Evaluating feature relevance XAI in network intrusion detection

Julian Tritscher[1], Maximilian Wolf[2], Andreas Hotho[1], and Daniel Schlör[1]

[1] University of Würzburg, Am Hubland, 97074 Würzburg, Germany
[2] Coburg University of Applied Sciences, 96450 Coburg, Germany
{tritscher, m.wolf, hotho, schloer}@informatik.uni-wuerzburg.de

Abstract. As machine learning models become increasingly complex, there is a growing need for explainability to understand and trust the decision-making processes. In the domain of network intrusion detection, post-hoc feature relevance explanations have been widely used to provide insight into the factors driving model decisions. However, recent research has highlighted challenges with these methods when applied to anomaly detection, which can vary in importance and impact depending on the application domain. In this paper, we investigate the challenges of post-hoc feature relevance explanations for network intrusion detection, a critical area for ensuring the security and integrity of computer networks. To gain a deeper understanding of these challenges for the application domain, we quantitatively and qualitatively investigate the popular feature relevance approach SHAP when explaining different network intrusion detection approaches. We conduct experiments to jointly evaluate detection quality and explainability, and explore the impact of replacement data, a commonly overlooked hyperparameter of post-hoc feature relevance approaches. We find that post-hoc XAI can provide high quality explanations, but requires a careful choice of its replacement data as default settings and common choices do not transfer across different detection models. Our study showcases the viability of post-hoc XAI for network intrusion detection systems, but highlights the need for rigorous evaluations of produced explanations.

Keywords: Anomaly detection · Feature relevance · Explainable AI.

1 Introduction

Explainable artificial intelligence (XAI) is a rapidly growing research field that has recently gained particular attention in the security-critical application domain of network intrusion detection [16]. In this domain, the increasing complexity of detection systems has led to a growing use of post-hoc explainability methods that can shed light on the decision-making process of trained machine learning models [1,10,12,15,18,20,25,26,32,33,34]. Despite the widespread use of post-hoc XAI for explaining network intrusion detection systems (NIDS), there is a lack of quantitative evaluation of the resulting explanations, as most studies are limited to small qualitative discussions of single data point explanations.

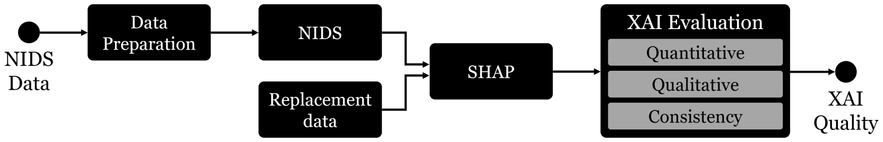


Fig. 1: Experimental setup for the XAI evaluation of SHAP within NIDS. We investigate XAI performance across multiple NIDS and across different choices of replacement data.

As intrusion attacks are usually rare and not always known ahead of time, a subfield of NIDS models the task as anomaly detection [3]. Within the field of anomaly detection, however, recent research finds that the application of popular post-hoc XAIs to machine learning models poses significant challenges [30]. These challenges arise from the selection of replacement data (also called reference, background or baseline data in literature), a common hyperparameter of popular post-hoc XAIs, that is used to contrast observed feature values with alternative observations. This hyperparameter is often overlooked in applications and can have domain-specific impact on the performance of various XAI techniques, vastly decreasing explanation quality if set inappropriately.

In this paper, we address the lack of quantitative evaluation of post-hoc XAIs when applied to NIDS by building an expert-annotated dataset. We conduct a comprehensive quantitative and qualitative evaluation of the impact of replacement data on the performance of SHAP [14], a commonly used post-hoc XAI method in anomaly-based NIDS [1,10,12,15,18,20,25,26,32,33,34]. We evaluate SHAP across multiple NIDS and multiple established choices of replacement values[3]. Our experimental setup is illustrated in Figure 1. Our study shows that the choice of replacement data is critical for obtaining good explanations and that the optimal selection strategy for replacement data not only depends on the application domain but also on the model being explained. We further demonstrate that commonly used replacement values do not always lead to good explanations, making quantitative evaluations of explanations an essential step in building new explainable NIDS.

In summary, our contributions are as follows: (1) We rigorously evaluate the popular post-hoc XAI method SHAP when applied to NIDS both qualitatively and quantitatively using ground truth explanations, finding that SHAP can indeed provide high-quality explanations for NIDS. (2) We systematically investigate the impact of replacement data and show that common choices do not always result in good explanations. This emphasizes the importance of this often overlooked hyperparameter and demonstrates the need for quantitative evaluation of XAI in practice.

