Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062801

ARTICLE

A New Cybersecurity Approach Enhanced by xAI-Derived Rules to Improve Network Intrusion Detection and SIEM

Federica Uccello[1,2], Marek Pawlicki[3,4], Salvatore D'Antonio[1], Rafał Kozik[3,4] and Michał Choraś[3,4,*]

[1]Centro Direzionale, Department of Engineering, University of Naples 'Parthenope', Isola C4, Napoli, 80133, Italy
[2]Department of Computer and Information Science, Software and Systems, Linköping University, Linköping, 58183, Sweden
[3]ITTI Sp. z o.o., Poznań, 61-612, Poland
[4]Faculty of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, Bydgoszcz, 85-796, Poland

*Corresponding Author: Michał Choraś. Email: chorasm@pbs.edu.pl

Received: 28 December 2024; Accepted: 05 March 2025; Published: 16 April 2025

ABSTRACT: The growing sophistication of cyberthreats, among others the Distributed Denial of Service attacks, has exposed limitations in traditional rule-based Security Information and Event Management systems. While machine learning–based intrusion detection systems can capture complex network behaviours, their "black-box" nature often limits trust and actionable insight for security operators. This study introduces a novel approach that integrates Explainable Artificial Intelligence—xAI—with the Random Forest classifier to derive human-interpretable rules, thereby enhancing the detection of Distributed Denial of Service (DDoS) attacks. The proposed framework combines traditional static rule formulation with advanced xAI techniques—SHapley Additive exPlanations and Scoped Rules - to extract decision criteria from a fully trained model. The methodology was validated on two benchmark datasets, CICIDS2017 and WUSTL-IIOT-2021. Extracted rules were evaluated against conventional Security Information and Event Management Systems rules with metrics such as precision, recall, accuracy, balanced accuracy, and Matthews Correlation Coefficient. Experimental results demonstrate that xAI-derived rules consistently outperform traditional static rules. Notably, the most refined xAI-generated rule achieved near-perfect performance with significantly improved detection of DDoS traffic while maintaining high accuracy in classifying benign traffic across both datasets.

KEYWORDS: Cybersecurity; explainable artificial intelligence; intrusion detection system; rule-based SIEM; distributed denial of service

1 Introduction

1.1 Context and Rationale

In recent years, the landscape of Artificial Intelligence (AI) and Machine Learning (ML) has expanded rapidly, empowering models with predictive capabilities that often exceed human ingenuity. However, these remarkable achievements frequently come at the expense of transparency, as the reasoning behind model outputs remains obscure [1,2]. Recognizing this issue, Explainable Artificial Intelligence (xAI) has emerged as a pivotal research area dedicated to elucidating the internal logic of ML models [3]. By providing insights into how features influence predictions, xAI holds the promise of bridging the gap between opaque algorithms and the practitioners who rely on them.

This pursuit for interpretability is especially important in cybersecurity, where the acceptance and efficacy of AI-driven solutions hinge on trustworthiness. Despite increasing adoption, many decision-makers in this field remain wary of AI's 'black-box' nature [4], further complicated by the legal and ethical implications of automated decisions [5]. As cyberthreats grow more sophisticated, ensuring that detection methods are both accurate and transparent becomes paramount. Here, xAI finds a compelling application, as trust and clarity underpin robust cybersecurity strategies [6–9].

Within the defensive arsenal of organizations, Security Information and Event Management (SIEM) systems have played a leading role by applying human-crafted, rule-based detection schemes [10] for a long time now [10–12]. By many, SIEM systems are believed to be indispensable [13]. While these rules are transparent and grounded in domain expertise, they may not fully capture the complexity and evolving nature of modern threats. Conversely, ML-powered Network Intrusion Detection Systems (NIDS) offer dynamic and data-driven capabilities to identify anomalous activities, yet their black-box models are often met with skepticism and a reluctance to fully integrate them into established workflows [14].

This research aims to unify these two paradigms—traditional rule-based SIEM and advanced ML-driven NIDS—into a synergistic solution. Specifically, this paper proposes a methodology to extract explainable and actionable detection rules directly from an ML model, leveraging xAI techniques. By doing so, the paper provides a conceptual and technological bridge between longstanding cybersecurity practices and the promise of cutting-edge ML algorithms.

While this paper primarily focuses on the application of xAI to enhance detection of DDoS attacks, the methodologies developed here have potential implications for a broader range of cyberthreats, suggesting a promising avenue for future research to expand the scope of these techniques in future work.

1.2 Major Contribution and Extension of This Paper

The principal contribution of this work comes in addressing the current gap in practical approaches leveraging xAI and ML-powered NIDS. Traditional SIEM systems primarily rely on predefined, static rules and patterns to detect anomalies, which often fail to adapt to the dynamic nature of modern cyberthreats. The reliance on pre-established thresholds and criteria can prevent the detection of low-and-slow attack variants, which do not initially trigger these thresholds but can escalate into severe disruptions. A significant drawback of relying on SIEM systems is the knowledge the user needs to possess to be able to update their rules reliably—knowledge which is expensive and hard to find [15,16]. This paper aims at addressing these shortcomings, overcoming some of the limitations of the use of SIEM systems. This is accomplished by augmenting conventional SIEM with xAI insights which distills precise, human-interpretable detection rules from trained models, boosting classification accuracy and maintaining operator confidence.

To the best of the authors' knowledge, this represents a novel and innovative approach, offering a previously unexplored pathway to harmonize established cybersecurity frameworks with the powerful, yet often opaque, potential of AI.

