Intrinsic and Post-Hoc Explainability for Anomaly Detection in Network Intrusion Detection Systems

Philipp Denzel [1], Marc Stadelmann [1], Yago Lizarribar Carillo [2], Gérôme Bovet [2], Frank-Peter Schilling [1], and Jasmina Bogojeska [1]
[1]*Centre for Artificial Intelligence (CAI)*, *Zurich University of Applied Sciences (ZHAW)*, Winterthur, Switzerland
{denp,stmd,scik,bogo}@zhaw.ch
[2]*Cyber-Defence Campus*, *armasuisse S+T*, Thun, Switzerland
{yago.lizarribarcarillo,gerome.bovet}@armasuisse.ch

***Abstract*—As cyber threats grow in sophistication, traditional rule-based defences increasingly fail to detect novel attacks, prompting a shift toward deep learning for anomaly-based intrusion detection systems. While transformer architectures have achieved state-of-the-art performance in detecting anomalous signatures in complex, high-dimensional network telemetry, the inherent opacity of neural networks hampers operational adoption, trust, auditability, and compliance. This paper addresses these challenges by presenting a rigorous evaluation of the Anomaly Transformer architecture as intrusion detection systems on modern benchmarks. Crucially, we introduce a purpose-built explainability framework for intrusion detection that unifies model-intrinsic and post-hoc explanations within a time-aware setting. By systematically contrasting the model's attention-based association mechanism with time-aware KernelSHAP explanations, we observe strong complementarity: while intrinsic mechanisms accurately localize anomalous events in time, post-hoc methods identify distinct feature attributions. Our findings highlight that relying on a single explanation modality is insufficient and motivate multifaceted explainability paradigms essential for delivering comprehensive, trustworthy, and actionable explanations in operational cybersecurity contexts. They are intended to support security analysts who require both temporal localization of anomalies and feature-level attribution for post-incident auditing and operational trust.**

***Index Terms*—Deep learning, Explainable AI, Network security, Intrusion detection, Anomaly detection, Time series analysis**

I. Introduction

The domain of cybersecurity has witnessed a fundamental paradigm shift necessitating the rapid evolution of Intrusion Detection Systems (IDS). As the global digital ecosystem expands to encompass the Internet of Things, cloud-native architectures, and high-speed 5G networks, the attack surface available to malicious actors has grown exponentially. Historically, network security has relied on signature-based IDS, which function on a deterministic model by comparing traffic against a database of known attack patterns. While effective against established threats, these rule-based approaches are inherently reactive and struggle to adapt to the dynamic nature of modern network traffic, making them ineffective against novel or evolving zero-day threats.

To address these systemic limitations, the research community has pivoted towards anomaly-based detection using Artificial Intelligence (AI) (cf. [1]). Unlike their signature-based counterparts, anomaly detection systems operate by establishing a baseline of "normal" behaviour and flagging deviations as potential intrusions. Recent advances in Deep Learning (DL) have significantly improved the accuracy of these systems (see [2, 3] for recent reviews). In its simplest form, DL-based anomaly detection employs auto-encoders [4, 5] trained exclusively on benign traffic to learn compressed latent representations of normal behaviour. During inference, anomalies are flagged when the reconstruction error exceeds a predefined threshold, as the model fails to accurately reconstruct patterns it has not learned. More broadly, DL models capture complex, non-linear representations of high-dimensional network traffic data, automating feature extraction. In particular, Transformer-based architectures have demonstrated state-of-the-art (SOTA) performance (see [6]), leveraging self-attention mechanisms to model long-range dependencies in longer packet sequences that traditional methods often miss.

However, the operational adoption of AI-driven IDS introduces a critical challenge: the "black box" problem. Deep neural networks are inherently opaque; a high-accuracy alert is of limited value if security analysts cannot quickly extract human-understandable reasoning to validate the threat and initiate remediation. Explainability is therefore not merely a feature but a requirement for fostering trust, auditability, and regulatory compliance in cybersecurity environments.

This paper addresses the interpretability gap by coupling the *Anomaly Transformer* (AT) architecture [7] with a novel dual-track explainability framework and conducting a rigorous, systematic evaluation of the approach for intrusion detection systems.

We contribute to the field in two primary ways: **(1)** *High-realism benchmarking*: we evaluate on BCCC-CIC-IDS-2017 [8], a corrected benchmark superseding legacy datasets such as KDD99 [9] and CIC-IDS-2017 [10], providing a realistic representation of modern attack vectors. **(2)** *Domain-specific explainability analysis*: we systematically contrast the AT's intrinsic association discrepancy with KernelSHAP attributions—a combination not previously explored in the IDS context.

Our findings reveal that these two explanation modalities are complementary rather than redundant. While intrinsic mechanisms excel at the temporal localization of anomalies, post-hoc methods provide distinct feature attributions, suggesting that

2026 IEEE Swiss Conference on Data Science and AI (SDS) | 979-8-3195-0600-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/SDS70563.2026.00023

2835-3420/26/$31.00 ©2026 IEEE
DOI 10.1109/SDS70563.2026.00023

a multifaceted approach is essential for auditor-ready AI.

