Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

# Beyond The Classifier: Dual-Path Explanation of A Hybrid CNN–Transformer–Autoencoder Intrusion Detection System

Imran Hussain[1], Dr. Pradeep Laxkar[2]

[1]Research Scholar, Department of CSE, ITM SLS Baroda University, Vadodara, Gujarat, India. Email ID: imran.hussain.rtm@gmail.com, ORCID ID: 0009-0006-0652-1933
[2]Professor, Department of CSE, ITM SLS Baroda University, Vadodara, Gujarat, India. Email ID: pradeep.laxkar@gmail.com, ORCID ID: 0000-0003-2996-1481

**ABSTRACT**
Hybrid intrusion detection architectures increasingly couple a discriminative classifier with a generative reconstruction branch, yet the explainability literature for network security still treats such detectors as if they had a single output. Post-hoc attribution is applied to the classifier, and the anomaly branch — the component that actually carries the burden of detecting unseen attacks — is left unexplained. This paper investigates that gap directly. Building on a previously published Hybrid Attentive Convolutional Autoencoder (HACA), we introduce X-HACA, a dual-path explanation framework comprising three components: Reconstruction Error Decomposition (RED), an exact, closed-form, zero-query attribution of the anomaly score over input features; Attention-Screened Shapley Estimation (ASSE), a two-stage estimator that uses an attention–gradient prior to reduce the effective player count of the Shapley game; and Dual-Path Consensus Explanation (DPCX), a log-pooled fusion of the two paths. Experiments on the official NSL-KDD and UNSW-NB15 partitions expose a blind spot that has gone unmeasured: on NSL-KDD, the top-10 features selected by KernelSHAP account for only 31.6% of the reconstruction evidence the detector actually acts upon, and LIME for 38.9%, whereas DPCX captures 96.5% while remaining competitive on classifier faithfulness. DPCX is the only method examined that is faithful to both heads on both datasets. We further report two deliberate negative results. Attention-guided importance sampling of Shapley coalitions collapses, retaining an effective sample size of 1.5%; and explanation-derived novelty signals — divergence between paths and conformity to class prototypes — do not improve zero-day detection over the reconstruction error itself, which reaches 89.47% AUROC against 78.79% for maximum softmax probability and 79.61% for the energy score. Finally, we show that 56.9% of the 3,750 genuinely unseen attack instances in KDDTest+ are misclassified as normal by the classifier, and that the anomaly branch recovers 46.0% of them at a 5% false-positive rate, quantifying the operational value of explaining both paths rather than one.

**Keywords:** Explainable artificial intelligence; intrusion detection; Shapley values; autoencoder; zero-day attack detection; faithfulness evaluation; NSL-KDD; UNSW-NB15

## INTRODUCTION

Deep learning has become the dominant modelling paradigm in network intrusion detection, and hybrid architectures in particular have proved effective at reconciling two objectives that pull in opposite directions. A discriminative classifier learns to separate traffic that resembles attack categories seen during training, while a generative reconstruction branch learns the manifold of ordinary traffic and flags departures from it. The parent study to this work proposed one such architecture, the Hybrid Attentive Convolutional Autoencoder (HACA), which combines one-dimensional convolutional layers for local feature extraction, a Transformer encoder for long-range dependencies, a fusion layer, and an autoencoder branch whose reconstruction error serves as an unsupervised anomaly score [1]. That study closed by identifying interpretability as its principal unresolved limitation. The present paper takes up that question.

The way the explainability literature has approached intrusion detection creates a specific and, we argue, consequential mismatch. Surveys of explainable AI for intrusion detection consistently report that SHAP and LIME are the two dominant techniques in the field [24], [25], and comparative evaluation frameworks such

as E-XAI have been built specifically to benchmark black-box explainers on network intrusion tasks [19]. Both methods, and essentially all of the work that applies them to security, explain a scalar classifier output: the probability that a flow belongs to an attack class. When the detector being explained is a hybrid with two heads, this convention silently restricts the explanation to one of them.

Why this matters becomes clear once the roles of the two heads are separated. A classifier trained on a fixed taxonomy of attacks can only assign traffic to categories it has already seen. Confronted with an attack type absent from training it must still emit one of its known labels, and it frequently chooses the benign one. The reconstruction branch has no such constraint; it can register that an input is unlike anything in the training distribution without needing a name for it. The division of labour is therefore not incidental. Known attacks are largely the classifier's responsibility and unknown attacks are largely the autoencoder's. An explanation that covers only the classifier explains the easy half of the problem.

This paper asks whether that omission is quantitatively significant, and what a two-path explanation would look like if built properly. We make the question concrete by measuring, for each attribution method, how much of the reconstruction error the detector actually incurs is accounted for by the features that method nominates as important. The answer on NSL-KDD is that KernelSHAP's top-10 features cover under a third of it. The explanation an analyst reads and the evidence the detector acts upon are, to a substantial degree, different things.

Our contributions are as follows.

(i) We formulate Reconstruction Error Decomposition (RED), an exact additive attribution of the autoencoder branch's anomaly score to individual input features. Because the anomaly score is a sum of per-feature squared errors, the decomposition is closed-form and requires no additional model queries — in contrast to the hundreds of perturbed forward passes consumed by KernelSHAP or LIME.

(ii) We propose Attention-Screened Shapley Estimation (ASSE), which uses an attention-rollout and gradient prior to screen the input down to a salient subset and then solves the Shapley regression on a reduced cooperative game whose remaining players are collapsed into a single block. We characterise both its advantage in the tight-query regime and the bias floor it incurs.

(iii) We define Dual-Path Consensus Explanation (DPCX), a log-pooled fusion of the discriminative and generative attributions, and show it is the only method among nine evaluated that is faithful to both heads of the detector on both benchmark datasets.

(iv) We report two negative results in full. Attention-guided importance sampling of Shapley coalitions suffers weight degeneracy, and explanation-derived novelty scores do not improve zero-day detection over the plain reconstruction error. Both are documented with the diagnostics that reveal why.

(v) We quantify the operational payoff of explaining the anomaly branch: on the natural zero-day protocol afforded by KDDTest+, the majority of unseen attacks are labelled benign by the classifier, and the anomaly branch recovers nearly half of those at a 5% false-positive rate.

The remainder of the paper is organised as follows. Section 2 reviews related work; Section 3 states the research gap; Section 4 develops the X-HACA framework; Section 5 describes the experimental protocol; Section 6 presents results; Section 7 discusses limitations and negative findings; Section 8 concludes.