This work is an invited extension of our published conference paper from the previous Asian Conference on Intelligent Information and Database Systems (ACIIDS) held in 2024 [17].

In the earlier study, the methodology and preliminary findings were introduced and validated using a single dataset.

Hereby, the experiments and evaluation are strengthened and broadened by incorporating another dataset, effectively doubling the scope of the evaluation and providing a more rigorous empirical validation of the proposed methodology. It is always important in machine learning, artificial intelligence and cybersecurity research and practice to generalize approaches and solutions on wide range datasets and

scenarios. Therefore, our further efforts in extending the previous work with additional experimental setups on another dataset.

This expanded evaluation underlines the method's scalability and generalizability, and its relevance for a wider range of cybersecurity environments.

1.3 Structure of the Paper

The paper is organized as follows: Section 2 overviews the related work, highlighting the lack of practical approaches; Section 3 delves into the enabling concepts and technologies of the proposed research; Section 4.1 provides details regarding the Datasets and methodology, presenting the correlation rules defined before and after the application of xAI; Section 5 presents the results of the experiments using both sets of rules; Section 6 analyzes the obtained results and provides a comparison between the discovered rules on two different benchmark datasets. Threats to validity and possible future directions are given in Section 7. Finally, Section 8 ends the paper with final remarks. The list of abbreviations used in the paper is enclosed in Table 1.

Table 1: List of abbreviations and their full forms

Abbreviation	Full form
xAI	Explainable Artificial Intelligence
SIEM	Security Information and Event Management
RF	Random Forest
RFM	Random Forest Model
IDS	Intrusion Detection System
IPS	Intrusion Prevention System
DDoS	Distributed Denial of Service
ML	Machine Learning
DT	Decision Tree
SVM	Support Vector Machine
LIME	Local Interpretable Model-Agnostic Explanations
SHAP	SHapley Additive exPlanations
ANCHORS	Rule-based local explanations (not an acronym, it is the name of an algorithm)
SMOTE	Synthetic Minority Over-sampling Technique
ANOVA	Analysis of Variance
CICIDS2017	Canadian Institute for Cybersecurity Intrusion Detection System 2017
WUSTL-IIOT-2021	Washington University in St. Louis Industrial Internet of Things 2021
IIoT	Industrial Internet of Things
PCAP	Packet Capture
MCC	Matthews Correlation Coefficient
BA	Balanced Accuracy
G-mean	Geometric Mean
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

(Continued)

Table 1 (continued)

Abbreviation	Full form
ACIIDS	Asian Conference on Intelligent Information and Database Systems
TCPRtt	TCP Round-Trip Time
Proto	Protocol
SrcRate	Source Rate

2 Related Work

This section reviews the existing literature on xAI applications in cybersecurity, particularly focusing on NIDS, to underscore the unique contribution of this research as related to the backdrop of previous work.

In the last few years, xAI has been gaining popularity in the cybersecurity domain, as it has become the subject of various reviews [6–9] and experiments.

For what concerns Intrusion Detection Systems (IDS) [6] introduces an advanced IDS utilizing ML ensemble methods like Decision Trees (DT), Random Forest (RF), and Support Vector Machines (SVM), achieving promising results. In [18], the critical role of trust management in IDS is explored, emphasizing the need for transparency in ML models utilized in the cybersecurity domain. It discusses the challenges of interpreting "black-box" AI models and the importance of xAI in enhancing trust by enabling human experts to comprehend the model's decision-making process. The paper [19] introduces an IDS employing ML techniques like DT, RF, and xAI on real-world Software-Defined Networking (SDN) data. The evaluation covers various intrusion scenarios, achieving high accuracy. By integrating Machine Learning and xAI using Local Interpretable Model-Agnostic Explanations (LIME), the research aims to boost intrusion detection accuracy and maintain data integrity. The objectives are to classify intrusions using DT and RF algorithms in SDN, compare their performance, and implement LIME to identify key network features contributing to intrusions. In [20], an approach for detecting DDoS attacks using xAI is proposed. The method focuses on identifying attack behaviours in network traffic flows without analyzing packet payloads. By leveraging autoencoders and xAI, it provides a better understanding of the most influential features for attack detection, and setting individual thresholds for such features to support attack detection. The work carried out in [21] aims at contributing to rigorous xAI-equipped IDS for DDoS attacks. It introduces an approach based on artificial immune systems, featuring a decision tree model. The paper details the process of mapping, combining, and merging to transform continuous features into boolean expressions, which are then simplified into prime implicants. The resulting prime implicants serve as rules for DDoS intrusion detection. Most of the research works in the application of xAI in ML-powered NIDS focus on opening the black box of AI, hoping to gather trust for the emerging technology. There is a very relevant gap in the utilisation of knowledge garnered by ML algorithms in NIDS, which would translate into immediately available, actionable intel for the security operatives. The research contained within this paper provides an innovative experimental study to employ xAI to enhance traditional rule-based SIEM systems and IDS, by using ML logic to support the definition of highly accurate discrimination criteria for attack detection. This bridges the gap between the immense benefits of the application of data-based algorithms and the apparent lack of trust in the emerging technology among the cybersecurity community. Despite using various search engines and bibliographic methods, no similar approaches have been found other than the aforementioned works. To the best of the auhtors'knowledge, the proposed research is a novel, innovative approach filling the research gap. In order to illustrate the landscape of current research on the application of xAI in cybersecurity, Table 2 provides a comprehensive overview of related works in the field.