The remainder of this paper is structured as follows. In Section 2, we first outline related work in terms of explainable NIDS. We then describe the methodology in Section 3, covering the dataset and preprocessing, anomaly detection

[3] Code and annotations are available under `https://professor-x.de/xai-nids`.

and XAI methodology, as well as our evaluation protocol. In Section 4, we outline our experimental setup and report and discuss intrusion detection performance, as well as XAI performance qualitatively and quantitatively before we conclude this work in Section 5.

2 Related work

The majority of research literature, as well as commercial NIDS, is based on two main approaches: misuse detection, and anomaly detection. Misuse detection is typically modeled as supervised classification task that detects known threats based on predefined patterns, making it the most commonly used approach in NIDS [4]. In contrast, anomaly detection-based NIDS is an unsupervised or semi-supervised task that identifies anomalies which deviate from a well-defined normal behavior.

A survey of current methods, challenges, and opportunities for explainable intrusion detection systems (X-IDS) has been conducted by Neupane et al. [16], providing a comprehensive overview of the state-of-the-art in a variety of different modeling approaches with respect to explainability technique and machine learning methodology. When developing X-IDS, a black-box approach is often recommended [16]. This is reflected in the significant amount of research that incorporates SHapley Additive exPlanations (SHAP) [14] in NIDS, allowing to explain arbitrary black-box models [1,10,12,15,18,20,25,26,32,33,34]. All of these approaches include SHAP as post-hoc XAI approach to generate explanations for misuse detection, which are then exemplarily discussed with respect to the specific features that the authors would expect to explain different types of attacks. In addition to these distinct interpretations of selected examples, none of these studies investigates the quality of explanations from a quantitative perspective.

The study by Dang [5] stands out as one of the few that quantitatively evaluates explainability for intrusion detection. It applies partial dependence plots and SHAP and compares pre- and post-explainability-based feature selection. In contrast, our work models NIDS as anomaly detection task and incorporates ground truth annotations to directly evaluate explainability of a specific model, instead of relying on an indirect evaluation that assesses whether relevant features give sufficient prediction quality with a new model. Although misuse detection models are common in studies on explainable NIDS, only few proposed works focus on explainability of anomaly detection-based NIDS. The Gradient-based Explainable Variational Autoencoder (GEE) [17] is a framework to detect and explain anomalies in network traffic, which analyzes the gradients contributed by each feature of the data point to explain anomalies. However, their evaluation of these gradient fingerprints as explanation is limited to a discussion of examples and their clustering. The most similar related work to our study is presented by Antwarg et al. [2]. In their study, an Autoencoder is used together with SHAP for explainable anomaly detection. They investigate the quality of explanations on an artificial dataset quantitatively, whereas the evaluation on real-world datasets is limited to the surrogate task of reducing anomality similar to [5] and expert interviews. Additionally, they highlight the potential influence of replacement data on explanation quality but leave evaluations as future work.

With our study, we extend the existing work in two aspects. First, in absence of an "application-grounded evaluation" with real humans and the real task [8], we collect ground truth explanations from three domain experts in the domain of network intrusion detection for quantitative evaluation. Second, we address the issue of replacement data as raised by Antwarg et al. [2] and systematically evaluate several approaches to select these replacement data for different anomaly detection models applied to the CIDDS-001 dataset [23] in a quantitative and qualitative evaluation.

3 Methodology

To study the performance of post-hoc feature relevance explanations in NIDS, we first obtain ground truth explanations through an annotation process with three domain experts. We train multiple machine learning models in an anomaly detection setting through a hyperparameter study, and apply post-hoc feature relevance explanations to the resulting best performing models. The ground truth explanations allow us to then follow the experimental setting of [31] for XAI evaluation. In the following, we present the used data, pre-processing steps, machine learning models, XAI approach, and XAI evaluation setup.

3.1 Data and Labeling Process

To validate explanations within multiple NIDS, we use the established CIDDS-001 dataset [24], which we additionally augment with ground truth explanations of attacks. CIDDS-001 features network traffic of a simulated computer network of virtual machines, where clients interact within the network via scripted normal actions or in different attack scenarios. Attacks included within the dataset are denial-of-service (dos) attacks that target available services within the local network, port scans that test for open ports of nodes within the local network, ping scans that sweep the local network to discover IP addresses in use, and brute force attacks that attempt to establish a password-protected ssh connection to an internal node by repeatedly trying passwords using a brute force algorithm.

The total dataset consists of 4 weeks of traffic, with attacks included in week 1 and week 2. In our experiments we use anomaly-based NIDS that train on the exclusively normal data of week 3 and 4 (train), validating hyperparameters on week 2 (valid) and testing on week 1 (test).