We acknowledge that network intrusion detection operates in an inherently adversarial environment. Consistent with best practices outlined by [11], our evaluation uses a realistic, corrected benchmark (BCCC-CIC-IDS-2017) to mitigate concept drift and data leakage concerns common in legacy IDS datasets. A full adversarial robustness certification is beyond the scope of this applied study and is deferred to future work.

The remainder of this article is structured as follows: Section II reviews related work in AI-based intrusion detection systems and explainability methods. Section III presents our methodology, detailing the AT architecture and our dual-track explainability framework combining intrinsic and post-hoc explanations. Section IV describes the experimental setup, including datasets and model configurations. Section V presents detection performance results and a systematic explainability analysis comparing intrinsic attention mechanisms with time-aware KernelSHAP attributions. Finally, Section VI concludes with a discussion of implications for operational deployment and directions for future work.

II. Related work

AI-based IDS have fragmented into distinct architectural paradigms, each addressing the high dimensionality and non-linearity of network traffic differently.

A. Deep learning approaches for IDS

Recurrent Neural Networks (RNNs) [12] and **Long Short-Term Memory** (LSTM) [13] networks are widely employed to capture the sequential nature of TCP/IP sessions, where the legitimacy of packets is context-dependent. For instance, *OmniAnomaly* [14] utilizes stochastic RNNs to model robust anomaly detection in server telemetry. Recent innovations include bi-directional LSTMs (BiLSTMs) [15, 16] that process flows in both directions to better infer context. Although RNN-based methods such as OmniAnomaly represent strong baselines, their sequential processing introduces substantial training and inference latency that makes direct comparison with Transformer-based models architecture-dependent.

Convolutional Neural Networks (CNNs) [17, 18] approach IDS by treating network flows as images or spatial matrices, using kernels to extract local corrections between adjacent bytes or flags, e.g. [19] or [20]. While efficient, standard CNNs lack long-term memory, often failing to detect "slow" attacks that unfold over extended duration. To increase detection sensitivity to anomalies at various time scales, some models, such as *TimesNet* [21], involve a multi-scale, multi-periodicity approach using different kernel sizes.

Transformers [22] have recently emerged as the leading architecture for multivariate time-series anomaly detection, surpassing CNN-LSTM hybrids by leveraging self-attention to model long-range dependencies simultaneously. Architectures such as *RTDIS* [23] and *IDS-MTran* [24] utilize stacked encoders to handle high-dimensional traffic. The evaluation of anomaly detection methods poses particular challenges due to class imbalance and threshold sensitivity. [25] propose *TranAD*, a deep transformer network for multivariate time-series anomaly detection that incorporates adversarial training and provides a systematic evaluation protocol including threshold-independent metrics, but has so far not been applied in a realistic IDS scenario. More broadly, threshold-independent metrics such as AUROC and AUPRC are increasingly recommended over threshold-dependent metrics [7], a practice we adopt in Section V.

B. Explainability in Network Security

Trust in AI-driven IDS relies on bridging the gap between detection accuracy and interpretability. Approaches in this domain are generally grouped into post-hoc or intrinsic mechanisms [26].

Post-hoc methods approximate black-box models with interpretable surrogates and attribute predictions or parts thereof to specific features; LIME [27] and SHAP [28] are among the most popular techniques in this category. The first provides linearized local approximations, while the latter utilizes cooperative game theory to attribute feature importance consistently. These methods are model-agnostic but can be computationally expensive and sensitive to background distribution choices in time-series contexts.

Intrinsic methods exploit model architecture for transparency. Graph Neural Networks (GNNs), for instance, provide a natural and direct representation of a network system's node-based topology, where hosts are modelled as nodes and communication flows as edges. Accordingly, [29] introduced *GNN-IDS* to provide such node-level explanations. For Transformers, attention weights offer an intrinsic mechanism for visualizing model focus. [30] demonstrated this with *Roulette*, a model that highlights relevant traffic features via attention maps. While recent work has explored latent-space interpretability [31] and counterfactuals [32], this paper focuses on unifying intrinsic attention-based signals with post-hoc SHAP analysis to provide a comprehensive audit trail for security analysts.

III. Methodology

This section describes our methodological framework consisting of two components: the AT architecture for detection, and a dual-track explainability framework combining intrinsic and post-hoc explanations.

A. Anomaly Transformer Architecture

The AT [7] architecture was selected based on three criteria. First, its *association discrepancy* mechanism provides a native, intrinsic explanation signal directly linked to the anomaly score, a property absent in most deep IDS architectures. Second, it achieves state-of-the-art results on established time-series benchmarks while remaining applicable to multivariate flow-level telemetry without architectural modification. Third, its unsupervised formulation is well-suited to the IDS setting, where ground-truth attack labels are scarce or delayed in operational environments.