Table 2: Overview of key studies on the application of xAI in cybersecurity, highlighting methodologies, focus areas, and contributions

Reference	Methodology	Focus	Key contributions
Patil et al., 2022	ML ensemble methods (DT, RF, SVM)	IDS	Enhanced model transparency, addressed adversarial attacks
Mahbooba, 2021	xAI applications	IDS trust management	Discussed the importance of transparency to enhance trust in IDS
Karna et al., 2021	ML techniques with LIME	IDS	Enhanced intrusion detection
Kalutharage, 2023	Autoencoders and xAI	DDoS attack detection	Improved attack detection by identifying key features and setting thresholds
Zhou et al., 2022	Artificial immune systems, decision trees	IDS for DDoS attacks	Developed boolean rules from continuous features for DDoS detection
Nwakanma et al., 2023	Review of xAI in NIDS	Review of xAI applications	Highlighted the application of xAI in improving NIDS in autonomous vehicles
Choras, 2020	Exploration of AI implications in cybersecurity	The need for xAI, fairness and security of AI	Discussed the implications of ML/AI in critical applications
Doshi, 2017	Theoretical approaches to xAI	Theoretical exploration of xAI	Discussed foundational methods and approaches in xAI

3 Background and Datasets

This section provides an overview of the enabling technologies and datasets employed in this study, providing the foundational concepts for understanding the contribution of this work.

3.1 Explainable Artificial Intelligence

The focus of xAI is on creating methods that render ML models more transparent and interpretable, ultimately enabling a human audience to comprehend their internal logic [22]. Various techniques fall under the umbrella of xAI, but they can generally be categorized into three main groups:

- **Feature importance:** Approaches that pinpoint the most influential features in a model's decision-making process, assisting practitioners in understanding which inputs are driving predictions.
- **Rule-based explanations:** Techniques that distil complex model logic into human-readable rules, thereby clarifying how certain outcomes are reached.
- **Counterfactual explanations:** Methods that determine the minimal input modifications needed to alter a model's output, shedding light on the underlying rationale and potential scenarios where different predictions could emerge.

By employing these xAI strategies, cybersecurity professionals can gain deeper insights into how ML models detect anomalies, recognize patterns, and differentiate between benign and malicious activities.

This enhanced understanding helps to identify potential biases, troubleshoot unexpected behaviours, and ultimately cultivate greater trust and confidence in the outputs of the models.

3.2 Random Forest Algorithm

The RF classifier is a robust ensemble learning method that integrates the strengths of multiple decision trees. By applying bootstrap aggregation (bagging), it constructs numerous trees on randomly selected subsets of the training data, fostering diversity in their learned patterns. This diversity enables the RF to uncover complex relationships between features and the target variable. Each tree serves as a "weak learner," and their collective predictions are aggregated through majority voting [23,24], effectively creating a strong, stable predictor. Owing to its ability to resist overfitting and model intricate interactions, the RF algorithm has proven to be highly proficient in network intrusion detection tasks [25].

The rationale for employing RF algorithm is in its ability to achieve classification outcomes comparable to those of more complex deep learning approaches on NIDS data. Crucially, in alignment with the objective of reducing barriers to entry in security technology, RF algorithm offers a more straightforward configuration process than their counterparts, such as neural networks. Being a supervised algorithm, it allows one to perform xAI operations for rule derivation, facilitating the approach proposed in this paper.

3.3 Datasets Description

The proposed methodology was tested using two different datasets: CICIDS2017 (D_1) [26], and WUSTL-IIOT-2021 [27]. CICIDS2017 is a well-established benchmark dataset specifically designed for the evaluation of IDSs and Intrusion Prevention Systems (IPSs) in the face of intricate and evolving network attacks. The selection of this dataset was driven by the attention given to addressing the limitations commonly found in other datasets. These limitations encompass outdated information, limited diversity, and the anonymization of vital data. CICIDS2017 stands out for its ability to offer a realistic representation of actual network traffic. It draws from Packet Capture (PCAPs) data and encompasses a wide spectrum of benign background traffic and various attack scenarios. These scenarios include Brute Force FTP, Brute Force SSH, Heartbleed, Web Attack, Infiltration, Botnet, and DDoS. In alignment with the research's focus, the DDoS attack subset, in conjunction with the benign traffic captured during that scenario, was leveraged for the experimental work. The WUSTL-IIOT-2021 dataset emulates real-world industrial environments and is specifically designed to address the security challenges faced by Industrial Internet of Things (IIoT) systems. It features a diverse range of network traffic data, including both benign operations and various types of malicious activities, representing common cyberattacks in IIoT settings. This dataset was created to support research on AI and machine learning-based IDS, with a focus on modern threats targeting IIoT networks. Its realistic and varied data provides a resource for evaluating AI-driven security models, particularly in areas like xAI and distributed AI, ensuring robustness in identifying cyber threats in industrial contexts.

Owing to the notable class imbalance present in the original datasets, a preliminary data preprocessing phase was performed. To mitigate the skew in the class distribution, the Synthetic Minority Over-sampling Technique (SMOTE) was adopted, following the approach outlined in [28]. Any records containing missing values or duplicates were subsequently removed.