For the XAI evaluation, 20 attacks are sampled randomly from each of the available dos, port scan, ping scan, and brute force attacks within the test set, obtaining 80 attack data points in total. To gain ground truth explanations, we conduct an annotation process with three intrusion detection experts. All three experts are given a brief introduction to the data, and are then tasked to independently annotate the 80 attacks. Annotations are created on a per-data basis, where experts assess each feature of a given attack data point regarding whether it is indicative of the underlying attack. Experts are tasked with marking all indicative features within the attack, creating a binary ground truth of relevant features. Overall, the three experts achieved an inter-annotation agreement of 88%, based on Krippendorff's alpha coefficient [11]. Differences in annotation were then discussed among the experts and unified to obtain the ground truth.

3.2 Pre-processing

Network data processed by NIDS commonly consists of both numerical (e.g. packet sizes) and categorical features (e.g. IP addresses and ports) [7].

Categorical features with small numbers of observed combinations, such as the type of network traffic, are commonly encoded using a one-hot representation. For categorical features such as IP addresses or ports, that may contain large numbers of observed values, direct one-hot encoding is undesirable to prevent an explosion in feature space. Common representations are one-hot encoding after replacing specific value groups with dummy tokens to reduce feature space (e.g., modeling all external IP addresses through one token [29]), modeling IP addresses and ports as numerical variables, bit-wise encoding, or learning of vector representations [22]. In this work, we utilize one-hot encoding for all categorical variables, using aggregation for IP addresses and ports to retain the categorical nature of these features while limiting the increase in feature space that would result from direct one-hot encoding. For IP addresses, we aggregate external IPs into one token as the CIDDS-001 dataset is focused on internal private network traffic. For ports, we focus on standardized ports, aggregating all ports above 1024 into one token, and additionally grouping all ports that occur less than 10 times within the training data.

Numerical features are commonly standardized to prevent machine learning models from showing sensitivity to feature value ranges. Possible standardization techniques include min-max scaling, z-score scaling, or quantization. In our experiments, we follow [31] by using quantization for all numerical features. We create buckets with equal value frequency according to the training data, and additionally limit the out-most buckets to only 1% of the data to capture outlier values and highlight them for the machine learning models.

Since we investigate anomaly-based NIDS that cannot model sequential dependencies within the data, we further add an additional feature that aggregates the number of flows that were registered within the last 10 minutes for a specific IP address and port combination, which is a common preprocessing approach [7]. Aggregating this information for source and destination IPs/ports gives detection systems access to a simple representation of usual traffic frequencies when learning the normal network behavior.

3.3 Intrusion Detection

In our evaluation we focus on anomaly-based NIDS, evaluating the explanation process of attacks detected by three well-established anomaly detection models that have been successfully used as NIDS [17,19,29].

Autoencoders (AEs) [9] are under-complete neural networks that learn a simplified data representation of normal behavior by reproducing their inputs at the output layer. Isolation Forests (IFs) [13] are ensemble-based models that use the concept of isolation to identify anomalies using multiple decision trees. One-class Support Vector Machines (OC-SVMs) [27] are maximum margin classifiers that detect anomalies by constructing a hyperplane which separates the given data from the origin in the feature space. For computational efficiency, the used OC-SVM is trained on a random slice of 0.1% of the available training data.

We evaluate the suitability of these methods through a parameter study, and report our results on multiple random seeds during our experiments to capture the statistical variation in the non-deterministic IF and AE, as well as the random training data sampling for the OC-SVM. Results are reported on the established area under the precision recall curve (PR) and area under the receiver operator characteristic curve (ROC) scores. We report all results on both metrics, but rely on the PR score to select hyperparameters, as it is known to be more suited to unbalanced settings such as anomaly detection [6].

3.4 XAI: SHAP

Kernel SHapley Additive exPlanations (SHAP) [14] is a model-agnostic post-hoc XAI approach, that assigns each feature a score which represents how much it contributed to a single model decision. These feature relevance explanations are obtained by repeatedly removing feature combinations from the input and monitoring the model output. Since many machine learning models can not handle missing feature values, SHAP instead replaces values using replacement data that may be chosen as hyperparameter.

For this replacement data, multiple choices exist in literature. Next to SHAP's default implementation using cluster center points of *k-means* clustering from training data, and the use of the *zero-vector* or overall *mean* of training data established in classification settings, replacements that are conditional on the data point to explain can be used in anomaly detection [30]. The latter replacement option may be chosen to prevent the creation of new anomalies when placing normal replacement values into the potentially unfitting context of the data point to explain. Possible options are the use of nearest neighbors (*NN*) from normal data, or gradient-based optimization procedures (*opt*) that generate a normal data point in the proximity of the point to explain [28].