## 2. RELATED WORK

### 2.1 Post-hoc attribution methods

Two model-agnostic families dominate practice. LIME fits a sparse linear surrogate to the model's behaviour in a neighbourhood of the instance being explained [3]. KernelSHAP recovers Shapley values [17] by solving a weighted least-squares problem whose sample weights are given by the Shapley kernel [2], and subsequent work has refined its estimator design and convergence behaviour [18]. Gradient-based alternatives are cheaper: Integrated Gradients accumulates gradients along a path from a baseline to the input and satisfies a completeness axiom [4], while gradient-times-input requires a single backward pass.

For Transformer models, attention weights are an appealing free signal, and attention rollout propagates them across layers while correcting for residual connections [6]. Whether attention constitutes an explanation has been contested at length: Jain and Wallace argued that attention distributions correlate poorly with gradient-based importance and can be substantially altered without changing predictions [7], while Wiegreffe and Pinter countered that this depends on what one asks of an explanation [8]. Our design takes a deliberately modest position on this dispute — attention is used as a sampling prior, never as the attribution itself.

### 2.2 Evaluating explanations

Because attributions cannot be validated against ground truth, faithfulness is assessed by perturbation. Deletion and insertion curves progressively remove or restore features in order of attributed importance and measure the resulting trajectory of the model output [10]; the ERASER benchmark formalised the related notions of comprehensiveness and sufficiency [9]. A separate line of work examines whether explanations are stable and trustworthy at all: Alvarez-Melis and Jaakkola showed that interpretability methods can be highly sensitive to imperceptible input changes [27], Adebayo et al. demonstrated that some saliency methods

are insensitive to model randomisation [29], and Slack et al. constructed adversarial models that conceal bias from LIME and SHAP [28]. Rudin has argued more broadly that post-hoc explanation of black boxes is the wrong response in high-stakes settings [30].

### 2.3 Explainable intrusion detection

Systematic reviews report SHAP and LIME as the prevailing methods in explainable intrusion detection, valued for producing both global feature rankings and per-alert justifications [24], [25]. Representative applications include LIME and SHAP applied to multi-layer perceptrons [20], SHAP-based analysis of XGBoost and TabNet on UNSW-NB15 [23], and SHAP integrated into federated detection for connected vehicles [26]. E-XAI provides an end-to-end framework for evaluating black-box explainers specifically in the intrusion detection setting [19]. What unites these studies is that the explained quantity is always a classifier output. We are not aware of prior work that attributes a hybrid detector's reconstruction-based anomaly score and quantifies the discrepancy between the two explanations.

### 2.4 Detecting unseen attacks

Reconstruction-based anomaly detection is a long-standing approach to unknown threats. Hindy et al. trained autoencoders on normal traffic and reported 92.96% zero-day detection accuracy on NSL-KDD [22], and two-stage LSTM-autoencoder designs have been applied to the same problem [21]. From the out-of-distribution literature, the standard reference points are the maximum softmax probability baseline [13], Mahalanobis distance in a learned feature space [14], and the energy score [15]. We evaluate our explanation-derived signals against all three.

## 3. RESEARCH GAP

Three gaps follow from the preceding review.

Single-head explanation of multi-head detectors. Hybrid architectures that fuse a classifier with a reconstruction branch are common, but the explanation methods routinely applied to them attribute only the classification output. Whether this omission is material has not been measured, because no metric in current use asks how much of the anomaly evidence an attribution covers.

Cost asymmetry between the two paths. KernelSHAP and LIME require several hundred perturbed forward passes per alert, a considerable expense at line rate. Whether the anomaly path admits a cheaper — or exact — attribution has not been examined, despite the additive structure of the reconstruction loss making this question natural.

Unverified claims about explanation-based novelty detection. It is intuitively attractive to suppose that explanations of unknown attacks look anomalous in explanation space, and hence that explanation statistics could serve as zero-day signals. This proposition is rarely tested against strong baselines. We test it and find it does not hold for this architecture.

## 4. THE X-HACA FRAMEWORK

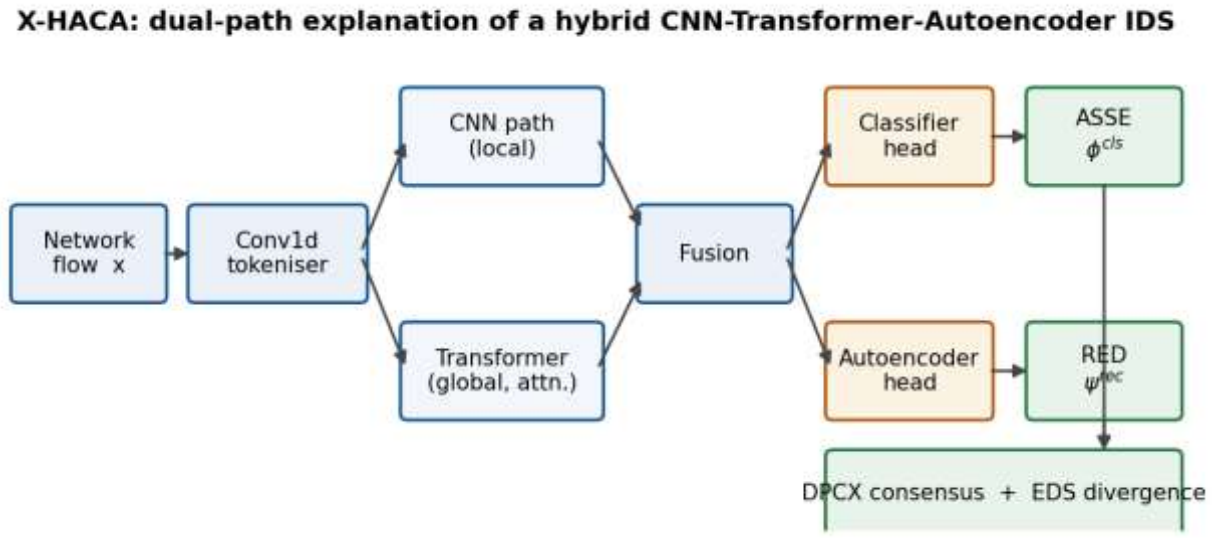



**Fig. 1. The X-HACA framework. The HACA backbone (left) feeds two heads. ASSE attributes the classification head, RED exactly decomposes the anomaly head, and DPCX fuses the two into a consensus explanation.**

### 4.1 Backbone

We retain the parent architecture's structure [1]. An input flow vector x ∈ R^D is tokenised by a strided one-dimensional convolution into T tokens of dimension d. Two parallel paths follow: a convolutional stack

capturing local structure, and a Transformer encoder [5] whose multi-head self-attention captures global dependencies. Mean-pooled outputs of both paths are concatenated and passed through a fusion layer producing z. Two heads consume z: a classification head emitting logits over C classes, and an autoencoder head that compresses z to a latent code and decodes it back to a reconstruction $\hat{x} \in R^D$. The anomaly score is

$$A(x) = \Sigma_i (x_i - \hat{x}_i)^2 .$$

Training minimises class-weighted cross-entropy plus reconstruction mean-squared error. Tokenisation before the Transformer keeps the attention sequence short, which matters because the explanation experiments require orders of magnitude more forward passes than training does.

