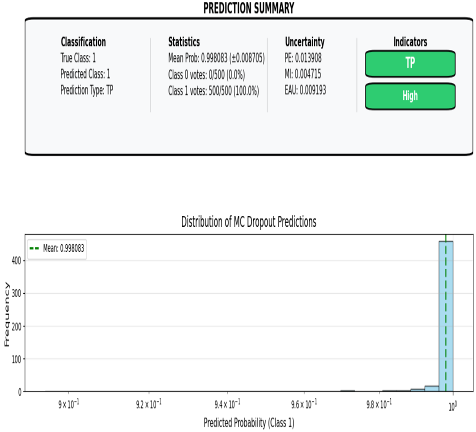


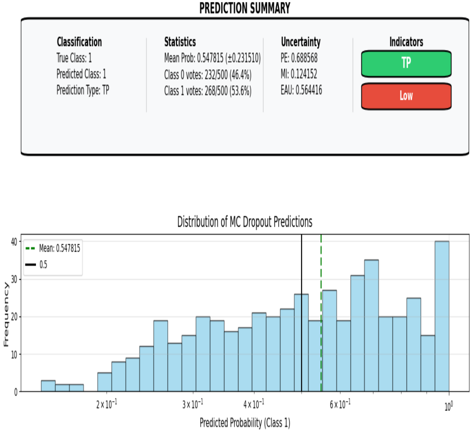






Uncertainty with Monte-Carlo Dropout





Explainable AI for Deep Learning-based IDS

Making Black-Box Security Decisions Transparent and Trustworthy

Why XAI is Critical

- ✓ **Trust & Compliance**: Transparency for analysts and regulations
- ✓ **Debugging**: Identify biases and failures
- ✓ **Knowledge Discovery**: Learn new attack patterns
- ✓ **FP/FN Analysis**: Understand misclassifications

Mathematical Formulation

Explanation method $g : (f, \mathbf{x}) \to \mathbf{r} \in \mathbb{R}^d$

- f: black-box classifier
- $\mathbf{x}$: d-dim feature vector
- $\mathbf{r}$: explanation vector
- $|r_i|$: feature importance
- $\text{sign}(r_i)$: contribution direction

XAI Methods for IDS

LIME

- Local linear approximations
- Perturbs input features around a baseline instance
- r_i = local model coefficient

SHAP

- Game theory-based
- Feature contribution scores w.r.t a baseline prediction
- r_i = Shapley value

Integrated Gradients

- Gradient-based attribution
- Path integration from baseline
- $r_i = (x_i - x_i') \times \int_0^1 \frac{\partial f}{\partial x_i} d\alpha$

Evaluating XAI Methods for IDS

Ensuring Reliable and Actionable Explanations for Security Analysts

Why Evaluate XAI?

- XAI explanations can **vary** between methods
- Analysts need **consistent** explanations

3 Evaluation Metrics

Intersection (Method Agreement)

- $IS = \frac{|R_i \cap R_j|}{k}$ where R_i, R_j are top-k features
- Measures *consensus* between XAI methods

Sparsity (Interpretability)

