IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Explainable AI models for cyber threat hunting in SEIM and cloud security platforms

Dinesh kollu

Sikkim Manipal University, Gangtok, Sikkim, India.

Abstract

With the growing use of machine learning in Security Information and Event Management (SIEMs) as well as cloud security platforms, automated threat detection has been greatly enhanced, and the problem of model decisions being interpretable is also a significant impediment to viable cyber threat hunting. This paper presents a framework of explainable AI, which uses post-hoc explainability and baseline detection models to enable behavioural interpretation and analyst-focused exploration of security alerts. Cybersecurity datasets are publicly available and are referred to as CICIDS-2017 and UNSW-NB15 and they are the proxy SIEM and cloud security telemetry to prove the applicability of the proposed methodology. The framework facilitates instance-level and generalisations of detected events, which allows the analysts to find the behavioural fingerprints that lead to alerts and to relate them to attack patterns of higher level. A qualitative analysis is performed to test the interpretability, behavioural relevance and usefulness in practice of the generated explanations to the workflow of threat-hunting. The findings suggest that the suggested framework increases the level of understanding among the analysts in addition to validating the alert as well as enabling proactive investigation. The article identifies explainability as one of the major facilitators of reliable and operationally efficient cyber threat hunting in current SIEM and cloud security settings.

Onions: Explainable AI (XAI), Cyber Threat Hunting, SIEM and Cloud Security, Analyst-Centric Investigation, Behavioural Attack Analysis.

1. Introduction

The growing size, diversity and pace of security telemetry data created by distributed and cloud-based systems have made Security Information and Event Management (SIEM) and cloud security platforms key elements of the modern enterprise cyber-defence infrastructure. Machine learning algorithms are being widely implemented in automated identification of threats and prioritisation of alerts due to the increasing number of alerts and sophisticated patterns of attack. The recent surveys and empirical research demonstrated that data-driven models could make a significant enhancement on the detection of malicious activities in the network and system monitoring conditions compared with the rule-based models [1], [2].

Nevertheless, the use of machine learning models in the operational adoption of Security Operations Centres (SOCs) is scarce despite their promising ability to find out the presence of threats. A lack of transparency in model decisions is one of the primary causes since it is hard to know why a certain alert is being generated in the case of analysts. Black-box learning systems can cause severe operational and trust-related problems in practical security settings, as pointed out by Sommer and Paxson [3], because analysts must justify response operations and root-cause analysis and validate alerts.

To enhance the understandability of machine learning models, Explainable Artificial Intelligence (XAI) has thus become a major research area. Post-hoc interpretations of complex models can be achieved using methods like Local Interpretable Model-Agnostic Explanations (LIME) [4], SHapley Additive exPlanations (SHAP) [5], and others. These methods have found extensive applications in various high stakes fields such as healthcare and finance, to enhance transparency and confidence to users.

Within the cybersecurity field, the application of XAI to explain intrusion detection model and anomaly detection model has been examined in a growing literature [6], [7]. The current research is mainly concerned with the visualisation of the importance of features or the description of the classification results in relation to individual detection problems. Although these attempts enhance the interpretability of models, they mostly discuss explainability as a side effect of justifying model predictions, but not as a part of operational cyber threat hunting.

In real world SOC operations, analysts do not just check alerts but also actively threat hunt, conduct behavioural correlation and investigational reasoning on two or more telemetry sources. Threat hunters must be capable of correlating low-level detection results with high-level adversary actions and strategies, e.g. reconnaissance, credential access and lateral movement as represented by the popular MITRE ATT&CK model. Nonetheless, current explainable cybersecurity research seldom incorporates outputs of explanation in analyst-friendly investigation activities or threat-hunting behaviour of behavioural logic. This paper fills this gap by advancing an explicable artificial intelligence-based framework of cyber threat hunting in SIEM and cloud security tools. As compared to the traditional detection-centric models, the proposed model merges the post-hoc explainability with baseline machine learning detectors, facilitating analyst-centric inquiry and behavioural elucidation of notifications. Cybersecurity datasets that are publicly available are used as proxy SIEM and cloud telemetry to show the relevance of the methodology. The framework provides instance-level and global explainable alerts and provides the ability to map behavioural indicators to higher-level adversarial activities, which in turn aids in the proactive and interpretable threat-hunting processes.