**4.2 Reconstruction Error Decomposition (RED)**

The anomaly score is already a sum of independent per-feature terms. Its attribution is therefore exact rather than approximate:

$$\psi_i(x) = (x_i - \hat{x}_i)^2, \quad \Sigma_i \psi_i(x) = A(x).$$

Two properties distinguish RED from sampling-based attribution. It is exact — the efficiency property holds identically, not in expectation — and it costs one forward pass, the same forward pass already performed to obtain the anomaly score. There is no query overhead whatsoever. The apparent triviality of this decomposition is precisely the point: the quantity is available for free, and the literature has not been using it.

**4.3 Attention-Screened Shapley Estimation (ASSE)**

For the classification head we require Shapley values, whose exact computation is exponential. KernelSHAP estimates them by sampling coalitions $S \subseteq \{1..D\}$ and solving a weighted regression, but the number of coalitions needed grows with D, and D = 122 for NSL-KDD.

ASSE proceeds in two stages. First, a prior is computed at the cost of one forward and one backward pass. Let $R \in \Delta^T$ be the residual-corrected attention rollout over tokens [6] and $g_i = |x_i \cdot \partial f_y/\partial x_i|$. Attention supplies token-level salience and the gradient resolves within a token:

$$a_i \propto R_\{t(i)\} \cdot g_i / \Sigma_\{j \in token(i)\} g_j ,$$

smoothed toward uniform so no feature is starved of coverage. Second, the k features with the largest prior form a salient set A; the remaining D − k features are collapsed into a single block player R. KernelSHAP is solved on this reduced game of k + 1 players, and the block's value is redistributed across R in proportion to the prior.

The estimate is the Owen value of the partition $\{a_1\},\ldots,\{a_k\}$, R rather than the Shapley value of the full game [16]. This is an approximation, and we characterise its error empirically rather than claiming it away: reducing the player count from 122 to k + 1 sharply improves conditioning at small budgets, but introduces a bias floor that a full-game estimator eventually surpasses.

**4.4 Dual-Path Consensus Explanation (DPCX) and divergence**

Normalising each attribution to a distribution over features, $p^{cls} \propto |\varphi|$ and $p^{rec} \propto \psi$, the consensus explanation is their log-pooled combination:

$$c_i \propto (p^{cls}_i)^\alpha (p^{rec}_i)^{1-\alpha} ,$$

with $\alpha = 0.5$ throughout. Log-pooling is chosen over linear mixing because it is zero-preserving: a feature that neither path implicates cannot enter the consensus, whereas an arithmetic mean would admit it on the strength of one path alone.

We additionally define the Explanation Divergence Score as the Jensen–Shannon divergence between the paths, $EDS(x) = JSD(p^{cls} \| p^{rec})$, and an Explanation Conformity Score measuring divergence from the mean explanation of the predicted class. Our hypothesis was that unseen attacks would exhibit distinctive values of these statistics. Section 6.5 reports that they do not, in the sense that matters operationally.

## 5. EXPERIMENTAL SETUP

**5.1 Datasets and protocol**

We use the official partitions of both datasets rather than random splits. For NSL-KDD [11] this is KDDTrain+ (125,973 flows) and KDDTest+ (22,544 flows); categorical fields are one-hot encoded against a shared vocabulary, giving D = 122, and features are min-max scaled using training statistics only. For UNSW-NB15 [12] we use the supplied training (82,332) and testing (175,341) partitions with D = 42, applying log1p to heavy-tailed magnitude features before scaling.

The choice of the official NSL-KDD partition is deliberate and consequential. KDDTest+ contains 3,750 instances drawn from 17 attack types that do not occur anywhere in KDDTrain+, listed in Table 7. This provides a natural zero-day evaluation with genuinely unseen attacks, rather than an artificial hold-out. It is also markedly harder than a random split of KDDTrain+, and we report both regimes to make the comparison explicit.

**5.2 Implementation**

Models are implemented in PyTorch and trained on a single CPU core, using AdamW with a one-cycle schedule, batch size 512, and class-weighted cross-entropy to address the severe imbalance of the R2L and

U2R families. The NSL-KDD backbone has 91,775 parameters and the UNSW-NB15 backbone 83,340. The reference Shapley vectors used to score the estimators are obtained from paired KernelSHAP at 24,576 coalitions per instance.

### 5.3 Evaluation metrics

Classification faithfulness uses deletion AUC, insertion AUC, comprehensiveness and sufficiency [9], [10], with features replaced by the training mean.

Anomaly faithfulness requires care. Background masking is not a valid deletion test for the reconstruction head, because replacing $x_i$ changes the reconstruction target itself; the score can rise rather than fall, and the curve becomes uninterpretable. We therefore use two valid alternatives. Anomaly mass coverage at k is the fraction of the realised reconstruction error carried by an attribution's top-k features. Repair AUC substitutes the top-k features with their own reconstructed values and tracks the residual anomaly score, for which lower is better.

Zero-day detection is scored by AUROC, average precision, false-positive rate at 95% true-positive rate, and true-positive rate at 1% false-positive rate. Stability is the Jaccard overlap of the top-10 features under Gaussian input perturbation.

## 6. RESULTS AND DISCUSSION

### 6.1 Backbone detection performance

Table 1 reports detection performance. On a held-out split of KDDTrain+ the backbone reaches 99.20% accuracy, consistent with the regime reported in the parent study [1]. On the official KDDTest+ partition the same model reaches 76.82%. The gap is not a defect of the model but a property of the benchmark, and its cause is visible in Table 2.