While the choice of replacement values was found to have great influence on the explanations of anomaly detection models [30], current works that employ SHAP in NIDS currently either do not mention replacement values at all [1,10,12,15,20,25,26,32] or use SHAP's default implementation [18], which motivates the investigation of replacement values for explaining NIDS.

3.5 XAI evaluation

For qualitative inspection of their explanations, SHAP provides visualizations that show the contribution of features to a model decision. While applications of SHAP in NIDS perform brief qualitative inspections using these plots [1,10,12,15,18,20,25,26,32], they do not conduct a rigorous quantitative evaluation, which proves difficult due to the lack of ground truth explanations.

Using the binary ground truth obtained through our expert labeling process described in Section 3.1, we are able to evaluate SHAP explanations with established metrics. This evaluation is based on the observation that well performing NIDS systems need to rely on features that are indicative of an attack to successfully separate attacks from normal behavior. Since the binary ground truth marks features that are indicative of an attack and detection performance can be assessed prior to explanation quality, established metrics can be used to assess whether indicative features are rated as more relevant than other features [31]. Following [31], we use two metrics to compare the feature relevance scores of a single data point with the ground truth explanations. ROC scores favor correctly identifying anomalous features within the highest ranking results over identifying all anomalies with decent scores. Cosine similarity (COS), on the other hand, favors a complete match of the entire ground truth explanation, showing how well the XAI highlighted all anomalous features. We report both metrics, but focus our evaluation on ROC scores, as machine learning models do not need to find all suspicious features to identify an attack.

Finally, we also conduct the consistency evaluation of [31] that aims to discover whether similar attacks are detected and explained in a consistent way. This may, for example, be used in practice to generate attack fingerprints based on common explanation patterns, as illustrated in [17]. To showcase the similarity of explanations, we remove all features with an impact of less than 25% of the most influencing feature to remove noise and calculate the Hamming distance between all fraudulent samples. The resulting distances may be visualized as a heatmap, where data points are ordered by their attack type.

4 Experiments

In this section, we discuss the anomaly detection results, as well as the results from an XAI perspective.

4.1 Anomaly detection results

To ensure that explanations are generated for models that are capable of detecting intrusion attacks, we conduct a parameter study through grid search. The investigated parameter sets are reported in Table 1. The best models are chosen through PR score on the *eval* split, and performance is reported on both *eval* and the independent *test* split in Table 2. All models score highly on attacks within the test dataset on both PR and ROC score, with little statistical fluctuation across different random seeds. This shows that the models are suited to detect attacks within the netflow data, and allows us to use a model with these hyperparameter settings for generating explanations.

Table 1: Hyperparameter grid with tested parameter sets for investigated NIDS.

approach	parameters
OC-SVM	kernel $\in$ [rbf], $\gamma \in$ [1e3, 1e2, 1e1, 1e0, 1e-1, 1e-2, 1e-3], $\nu \in$ [0.2, 0.4, 0.6, 0.8]
AE	neurons $\in$ {[32, 16, 8, 16, 32], [64, 32, 16, 32, 64], [128, 64, 32, 64, 128], [64, 32, 16, 8, 16, 32, 64], [128, 64, 32, 16, 32, 64, 128], [256, 128, 64, 32, 64, 128, 256], [128, 64, 32, 16, 8, 16, 32, 64, 128], [256, 128, 64, 32, 16, 32, 64, 128, 256], [512, 256, 128, 64, 32, 64, 128, 256, 512]}, learning rate $\in$ [1e-2, 1e-3, 1e-4], batch size $\in$ [2048]
IF	trees $\in$ [16, 32, 64, 128], max samples $\in$ [0.4, 0.6, 0.8, 1.0], max features $\in$ [0.4, 0.6, 0.8]

4.2 XAI results

For our explanation evaluation, we run SHAP on all anomaly detection models, explaining the 80 attack data points labeled as described in Section 3.1. To capture the impact of different replacement data choices within SHAP, we use SHAP with the replacement options discussed in Section 3.4, namely the *zero-vector*, *k-means* cluster centers, overall *mean*, *NN*, and *opt*. All replacement methods that require data were fitted only on the training data. We additionally use the gradient-based optimization process to explain AE, as it is the only differentiable architecture. To contrast SHAP's XAI results, we calculate baseline explanations. As baselines we report the explanation scores of uniform random noise sampled from $[-1, 1]$, as well as random noise multiplied with the input anomaly (noise×input). Further, for AE we report the scores obtained through using the reconstruction error of individual features as explanation, which is another option to extract explanations from AE [21].