Following the cleaning process, dimensionality reduction was undertaken using the SelectKBest method from the scikit-learn library. This procedure employs a univariate statistical test based on the Analysis of Varience (ANOVA) F-value to gauge how features relate to their labels. Analysing the feature importance scores revealed that, beyond the first 14 features, subsequent scores display a substantial decline. Correlation checks were also conducted to eliminate overlaps, thereby retaining only those features with the most representative information.

4 Proposed Method for SIEM Rules Enhancement

This section outlines the innovation of the proposed method to enhance SIEM rules using xAI, detailing the procedural steps and the rationale behind the integration of ML models and xAI techniques. It introduces the novel strategy for deriving SIEM rules through explainability. The process begins with defining a set of traditional static rules. It then proceeds with mining an enhanced rule set by examining the decision trees empirically, then by using SHAP [29], and ANCHORS [30]. All resulting rules from both approaches have been validated with an RF classifier, and the corresponding experimental outcomes are detailed in Section 5. The pipeline of rule derivation is showcased in Fig. 1.

Figure 1: The pipeline of deriving rules from ML classifiers with xAI

4.1 Definition of Static SIEM Rules

The following subsection delineates how the static SIEM rules were formulated.

Let$X$ represent the set of features in the original dataset D, and let $y = \{DDoS, BENIGN\}$ be the corresponding labels. The aim is to identify x_{rule} and use it as discrimination criteria between DDoS and BENIGN samples. As shown in Eq. (1), the difference between the original set of labels and the SIEM-labelled one must be minimized, in order to minimize the number of mislabelled samples.

$$x_{rule} \in X : min_{x_{rule}} |y_{siem} \backslash y| \tag{1}$$

Within the present research, four different correlation rules have been generated for D_1. The correlation criteria have been defined through a linear correlation analysis between the dataset features and the DDoS label. Let C be the correlation index and x_i be the ith feature in the dataset. The linear correlation analysis selects the feature x_i with the highest correlation index, as shown in Eq. (2):

$$j = argmax_i(C(f_i, DDoS)) \tag{2}$$

After this analysis, the $FlowDuration$, $BwdPacketLengthMean$, and $FwdPacketLengthMean$ features have been selected and used as a foundation to build discrimination criteria, as shown in the following. The four versions are the results of the correlation rules shown in Eqs. (3)–(6). In the equations, the features $BwdPacketLengthMean$ and $FwdPacketLengthMean$ are represented as $BPLengthMean$ and $FPLengthMean$, respectively.

$$y_{s1} = \begin{cases} DDoS & if \quad FlowDuration \geq T_{FlowDuration} \\ BENIGN & otherwise \end{cases} \tag{3}$$

$$y_{s2} = \begin{cases} DDoS & if \quad BPLengthMean \geq T_{BPLengthMean} \\ BENIGN & otherwise \end{cases} \tag{4}$$

$$y_{s3} = \begin{cases} DDoS & if \quad BPLengthMean \geq T_1 \\ & \quad\quad and \\ & FPLengthMean \leq T_2 \\ BENIGN & otherwise \end{cases} \tag{5}$$

$$y_{s4} = \begin{cases} DDoS & if \quad BPLengthMean \geq TM \times FPLengthMean \\ BENIGN & otherwise \end{cases} \tag{6}$$

The thresholds $T_{FlowDuration}$ and $T_{BwdPacketLengthMean}$ have been set equal to the mean value of the selected feature for all the samples of the original dataset. The thresholds T_1 and T_2 have been obtained as shown in Eq. (7).

$$T = k \cdot \sigma \tag{7}$$

where σ is defined as shown in Eq. (8), considering individual standard deviations for the entire dataset, DDoS, and BENIGN instances. The k parameter represents the coefficient used to adjust the thresholds based on standard deviation. Its value has been determined through an empirical trial-and-error method.

$$\sigma = \sqrt{\frac{(\sigma_{total})^2 + (\sigma_{DDoS})^2 + (\sigma_{BENIGN})^2}{3}} \tag{8}$$

The same logic has been applied considering $BwdPacketLengthMean$ and $FwdPacketLengthMean$ respectively for T_1 and T_2. The $ThresholdMultiplier$ (TM) parameter has been obtained by calculating the average $BwdPacketLengthMean$ to $FwdPacketLengthMean$ ratio for DDoS instances.

To further validate the proposed methodology, an additional baseline rule was extracted using the same approach on D_2. As a result, the following rule was defined (Eq. (9)), where $SrcRate$ represents the source packets per second:

$$y_{s5} = \begin{cases} DDoS & if \quad SrcRate \geq T_{SrcRate} \\ BENIGN & otherwise \end{cases} \tag{9}$$

Similarly to the previous set of rules, the threshold $T_{SrcRate}$ has been set equal to the mean value of the selected feature for all the samples of the original dataset.

4.2 Novel Method for SIEM Rules Mining via xAI

The application of xAI on the RF for the D_1 has revealed a set of complex nested rules.

All the extracted rules are used for binary classification, dividing network traffic into two classes: $Class : 0$ (benign) and $Class : 1$, which represents DDoS attacks. The decision rules in the trees are based on network traffic features.

Many of the conditions in the trees involve numerical thresholds. For example, a common pattern is to check if a particular feature is greater than or less than a certain threshold value. If the condition is met, the traffic is classified as benign ($Class : 0$); otherwise, it proceeds to the next condition. While some conditions are simple comparisons with single features, others are more complex.