**Table 1. Backbone detection performance on the official test partitions.**

| Metric | NSL-KDD (KDDTest+) | UNSW-NB15 |
|---|---|---|
| Accuracy (%) | 76.82 | 88.85 |
| Macro precision (%) | 69.61 | 86.85 |
| Macro recall (%) | 58.81 | 91.16 |
| Macro F1 (%) | 57.26 | 88.01 |
| Weighted F1 (%) | 74.04 | 89.16 |
| Detection rate (%) | 69.44 | 84.76 |
| False positive rate (%) | 7.32 | 2.44 |
| Parameters | 91,775 | 83,340 |

**Table 2. Per-class recall on KDDTest+. R2L collapses because most of its test instances belong to attack types absent from training.**

| Class | Normal | DoS | Probe | R2L | U2R |
|---|---|---|---|---|---|
| Recall (%) | 92.68 | 83.05 | 69.35 | 14.62 | 34.33 |

R2L recall of 14.62% is the signature failure mode of KDDTest+: the majority of R2L instances in the test partition are attack types the classifier has never seen, and it assigns them to known classes — usually Normal. This is exactly the regime in which the anomaly branch is supposed to earn its place, and Section 6.5 measures whether it does.

### 6.2 Estimator efficiency and its limits

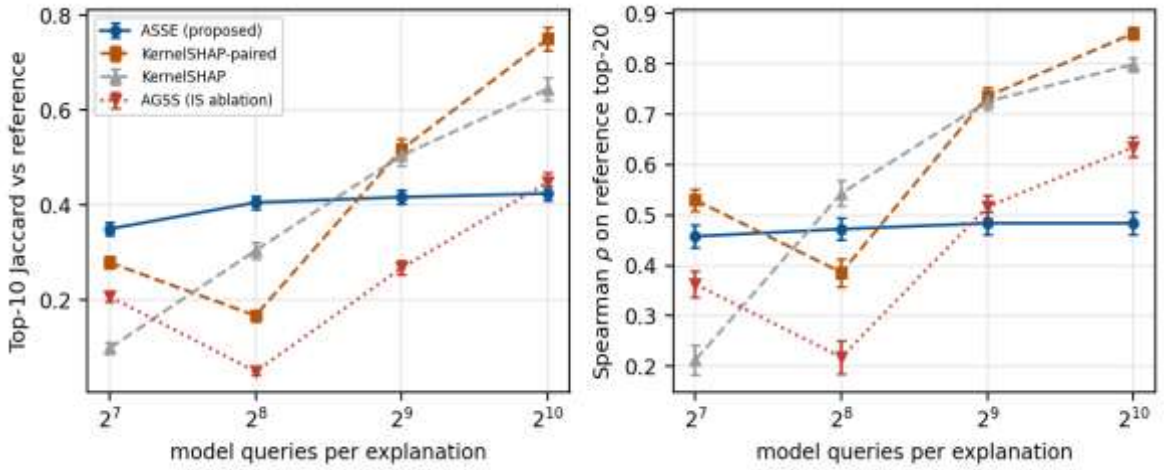



**Fig. 2. Fidelity to a 24,576-coalition reference as a function of query budget. ASSE leads in the tight-budget regime and then saturates; paired KernelSHAP overtakes it once the budget exceeds a few hundred queries.**

**Table 3. Top-10 Jaccard agreement with the high-budget reference, by query budget. Best per column in bold.**

| Method | 128 | 256 | 512 | 1024 |
|---|---|---|---|---|
| ASSE (proposed) | 0.350 | 0.405 | 0.417 | 0.424 |
| KernelSHAP | 0.099 | 0.305 | 0.505 | 0.645 |
| KernelSHAP-paired | 0.279 | 0.168 | 0.517 | 0.750 |
| AGSS (IS ablation) | 0.208 | 0.050 | 0.268 | 0.448 |

At 128 queries ASSE achieves 0.350 top-10 Jaccard against 0.099 for KernelSHAP, a 3.5-fold improvement, and remains ahead of the paired variant. It then saturates near 0.42 while paired KernelSHAP continues improving to 0.750 at 1,024 queries. This is the bias floor predicted in Section 4.3: screening buys conditioning at the cost of a ceiling determined by the prior's ability to place the truly important features inside the salient set. A sweep over the screening width confirms the trade-off, with the optimum shifting from $k \approx 16$ at 128 queries to $k \approx 48$ at 512.

The honest reading is that ASSE is useful under a hard per-alert query constraint and should not be preferred otherwise. We report it on those terms.

### 6.3 The blind spot: dual-path faithfulness

This is the central result. Tables 4 and 5 evaluate nine attribution methods against both heads.

**Table 4. Faithfulness on NSL-KDD (n = 150). The first four columns score the classification head, the last two the anomaly head.**

| Method | Del.AUC ↓ | Ins.AUC ↑ | Comp. ↑ | Suff. ↓ | AnoCov@10 ↑ | Repair ↓ |
|---|---|---|---|---|---|---|
| Random | 0.730 | 0.747 | 0.180 | 0.370 | 0.067 | 0.607 |
| Attention-only | 0.637 | 0.952 | 0.302 | 0.060 | 0.371 | 0.330 |
| Gradient × Input | 0.662 | 0.922 | 0.327 | 0.156 | 0.670 | 0.235 |
| Integrated Gradients | 0.656 | 0.926 | 0.327 | 0.156 | 0.631 | 0.239 |
| LIME | 0.605 | 0.964 | 0.406 | 0.062 | 0.389 | 0.394 |
| KernelSHAP | 0.613 | 0.934 | 0.373 | 0.100 | 0.316 | 0.452 |
| ASSE (proposed) | 0.666 | 0.837 | 0.327 | 0.238 | 0.416 | 0.373 |
| RED (proposed) | 0.686 | 0.799 | 0.289 | 0.169 | 0.990 | 0.155 |
| DPCX (proposed) | 0.661 | 0.794 | 0.292 | 0.251 | 0.965 | 0.158 |