Quantitative results. XAI results are reported through ROC and COS scores for each individual attack, as discussed in Section 3.5. We report the mean and standard deviation of these scores across all 80 labeled attacks for our baselines in Table 3, and for the SHAP explanations in Table 4. While almost all combinations of detection models and replacement values are able to surpass the random baselines, we observe large variation across models and replacements. The highest explanation scores both in ROC and COS are similar for all models, suggesting that explaining NIDS through SHAP is feasible. However, we observe that no coherent best replacement choice exists across anomaly detection models. Additionally, we find that *NN* replacements perform poorly across all

Table 2: Best results of each approach on evaluation and test set, reporting mean and standard deviation of scores across 10 random seeds.

model	PR^{eval}	PR^{test}	ROC^{eval}	ROC^{test}
OC-SVM	$\mathbf{98.6 \pm 0.2}$	$\mathbf{99.5 \pm 0.1}$	$\mathbf{99.9 \pm 0.0}$	$\mathbf{99.9 \pm 0.0}$
AE	98.4 ± 1.1	99.2 ± 0.5	99.8 ± 0.1	$\mathbf{99.9 \pm 0.0}$
IF	98.0 ± 1.4	99.3 ± 0.5	99.7 ± 0.2	99.9 ± 0.1

Table 3: Baselines for quantitative explanation scores.

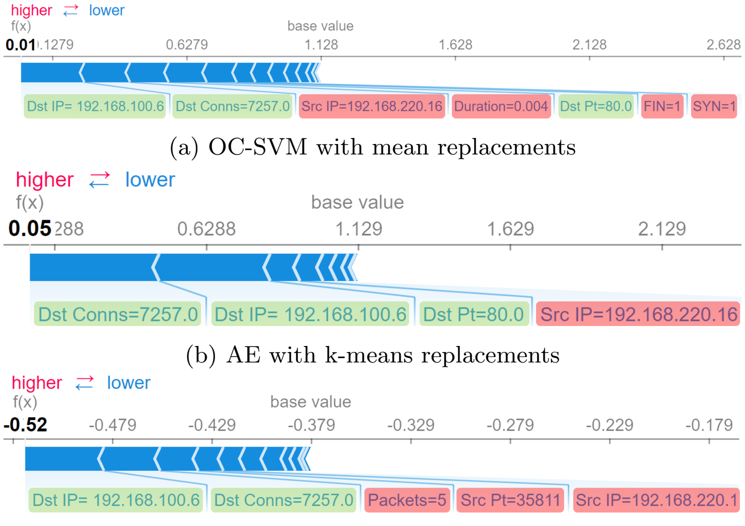
Baseline	ROC$_{XAI}$	COS$_{XAI}$
uniform noise	49.4 ± 21.7	-1.7 ± 38.1
noise×input	49.8 ± 19.2	-0.2 ± 26.4
AE reconstruction	$\mathbf{87.9 \pm 10.3}$	$\mathbf{72.7 \pm 14.8}$

models, which may be caused by its use of proximity. When replacement features are identical to the features they are replacing, by construction, SHAP assigns a feature relevance score of zero. Since *NN* selects data points that are close to the anomaly, and may therefore share feature values with the anomaly, these features can not be recognized by SHAP. In application scenarios where anomalies are not easily identified through simple distance measures, this can introduce a bias into SHAP explanations, leading to poor performance.

For OC-SVM, all replacement values beside *NN* work comparably well, with highest scores achieved through *mean* replacement. For AE, *k-means* and *opt* show high explanation scores, while *mean* performs decently. The *zero-vector* replacements, on the other hand, show problematic behavior. The *zero-vector* itself obtains higher anomaly scores for AE than attacks within the dataset. This causes issues, as SHAP constructs its explanations relative to the model prediction obtained on the replacement data. When the replacement is rated as more anomalous than the data point to explain, SHAP scores highlight features that change the model output away from the anomalous replacement values, as opposed to showing the desired features that move the model output away from normal behavior. As a consequence, the resulting explanations are entirely unsuitable for identifying anomalous feature values, as reflected in the obtained ROC and COS scores. Low performance on *mean* can be explained by the same observation in a weakened form, as the mean data point obtains similar anomaly scores to those of the attacks within the test set. The replacements drawn from the multiple *k-means* cluster centers, however, appear to stabilize this behavior, being scored by AE as less abnormal than attacks and producing high explanation scores. IF exhibits similar behavior on different replacements, obtaining poor explanations when using *k-means*, *mean*, and *NN* replacements. Again, we find that these replacements obtain anomaly scores from IF that are in the same

Table 4: Quantitative explanation scores using different reference values.