Numerous conditions combine multiple features or rely on logical operators such as “and” to form compound criteria. In certain instances, these conditions include both “greater than” and “less than” thresholds, which increases the complexity of the decision logic. Nodes within the trees frequently adopt an “If...Then...Else” structure, where the tree follows the “Then” branch and assigns a traffic label if a condition is met, or proceeds to the “Else” branch for further evaluation if it is not.

A prominent observation is that most decision pathways culminate in a benign ($Class : 0$) classification, indicating a strong emphasis on identifying benign or non-malicious traffic. Despite this shared objective, the exact conditions and threshold values for defining benign traffic differ from one tree to another.

The decision tree outputs were initially examined to select a subset of rules suitable for conversion into a nested SIEM rule. This consolidated rule is presented in Eq. (10) in a streamlined format to enhance clarity. The thresholds within this rule are highly precise, having been extracted from recurrent values observed in the forest. These thresholds are denoted as T_p parameters, where p corresponds to the relevant feature. Where possible, the features themselves are abbreviated to maintain a concise representation.

Subsequently, the SHAP values have been extracted and analyzed, leading to the following SIEM rule (Eq. (11)). The features with the highest SHAP values have been considered and employed in the definition of discriminatory criteria. In the Equation, the threshold T_{TPL} used for the $TotalBackwardPackets$ feature has been set equal to the mean value of the feature for the entire dataset. The threshold for $FlowDuration$ is the same employed in Eq. (3). The $Whitelist$ and $Blacklist$ have been derived by checking the exclusive values of $DestinationPort$ for BENIGN and DDoS labes, respectively.

$$y_{rfm} = \begin{cases} BENIGN \quad if \quad FlowDuration \leq T_{FL} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad ACKFlagCount \leq T_{ACK} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad PLengthMean \leq T_{PL} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad SYNFlagCount > T_{SYN} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad IdleMean \leq T_{IM} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad DestPort \leq T_{DP} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad BPLengthMean \leq T_{BP} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad MinPacketLength \leq T_{MP} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad FPLengthMean > T_{FP} \\ DDoS \qquad\qquad otherwise \end{cases} \tag{10}$$

$$y_{rfm} = \begin{cases} BENIGN \quad if \quad DestPort \in Whitelist \\ DDoS \quad if \quad DestPort \in Blacklist \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad TotBWDP \leq T_{TPL} \\ \qquad\qquad\qquad\quad or \\ \qquad\qquad\quad FlowDuration > T_{FlowDuration} \\ BENIGN \qquad\qquad otherwise \end{cases} \tag{11}$$

Finally, a third rule has been derived using explainability features provided by ANCHORS. The rule is formalized in Eq. (12), where T_1, T_2, and T_PL have been defined previously, while a and b have been set equal

to the minimum and maximum values of the feature for BENIGN samples, respectively:

$$y_{rfm} = \begin{cases} BENIGN & if \quad a \geq FPLengthMean > b \\ DDoS & if \quad BPLengthMean \geq T_1 \\ & \quad or \\ & \quad FPLengthMean < T_2 \\ & \quad or \\ & \quad PLengthMean > T_{PL} \\ BENIGN & \quad otherwise \end{cases} \tag{12}$$

Additional experiments were conducted on D_2. In particular, SHAP and ANCHORS were employed to mine the following rule (Eq. (13)). In the Equation, $Protot$ denotes the network protocol, $SrcLoss$ and $DstLoss$ the Source/Destination packets respectively that are re-transmitted and/or dropped, while $TcpRtt$ indicates the TCP connection setup round-trip time.

$$y_{rfm} = \begin{cases} BENIGN & if \quad Proto \in Whitelist \\ DDoS & if \quad SrcLoss > DstLoss \\ & \quad or \\ & \quad TcpRtt < \varepsilon \\ Benign & \quad otherwise \end{cases} \tag{13}$$

The $Whitelist$ parameter has been derived by checking the exclusive values of the $Prot$ feature for Benign and DDoS labels, respectively. Finally, the ε parameter is defined as shown below (Eq. (14)):

$$\varepsilon \to 0 \tag{14}$$

This value has been selected as the threshold for the $TcpRtt$ feature, as it has been observed that this feature was close to zero, on average, for DDoS samples only.

5 Experimental Setup and Results

The following section presents the setup of the performed experiments and the detailed results obtained, as the empirical evidence of the effectiveness of the proposed approach.

The results of the experimental approach are summarized in Tables 3 and 4. The metrics in Table 3 have been extracted by training the ML classifier using the SIEM-labelled datasets, where the labels were assigned through the correlation rules described previously, and using the ground truth dataset for testing. The ones in Table 4 were obtained using the SIEM-labelled dataset obtained from the rules extracted through xAI from the RF Model (RFM) for training, and the ground truth dataset for testing.