The rest of this paper is structured in the following way. Section 2 consists of reviews of literature on machine-learning-based security analytics and explainable cybersecurity. Section 3 shows the suggested explainable threat-hunting framework. Section 4 explains the qualitative assessment plan and methodology. The section 5 is the qualitative evaluation of the findings and practical implications on the SOC activities. Lastly, the paper is in a conclusion in Section 6, which presents future research directions.

2. Literature Review

2.1 Machine-Learning-Based SIEM and SOC Environment Security Analytics.

Intrusion detection and security monitoring in an enterprise setting has been widely studied using machine learning techniques. The initial and popular survey research has shown that anomaly detection and classification-based methods have the potential to identify malicious behavior in network and system telemetry [8], [9]. These publications formed the principles of using data-driven techniques to huge-scale monitoring settings analogous to those operated by SIEM platforms.

A number of literature have also studied the application of the intrusion detection system and security analytics pipeline in practice. Liao et al. [10] gave an extensive survey of intrusion detection system and pointed out the increasing trend of the shift of signature based system towards learning based detection system. Still more current research focuses on feature engineering, traffic aggregation and scalable learning architecture to be deployed in the real-world [11], [12].

Accessibility to realistic cybersecurity data has also contributed to the expedited security research using machine learning. Sharafaldin et al. presented CICIDS-2017 to help assess current intrusion detection methods [13], whereas Ring et al. examined the shortcomings and representativity of available intrusion detection datasets to operate with [14]. These investigations suggest that community datasets may be decent proxies of security monitoring telemetry when the actual SOC data are not available.

Nevertheless, even though the detection accuracy is rising and reported in the literature, there is a recent empirical research questioning the practicality of entirely detection-based machine learning systems. Apruzzese et al. showed that practical security benefits do not always correspond with high predictive performance especially when the detection systems are implemented in dynamic and adversarial environments [15]. The findings of their study point to the necessity to have security analytics solutions that assist the process of decision-making of the analysts instead of paying attention to the accuracy of the classification only.

2.2 Explainable Artificial Intelligence to Cybersecurity Analytics.

The research has seen the emergence of explainable Artificial Intelligence as a significant research direction to enhance the transparency and reliability of machine learning systems. Arrieta et al. have given an extensive description of explainable AI as a concept within the framework of trustful machine learning and determined interpretability as one of the main conditions of the high-stakes decision-making context [16]. The same was also reported by Samek et al., who indicated the significance of quality of explanation and user-centred assessment in explainable learning systems [17].

In cybersecurity, there have been a number of studies that have begun to investigate the explainability methods to interpret intrusion detection and malware detection models. Das and Rad examined possibilities and issues of explaining the AI techniques used in cybersecurity analytics and emphasized the challenge of communicating feature-level explanations into the form of meaningful security information to practitioners [18]. Warnecke et al. suggested the methods of explaining the anomaly-based intrusion detection systems and demonstrated that explainability could be helpful in the interpretation of unusual traffic patterns [19].

Hind et al. proposed frameworks of explanations to enhance the transparency and trust of users in deployed machine learning systems and highlighted the importance of domain-related explanations towards operational settings [20]. On the same note, Begoli et al. explained the practical needs of explainable AI systems and opined that explanations should be based on user goals and workflows and not necessarily based on technical measures [21].

Though such studies show that it is possible to explain security models, the majority of existing literature considers explainability an after-processing step to verify the results of prediction. The explanation results of individual classification are normally given in isolation without reflecting how these data can be used to drive on-going investigation, correlation of alerts and reasoning by analysts in SOC environments.

2.3 elaboration and the disconnect between Cyber Threat Hunting Workflows.

Sarker et al. introduce a multi-layer cybersecurity analytics architecture, which incorporates security data collection, data preparation, security modelling based on machine-learning and incremental learning into one architecture (Figure 1) [7]. The framework focuses on the heterogeneous telemetry sources like network, host and application logs and on the relevance of continuous learning to adhere to the changing cyber threats. Although the layered architecture offers a more detailed basis of automated security analytics, it mainly concerns itself with data processing and model-based detection. The framework fails to clearly consider how the results of detection can be interpreted by the analysts and how explanation results can be integrated into the operation investigation procedures.

