

An Evaluation of Explainable Machine Learning Techniques for Network Intrusion Detection and Digital Forensic Incident Reconstruction

Keval Mahendra Motivaras

MSc Cybersecurity and Digital Forensics · University of Wales Trinity Saint David

CICIDS2017

XGBoost

SHAP + LIME

Forensic Evidence

Problem

Interpretability becomes an issue for many of the intrusion-detection models that perform well. The consequences of false positives and negatives are still important despite the accuracy level of the model.

Aim

Evaluate if explainable machine learning can improve understanding of decisions on network intrusion and support evidence-based digital forensics investigation.

Scope

DoS attack traffic from CICIDS2017 dataset for Benign/Attack Binary Classification, having 691,406 flows with 78 numerical features.

Research questions

RQ1

How effectively do DT, RF and XGBoost classify the traffic?

RQ2

Which features influence XGBoost predictions?

RQ3

Are SHAP and LIME explanations consistent and stable?

RQ4

How far can XAI support forensic-oriented reconstruction?

Domain Understanding

Three domains connected in one research problem

1

Network Intrusion Detection

- Supervised flow classification
- Benign vs Attack
- False positives / false negatives

2

Explainable AI (XAI)

- SHAP: global + local attribution
- LIME: local surrogate explanations
- Agreement and stability evaluation

3

Digital Forensics

- Evidence traceability and provenance
- Link prediction back to source flow
- Timeline requires timestamps + identities

Core concept: Detection accuracy $\neq$ Explanation reliability $\neq$ Forensic truth

Methodology

Controlled pipeline from data preparation to forensic-oriented interpretation



Final workflow: Detect → Explain → Compare → Test Stability → Analyse Errors → Link to Evidence

80/20 stratified split · random_state = 42 · same experimental conditions for model comparison

Environment

Python · Google Colab ·
Pandas · NumPy · scikit-learn ·
XGBoost

Explainability

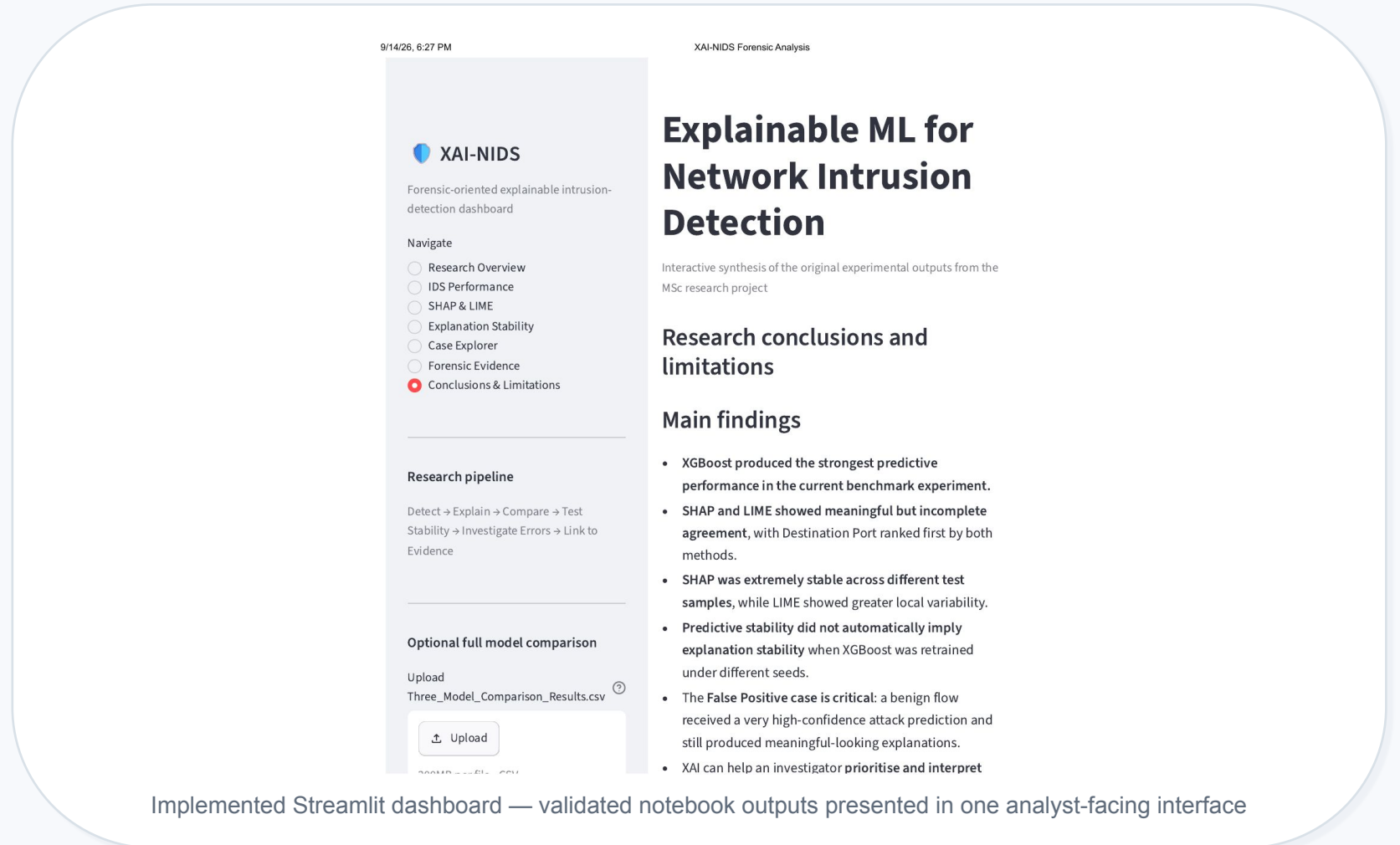
SHAP TreeExplainer · LIME
Tabular Explainer · same TP /
FP / FN cases