In a standard Transformer, self-attention computes pairwise associations between all time steps, allowing each position

to attend freely to any other position in the sequence. This flexibility is powerful for capturing long-range dependencies but provides no inherent mechanism for distinguishing normal from anomalous patterns. The AT modifies this mechanism by introducing a second, parallel association branch. The key insight is that anomalies, being rare and contextually isolated, tend to exhibit concentrated, local associations with nearby points; they lack the recurring context that would connect them to distant points. Normal points, on the other hand, participate in broader patterns (e.g., periodic behaviours, trends) and thus maintain associations across the entire series.

To operationalize this insight, the model architecture employs an *anomaly-attention* mechanism with two branches:

- a *prior-association* $\mathcal{P}$ computed using a learnable Gaussian kernel, encoding the assumption that temporally adjacent points *should* be more strongly associated.
- a *series-association* $\mathcal{S}$ computed via standard self-attention weights, capturing dynamic patterns (e.g., periods, trends) across the series.

The association discrepancy AD between these distributions is quantified using symmetrized KL divergence, forming the basis of the anomaly score AS, defined as:

$$\mathrm{AS}(X) = \mathrm{softmax}(-\mathrm{AD}(\mathcal{P}, \mathcal{S}; X)) \odot \left[||X_{i,:} - \hat{X}_{i,:}||_2 \right]_i \quad (1)$$

where $X \in \mathbb{R}^{T \times d}$ denotes the input series and $\hat{X}$ its reconstruction in dimensionality d and window size T. Anomalies are expected to have smaller discrepancies as they are rare and build temporally local associations, while normal points have larger discrepancy as they can build global associations across the whole time series sequence. The softmax term gives higher weight to points with smaller discrepancies which are the potential anomalies and this is multiplied with the reconstruction error that is larger for anomalies.

The optimization employs a minimax optimization framework: In the *minimize phase*, the prior-association is adjusted to approximate the series-association, ensuring adaptability to diverse temporal patterns. In the *maximize phase*, the series-association is optimized to enlarge the discrepancy under reconstruction loss constraint, forcing anomalies to exhibit even more concentrated local associations. This interplay amplifies the distinction between normal and abnormal points. Points with anomaly scores above a threshold are labelled as anomalies.

A notable practical advantage of the AT is its training efficiency. The incorporation of the Gaussian prior as an inductive bias effectively reduces the number of transformer layers and constrains the learning problem, guiding the model toward meaningful associations without requiring exhaustive exploration of the parameter space.

B. Explainability Framework

We propose a dual-track explainability framework that unifies model-intrinsic signals with post-hoc explanations, providing complementary evidence for root-cause triage.

1) Intrinsic Explanations: While [7] demonstrated that the association discrepancy and reconstruction terms can be visualized separately for model analysis, they did not develop a systematic explainability framework. We extend their work by building three native explanation modalities.

Prior Association Heatmaps: For each window containing an anomaly, we visualize the prior-association kernels as time $\times$ time matrices

$$\mathcal{P}_{ij} = \mathrm{Rescale}\left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{|X_{j,:} - X_{i,:}|^2}{2\sigma_i^2} \right) \right], \quad \mathcal{P} \in \mathbb{R}^{T \times T} \quad (2)$$

where σ_i^2 is the variance of the Gaussian at the i-th time point. Anomalous points tend to show narrower, near-diagonal associations compared to the broader associations of normal points. The corresponding σ-curves (Gaussian kernel width) reveal characteristic dips at anomalous time steps.

Series Association (Attention) Maps: Self-attention weights (for each layer and attention head) are computed as:

$$\mathcal{S}_{ij} = \mathrm{softmax}\left(\frac{Q_{im} K_{mj}}{\sqrt{d_k}} \right), \quad \mathcal{S} \in \mathbb{R}^{T \times T} \quad (3)$$

where $Q, K \in \mathbb{R}^{T \times d_k}$ are query and key matrices. Row-wise interpretation reveals which time steps influence a given query, while column-wise analysis identifies globally salient keys.

Feature Attribution via Attention Weighting: Per-feature contributions are derived by convolving attention weights with the input tensor:

$$\tau_{tf} = W_{ti} X_{if} \quad (4)$$

where W_{ti} are weight matrices and X_{if} the input series. This yields importance scores τ_{tf} for each feature f at time step t.

2) Post-hoc Explanations with time-aware KernelSHAP: To complement intrinsic signals, we employ KernelSHAP [28], which attributes predictions to input features by estimating each feature's marginal contribution under coalitions of present/absent features.

