SHORT-PAPER

# Multi-Viewpoint Evaluation of Explanation Quality in X-IDS Using Aggregated and Consensus Metrics

**MOHAMMED ALQULITI**, University of Southampton, Southampton, Hampshire, U.K.

**ERISA KARAFILI**, University of Southampton, Southampton, Hampshire, U.K.

**BOOJOONG KANG**, University of Southampton, Southampton, Hampshire, U.K.

**Open Access Support** provided by:

**University of Southampton**



**Published:** 10 September 2025

**Citation in BibTeX format**

SIGCOMM '25: ACM SIGCOMM 2025 Conference
*September 8 - 11, 2025*
*Coimbra, Portugal*

**Conference Sponsors:**
SIGCOMM


https://doi.org/10.1145/3744969.3748440
ISBN: 9798400720260

# Poster: Multi-Viewpoint Evaluation of Explanation Quality in X-IDS Using Aggregated and Consensus Metrics

Mohammed Alquliti
University of Southampton
Southampton, United Kingdom
M.H.Alquliti@soton.ac.uk

Erisa Karafili
University of Southampton
Southampton, United Kingdom
E.Karafili@soton.ac.uk

BooJoong Kang
University of Southampton
Southampton, United Kingdom
B.Kang@soton.ac.uk

## ABSTRACT

Explainable intrusion detection systems (X-IDS) typically provide a set of explanations for each alert, while provided explanations may not be sufficient for security analysts to make time-critical decisions. Existing evaluation methods only consider such cases with a single set of explanations, as a result, limiting the scope of evaluation. This work expands a set of explanation evaluation metrics, our earlier work, to extend the scope of evaluation covering X-IDS providing multiple sets of explanations. Additional intermediate metrics are proposed to capture characteristics of multiple sets of explanations so the evaluation metrics can be computed for the multiple sets of explanations. The experimental results show the proposed metrics reveal more insights.

## CCS CONCEPTS

• **Security and privacy** → **Intrusion detection systems**; • **Computing methodologies** → **Machine learning**; **Artificial intelligence**.

## KEYWORDS

Explainability, XAI, Explanation Evaluation, IDS.

**ACM Reference Format:**

Mohammed Alquliti, Erisa Karafili, and BooJoong Kang. 2025. Poster: Multi-Viewpoint Evaluation of Explanation Quality in X-IDS Using Aggregated and Consensus Metrics. In *ACM SIGCOMM 2025 Posters and Demos (SIGCOMM Posters and Demos '25), September 8–11, 2025, Coimbra, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3744969.3748440

## 1 INTRODUCTION

Nowadays, machine learning and deep learning intrusion detection systems (ML/DL-based IDSs) can detect and analyse complex attacks and patterns [1–3, 5, 7, 9]. However, ML/DL-based IDSs need explainable artificial intelligence (XAI) to explain the reasoning behind their decisions [2].

ML/DL-based IDSs equipped with explainable AI (X-IDSs) are increasingly used to help security analysts understand why an attack is detected [2, 3]. However, current evaluation metrics assess the quality of only one explanation per a predicted attack class [1–3]. In practice, IDSs can yield the same prediction from different training subsets [10], each accompanied by multiple explanations (viewpoints). Relying on one viewpoint can hide important domain indicators, while presenting all viewpoints simultaneously can overwhelm analysts. Currently, there is no established way to quantify the quality of multiple explanations regarding additional coverage provided, and the increased confidence in indicators revealed when multiple explanations agree. Therefore, this gap needs to be addressed for using multiple explanations produced by X-IDS that analysts can trust.

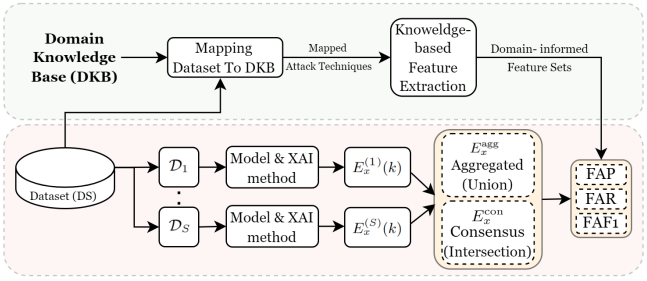


**Figure 1: Evaluation pipeline. Each of the $S$ data subsets produces its top-$k$ explanation $E_x^{(s)}(k)$. We form an Aggregated (union) set and a Consensus (intersection) set, then compute FAP, FAP, and FAF1 against the domain-informed feature sets.**