**Table 5. Faithfulness on UNSW-NB15 (n = 150).**

| Method | Del.AUC ↓ | Ins.AUC ↑ | Comp. ↑ | Suff. ↓ | AnoCov@10 ↑ | Repair ↓ |
|---|---|---|---|---|---|---|
| Random | 0.704 | 0.708 | 0.171 | 0.426 | 0.244 | 0.655 |
| Attention-only | 0.670 | 0.861 | 0.290 | 0.346 | 0.227 | 0.678 |
| Gradient × Input | 0.671 | 0.818 | 0.232 | 0.437 | 0.601 | 0.468 |
| Integrated Gradients | 0.682 | 0.811 | 0.170 | 0.423 | 0.613 | 0.460 |
| LIME | 0.441 | 0.955 | 0.627 | 0.111 | 0.319 | 0.701 |
| KernelSHAP | 0.495 | 0.958 | 0.504 | 0.079 | 0.303 | 0.680 |
| ASSE (proposed) | 0.551 | 0.926 | 0.477 | 0.128 | 0.349 | 0.634 |
| RED (proposed) | 0.721 | 0.862 | 0.261 | 0.267 | 0.943 | 0.330 |
| DPCX (proposed) | 0.683 | 0.884 | 0.330 | 0.186 | 0.892 | 0.374 |

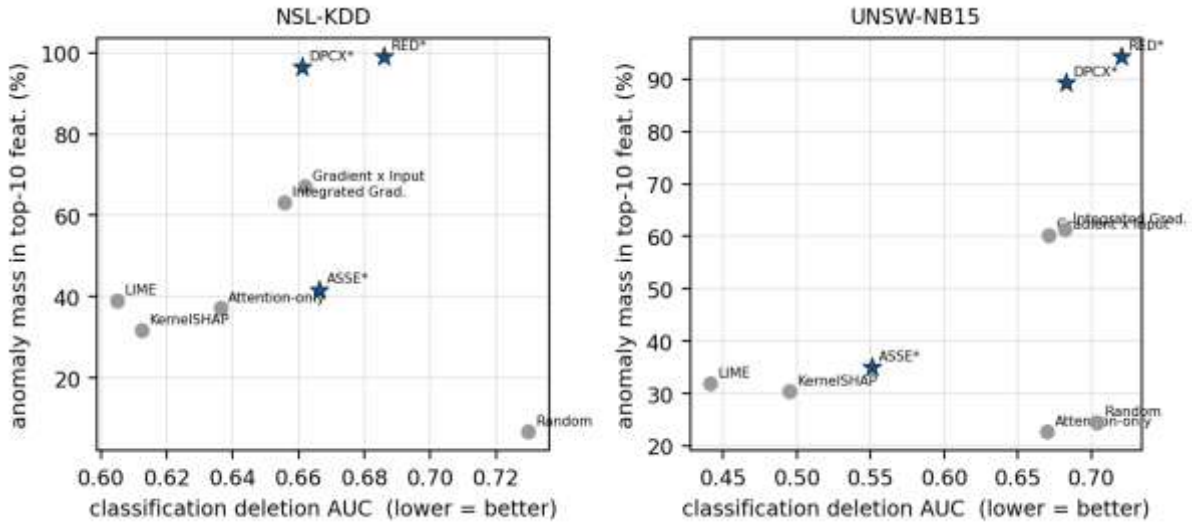



**Fig. 3. Classification faithfulness against anomaly coverage. The desirable region is lower-left to upper-left. Only the proposed dual-path methods occupy the upper region on both datasets.**

The pattern is consistent across both datasets. Classifier-oriented methods win on classifier metrics: LIME attains the best deletion AUC on NSL-KDD (0.605) and UNSW-NB15 (0.441), with KernelSHAP close behind. On the anomaly head they fail. KernelSHAP's ten most important features account for 31.6% of the reconstruction error on NSL-KDD and 30.3% on UNSW-NB15; LIME reaches 38.9% and 31.9%. On NSL-KDD, KernelSHAP's coverage of 0.316 is barely above what a random ranking achieves once feature count is accounted for, and its repair AUC of 0.452 is worse than that of a single backward pass of gradient-times-input (0.235).

RED, being exact, captures 99.0% and 94.3%. DPCX captures 96.5% and 89.2% while recovering most of the classifier faithfulness that RED alone gives up — on NSL-KDD its deletion AUC of 0.661 sits between gradient-times-input and ASSE, and well clear of random. No single-path method occupies both regions. That is the finding: an analyst reading a SHAP explanation of a hybrid detector is reading an account of one head, and the omitted head is the one that generalises.

A secondary observation concerns Attention-only, which posts a strong insertion AUC (0.952) alongside a mediocre deletion AUC. Insertion rewards attributions that concentrate on features sufficient to restore a confident prediction, which coarse token-level attention can achieve without localising the features whose removal actually damages the decision. This asymmetry is a reason to read the two curves together rather than either alone, and is consistent with the cautions raised about attention-as-explanation [7], [8].