For time-series data, the model output is an anomaly score over a window $X \in \mathbb{R}^{T \times D}$. We define the black-box function $f(X)$ as the detector's scalar score (Eq. 1). SHAP returns local attributions $\phi_{t,d}$ for each time step t and feature d, satisfying:

$$f(X) \approx \phi_0 + \sum_{t,d} \phi_{t,d} \quad (5)$$

To preserve temporal structure, we employ structured coalitions: (i) contiguous temporal masks, (ii) feature-wise masks across the window, or (iii) feature-by-time tiles. "Feature dropping" is implemented by replacing with samples from a background distribution reflecting normal behaviour, randomly drawn benign windows from the training burn-in period.

Aggregating $\phi_d = \sum_t |\phi_{t,d}|$ yields per-feature importance; conversely, aggregating along the feature axis highlights influential time steps. We utilize the OmniXAI library's[1] time-series explainer for implementation.

[1]https://github.com/salesforce/OmniXAI

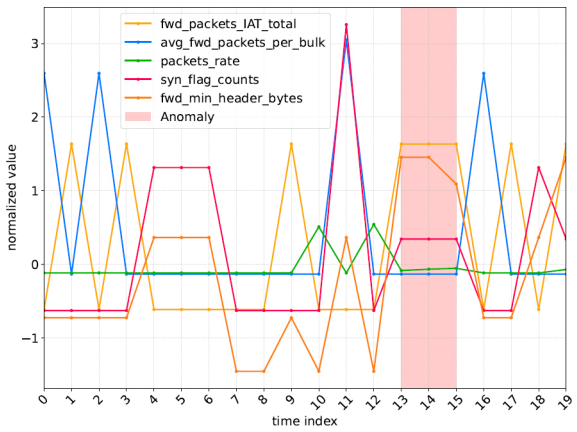


Fig. 1: Raw BCCC-CIC-IDS-2017 time-series excerpt (normalized) for a representative 20-step window containing a confirmed Port Scan attack (ground-truth anomaly region shaded in red, steps 13–15). Note the spike in `fwd_packets_IAT_total` (yellow) and the simultaneous drop in `avg_fwd_packets_per_bulk` (blue) that co-occur with the labelled attack; patterns that are subsequently recovered by both the attention and SHAP feature attributions (see Fig. 4a and 4b).

IV. Experimental Setup

A. Datasets

ServerMachineDataset (SMD) [14] was used for preliminary validation. This five-week multivariate time series dataset monitors 28 server entities through 38 metrics (CPU load, memory usage, network traffic, disk operations) at one-minute intervals, with 708,405 training samples and 708,420 test samples (4.16% anomaly ratio). Feature correlation analysis reveals clusters corresponding to CPU load metrics and network interface activity, while other features remain relatively independent. Unlike the IDS dataset described below, anomalies in SMD vary substantially in duration and often manifest as subtle deviations requiring domain expertise to identify through manual inspection. However, it should be noted that SMD represents server-side telemetry rather than network traffic data, making it less representative of real-world intrusion detection scenarios of an IDS dataset, which captures actual network flows with diverse, labelled attack vectors.

BCCC-CIC-IDS-2017 [8] serves as our primary benchmark. This high-fidelity, multivariate time series dataset is a refined version of CIC-IDS-2017 [10], re-processed using NTLFlowLyzer [8], a novel network traffic analyzer, to address inconsistencies in TCP flow termination and timing. Key improvements include an expanded feature set (114 features vs. 80) and corrected flow construction for reliable ground-truth labelling. Each sample represents a network flow characterized by features spanning multiple categories: flow identifiers (source/destination ports, protocol), payload statistics (byte counts, min/max/mean/variance for forward and backward directions), header information (total, mean, and per-direction header bytes), bulk transfer statistics, TCP flag counts, and other flow properties; an example window is shown in Fig. 1. The dataset includes diverse attack scenarios: DoS/DDoS (Hulk, GoldenEye, Slowloris), web attacks (SQL Injection, XSS, Brute Force), exploitation (Heartbleed, infiltration), and access attacks (SSH/SFTP brute-force, Botnet).

We split the dataset as follows: the training set comprises the first 80% (1,950,441 samples) with a "burn-in phase" without anomalous events in the first quarter; the test set contains the remaining 20% (487,611 samples) with an anomaly ratio of 26.74%. Features were normalized using standard scaling, and multi-class labels were collapsed into binary (benign/anomaly).

B. Models and Training

The AT [7] serves as our primary model of interest. We conducted an extensive hyper-parameter sweep across kernel size (for the Gaussian prior association), anomaly ratio (which calibrates the threshold for flagging anomalies), sliding window size, window stride, learning rate, and batch size. For the BCCC-CIC-IDS-2017 dataset, the input channel dimension was set to 114 to accommodate the full feature set. Our experiments revealed several notable behaviours: the model trains efficiently, achieving precision ≥ 0.85 after just a single epoch, and benefits from larger window sizes with approximately 20% overlap (smaller window stride). Smaller Gaussian prior kernel sizes improved recall, while the anomaly threshold of approximately 0.47 provided optimal scoring. The final configuration uses kernel size 3, anomaly ratio 2, window size 20, stride 10–20, learning rate 1.5×10^{-5}, batch size 256, and early stopping patience of 3 epochs; training was conducted over 10 epochs. Features were normalized using standard scaling, and multi-class attack labels were collapsed into binary classes (benign/anomaly).