To address this gap, our work builds on the feature-alignment metrics introduced in our earlier study [1], which quantified how well an X-IDS that produces a single explanation aligned with domain-informed feature sets derived from expert knowledge bases. In this work, we extend that foundation to a multi-viewpoint setting. For each network traffic instance, we generate several distinct explanations produced from separate data subsets and evaluate them collectively. We evaluate these viewpoints through: an *aggregated* (union) set that combines all features highlighted by any explanation, and a *consensus* (intersection) set that captures only the features selected by all explanations. These perspectives enable quantifying the potential coverage gained from the additional explanations and the high-confidence features that occur in every explanation.

The following sections describe our metrics and present some preliminary results.

## 2 METHODOLOGY

Our previous work introduced evaluation metrics that quantify how well an X-IDS that produces a *single* top-$k$ explanation aligns with a domain-informed knowledge of an attack [1]. For each feature


*SIGCOMM Posters and Demos '25, September 8–11, 2025, Coimbra, Portugal*

ACM ISBN 979-8-4007-2026-0/25/09.
https://doi.org/10.1145/3744969.3748440

vector of a test instance $x$, an XAI method $g(f, x)$ yields an ordered set of features $E_x(k)$ where $k$ is the specified number of most influential features. Then, we computed Feature-Alignment Precision (FAP), Feature-Alignment Recall (FAR), and their harmonic mean (FAF1) by comparing the XAI method's top-$k$ feature set $E_x(k)$ with the domain-informed feature set associated with the predicted class $F_c$. These metrics revealed the explanation's correctness (FAP), and completeness (FAR) at instance, class, and dataset levels.

We now extend that work to scenarios in which an X-IDS produces multiple explanations for the same sample (e.g., by training on different data subsets). Considering these multiple explanations can help security analysts by revealing additional domain-informed features that might be missed when relying on a single explanation. At the same time, recurring important features across multiple explanations can provide a robust basis for efficient and reliable triage.

Accordingly, let $S$ denote the number of distinct explanations generated for the same feature vector of an $x$. For each explanation, we obtain a top-$k$ feature set $E_x^{(1)}(k), \ldots, E_x^{(S)}(k)$ for the same feature vector of an instance $x$. We combine these sets to capture two alternative viewpoints:

***Aggregated Set (Union):*** This set maximises coverage of potentially domain-informed features by collecting every feature flagged as important by *any* of the $S$ explanations. We define the *aggregated* explanations for instance $x$ at cutoff $k$ as the union of all $S$ individual top-$k$ sets:

$$E_x^{\text{agg}}(k) = \bigcup_{s=1}^{S} E_x^{(s)}(k). \tag{1}$$

***Consensus Set (Intersection):*** This set represents high-confidence features that all the explanations deem important for the same instance. The *consensus* set keeps only the features that every single explanation agrees on:

$$E_x^{\text{con}}(k) = \bigcap_{s=1}^{S} E_x^{(s)}(k). \tag{2}$$

***Union-Based Metrics:*** $\text{FAP}_I^{\text{agg}}(k)$ measures the fraction of features in the aggregated set that are present in the domain-informed feature set $F_c$, while $\text{FAR}_I^{\text{agg}}(k)$ measures the fraction of the domain-informed features that appear in the aggregated set:

$$\text{FAP}_I^{\text{agg}}(k) = \frac{|E_x^{\text{agg}}(k) \cap F_c|}{|E_x^{\text{agg}}(k)|}, \quad \text{FAR}_I^{\text{agg}}(k) = \frac{|E_x^{\text{agg}}(k) \cap F_c|}{|F_c|} \tag{3}$$

***Consensus-Based Metrics:*** $\text{FAP}_I^{\text{con}}(k)$ quantifies how many features in the consensus set that belong to the domain-informed feature set, and $\text{FAR}_I^{con}(k)$ indicates how much of the domain-informed feature set the consensus set covers:

$$\text{FAP}_I^{\text{con}}(k) = \frac{|E_x^{\text{con}}(k) \cap F_c|}{|E_x^{\text{con}}(k)|}, \quad \text{FAR}_I^{\text{con}}(k) = \frac{|E_x^{\text{con}}(k) \cap F_c|}{|F_c|} \tag{4}$$