### 6.4 Global explanations

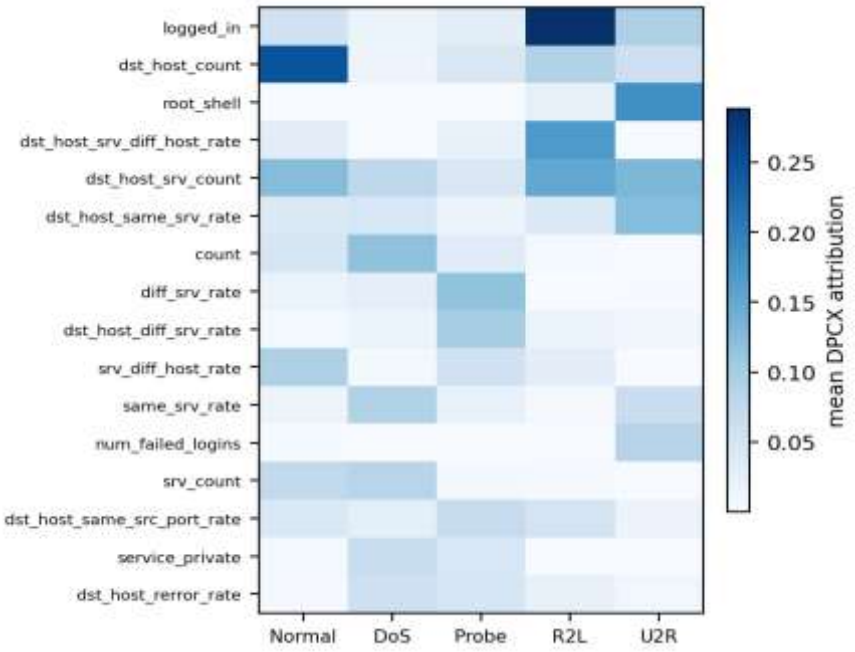



**Fig. 4. Mean DPCX attribution per predicted family on NSL-KDD, restricted to the sixteen most-attributed features.**

**Table 6. Class-level DPCX prototypes on NSL-KDD.**

| Predicted family | Top-5 DPCX features (mean attribution) |
|---|---|
| Normal | dst_host_count (0.249); dst_host_srv_count (0.124); srv_diff_host_rate (0.093); srv_count (0.075); logged_in (0.055) |
| DoS | count (0.118); same_srv_rate (0.089); srv_count (0.084); dst_host_srv_count (0.078); service_private (0.067) |
| Probe | diff_srv_rate (0.116); dst_host_diff_srv_rate (0.098); dst_host_same_src_port_rate (0.069); srv_diff_host_rate (0.055); dst_host_rerror_rate (0.046) |
| R2L | logged_in (0.288); dst_host_srv_diff_host_rate (0.169); dst_host_srv_count (0.152); dst_host_count (0.090); dst_host_same_src_port_rate (0.050) |
| U2R | root_shell (0.181); dst_host_srv_count (0.132); dst_host_same_srv_rate (0.124); logged_in (0.093); num_failed_logins (0.085) |

The prototypes recover textbook attack semantics without supervision of any kind. The DoS signature is dominated by connection-rate and same-service statistics together with rejected-flag and private-service indicators, which is the expected profile of flooding. Probe is characterised by service-diversity features — diff_srv_rate and dst_host_diff_srv_rate — and by error rates, precisely the trace a scanner leaves as it sweeps ports. R2L is dominated by logged_in with strong contributions from host-service diversity and ftp_data, consistent with credential-abuse over remote services. U2R is led by root_shell and num_failed_logins, the canonical markers of privilege escalation.

This agreement is qualitative and should be treated as such; it is a sanity check that the consensus explanation is not arbitrary, not evidence of faithfulness, which Section 6.3 measures directly.

### 6.5 Zero-day detection: a negative result

We hypothesised that unseen attacks would leave a distinctive trace in explanation space — either as divergence between the two paths (EDS) or as non-conformity to the prototype of the assigned class. Table 7 tests this against established out-of-distribution baselines on the 3,750 genuinely unseen instances in KDDTest+.

**Table 7. Zero-day detection on KDDTest+ (3,750 unseen-attack instances of 22,544). All values are percentages.**

| Novelty score | AUROC | AUPR | FPR@95TPR ↓ | TPR@1%FPR |
|---|---|---|---|---|
| Reconstruction error (RED) | 89.47 | 66.45 | 51.28 | 27.17 |

| Novelty score | AUROC | AUPR | FPR@95TPR ↓ | TPR@1%FPR |
|---|---|---|---|---|
| RED + Mahalanobis (fused) | 88.96 | 61.99 | 53.12 | 16.24 |
| Mahalanobis (fused rep.) [14] | 85.45 | 52.01 | 58.03 | 10.27 |
| Energy score [15] | 79.61 | 37.37 | 69.72 | 1.17 |
| Predictive entropy | 78.86 | 35.48 | 59.17 | 2.61 |
| Max softmax probability [13] | 78.79 | 35.57 | 58.31 | 2.53 |
| Mahalanobis (explanation space) | 76.89 | 29.20 | 66.84 | 0.35 |
| EDS (proposed, negative result) | 73.15 | 33.30 | 97.74 | 2.59 |
| Prototype conformity (negative) | 67.80 | 25.86 | 77.65 | 0.67 |

The hypothesis fails. EDS reaches 73.15% AUROC and prototype conformity 67.80%, both far below the 89.47% obtained by the reconstruction error alone. Two further checks confirm this is not a matter of sign or calibration. EDS correlates at −0.65 with the entropy of the reconstruction attribution, indicating that much of what it measures is attribution concentration rather than genuine cross-path disagreement. And when the two signals are combined under a fusion trained only on known classes — so that no novel-class information leaks — the result (88.65%) is slightly worse than the reconstruction error by itself. Explanation-derived statistics add nothing here.

We report this because the opposite claim is intuitive, is easy to assert without a strong baseline, and would be wrong. The proper conclusion is that the anomaly branch's own scalar output is already close to the best available signal from this architecture, and that the contribution of explanation lies in describing why an alert fired rather than in deciding whether it should.

What is worth emphasising is how strong that baseline is. The reconstruction error outperforms maximum softmax probability by more than ten AUROC points and the energy score by nearly ten, which is a substantive validation of the parent architecture's autoencoder branch [1] and consistent with prior autoencoder-based zero-day work [21], [22].

### 6.6 What the anomaly branch recovers

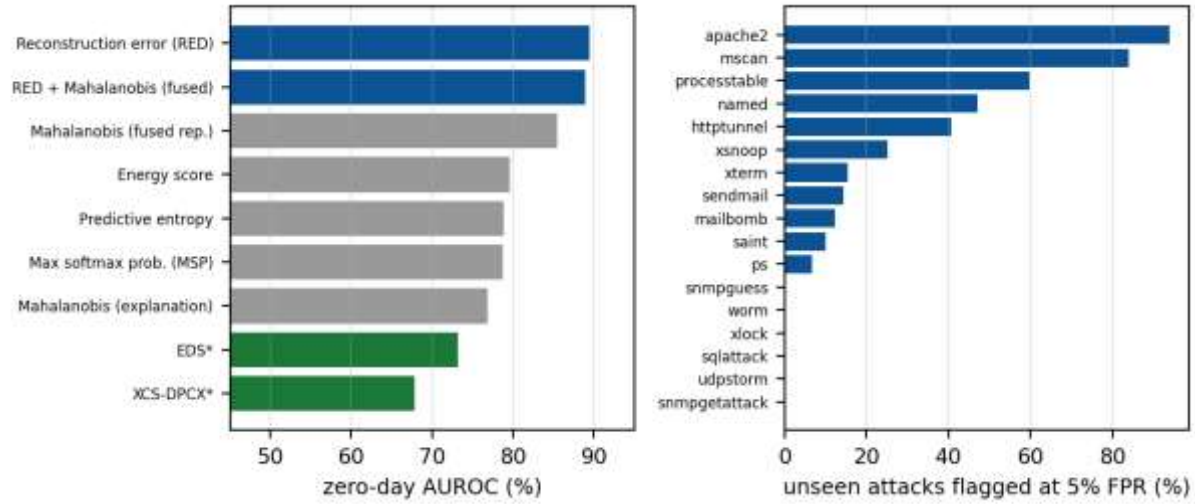



**Fig. 5. Left: zero-day AUROC by novelty score. Right: proportion of each unseen attack type flagged by the anomaly branch at a 5% false-positive rate.**

Of the 3,750 unseen-attack instances, 2,132 — 56.9% — are classified as Normal by the classification head. These are the flows a classifier-only system, and a classifier-only explanation, would miss entirely. Thresholding the anomaly branch at a 5% false-positive rate on normal traffic recovers 981 of them, or 46.0%.