Threat hunting is not equal to traditional alert-based detection in an actual security operation. It is a proactive search of telemetry, hypothesis-based analysis and behavioural correlation on a variety of data sources to discover stealthy or emerging attacks. Structured adversary knowledge bases, including the MITRE ATT&CK framework, usually inform behavioural reasoning when threat hunting, and structured attacker behaviour is organised into tactics and techniques that facilitate investigative reasoning and reporting [22].

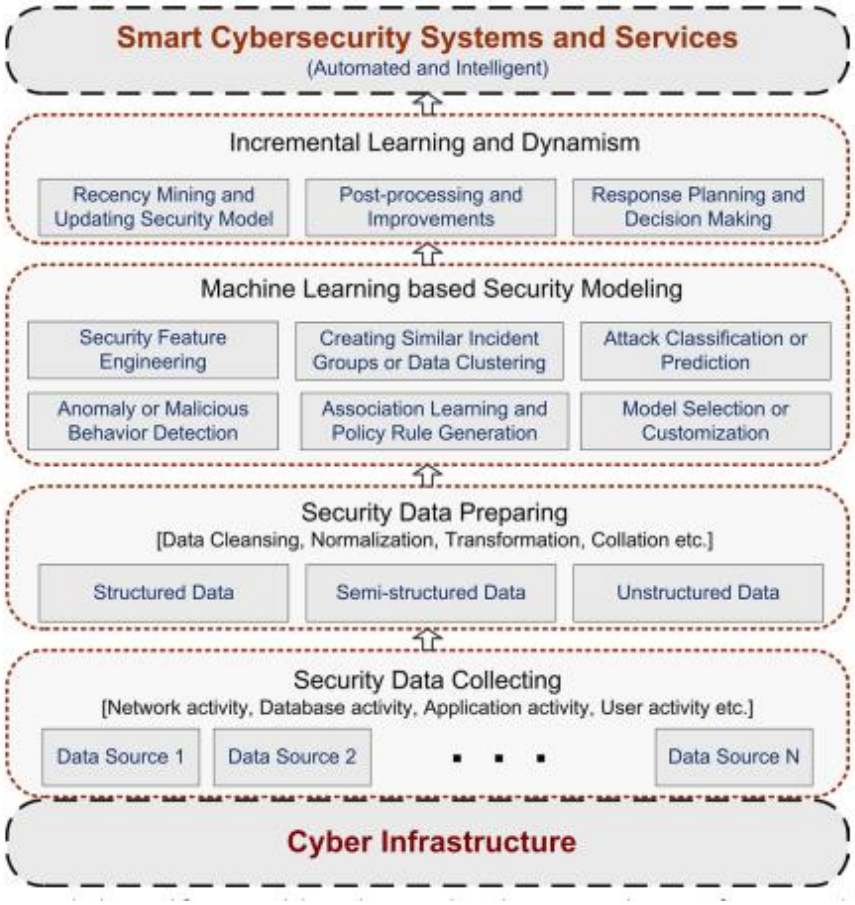

Figure 1: Multi-layer cybersecurity analytics framework illustrating security data collection, data preparation, machine-learning-based modelling and incremental learning for decision support [7]

Nevertheless, the available machine-learning-driven cybersecurity models and explainable intrusion detection literature are detection-oriented. Usually, only feature importance or local model interpretation is explained, and they are normally used to justify individual predictions. They are seldom incorporated in analyst operations that facilitate behavioural association, investigative hypothesis development and mapping of signals to superior adversary operations.

In addition, previous research on explainable artificial intelligence has stressed that the explanations should be crafted to aid human reasoning and contextualisation, and not just reveal internal model behaviour [23]. This need is evident especially in SOC settings where analysts use behavioural narratives and contextual evidence to make judgements about whether or not an alert is a real security incident and to make decisions about what to do as a response.

In general, there is apparent discrepancy between explicable machine-learning methods and functioning cyber threat-hunting patterns. Though existing frameworks offer efficient channels of data collection and automated detection, they do not provide much assistance to the interpretative behaviour analyst and investigative processes in line with adversary behaviour models. Such a gap drives the rationale of explaining, analyst-friendly threat-hunting framework that clearly incorporates post-hoc explainability with SIEM and cloud security analytics to assist with a practical investigation and decision making.

2.4 Research Gap and Motivation

In accordance with the analyzed literature, the absence of methodological frameworks that incorporate explainable AI into the system of cyber threat-hunting in SIEM and cloud security within the framework of SIEM and cloud security is obvious. The available literature is mainly dedicated to the explanation of model predictions, yet, it does not offer much assistance to behavioural interpretation, investigation arguments and hypothesis development by analysts.