Table 3: Summary of classification reports for SIEM detection implemented using initial correlation rules

Rules	Precision		Recall		F1-score		Accuracy	BA	MCC	G-mean
	BENIGN	DDoS	BENIGN	DDoS	BENIGN	DDoS				
Eq. (3)	0.79	0.62	0.46	0.88	0.58	0.73	0.67	0.66	0.36	0.62
Eq. (4)	0.73	0.98	0.98	0.63	0.84	0.76	0.77	0.8	0.66	0.79
Eq. (5)	0.72	0.91	0.94	0.63	0.81	0.75	0.79	0.78	0.59	0.77
Eq. (6)	0.73	0.99	1	0.63	0.84	0.77	0.81	0.81	0.67	0.79
Eq. (9)	0.62	0.99	0.99	0.49	0.77	0.57	0.70	0.74	0.78	0.74

Table 4: Classification report for SIEM detection implemented using RF rules derived through xAI

Rules	Precision		Recall		F1-score		Accuracy	BA	MC	G-mean
	BENIGN	DDoS	BENIGN	DDoS	BENIGN	DDoS				
Eq. (10)	1	0.82	0.78	1	0.88	0.9	0.91	0.89	0.8	0.88
Eq. (11)	1	0.93	0.92	1	0.96	0.96	0.96	0.96	0.92	0.96
Eq. (12)	1	0.96	0.96	1	0.98	0.98	0.98	0.98	0.96	0.98
Eq. (13)	1	0.99	0.99	1	0.99	0.99	0.99	0.99	0.99	0.99

The tables show a summary of the classification report based on Precision (15), Recall (16), F1-Score (17), Accuracy (18), Balanced Accuracy (BA) (19), Matthews Correlation Coefficient (MCC) (20), and Geometric Mean (G-mean) (21). In the following equations, for the sake of brevity, True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are denoted using the acronyms.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{18}$$

$$\text{BA} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{19}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FP})}} \tag{20}$$

$$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{21}$$

According to the results, the rules derived from RF through xAI outperform SIEM detection rules in multiple aspects. As shown in Table 4, the rule defined in Eq. (12) achieved a Precision of 1.0 for the BENIGN class, and 0.96 for the DDoS class, Recall of 0.86 for the BENIGN class, and 1.0 for the DDoS class, F1-score of 0.98 for both classes, Accuracy, BA and G-mean of 0.98, and MCC of 0.96. The best SIEM detection rule (Eq. (6)), on the other hand, achieved a Precision of 0.73 for the BENIGN class, and 0.99 for the DDoS class, Recall of 1.0 for the BENIGN class, and 0.63 for the DDoS class, F1-score of 0.84 for the BENIGN class, and 0.77 for the DDoS class, and overall Accuracy of 0.81. The same performance trends were observed when using the dataset *D*2, with near-perfect results using the xAI-mined rule, reaching a Precision of 1.0 for the BENIGN class and 0.99 for the DDoS class, along with a Recall of 0.99 for the BENIGN class and 1.0 for the DDoS class. This led to an F1-score of 0.99 for both classes, an overall Accuracy and BA of 0.99, an MCC of 0.99, and a G-mean of 0.99. These results surpass the best-performing SIEM detection rule (Eq. (9)), which had a Precision of 0.62 for the BENIGN class and 0.99 for the DDoS class, a Recall of 0.99 for the BENIGN class and 0.49 for the DDoS class, and a lower F1-score of 0.77 and 0.57 for BENIGN and DDoS classes, respectively. The xAI-derived rules demonstrate superior performance, but also highlight the ability of ML models to extract more nuanced patterns in IIoT data, resulting in significantly improved detection rates for cyberattacks. The classification reports show that the usage of xAI to derive discriminatory criteria from RFM holds the potential to improve classic SIEM detection rules. The higher F1-score also indicates a better

balance between Precision and Recall. In comparison with traditional static rules, the nested rule brings the knowledge of more complex relationships between features.

6 Discussion and Lessons Learnt

This section analyzes the outcomes, discusses the scalability and generalizability of the proposed method, and reflects on the lessons learned from the application of xAI in enhancing cybersecurity practices.

The experimental findings demonstrate that SIEM rules generated through the RF model using xAI methods can surpass the performance of conventional SIEM detection rules. As shown in Table 4, the rule specified in Eq. (10) attains considerably higher Precision and Recall scores in classifying both BENIGN and DDoS samples. Specifically, a Precision of 1.0 for the BENIGN class reflects near-perfect accuracy in identifying benign traffic, while a Recall of 1.0 for the DDoS class indicates the detection of nearly all actual DDoS instances. This equilibrium between Precision and Recall is also evident in a high F1-score, underscoring the rule's effectiveness in accurately classifying both classes. Moreover, the elevated BA, MCC, and G-mean suggest that the rule remains robust even under imbalanced data conditions.

Subsequent refinement of the SIEM rule, shown in Eq. (11) and informed by SHAP values, yields further improvements, notably in terms of DDoS Precision and BENIGN Recall, as well as F1-score for both classes. This rule also achieves higher overall Accuracy, BA, MCC, and G-mean. The third rule, derived through ANCHORS and presented in Eq. (12), exhibits near-optimal performance across all metrics, thereby surpassing the preceding two.

By contrast, the strongest of the initial SIEM rules (Eq. (6)) maintains a high Precision for the DDoS class but suffers from a considerably reduced Recall for that class. Consequently, although the rule precisely identifies DDoS when triggered, it misclassifies a non-negligible number of malicious instances as benign, leading to a diminished F1-score and lower Accuracy. Despite achieving the best BA, MCC, and G-mean among the original SIEM rules, it still performs markedly below the xAI-derived rules in these measures.