	OC-SVM		AE		IF	
Replacement	ROC$_{XAI}$	COS$_{XAI}$	ROC$_{XAI}$	COS$_{XAI}$	ROC$_{XAI}$	COS$_{XAI}$
k-means	88.1 ± 8.7	68.6 ± 12.6	$\mathbf{91.3 \pm 8.7}$	$\mathbf{71.8 \pm 15.9}$	73.5 ± 19.3	52.8 ± 23.1
zeros	85.2 ± 9.2	48.1 ± 13.9	51.9 ± 11.7	-20.7 ± 16.6	$\mathbf{87.2 \pm 8.7}$	$\mathbf{65.9 \pm 13.2}$
mean	$\mathbf{90.1 \pm 7.5}$	$\mathbf{71.0 \pm 12.9}$	82.6 ± 12.7	57.4 ± 19.6	65.8 ± 18.7	33.4 ± 25.7
k-NN	71.3 ± 11.9	56.4 ± 18.1	75.4 ± 11.6	61.1 ± 15.7	68.8 ± 11.7	43.8 ± 22.7
opt	N/A	N/A	88.5 ± 8.7	68.9 ± 14.3	N/A	N/A



(a) OC-SVM with mean replacements

(b) AE with k-means replacements

(c) IF with zero-vector replacements

Fig. 2: SHAP explanations for each NIDS with best performing replacement values on a single non-cherry-picked dos attack. Bar width corresponds to feature influence, and features that are indicative of an attack (green) or non-indicative (red) are highlighted according to ground truth. OC-SVM distributes relevance to many different features. AE and IF highlight relevant features well, while IF additionally highlights irrelevant packet size.

value range as the attacks within the dataset. IF only achieves good explanation performance using the *zero-vector* as replacement, which for IF obtains smaller anomaly scores compared to the attacks.

Qualitative results. To gain a more in-depth understanding of the observed explanation behavior, we investigate the individual explanations obtained by different models and replacement values in detail. We utilize SHAP plots to visualize and closely inspect the annotated explanations. Exemplary SHAP plots for different NIDS on a dos attack can be seen in Figure 2.

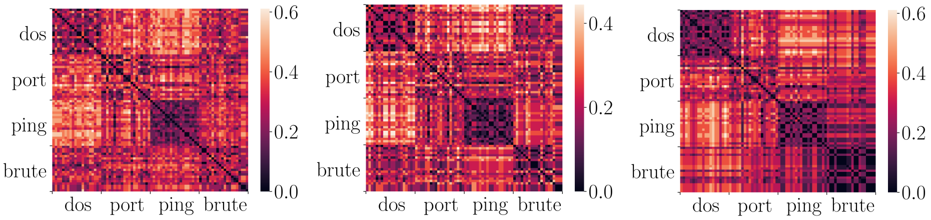
Across all replacement options we make the following observations. We find that for dos attacks OC-SVM correctly identifies high connection frequencies of unusual IP addresses, but also highlights flow duration. Port scans are mainly detected over suspicious IPs and ports, with anomalously high connection frequencies of the attacker also highlighted in some data points. In ping scans, the atypical ICMP protocol in combination with the two IP addresses is found, but not without highlighting some irrelevant packet size related features. Finally, brute force attacks are explained through the anomalous combination of the port used by ssh, together with the IP addresses, as well as through frequent connections to the ssh port. Overall, we find OC-SVM explanations to assign relevance to many different features, a similar behavior to that observed in [31].

While OC-SVM assigns a large amount of relevance to truly anomalous features, this also produces a lot of noise, since many irrelevant features obtain a smaller but not negligible amount of relevance.

AE reliably highlights dos attacks through high connection frequencies and corresponding IP address and port of the attack victim. Port scans are detected mainly through the unusual combination of IP addresses and the victim's port. On ping scans, explanations successfully highlight uncommon ICMP traffic, but do not highlight further indicative features. Brute force attacks are identified by the atypical combination of ssh port and IP addresses, but also contain some noise with highlighted irrelevant TCP flags.

IF finds the anomalous combination of high connection frequency and IP address for the dos victim, but also highlights some irrelevant TCP flags and packet sizes. It successfully identifies port scans through high usage frequency of the attacker's port in combination with the corresponding attacker IP, but also highlights packet sizes. Ping scans are detected through identifying unusual ICMP traffic between attacker and victim IPs and ports. Finally, it identifies brute force attacks through the combination of ssh port and IP addresses, while also highlighting few TCP flags. Overall, IF using *zero-vector* replacement successfully identifies many relevant attack characteristics, and shows a slight preference for highlighting irrelevant TCP flags and packet sizes. Other replacements result in highlighting many irrelevant TCP flags and irrelevant IP address entries.