This paper fills this gap by presenting an elucidable AI-based structure that expressly combines post-hoc explainability as well as baseline detectives models in supporting analyst-focused threat hunting. The suggested methodology focuses on behavioural interpretation of alerts and is supportive in mapping out the

explanation outputs to adversarial activities and therefore allow more effective and transparent cyber threat-hunting workflows.

3. Suggested Explainable Threat-Hunting Framework.

In order to provide a clear description of the general workflow and building blocks of the offered solution, the architectural diagram of the explicable threat-hunting framework is included in Figure 2. The illustration signifies the combination of SIEM and cloud telemetry, engineering of behavioural features, a bottom-up identification of threats, the post-hoc explainability, and threat hunting that is configured with a single working pipeline.

3.1 Framework Overview

This paper suggests an explainable AI-based system hunting of cyber threats in SIEM and cloud security systems. The architecture combines a conventional machine-learning-driven detection system with a post-hoc elucidation enhancer so as to guide analyst-friendly investigation and behavioural interpretation of identified security occasions.

The general model comprises of five steps:

The elements proposed include (i) acquisition of SIEM-equivalent and cloud telemetry,

(ii) aggregation of behavioural features and engineering,

(iii) machine-learning classification-based detection of threat with existing machine-learning classifiers,

Explainability analysis: (iv) post-hoc explainability analysis, and

(v) threat hunting and behavioural validation of an analyst.

The key aim of the suggested framework is closing the process between automated detection and human-based investigation by giving interpretable information regarding the behavioural factors that lead to each security alert.

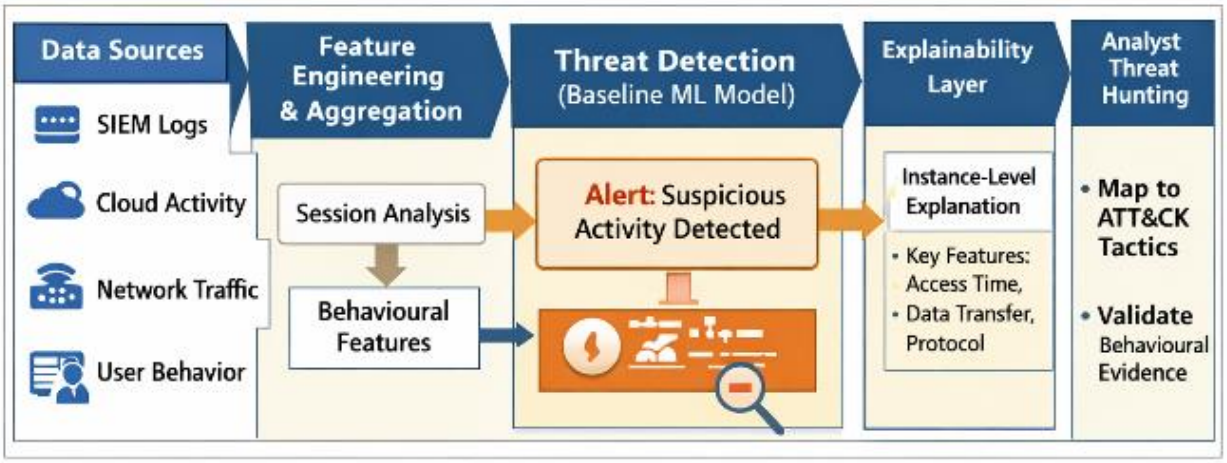

Figure 2. Proposed explainable threat-hunting framework for SIEM and cloud security platforms

3.2 Data Ingestion and Telemetry Layer

The framework takes into account heterogeneous security telemetry emitted by network and system monitoring components that are generally present in SIEM and cloud security solutions. These telemetry streams are events that are aggregated in terms of network connections, session activities and communication patterns. This work assumes the publicly available cybersecurity datasets as proxy SIEM and cloud monitoring logs.

3.3 Engineering Behavioural Features.

Raw records are converted to behaviour-oriented indicators to make them easily interpreted by the analysts. The features extracted capture session level and flow level features like communicative intensity, duration statistics, protocol usage patterns and volume distributions. This representation allows the explainability layer to generate meaningful explanations that are associated with operational security indicators as opposed to low level packet attributes.