The initial SIEM rules and the RF rules extracted via xAI, used for the definition of the SIEM rules shown in Eqs. (10) and (11), present both similarities and differences. The main similarity is the core logic: conditions are set to determine the final decision, following a binary classification. On the other hand, the RF rules have a more complex structure with multiple conditions and branches, whereas the SIEM rules are simpler and typically involve a single feature comparison for each rule. In the RF rules, conditions are often based on specific feature values, while the SIEM rules use wider and more generic threshold values to make decisions. The SIEM rules have more human-readable conditions, making it easier for someone to understand the logic behind the rules. Another key difference is that the RF rules prioritize the detection of benign traffic, classifying as anomalous all the samples that do not follow the pattern considered innocuous. On the other hand, the traditional SIEM rules set conditions to detect malicious traffic, labelling as benign every sample that does not follow the pattern considered anomalous. Additionally, unlike certain SIEM rules, specifically, Eq. (6), in RF rules features are never compared between each other. Another remarkable difference is that the SIEM rules consider a small subset of features among the most relevant ones, while the RF rules have a wider variety of relevant features taken into consideration. The SIEM rules derived from xAI look to find the common ground between the two sets, by implementing a more complex and accurate correlation logic, while still maintaining human readability and enabling practical implementation in common rule-based SIEM systems. As shown previously, this approach holds the potential to improve traditional SIEM rules' reliability in detecting network anomalies.

The results using the *D2* dataset further confirm the effectiveness of the xAI-derived rules. In particular, the rule from Eq. (13) achieved near-perfect performance, with Precision, Recall, and F1-score all consistently

high, reaching 0.99 or higher for both BENIGN and DDoS classes. This shows the rule's capability to accurately identify both benign and malicious traffic in a complex IIoT environment. Compared to the traditional SIEM rule (Eq. (9)), which struggled with significantly lower Recall for the DDoS class (0.49), the xAI-derived rule demonstrated superior balance between detecting threats and minimizing false positives. The improved results from $D2$ reinforce the advantages of using xAI to refine detection rules in IIoT systems, especially in scenarios where traditional approaches fall short.

7 Threats to Validity and Possible Future Works

While the current study effectively demonstrates the application of xAI-enhanced SIEM systems specifically for DDoS attacks, it acknowledges certain limitations regarding the scope of dataset diversity and validation under dynamic network conditions. Furthermore, the primary focus has been to explore the novel integration of xAI within the context of NIDS, emphasizing the utility of xAI-derived rules not for their intrinsic explainability to end-users but for their capability to distill complex ML insights into actionable, reliable rules within SIEM systems. The goal is to enhance detection capabilities and operational efficiency, even if the security analysts may not fully comprehend the underlying complexities of the ML model. This approach recognizes that in operational environments, the practical applicability and reliability of detection rules often take precedence over the detailed understanding of their derivation, aligning with the primary needs of cybersecurity professionals. Future work will aim to address these limitations by extending the validation of our approach to more varied real-world environments and incorporating a broader spectrum of cyberthreats beyond DDoS. This will involve testing the xAI-enhanced SIEM system across multiple datasets that reflect a wider range of network conditions and attack scenarios, thereby strengthening the generalizability and robustness of our findings. Furthermore, comparative analysis with other cutting-edge methods will also be pursued to benchmark the effectiveness and efficiency of the proposed xAI application in cybersecurity.

8 Conclusions

This paper is the significant extension of our previous work presented at ACIDS conference [17]. This study investigates the integration of xAI and RFM to enhance traditional SIEM detection rules, with an emphasis on classifying network traffic. A new rule-mining methodology has been introduced, alongside a proof-of-concept implementation for detecting DDoS attacks using two distinct benchmark datasets. The evaluation compares SIEM rules generated through xAI with conventional static rules, highlighting the advantages of the proposed framework. The results indicate that the xAI-derived rules achieve notably superior performance metrics compared to traditional rules. These advanced rules integrate multiple conditions and branches, offering more detailed and refined insights. Such higher complexity allows them to capture intricate relationships between features, leading to more accurate and sophisticated detection criteria. The outcomes of this research underline the potential of xAI and RFM as effective tools in cybersecurity, particularly for anomaly detection and security monitoring. Incorporating these techniques into current practices can strengthen organizations' abilities to identify and address network anomalies, thereby safeguarding their digital infrastructure and assets.

Acknowledgement: The authors would like to thank all supporting institutions for enabling this research. Any opinions, findings, and conclusions expressed in this article are those of the authors and do not necessarily reflect the views of the affiliated organizations.

Funding Statement: This work is funded under the Horizon Europe AI4CYBER Project, which has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101070450.

Author Contributions: Conceptualization, Federica Uccello and Marek Pawlicki; methodology, Federica Uccello, Marek Pawlicki, and Salvatore D'Antonio; software, Federica Uccello and Marek Pawlicki; validation, Federica Uccello and Marek Pawlicki; formal analysis, Federica Uccello and Marek Pawlicki; investigation, Federica Uccello; resources, Michał Choraś and Rafal Kozik; data curation, Federica Uccello; writing—original draft preparation, Federica Uccello; writing—review and editing, Marek Pawlicki and Michał Choraś; visualization, Federica Uccello and Marek Pawlicki; supervision, Salvatore D'Antonio and Michał Choraś; project administration, Michał Choraś; funding acquisition, Michał Choraś and Marek Pawlicki. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets employed in this study (CICIDS2017 and WUSTL-IIOT-2021) are publicly available benchmark datasets. Interested researchers can obtain the data from their respective official repositories under the terms specified by the dataset publishers.