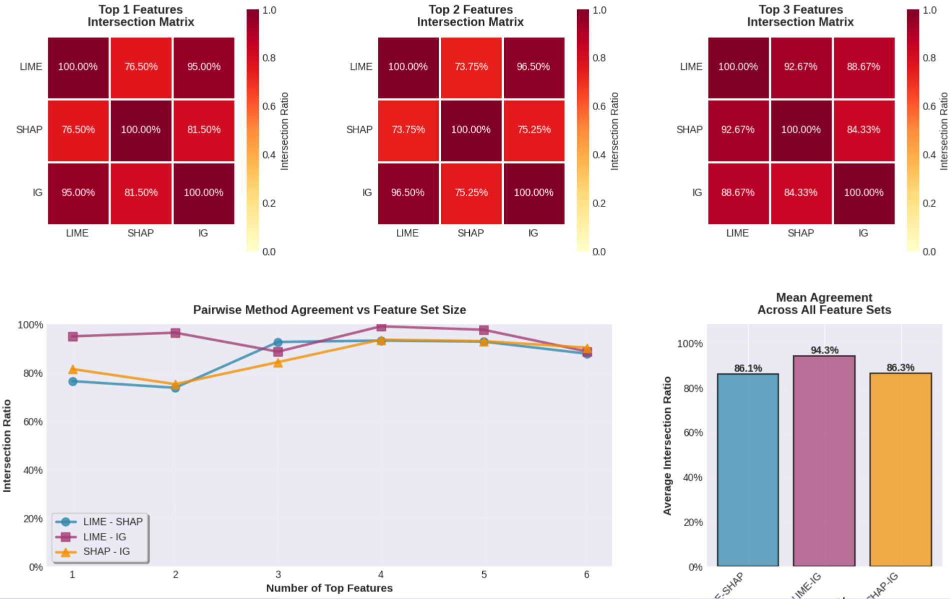
- Measures concentration of importance in *few key* features
- $\text{Sparsity}(k) = \frac{\sum_{i=1}^{k} |r_{(i)}|}{\sum_{j=1}^{d} |r_j|}$, where $|r_{(1)}| \geq |r_{(2)}| \geq \dots \geq |r_{(d)}|$

Stability (Reproducibility)

- Consistency across n independent runs
- Addresses *stochastic* nature of XAI (LIME and SHAP)

Reliable XAI

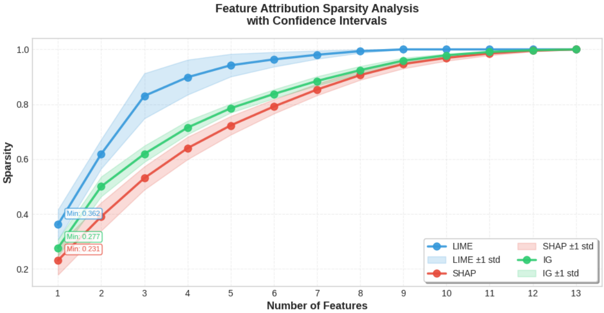
Reliable XAI = High Intersection ∩ Appropriate Sparsity ∩ Strong Stability

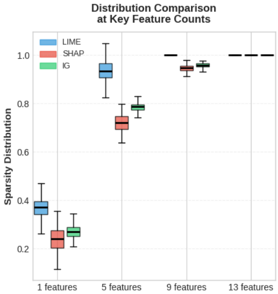


Top 1 Features
Intersection Matrix
LIME
SHAP
IG
100.00% 76.50% 95.00%
76.50% 100.00% 81.50%
95.00% 81.50% 100.00%
Top 2 Features
Intersection Matrix
100.00% 73.75% 96.50%
73.75% 100.00% 75.25%
96.50% 75.25% 100.00%
Top 3 Features
Intersection Matrix
100.00% 92.67% 88.67%
92.67% 100.00% 84.33%
88.67% 84.33% 100.00%
Intersection Ratio
0.0
0.2
0.4
0.6
0.8
1.0
Pairwise Method Agreement vs Feature Set Size
Intersection Ratio
Number of Top Features
0%
20%
40%
60%
80%
100%
1
2
3
4
5
6
LIME - SHAP
LIME - IG
SHAP - IG
Mean Agreement
Across All Feature Sets
Average Intersection Ratio
86.1%
94.3%
86.3%