Consistency results. As a final step in our evaluation, we investigate whether the obtained SHAP explanations show consistent patterns across specific attack scenarios, which might allow for pattern matching and detection of common attack scenarios through explanations. We construct heatmaps as described in Section 3.5 for each anomaly detection model using their best performing replacement values, and visualize the heatmaps in Figure 3. OC-SVM consistently highlights similar features across ping scan attacks, with some clear patterns showing on dos attacks. On port scans and brute force attacks, patterns are less pronounced with multiple points showing very low similarity of their most relevant features. Compared to the other models, OC-SVM's similarities are also lower across all groups of attacks, which again highlights OC-SVM's behavior to split relevance between many individual features, causing noisy explanations.



(a) OC-SVM *mean* (b) AE *k-means* (c) IF *zero-vector*

Fig. 3: Heatmaps showing the similarity of feature relevance explanations.

Both AE and IF explanations show similar patterns with clear similarity between dos, port scan, and brute force attacks, with highlighted features varying more on port scans and IF explanations being slightly more consistent throughout.

Discussion. Overall, we observe that the direct application of SHAP to popular anomaly detection techniques can indeed provide strong explanations within NIDS, validating the popular use of SHAP within this domain. However, we find that the selection of replacement values within SHAP is critical for explanation quality, and there is no apparent replacement that consistently performs well across models. This highlights the need for incorporating quantitative XAI evaluations into the development of explainable NIDS, as the use of common default values for SHAP replacement data does not guarantee high quality explanations.

5 Conclusion

In this paper we constructed ground truth explanations for multiple attack scenarios within an established network intrusion detection dataset through expert annotation. We used these annotations to conduct an in-depth quantitative evaluation of multiple anomaly-based approaches for NIDS when explained by the popular SHAP post-hoc XAI approach and specifically investigated the impact of choosing different replacement strategies.

Our findings indicate that SHAP can produce high-quality explanations for all investigated detection models, but the choice of replacement values significantly impacts the quality of the resulting explanations. Our findings also emphasize the importance of considering the selection of replacement values during the design of explainable NIDS, as well as the systematic evaluation of post-hoc XAI techniques used in the pipeline, since we have demonstrated that SHAP's default replacement choice may not always produce satisfactory explanations for all models.

While this paper yields a positive result for the use of SHAP in NIDS, it is subject to several limitations that provide potential opportunities for future work. Beyond the investigation of XAI quality across multiple NIDS and replacement data, evaluations could be extended to investigate the impact of data preprocessing schemes and different XAI approaches. Additionally, while we quantitatively validate that SHAP successfully identifies anomalous features within NIDS attacks when using appropriate replacement data, the benefit of SHAP from a user perspective is also worthy of investigation. Finally, our work highlights that a consistent choice for replacement values is desirable for anomaly-based NIDS, marking the construction of model-invariant replacement values as relevant future work.

References

1. Alani, M.M., Miri, A.: Towards an Explainable Universal Feature Set for IoT Intrusion Detection. Sensors **22**(15), 5690 (Jul 2022). https://doi.org/10.3390/s22155690
2. Antwarg, L., Miller, R.M., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using Shapley Additive Explanations. Expert Systems with Applications **186**, 115736 (Dec 2021). https://doi.org/10.1016/j. eswa.2021.115736
3. Buczak, A.L., Guven, E.: A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials **18**(2), 1153–1176 (2016). https://doi.org/10.1109/COMST.2015.2494502
4. Casas, P., Mazel, J., Owezarski, P.: Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge. Computer Communications **35**(7), 772–783 (Apr 2012). https://doi.org/10.1016/j. comcom.2012.01.016
5. Dang, Q.V.: Improving the performance of the intrusion detection systems by the machine learning explainability. International Journal of Web Information Systems **17**(5), 537–555 (Sep 2021). https://doi.org/10.1108/IJWIS-03-2021-0022
6. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd int. conf. on Machine learning. pp. 233–240 (2006)
7. Davis, J.J., Clark, A.J.: Data preprocessing for anomaly based network intrusion detection: A review. Computers & Security **30**(6), 353–375 (Sep 2011)
8. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (Mar 2017)
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`
10. Houda, Z.A.E., Brik, B., Khoukhi, L.: "Why Should I Trust Your IDS?": An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks. IEEE Open Journal of the Communications Society **3**, 1164–1176 (2022). https://doi.org/10.1109/OJCOMS.2022.3188750
11. Krippendorff, K.: Content Analysis : An Introduction to Its Methodology, pp. 145–154. Beverly Hills : Sage Publications (1980)
12. Le, T.T.H., Kim, H., Kang, H., Kim, H.: Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method. Sensors **22**(3), 1154 (Feb 2022). https://doi.org/10.3390/s22031154
13. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE Int. Conf. on Data Mining. pp. 413–422. IEEE (2008)
14. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions. CoRR **abs/1705.07874** (2017)
15. Mane, S., Rao, D.: Explaining network intrusion detection system using explainable ai framework. arXiv preprint arXiv:2103.07110 (2021)
16. Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., Seale, M.: Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities (Jul 2022)
17. Nguyen, Q.P., Lim, K.W., Divakaran, D.M., Low, K.H., Chan, M.C.: GEE: A Gradient-based Explainable Variational Autoencoder for Network Anomaly Detection. In: 2019 IEEE Conference on Communications and Network Security (CNS). pp. 91–99 (Jun 2019)
18. Oseni, A., Moustafa, N., Creech, G., Sohrabi, N., Strelzoff, A., Tari, Z., Linkov, I.: An Explainable Deep Learning Framework for Resilient Intrusion Detection in IoT-Enabled Transportation Networks. IEEE Transac-