Isolation Forest [33] serves as our traditional machine learning baseline. Unlike deep learning approaches that model normal behaviour, Isolation Forest detects anomalies by recursively partitioning data using random splits. The assumption is that anomalies require fewer partitions to isolate and thus exhibit shorter path lengths in the resulting tree structure. An ensemble of such isolation trees yields an anomaly score based on normalized path lengths. This algorithm remains relevant due to its linear time complexity and low memory footprint. We optimized across the number of estimators, maximum samples per estimator, maximum features, and contamination ratio. The final configuration uses 50 estimators, maximum samples of 64, maximum features of 1, and contamination of 0.075.

TimesNet [21] represents a CNN-based alternative that addresses time-series anomaly detection through a fundamentally different approach. Rather than using attention mechanisms, TimesNet leverages multi-periodicity by transforming 1D time series into 2D representations. It first identifies dominant periods using Fast Fourier Transform, then reshapes the data into 2D tensors where columns capture intra-period variations and rows capture inter-period variations. These tensors

TABLE I: Detection performance on the benchmark datasets

(a) SMD

Model	Accuracy	Precision	Recall	F1
AT	**0.989**	**0.894**	**0.955**	**0.923**
TimesNet	0.844	0.818	0.913	0.863
Isolation Forest	0.532	0.423	0.733	0.536

(b) BCCC-CIC-IDS-2017

Model	Accuracy	Precision	Recall	F1	AUROC
AT	**0.972**	**0.987**	0.960	**0.973**	**0.970**
TimesNet	0.859	0.955	0.772	0.854	0.919
Isolation Forest	0.565	0.552	**0.992**	0.709	-

are processed using parameter-efficient inception blocks with multi-scale 2D convolution kernels. Due to computational constraints, hyperparameter optimization was limited compared to the AT. The final configuration uses anomaly ratio 2, window size 100, and learning rate 10^{-4}. Notably, TimesNet exhibits substantially higher training latency and inference times compared to the AT, which may limit its applicability in real-time detection scenarios.

V. Results and Analysis

A. Detection Performance

The model performance benchmarks are detailed in Table I and highlight clear and consistent differences in the precision-recall trade-off across models and datasets. For both benchmarks, the AT yields the strongest overall balance, achieving the top F1-scores (0.923 on SMD and 0.973 on BCCC-CIC-IDS-2017), which indicates that it maintains high sensitivity to anomalous events without substantially inflating false alarms. The contrast to the baselines is informative: TimesNet attains comparatively high precision on both datasets (0.818 on SMD and 0.955 on BCCC-CIC-IDS-2017) but systematically lower recall than the AT, suggesting that its detections are conservative; it triggers fewer alerts but misses a larger share of true anomalies, particularly on the higher-dimensional IDS benchmark (114 features). Isolation Forest exhibits the opposite behaviour: it reaches very high recall on BCCC-CIC-IDS-2017 (0.992), but with much lower precision (0.552), implying a large number of false positives that would likely overwhelm analysts in an operational setting.

Importantly, these patterns should be interpreted in light of dataset characteristics: SMD has a low anomaly ratio ($\approx$4.16%), whereas BCCC-CIC-IDS-2017 contains a much larger fraction of anomalies in the test split ($\approx$26.74%), so accuracy alone is less diagnostic under imbalance. Given the class imbalance in both datasets, threshold-dependent metrics such as Accuracy and F1 are sensitive to threshold selection. We therefore treat the Area Under the Receiver Operating Characteristic (AUROC) and Average Precision (AP) as the primary ranking metrics, as they evaluate separation quality across all thresholds. Consistent with the F1-scores, the AT also achieves excellent ranking performance on BCCC-CIC-IDS-2017 (AUROC $= 0.97$, AP $= 0.97$), supporting the conclusion that it separates benign from malicious flows robustly across thresholds. The thresholds used for Table I were selected via the model's built-in anomaly ratio hyperparameter (set to the dataset's empirical anomaly rate), avoiding manual post-hoc tuning.

Overall, Table I suggest that transformer-based anomaly detection is not only consistently more accurate, but also operationally better to calibrate than the CNN-based and traditional baselines: it reduces missed detections relative to TimesNet while avoiding the excessive false-positives seen with Isolation Forest.

In addition, we observed that (i) anomaly threshold selection primarily trades precision against recall, (ii) windowing matters substantially — larger windows with $\approx$20% overlap benefit the AT, (iii) smaller Gaussian prior kernels tend to improve recall, and (iv) the AT reaches useful precision early in training (precision $\geq$0.85 after one epoch), whereas TimesNet incurs noticeably higher training and inference latency.