Ethics Approval: This research did not involve human participants or animal testing. Ethical review and approval were therefore waived.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; New York, NY, USA. p. 1135–44.
2. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. 2017.
3. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning-a brief history, state-of-the-art and challenges. In: Joint European conference on Machine Learning and Knowledge Discovery in Databases; 2020; Cham, Switzerland: Springer. p. 417–31.
4. Choraś M, Pawlicki M, Puchalski D, Kozik R. Machine Learning-the results are not the only thing that matters! What about security, explainability and fairness? In: Computational Science-ICCS 2020: 20th International Conference; 2020 Jun 3–5; Amsterdam, The Netherlands. p. 615–28.
5. Choraś M, Pawlicka A, Jaroszewska-Choraś D, Pawlicki M. Not only security and privacy: the evolving ethical and legal challenges of e-commerce. In: Katsikas S, Cuppens F, Cuppens-Boulahia N, Lambrinoudakis C, Garcia Alfaro J, Navarro-Arribas G et al., editors. Computer Security. ESORICS, 2023 International Workshops; 2024; Cham: Springer Nature Switzerland. p. 167–81.
6. Patil S, Varadarajan V, Mazhar SM, Sahibzada A, Ahmed N, Sinha O, et al. Explainable artificial intelligence for intrusion detection system. Electronics. 2022;11(19):3079. doi:10.3390/electronics11193079.
7. Islam MU, Mozaharul Mottalib M, Hassan M, Alam ZI, Zobaed S, Fazle Rabby M. The past, present, and prospective future of XAI: a comprehensive review. In: Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence; 2022; Cham, Switzerland: Springer. p. 1–29.
8. Mendes C, Rios TN. Explainable artificial intelligence and cybersecurity: a systematic literature review. arXiv:230301259. 2023.
9. Nwakanma CI, Ahakonye LAC, Njoku JN, Odirichukwu JC, Okolie SA, Uzondu C, et al. Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: a review. Appl Sci. 2023;13(3):1252. doi:10.3390/app13031252.
10. González-Granadillo G, González-Zarzosa S, Diaz R. Security information and event management (SIEM): Analysis, trends, and usage in critical infrastructures. Sensors. 2021;21(14):4759. doi:10.3390/s21144759.

11. Vielberth M. Security Information and Event Management (SIEM). In: Jajodia S, Samarati P, Yung M, editors. Security Information and Event Management (SIEM); 2021; Berlin/Heidelberg: Springer Berlin Heidelberg. p. 1–3. doi:10.1007/978-3-642-27739-9_1681-1.
12. Zhang Y. Design and implementation of network security management system based on k-means algorithm. In: 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA). Chongqing, China; 2023. 1010–5.
13. Podzins O, Romanovs A. Why SIEM is irreplaceable in a secure IT environment? In: 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream); 2019; Vilnius, Lithuania. p. 1–5. doi:10.1109/eStream.2019.8732173.
14. Jacobs AS, Beltiukov R, Willinger W, Ferreira RA, Gupta A, Granville LZ. Ai/ml for network security: the emperor has no clothes. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security; 2022; New York, NY, USA. p. 1537–51.
15. Sun N, Ding M, Jiang J, Xu W, Mo X, Tai Y, et al. Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives. IEEE Commun Surv Tutorials. 2023;25(3):1748–74. doi:10.1109/COMST.2023.3273282.
16. Lu Y, Liu T, Zheng H, Zhu X. Research on constructing a network security event knowledge network based on multi-source data. In: 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC); 2024; New York, NY, USA. p. 749–59.
17. Uccello F, Pawlicki M, D'Antonio S, Kozik R, Choraś M. A novel approach to the use of explainability to mine network intrusion detection rules. In: Nguyen NT, Chbeir R, Manolopoulos Y, Fujita H, Hong TP, Nguyen LM et al., editors. Intelligent Information and Database Systems; 2024; Singapore: Springer Nature Singapore. p. 70–81.
18. Mahbooba B, Timilsina M, Sahal R, Serrano M. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity. 2021;11:6634811. doi:10.1155/2021/6634811.
19. Karna SK, Paudel P, Saud R, Bhandari M. Explainable prediction of features contributing to intrusion detection using ML algorithms and LIME. Medicon Eng Themes. 2023;5(3):6–15.
20. Kalutharage CS, Liu X, Chrysoulas C, Pitropakis N, Papadopoulos P. Explainable AI-based DDOS attack identification method for IoT networks. Computers. 2023;12(2):32. doi:10.3390/computers12020032.
21. Zhou Q, Li R, Xu L, Nallanathan A, Yang J, Fu A. Towards explainable meta-learning for DDoS detection. arXiv:2204.02255. 2022.
22. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv:2006.11371. 2020.
23. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
24. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995; IEEE. Vol. 1, p. 278–82. doi:10.1109/ICDAR.1995.598994.
25. Mihailescu ME, Mihai D, Carabas M, Komisarek M, Pawlicki M, Hołubowicz W, et al. The proposition and evaluation of the roedunet-SIMARGL2021 network intrusion detection dataset. Sensors. 2021;21(13):4319. doi:10.3390/s21134319.
26. Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy ICISSP. Funchal, Madeira, Portugal; 2018. Vol. 1, p. 108–16.
27. Zolanvari M, Teixeira MA, Gupta L, Khan KM, Jain R. WUSTL-IIOT-2021 dataset for IIoT cybersecurity research [Internet]. MO, USA: St. Louis; [cited 2025 Mar 4]. Available from: http://www.cse.wustl.edu/$\sim$jain/iiot2/index.html.
28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. doi:10.1613/jair.953.
29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems; 2017; Redwood City, CA, USA. p. 30.
30. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2018; Menlo Park, CA, USA. Vol. 32.