tions on Intelligent Transportation Systems **24**(1), 1000–1014 (Jan 2023). https://doi.org/10.1109/TITS.2022.3188671

19. Patel, D., Srinivasan, K., Chang, C.Y., Gupta, T., Kataria, A.: Network Anomaly Detection inside Consumer Networks—A Hybrid Approach. Electronics **9**(6), 923 (Jun 2020)
20. Pawlicki, M., Zadnik, M., Kozik, R., Choraś, M.: Analysis and Detection of DDoS Backscatter Using NetFlow Data, Hyperband-Optimised Deep Learning and Explainability Techniques. In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J.M. (eds.) Artificial Intelligence and Soft Computing, vol. 13588, pp. 82–92. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-23492-7_8
21. Ravi, A., Yu, X., Santelices, I., Karray, F., Fidan, B.: General Frameworks for Anomaly Detection Explainability: Comparative Study. In: 2021 IEEE International Conference on Autonomous Systems (ICAS). pp. 1–5 (Aug 2021)
22. Ring, M., Schlör, D., Landes, D., Hotho, A.: Flow-based Network Traffic Generation using Generative Adversarial Networks. Computers & Security **82**, 156–172 (May 2019)
23. Ring, M., Wunderlich, S., Grüdl, D., Landes, D., Hotho, A.: Creation of flow-based data sets for intrusion detection. Journal of Information Warfare **16**, 40–53 (2017)
24. Ring, M., Wunderlich, S., Grüdl, D., Landes, D., Hotho, A.: Flow-based benchmark data sets for intrusion detection. In: Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS), pp. 361–369. ACPI (2017)
25. Sarhan, M., Layeghy, S., Portmann, M.: Evaluating Standard Feature Sets Towards Increased Generalisability and Explainability of ML-based Network Intrusion Detection (Aug 2021)
26. Sauka, K., Shin, G.Y., Kim, D.W., Han, M.M.: Adversarial Robust and Explainable Network Intrusion Detection Systems Based on Deep Learning. Applied Sciences **12**(13), 6451 (Jun 2022). https://doi.org/10.3390/app12136451
27. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation **13**(7), 1443–1471 (Jul 2001)
28. Takeishi, N., Kawahara, Y.: On anomaly interpretation via shapley values. arXiv preprint arXiv:2004.04464 (2020), `https://arxiv.org/pdf/2004.04464.pdf`
29. Torabi, H., Mirtaheri, S.L., Greco, S.: Practical autoencoder based anomaly detection by using vector reconstruction error. Cybersecurity **6**(1), 1 (Jan 2023)
30. Tritscher, J., Krause, A., Hotho, A.: Feature relevance XAI in anomaly detection: Reviewing approaches and challenges. Frontiers in Artificial Intelligence **6** (2023)
31. Tritscher, J., Schlör, D., Gwinner, F., Krause, A., Hotho, A.: Towards Explainable Occupational Fraud Detection. In: Machine Learning and Principles and Practice of Knowledge Discovery in Databases. pp. 79–96. Communications in Computer and Information Science, Springer Nature Switzerland, Cham (2023)
32. Wali, S., Khan, I.: Explainable AI and Random Forest Based Reliable Intrusion Detection system (Dec 2021). https://doi.org/10.36227/techrxiv.17169080.v1
33. Wang, M., Zheng, K., Yang, Y., Wang, X.: An explainable machine learning framework for intrusion detection systems. IEEE Access **8**, 73127–73141 (2020)
34. Zebin, T., Rezvy, S., Luo, Y.: An Explainable AI-Based Intrusion Detection System for DNS Over HTTPS (DoH) Attacks. IEEE Transactions on Information Forensics and Security **17**, 2339–2349 (2022). https://doi.org/10.1109/TIFS.2022.3183390