**Table 8. Per-type detection of the 17 unseen attack families by the anomaly branch.**

| Unseen attack | n | Flagged @5% FPR (%) | Unseen attack | n | Flagged @5% FPR (%) |
|---|---|---|---|---|---|
| apache2 | 737 | 94.0 | mailbomb | 293 | 12.3 |
| mscan | 996 | 83.9 | saint | 319 | 10.0 |
| processtable | 685 | 59.7 | ps | 15 | 6.7 |

| Unseen attack | n | Flagged @5% FPR (%) | Unseen attack | n | Flagged @5% FPR (%) |
|---|---|---|---|---|---|
| named | 17 | 47.1 | snmpguess | 331 | 0.3 |
| httptunnel | 133 | 40.6 | snmpgetattack | 178 | 0.0 |
| xsnoop | 4 | 25.0 | xlock | 9 | 0.0 |
| xterm | 13 | 15.4 | worm | 2 | 0.0 |
| sendmail | 14 | 14.3 | udpstorm / sqlattack | 4 | 0.0 |

Table 8. Per-type detection of the 17 unseen attack families by the anomaly branch.

The breakdown is more informative than the aggregate. Volumetric and scanning attacks are caught reliably — apache2 at 94.0% and mscan at 83.9% — because they distort flow statistics far from the normal manifold. The failures are systematic rather than random: snmpgetattack and snmpguess are detected at essentially zero rate. These are protocol-mimicry attacks whose flow-level statistics are almost indistinguishable from legitimate SNMP traffic, so a reconstruction-based detector operating on flow features has nothing to register. No amount of explanation repairs this; it is a limitation of the feature representation, and it identifies payload-level or sequence-level features as the necessary remedy.

### 6.7 Stability and computational cost

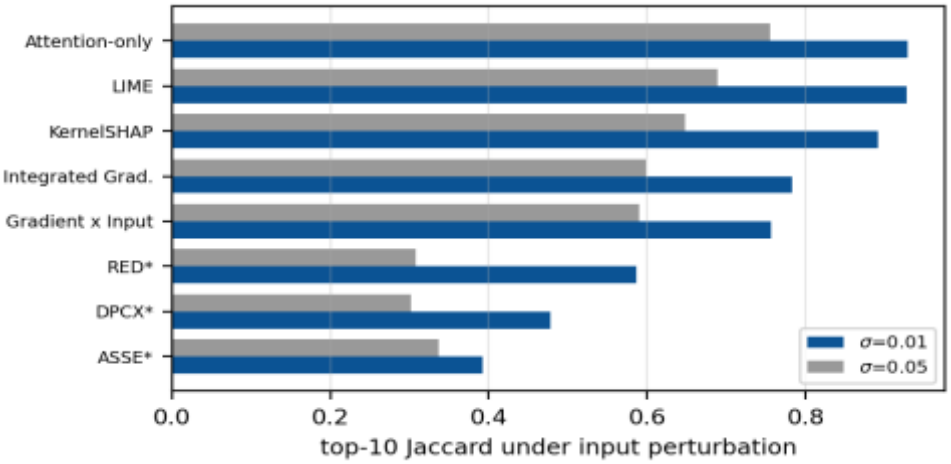



**Fig. 6. Top-10 Jaccard stability under Gaussian input perturbation.**

**Table 9. Explanation stability and query cost per explained alert.**

| Method | Stability σ=0.01 | Stability σ=0.05 | Model queries |
|---|---|---|---|
| Attention-only | 0.930 | 0.756 | 1 |
| LIME | 0.928 | 0.690 | 512 |
| KernelSHAP | 0.891 | 0.648 | 512 |
| Integrated Gradients | 0.783 | 0.599 | 32 |
| Gradient × Input | 0.756 | 0.590 | 1 |
| RED (proposed) | 0.587 | 0.308 | 1 |
| DPCX (proposed) | 0.478 | 0.303 | 257 |
| ASSE (proposed) | 0.392 | 0.337 | 256 |

Here the proposed methods do worst, and we state this plainly. ASSE is the least stable of the eight because a small input perturbation can reorder the prior and therefore change which features enter the salient set, a discrete decision that amplifies small changes. RED inherits the sensitivity of the autoencoder's reconstruction, and DPCX inherits from both. The stability–faithfulness tension this reveals matches the concerns Alvarez-Melis and Jaakkola raised about interpretability methods generally [27]: the methods that track the model most closely also track its local irregularities.

On cost, RED requires a single forward pass — one already performed to compute the anomaly score — against 512 for KernelSHAP and LIME. For a detector processing alerts at line rate this difference is decisive, and it is available for the half of the explanation that current practice omits.

## 7. LIMITATIONS AND THREATS TO VALIDITY

Screening bias. ASSE estimates Owen values under a coarse partition, not Shapley values of the full game. Its advantage is confined to budgets below roughly 256 queries, and paired KernelSHAP should be preferred above that.

Two negative results stand. Attention-guided importance sampling of Shapley coalitions retains an effective sample size of only 1.5% at 2,048 draws, with a log-weight spread of 61 nats; the Shapley kernel and any attention-shaped proposal are too dissimilar in 122 dimensions for self-normalised importance weighting to survive. And explanation-derived novelty signals do not beat the reconstruction error. Both are reported in full rather than omitted.

Stability. The proposed attributions are less stable under input perturbation than LIME or KernelSHAP. Ensembling explanations over a small perturbation neighbourhood is the obvious mitigation and is left to future work.

Scope. Experiments cover two flow-level benchmark datasets and one architecture. NSL-KDD in particular is dated, and although the unseen-attack structure of KDDTest+ makes it uniquely suitable for the zero-day protocol used here, generalisation to modern encrypted and IoT traffic requires separate validation. CIC-IDS2017, evaluated in the parent study [1], was not available in this environment.

Faithfulness metrics. Deletion and insertion rely on a background substitution that pushes inputs off the data manifold, a criticism that applies to all perturbation-based faithfulness evaluation. Anomaly coverage does not share this problem, being computed on the unperturbed input, which is part of why we prefer it for the anomaly head.