Evaluation

Accuracy · Precision · Recall ·
F1 · Jaccard · Spearman ·
Kendall · Cosine

Forensic layer

Source-record linkage ·
SHA-256 provenance ·
capability assessment



Limitations & Research Gap

Why the study goes beyond accuracy and explanation plots

Research gap

- Conclusion of most of the XAI-NIDS systems ends up with high levels of accuracy alongside SHAP/LIME visualization.
- Little consideration is done in testing of cross-method consistency and stability of explanation.
- Limited connection between model explanation and forensic evidence left for investigation.
- Prediction explanation is often confused with incident explanation.

Study limitations

- CICIDS2017 dataset with a DoS bias; binary classification task
- Flow level random split of train/test data; benchmark-dependent outcome
- SHAP and aggregated LIME use different sampling methods
- No presence of Source IP, Destination IP, Timestamp, Flow ID, or payload
- Deep investigation limited to select true positives, false positives, and false negatives

Contribution: evaluate explanations, stress-test stability, inspect model errors and link outputs back to source evidence.

Results — Key Findings

Headline results from the completed experimental pipeline

99.9747%

Accuracy · XGBoost

99.9444%

Precision · XGBoost

99.9861%

Recall · XGBoost

99.9652%

F1-score · XGBoost

Cross-XAI agreement

Top-10 common features: 7/10
Jaccard: 0.5385
Spearman: 0.6458
Destination Port ranked #1 by both

Explanation stability

SHAP global Top-10 Jaccard: 1.0000
LIME local Jaccard: TP 0.5824 · FP
0.7906 · FN 0.8404
Stable predictions ≠ identical
explanations

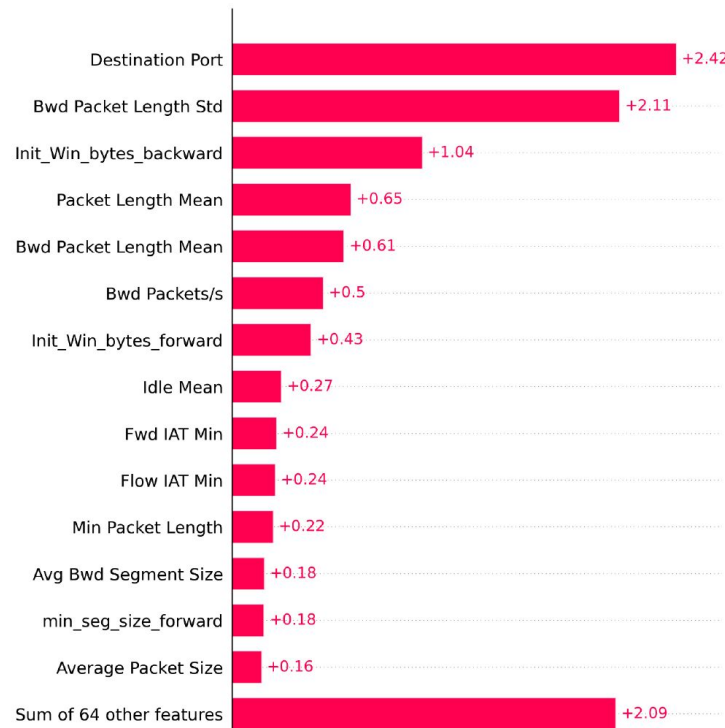
Critical error case

False Positive: BENIGN → Attack
Attack probability: 99.7936%
SHAP + LIME still produced plausible
explanations
Explanation ≠ correctness