B. Explainability Analysis

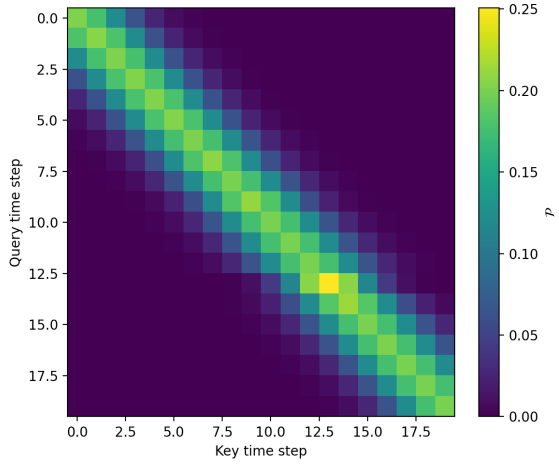
Beyond detection accuracy, the AT's architecture offers multiple avenues for explaining its decisions. The following analysis (on the primary BCCC-CIC-IDS-2017 dataset) systematically compares intrinsic and post-hoc explanations to assess their complementarity for root-cause resolution.

1) Prior Association Analysis: The prior association mechanism (Eq. 2) reveals distinct patterns for anomalous versus benign windows, visualized in Fig. 2a and 2b. For windows containing anomalies, the Gaussian prior exhibits characteristic narrowing (lower σ values) at anomalous time steps, indicating concentrated local associations. In contrast, benign windows maintain relatively uniform prior associations across all time steps. The predominantly diagonal structure of the prior association heatmap reflects the Gaussian kernel's built in locality bias: the prior association concentrates weight on temporally adjacent steps. This locality is precisely the intended property that makes the prior association useful for anomaly localization—anomalies produce a *sharper* diagonal peak (lower σ) compared to normal points, which distribute weight more broadly.

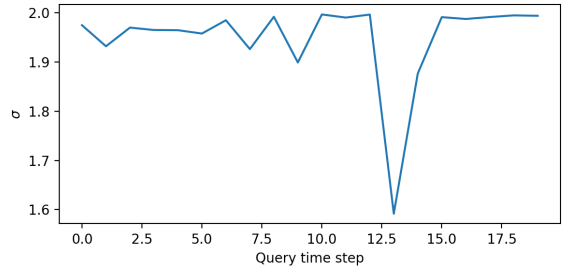
For attention heads, we observe that generally half the heads in each layer clearly exhibit prior peaks at anomalous events, suggesting head specialization in the learned representations. The 1D σ-curves provide a concise diagnostic. A noticeable decrease in σ at specific time steps reliably indicates anomaly candidates.

2) Attention Map Interpretation: Self-attention maps (Eq. 3) provide complementary temporal insights, as shown in Fig. 3. For anomalous windows, row-wise interpretation shows that key weights from earlier time steps contribute to triggering the anomalous event, the model learns temporal precursors. Column-wise analysis indicates that most queries induce high responses at anomalous time steps, manifesting as vertical

Fig. 2: Prior association visualizations (association heatmap 2a and corresponding σ curve 2b) for an anomaly located at time step 13 in a window.



(a) **Prior association heatmap** (of attention head 3). The diagonal clearly exhibits a prior peak at the anomalous event.



(b) 1D σ**-curve** of the prior association discrepancy corresponding to Fig. 2a. There is a noticeable decrease in σ at the anomalous event, corresponding to a peak in the Gaussian (becoming narrower).

bright stripes in the attention matrix. For benign windows, attention patterns are more diffuse, with no single time step attracting disproportionate attention. This contrast enables visual discrimination between normal and anomalous model behaviour.

3) Feature Attribution Comparison: We computed feature attributions using both intrinsic attention weighting (Eq. 4; in Fig. 4a) and KernelSHAP (Eq. 5; in Fig. 4b) for representative anomalous windows. *Intrinsic attributions* at anomalous time steps highlight features such as `fwd_packets_rate`, `avg_fwd_packets_per_bulk`, `avg_fwd_bulk_rate`, and `packets_rate`. At non-anomalous time steps within the same window, different features dominate, such as flag counts and header byte statistics. *SHAP attributions* reveal a distinct feature set. Top contributors include `fwd_packets_IAT_std`, `*_syn_flag_counts`, and `total_payload_bytes`. SHAP emphasizes packet statistics and flag families (inter-arrival statistics, counts, rate metrics), whereas intrinsic attri-