Adversarial robustness. Explanations can themselves be manipulated [28], and adversarial pressure on intrusion detection is well documented. We have not evaluated X-HACA under attack on the explanation mechanism.

## 8. CONCLUSION AND FUTURE WORK

This paper asked whether explaining only the classifier of a hybrid intrusion detector leaves something important unexplained, and answered it quantitatively. On NSL-KDD, the ten features nominated by KernelSHAP account for 31.6% of the reconstruction evidence the detector acts upon; LIME accounts for 38.9%. Since SHAP and LIME are the two methods that dominate explainable intrusion detection practice, this shortfall applies to most deployed explanations of hybrid detectors.

The remedy proposed is deliberately simple. Reconstruction Error Decomposition attributes the anomaly score exactly, in closed form, at zero query cost, and Dual-Path Consensus Explanation fuses it with a discriminative attribution. DPCX was the only method of nine that remained faithful to both heads on both datasets, capturing 96.5% and 89.2% of anomaly evidence while retaining competitive classifier faithfulness. Equally, the paper records what did not work. Attention-guided importance sampling degenerates. Explanation-space novelty signals do not improve on the reconstruction error, which at 89.47% AUROC is itself a strong zero-day detector and outperforms the standard out-of-distribution baselines by a wide margin. Reporting these plainly is what allows the positive claims to carry weight.

The operational finding is the one we would emphasise to practitioners. More than half of genuinely unseen attacks in KDDTest+ are labelled benign by the classifier, and the anomaly branch recovers 46% of them at a 5% false-positive rate. A hybrid detector's value against unknown threats resides substantially in the branch that current explanation practice ignores.

Future work follows from the limitations. Perturbation-ensembled attributions should recover stability without sacrificing faithfulness. The systematic failure on protocol-mimicry attacks such as snmpgetattack motivates payload-aware and sequence-level features, since the deficiency lies in the representation rather than the explanation. Extending the dual-path formulation to variational and contrastive anomaly heads, evaluating on contemporary encrypted-traffic benchmarks, and testing the framework against adversarial manipulation of the explanation mechanism are the natural next steps. A user study with security analysts would establish whether the measured faithfulness gains translate into faster or more accurate triage, which is ultimately the standard by which an explanation should be judged.

## REFERENCES

1. I. Hussain and P. Laxkar, “A Robust and Adaptive Deep Learning-Powered Intrusion Detection Framework for Real-Time Cyber Attack Detection,” ITM SLS Baroda University, Vadodara, India. (Parent study.)
2. S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017, pp. 4765–4774.
3. M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
4. M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” in Proc. 34th Int. Conf.

Machine Learning (ICML), 2017, pp. 3319–3328.
5. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017, pp. 5998–6008.
6. S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4190–4197.
7. S. Jain and B. C. Wallace, "Attention is not Explanation," in Proc. NAACL-HLT, 2019, pp. 3543–3556.
8. S. Wiegreffe and Y. Pinter, "Attention is not not Explanation," in Proc. EMNLP-IJCNLP, 2019, pp. 11–20.
9. J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "ERASER: A Benchmark to Evaluate Rationalized NLP Models," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4443–4458.
10. V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," in Proc. British Machine Vision Conference (BMVC), 2018.
11. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in Proc. IEEE Symp. Computational Intelligence for Security and Defense Applications (CISDA), 2009, pp. 1–6.
12. N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," in Proc. Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1–6.
13. D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," in Proc. Int. Conf. Learning Representations (ICLR), 2017.
14. K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," in Advances in Neural Information Processing Systems 31 (NeurIPS), 2018, pp. 7167–7177.
15. W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based Out-of-distribution Detection," in Advances in Neural Information Processing Systems 33 (NeurIPS), 2020, pp. 21464–21475.
16. G. Owen, "Values of Games with a Priori Unions," in Mathematical Economics and Game Theory, Berlin: Springer, 1977, pp. 76–88.
17. L. S. Shapley, "A Value for n-Person Games," in Contributions to the Theory of Games II, Princeton, NJ: Princeton Univ. Press, 1953, pp. 307–317.
18. I. Covert and S.-I. Lee, "Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression," in Proc. 24th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2021, pp. 3457–3465.
19. O. Arreche, T. R. Guntur, J. W. Roberts, and M. Abdallah, "E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection," IEEE Access, vol. 12, pp. 23954–23988, 2024.
20. D. Gaspar, P. Silva, and C. Silva, "Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron," IEEE Access, vol. 12, pp. 30164–30175, 2024.
21. V. Hnamte, H. Nhung-Nguyen, J. Hussain, and Y. Hwa-Kim, "A Novel Two-Stage Deep Learning Model for Network Intrusion Detection: LSTM-AE," IEEE Access, vol. 11, pp. 37131–37148, 2023.
22. H. Hindy, R. Atkinson, C. Tachtatzis, J.-N. Colin, E. Bayne, and X. Bellekens, "Utilising Deep Learning Techniques for Effective Zero-Day Attack Detection," Electronics, vol. 9, no. 10, art. 1684, 2020.
23. M. Mohale and I. C. Obagbuwa, "Evaluating Machine Learning-Based Intrusion Detection Systems with Explainable AI: Enhancing Transparency and Interpretability," Frontiers in Computer Science, vol. 7, art. 1520741, 2025.
24. "A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems to Enhance Transparency and Interpretability in Cybersecurity," Frontiers in Artificial Intelligence, vol. 8, art. 1526221, 2025.
25. "Explainable AI-Based Intrusion Detection Systems for Industry 5.0 and Adversarial XAI: A Systematic Review," Information, vol. 16, no. 12, art. 1036, 2025.
26. R. Taheri, R. Jafari, A. Gegov, F. Arabikhan, and A. Ichtev, "Explainable AI for Federated Learning-Based Intrusion Detection Systems in Connected Vehicles," Electronics, vol. 14, no. 22, art. 4508, 2025.
27. D. Alvarez-Melis and T. S. Jaakkola, "On the Robustness of Interpretability Methods," in Proc. ICML Workshop on Human Interpretability in Machine Learning (WHI), 2018.
28. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2020, pp. 180–186.
29. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in Advances in Neural Information Processing Systems 31 (NeurIPS), 2018, pp. 9505–9515.
30. C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
31. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López,

D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.

32. S. Ennaji, F. De Gaspari, D. Hitaj, A. Kbidi, and L. V. Mancini, "Adversarial Challenges in Network Intrusion Detection Systems: Research Insights and Future Prospects," arXiv:2409.18736, 2024.