3.4 Threat Detection Component Base.

The detection component also uses one or more established supervised learning classifiers as a baseline detector to differentiate benign and malicious activities. The present study does not suggest any new detection algorithm. Rather, baseline classifiers are employed in order to produce detection results that are needed to provide the following explainability analysis. This design will enable the suggested framework to be model-agnostic and easily scalable to other detection engines in a real-world SOC setting.

3.5 Explainability, Analyst-Oriented Interpretation.

The trained baseline detection models have a post-hoc explainability layer that is used to produce both instance-level and global explanations. The most powerful behavioural characteristics are found to be identified as the instance level explanations to each alert detected as well as general level to describe the general behavioural patterns that have been learnt by the detection models. These interpretations are made in a manner that can be comprehended by the SOC analysts about the rational explanation of behavioural alerts so that it can be used to facilitate investigation and decision-making.

3.6 Threat Hunting and Behavioural Validation.

The last phase of the framework is on threat hunting by an analyst. The most significant identified features of the explainability module are interpreted as the higher level behavioural indicators scanning behaviour, abnormal communication patterns, repeated access attempts and traffic anomaly. This step will allow analysts to authenticate the alerts through matching the outcomes of detection and meaningful behavioural evidence that will enhance the efficiency of the investigation and prioritisation of alerts.

4. Methodological Setup

Figure 3 shows the Methodological Setup followed by this research, which comprises dataset preparation, feature engineering, training the baseline model, analyze the explainability and the threat-hunting-oriented evaluation. This Setup is an overview of the sequence of steps that will be employed in proving the proposed framework on public SIEM-equivalent datasets..

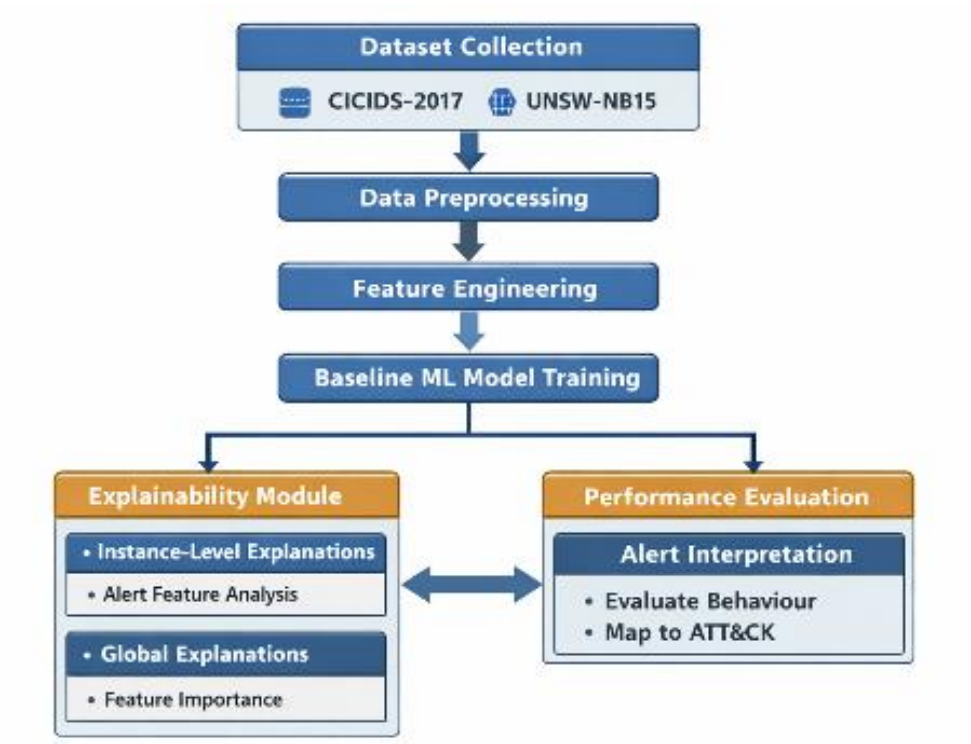

Figure 3. Methodological Setup for baseline threat detection and explainability-driven threat hunting

4.1 Datasets

In order to simulate realistic SIEM and cloud security telemetry two publicly available cybersecurity datasets are used.

The main data source is the CICIDS-2017 published by the Canadian Institute of Cybersecurity. It has labeled network traffic that depicts benign as well as various attack conditions such as brute-force attacks, denial-of-service attacks, web-based attacks and infiltration activities.