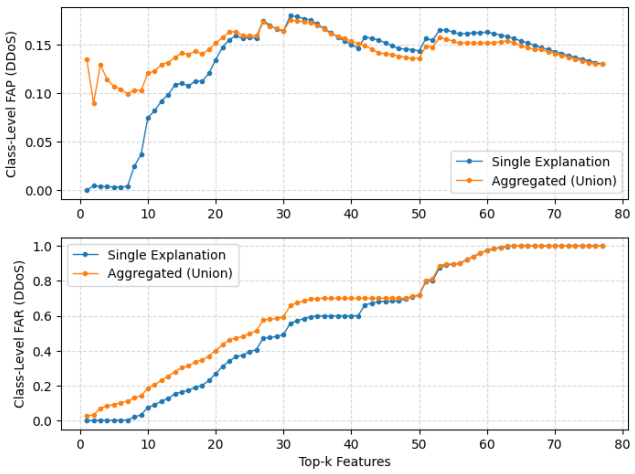


**Figure 2: Comparison of single- and multi-explanation results for DNN-based IDS on DDoS/DoS class. Top: class-level FAP. Bottom: class-level FAR. Aggregated union sets are obtained from two explanations ($S = 2$).**

For each instance, we compute the union- and consensus-based metrics. Then, we average them at two levels: **Class-level evaluation** shows explanation performance per attack class. **Dataset-level evaluation** provides a single score that reflects overall coverage and confidence over the dataset [1].

## 3 PRELIMINARY RESULTS

To evaluate our approach, we construct two $S = 2$ disjoint balanced training subsets from CICIDS-2017 dataset [7, 9]. We train a deep neural network (DNN)-based IDS with, hyperparameters from their original works [2], on the training subsets to produce two top-$k$ explanations per test instance. We use SHAP to produce these explanations [8]. We derive domain-informed feature sets from MITRE ATT&CK and D3FEND [4, 6] as described in our previous work [1]. Now, we compute aggregated and consensus metrics and show the results for class-level evaluation.

*First Attempts At Validation.* We ran the new metrics on the DNN-based IDS using multiple explanations ($S = 2$) for every instance in the DoS/DDoS class. We computed class-level FAP and FAR for both the single explanation and the aggregated union across all $k$ values. Figure 2 shows improvement at the union FAP and FAR already compared to single explanation. These early observations indicate that adding just one additional explanation (viewpoint) reveals domain-informed features with better correctness, validating the practical benefit of the proposed metrics.

## 4 FUTURE WORK

In future work, we plan to extend our evaluation to additional attack types and further investigate how the proposed metrics correlate with real-world analyst decision-making. We also aim to explore the computational cost of multiple explanation setups and improve scalability through efficient explanation aggregation.

# REFERENCES


[1] Mohammed Alquliti, Erisa Karafili, and BooJoong Kang. 2025. Evaluating Explanation Quality in X-IDS Using Feature Alignment Metrics. *arXiv preprint arXiv:2505.08006* (2025). https://arxiv.org/abs/2505.08006

[2] Osvaldo Arreche, Tanish Guntur, and Mustafa Abdallah. 2024. XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems. *Applied Sciences* 14, 10 (2024), 4170.

[3] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Claudio Stanzione. 2022. Explainable Artificial Intelligence in Cybersecurity: A Survey. *IEEE Access* 10 (2022), 93575–93600.

[4] Cybersecurity and Infrastructure Security Agency (CISA). 2023. *Best Practices: MITRE ATT&CK® Mapping*.

[5] Huiyao Dong and Igor Kotenko. 2025. Cybersecurity in the AI era: analyzing the impact of machine learning on intrusion detection. *Knowledge and Information Systems* (2025), 1–52.

[6] Peter E. Kaloroumakis and Michael J. Smith. 2021. *Toward a Knowledge Graph of Cybersecurity Countermeasures.* Technical Report. The MITRE Corporation. https://d3fend.mitre.org/resources/D3FEND.pdf

[7] Kahraman Kostas. 2018. *Anomaly Detection in Networks Using Machine Learning.* Master's thesis. University of Essex, Colchester, UK.

[8] Scott M. Lundberg and SuIn Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NeurIPS 2017*. 4765–4774.

[9] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *ICISSP 2018*. 108–116.

[10] Zichen Zhang, Shanshan Kong, Tianyun Xiao, and Aimin Yang. 2024. A Network Intrusion Detection Method Based on Bagging Ensemble. *Symmetry* 16, 7 (2024), 850.