XGBoost errors: 28 false positives · 7 false negatives

Results — Proof from XAI Outputs

Actual figures produced from the experimental analysis



SHAP global feature importance — Destination Port ranked highest

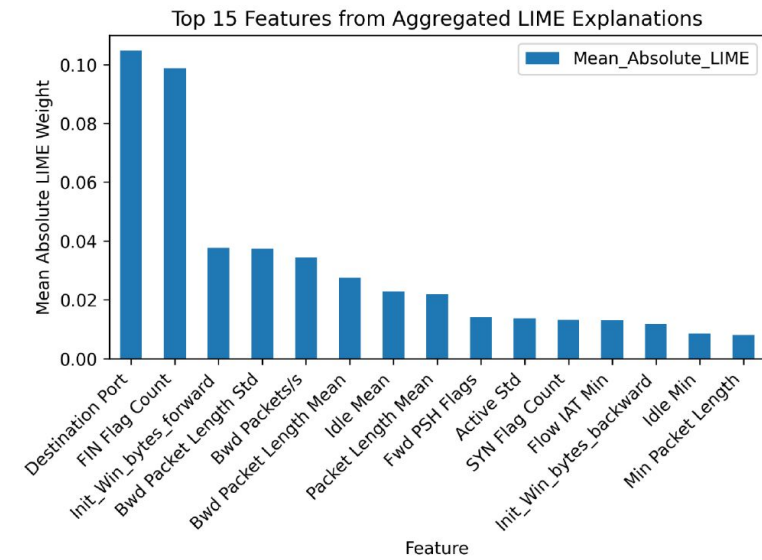


Figure 4: Aggregated LIME Feature Importance

Aggregated LIME importance — Destination Port also ranked highest

Key evidence • Meaningful overlap, but not identical rankings • 7 common Top-10 features • Method choice affects explanation emphasis

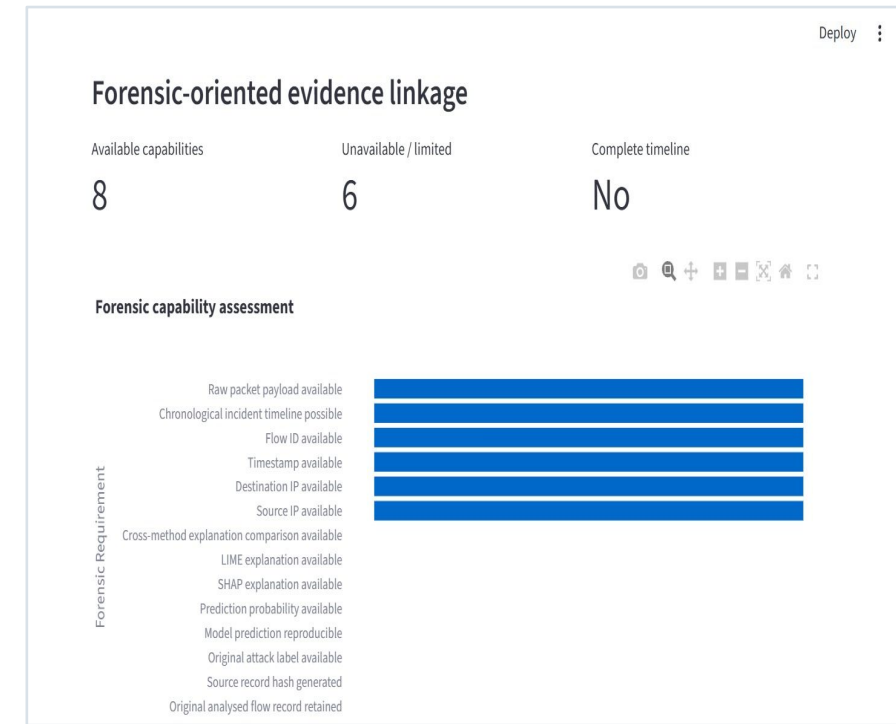
Results — Forensic Proof & Practical Output

Evidence linkage is useful, but full chronological reconstruction is not supported

	Forensic Requirement	Available
0	Original analysed flow record retained	Yes
1	Source record hash generated	Yes
2	Original attack label available	Yes
3	Model prediction reproducible	Yes
4	Prediction probability available	Yes
5	SHAP explanation available	Yes
6	LIME explanation available	Yes
7	Cross-method explanation comparison available	Yes
8	Source IP available	No
9	Destination IP available	No
10	Timestamp available	No
11	Flow ID available	No
12	Chronological incident timeline possible	No
13	Raw packet payload available	No - not present in ML CSV

Figure 5: Forensic capability assessment from Notebook 11.

This separation represents an important border in the experiment findings: Explainable AI
Forensic capability assessment — what evidence is available vs missing



Dashboard integrates model performance, XAI, stability, errors and forensic evidence

False Positive lesson: a detailed explanation can explain the model decision without proving that the decision is true.

Future work

- Expand to multi-class attack settings and use extra IDS datasets.
- Use temporal or scenario-based dataset partitioning and perform cross-dataset validation.
- Maintain more evidence information, such as IP addresses, timestamps, Flow IDs, PCAPs, and host logs.
- Evaluate faithfulness, robustness, and other error types.
- Create an automated near-real-time analyst process.

Conclusion

- XGBoost provided the best results according to the current testing criteria.
- SHAP and LIME help ensure greater transparency but need assessment.
- Explanation stability and prediction stability mean different things.
- Explainable AI can be used for forensic flow analysis; it is not evidence.
- High accuracy is not enough for forensic-based intrusion detection.

Detect → Explain → Compare → Test Stability → Analyse Errors → Link to Evidence

THANK YOU

Questions?

Core research message

Detective accuracy is not enough. The explanations have to be tested against the evidence before applying them to the forensics process.