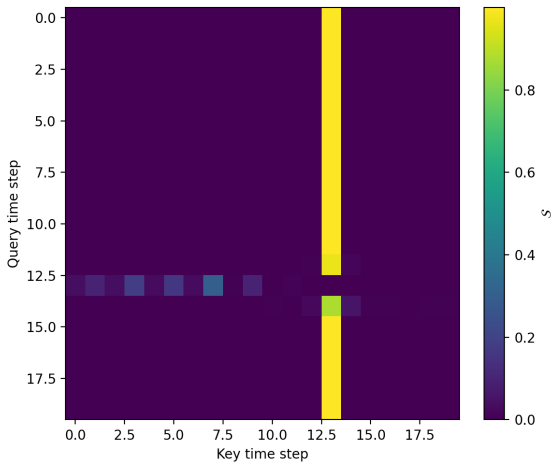
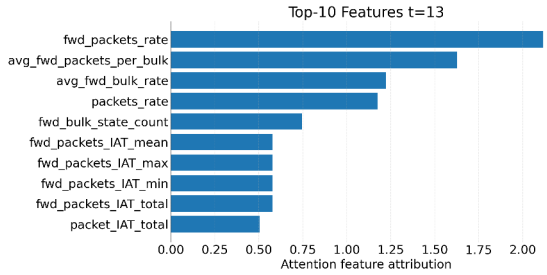
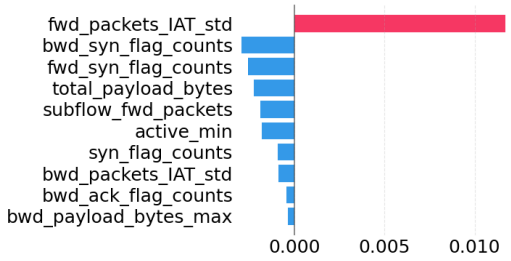


Fig. 3: **Attention map** (key vs. query of attention head 3) for an anomaly located at time step 13 in the window (same anomaly as in Fig. 2a and 2b). Row-wise attention weights indicate temporal precursors leading to the anomaly, whereas column-wise weights reflect its salience across queries.

Fig. 4: Feature attributions evaluated through attention weights 4a and KernelSHAP 4b.



(a) Attention feature attribution.



(b) KernelSHAP feature attribution.

butions more frequently elevate payload/header size variables. Regarding the complementarity of the explanation modalities discussed above, a critical finding is the **minimal overlap** between intrinsic and SHAP feature attributions. This divergence is expected given their different targets and assumptions:

- **SHAP** decomposes the final score $f(x)$ relative to a benign baseline and counterfactual coalitions, answering: "Which features, if absent, would most reduce the anomaly score?"
- **Intrinsic attributions** trace representation salience inside the anomaly-attention mechanism, answering: "Where did the model look to form associations?"

This complementarity has important practical implications. Prior associations and time-wise SHAP identify *when* in the window the score was elevated, aligning well with ground-truth anomaly locations. Feature-wise SHAP identifies *which raw inputs* pushed the score, useful for understanding attack signatures. Intrinsic attention clarifies *how the model reasoned*, revealing learned temporal dependencies and precursor patterns.

Relying on a single explanation modality is therefore insufficient. We recommend a multifaceted approach combining both perspectives. Intrinsic explanations for model validation, and temporal reasoning, and SHAP for feature-level audit trails aligned with domain expertise.

VI. Conclusion

This article presented an empirical evaluation of the Anomaly Transformer for network intrusion detection, with a focus on dual-track explainability.

On the BCCC-CIC-IDS-2017 benchmark, the model achieves state-of-the-art detection performance ($\mathrm{F1} = 0.97$, $\mathrm{AUROC} = 0.97$), outperforming CNN-based and traditional baselines.

More importantly, our systematic comparison of intrinsic (prior association, attention maps) and post-hoc (KernelSHAP) explanations reveals strong complementarity: intrinsic mechanisms excel at temporal localization, while SHAP provides feature-level attribution aligned with attack signatures. No single modality is sufficient; security analysts benefit from a multifaceted approach for operational IDS deployments, where intrinsic explanations support model validation and SHAP supports offline forensic audit trails. This is essential for building trust, supporting compliance, and enabling effective human-AI collaboration in cybersecurity operations.

Limitations and Future Work: Extending the framework to GNN-based IDS, incorporating quantitative XAI evaluation metrics (fidelity, stability), exploring lightweight SHAP approximations for near-real-time use, and validating robustness under adversarial conditions are natural next steps.

References

[1] S. Alam and Z. Altiparmak, "XAI-CF – examining the role of explainable artificial intelligence in cyber forensics," *CoRR*, 2024. [Online]. Available: http://arxiv.org/abs/2402.02452v2

[2] M. L. Ali, K. Thakur, S. Schmeelk, J. Debello, and D. Dragos, "Deep learning vs. machine learning for intrusion detection in computer networks: A comparative study," *Applied Sciences*, vol. 15, no. 4, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/4/1903

[3] R. Kimanzi, P. Kimanga, D. Cherori, and P. K. Gikunda, "Deep learning algorithms used in intrusion detection systems – a review," 2024. [Online]. Available: https://arxiv.org/abs/2402.17020