Evaluating XAI Methods for IDS

Sparisity

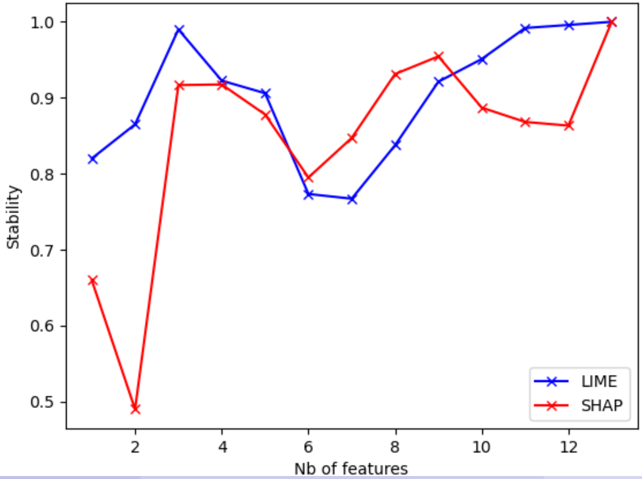




Figure

Evaluating XAI Methods for IDS

Stability



XAI-Guided Adversarial Attacks on IDS

Exploiting Explanations to Evade Detection

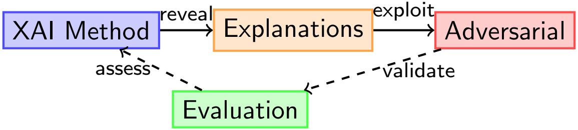
⚠ Vulnerability due to XAI

- XAI reveals **model's decision logic**
- Attackers can **strategically manipulate** traffic to **evade detection**.

Dual Purpose: Attack & Evaluation

The success of XAI-guided attacks validates explanation quality

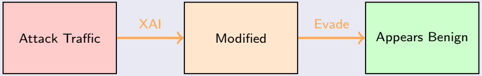
- ✓ **High evasion rate** ⇒ **High fidelity** explanations
- ✓ Poor attacks ⇒ Unreliable XAI
- ✓ Serves as **fidelity metric** for XAI methods



Attack Methodologies: TP-based vs FN-based

Strategic Manipulation of Network Traffic Using XAI

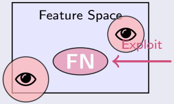
TP-based Attack



Strategy: Diminish attack signatures

- Aggregate TP explanations
- Target top-k attack indicators
- Shift features toward baseline
- Baseline is chosed s.t. $y_{\text{baseline}} \approx 0.5$

FN-based Attack

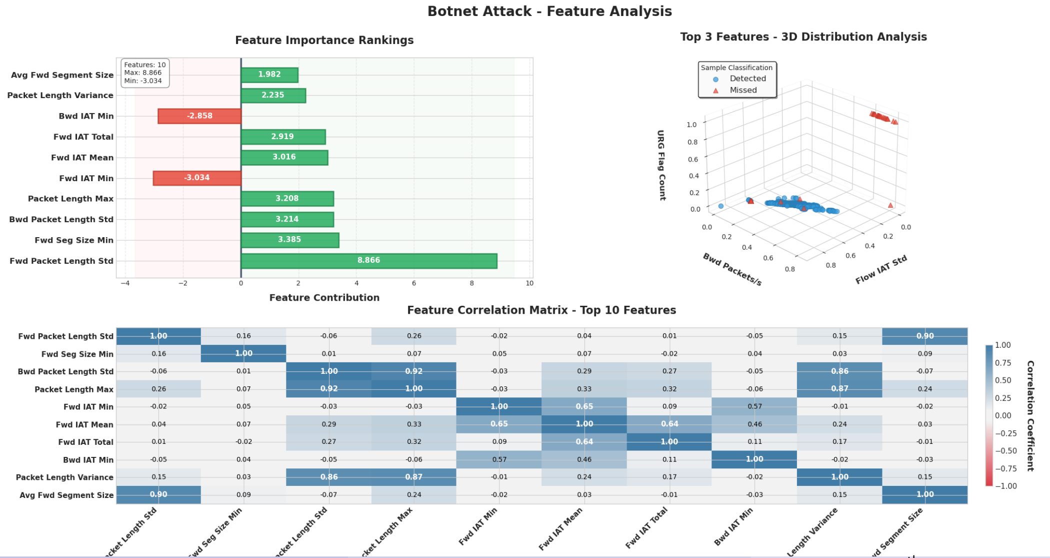


Strategy: Exploit model blind spots

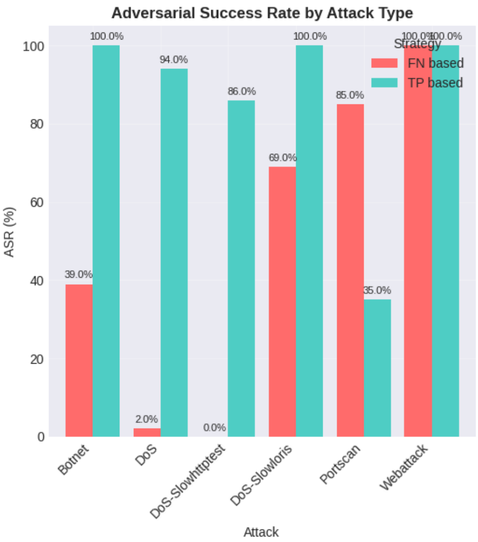
- Map FN explanations to find patterns
- Cluster vulnerable feature regions
- Craft attacks mimicking FN profile:

Successful attacks validate XAI's ability to identify decision boundaries and model vulnerabilities

XAI-driven adversarial examples



XAI-driven adversarial examples



Adversarial Attack Constraints: Ensuring Realistic Evasion

Balancing Attack Effectiveness with Practical Feasibility and Stealth

Constraint

Principle: Minimal realistic perturbation

- **Minimize** number of modified features
- **Correlation Independence:** Avoid features correlated with top-k important features
- **Backward Features Restriction** Only manipulate attacker-controllable features (e.g. we do not perturb network response features)
- **Benefits:**
  - ✓ Prevents cascading effects
  - ✓ Maintains feature independence
  - ✓ Reduces implementation complexity

Main Direction To Improve

Generate attacks in problem space.

Predictive Entropy as Adversarial Fingerprint

How Adversarial Examples Reveal Themselves Through Uncertainty Patterns

Entropy Behavior in Adversarial Examples

Key Finding: Adversarial examples typically exhibit **higher** Predictive Entropy (PE)

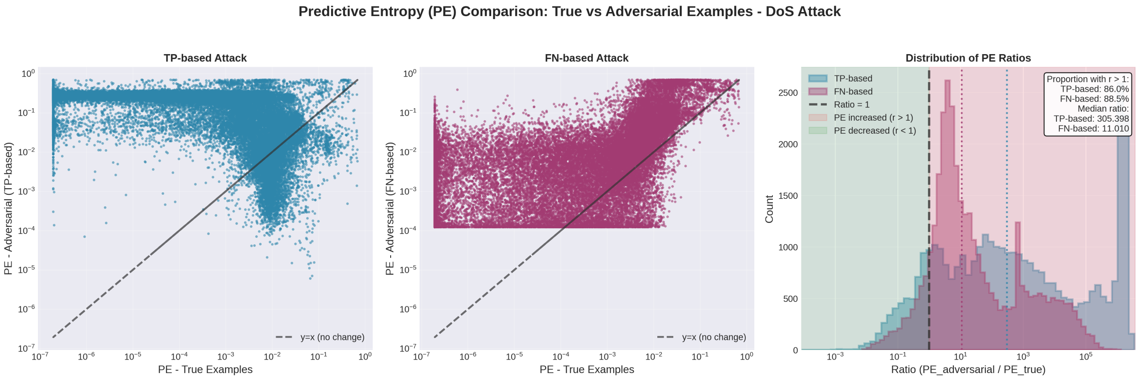
Why This Happens:

- Perturbations push inputs toward **decision boundaries**
- Model predictions become **inherently uncertain**

⚠ Points to consider

- **Not Universal:** Some adversarial examples show **decreased** PE
- Compare the attack and defense strategy with classical adversarial methods

PE as adversarial fingerprint



Implications for Defense

Detection Opportunity:

- PE serves as **statistical fingerprint**
- Quantitative signal for detection

Detection Challenge:

- Not all adversarial examples have high PE
- Need multi-signal detection approach

Conclusion: Towards Trustworthy DL-IDS

Key Contributions

- **Uncertainty Quantification:** Integrating MC-Dropout into DL-IDS with minimal computational overhead.
- **XAI Evaluation Framework:** Computing evaluation metrics (Intersection, Sparsity, Stability) to ensure reliable explanations.
- **Dual Purpose Adversarial Analysis:** XAI-guided attacks serve as both vulnerability assessment and XAI fidelity validation
- **Entropy-Based Defense:** Identified predictive entropy as statistical fingerprint for adversarial detection

Future Directions

- Extend to other NN architectures (Auto-encoders, transformers..)
- Create unified trustworthiness framework
- Real-time deployment

Thank You!