[4] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological Cybernetics*, vol. 59, no. 4-5, pp. 291–294, 1988. [Online]. Available: http://dx.doi.org/10.1007/BF00332918

[5] F. Farahnakian and J. Heikkonen, "A deep auto-encoder based approach for intrusion detection system," in *2018 20th International Conference on Advanced Communication Technology (ICACT)*, 2 2018, pp. 178–183. [Online]. Available: http://dx.doi.org/10.23919/ICACT.2018.8323688

[6] M. Ma, L. Han, and C. Zhou, "Research and application of transformer based anomaly detection model: a literature review," *CoRR*, 2024. [Online]. Available: http://arxiv.org/abs/2402.08975v1

[7] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=LzQQ89U1qm_

[8] M. Shafi, A. H. Lashkari, and A. H. Roudsari, "NTLFlowLyzer: Towards generating an intrusion detection dataset and intruders behavior profiling through network and transport layers traffic analysis and pattern extraction," *Computers & Security*, vol. 148, p. 104160, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404824004656

[9] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 7 2009, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/cisda.2009.5356528

[10] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, *A Detailed Analysis of the CICIDS2017 Data Set*, ser. Communications in Computer and Information Science. Springer International Publishing, 2019, pp. 172–188. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-25109-3_9

[11] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck, "Dos and don'ts of machine learning in computer security," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 3971–3988. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/arp

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. [Online].

Available: http://dx.doi.org/10.1038/323533a0

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[14] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 7 2019, pp. 2828–2837. [Online]. Available: http://dx.doi.org/10.1145/3292500.3330672

[15] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *CoRR*, 2015. [Online]. Available: https://arxiv.org/abs/1508.01991

[16] C. Zhang, J. Li, N. Wang, and D. Zhang, "Research on intrusion detection method based on transformer and CNN-BiLSTM in internet of things," *Sensors*, vol. 25, no. 9, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/9/2725

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [Online]. Available: http://dx.doi.org/10.1109/5.726791

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. [Online]. Available: http://dx.doi.org/10.1145/3065386

[19] M. Gao, L. Ma, H. Liu, Z. Zhang, Z. Ning, and J. Xu, "Malicious network traffic detection based on deep neural networks and association analysis," *Sensors*, vol. 20, no. 5, p. 1452, 2020. [Online]. Available: http://dx.doi.org/10.3390/s20051452

[20] P. Zhao, Z. Fan*, Z. Cao, and X. Li, "Intrusion detection model using temporal convolutional network blend into attention mechanism," *International Journal of Information Security and Privacy*, vol. 16, no. 1, pp. 1–20, 2021. [Online]. Available: http://dx.doi.org/10.4018/IJISP.290832

[21] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2d-variation modeling for general time series analysis," 2023. [Online]. Available: https://arxiv.org/abs/2210.02186

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762v7

[23] Z. Wu, H. Zhang, P. Wang, and Z. Sun, "RTIDS: a robust transformer-based approach for intrusion detection system," *IEEE Access*, vol. 10, pp. 64 375–64 387, 2022. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2022.3182333

[24] C. Xi, H. Wang, and X. Wang, "A novel multi-scale network intrusion detection model with transformer," *Scientific Reports*, vol. 14, no. 1, p. 23239, 2024. [Online]. Available: http://dx.doi.org/10.1038/s41598-024-74214-w

[25] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," *CoRR*, 2022. [Online]. Available: http://arxiv.org/abs/2201.07284v6

[26] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques," *CoRR*, 2019. [Online]. Available: http://arxiv.org/abs/1909.03012v2

[27] M. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the predictions of any classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016, pp. 97–101. [Online]. Available: http://dx.doi.org/10.18653/v1/N16-3020

[28] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[29] Z. Sun, A. M. Teixeira, and S. Toor, "GNN-IDS: Graph neural network based intrusion detection system," in *Proceedings of the 19th International Conference on Availability, Reliability and Security*, 2024, pp. 1–12. [Online]. Available: http://dx.doi.org/10.1145/3664476.3664515

[30] G. Andresini, A. Appice, F. P. Caforio, D. Malerba, and G. Vessio, "Roulette: a neural attention multi-output model for explainable network intrusion detection," *Expert Systems with Applications*, vol. 201, p. 117144, 2022. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2022.117144

[31] S. Piaggesi, R. Guidotti, F. Giannotti, and D. Pedreschi, "Explanations go linear: Interpretable and individual latent encoding for post-hoc explainability," *CoRR*, 2025. [Online]. Available: http://arxiv.org/abs/2504.20667v1

[32] A. Srinivasan, V. S. Ravi, J. C. Andresen, and A. Holst, "Counterfactual explanation for auto-encoder based time-series anomaly detection," *PHM Society European Conference*, vol. 8, no. 1, p. 9, 2024. [Online]. Available: http://dx.doi.org/10.36001/phme.2024.v8i1.4087

[33] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2008.17