In order to assess the generalisability of the proposed framework, the independent validation dataset is an independent dataset called UNSW-NB15, and it is provided by the University of New South Wales. This data reflects the current synthetic enterprise traffic and comprises multiple types of attacks like reconnaissance, activity of exploits and malware.

In the paper, both datasets are considered as proxy SIEM and cloud security telemetry, where each record is a collection of aggregated security events produced by monitoring parts.

4.2 Data Pre-processing

The data sets are purified with deleting missing records and non informative identifiers. Label encoding codes the categorical features and the normalisation of numerical features is done using min maximum scaling.

Only behaviourally meaningful attributes are preserved in order to maintain the ensuing explanations to be interpreted by security analysts. The problem of class imbalance is treated with stratified sampling when training and testing.

4.3 SIEM Representation of Features.

In order to capture operational SOC practices, the data is described as aggregated session and flow behavioural indicators. The characteristic set that is obtained reflects the abnormal frequencies of communication, the deviant traffic volume, the variation in the protocol usage and the characteristics of durations.

Such a representation enables successful behavioural interpretation when threat hunting.

4.4 Threat Detection Models Baseline.

No new detection algorithm is being developed in this study. The predictions of alerts on SIEM-equivalent telemetry are produced by well-trained and popular machine-learning classifiers as base-line detection models on SIEM-equivalent telemetry.

The point of such a baseline models usage is to provide a reasonable detection capability in order to allow the proposed explainability-motivated threat-hunting framework to be evaluated in a systematic way. The qualitative contribution of the work is the explainable analysis and analyst-based interpretation as opposed to optimising detection performance.

4.5 Explainability Module

Explainability Post-hoc explainability modules are disseminated on the trained baseline classifiers to provide instance and global clarifications. The instance-level explanations point at the indicators of behaviour that cause individual alerts, whereas the global ones emphasize the most significant ones in the whole dataset. This explainability on two levels can be useful in real time investigation as well as in the long term analysis of behaviour. This paper will use SHapley Additive exPlanations (SHAP) as a post-hoc explainability method in order to measure the impact of each behavioural feature on the predictions of the baseline detection models.

4.6 Qualitative Evaluation Protocol.

The suggested framework is considered in terms of its quality as interpreted, behaviourally relevant, and useful to the analyst, the generated explanation. The assessment is concerned with the analysis of the representative cases of alerts generated by the detection models based on the baseline and investigating the degree to which the observed behavioural features enhance threat-hunting and investigative reasoning. This evaluation does not aim at benchmarking the detection accuracy, but it investigates the usefulness of the explainability layer in facilitating behavioural interpretation and analyst-directed threat-hunting piping.

4.7 Threat Hunting and Behavioural Analysis.

Each detected alert has the most significant features generated by the explainability module that will be analysed and interpreted as behavioural evidence. These indicators are employed in correlating the results of detection to the higher level attacker actions including abnormal scanning, repeated accesses and anomalous communication behaviour.

Through this experimental analysis we will see how the proposed framework will assist in the practical hunter of cyber threats since it allows behavioural validation of alerts instead of using model confidence scores alone.

5. Qualitative Evaluation and Discussion.

This part is based on a qualitative assessment of the proposed explainable threat-hunting model on publicly accessible SIEM-equivalent datasets. The evaluation will be aimed at evaluating the interpretability of the generated explanations, behavioural relevance, and usefulness to the analyst, and not at comparing the performance of detection. Based on this, the discussion is based on the quality of the explanation, the stability of behaviour and investigative support of the threat-hunting processes.

5.1 Baseline Detection Playground in the Proposed Framework.

Only baseline detection models are implemented to produce alert forecasts that are needed to conduct further explainability analysis. The detection step is considered as enabling element of the proposed framework and the main contribution of this study is a study of how explainability layer can help behavioural understanding and threat hunting by analysts.

The design could be viewed as a realistic representation of SOC settings, where detection engines already run and the key issue is in the understanding, verifying and contextualising the issued alerts.

5.2 Analysis of Instance-Level Explanation.

To illustrate the answer to the question, the SHAP-based explainability layer shows a parsimonious set of behaviour-oriented features along with their relative importance to the detection result. As an analyst, such explanations give instant understanding of behavioural antecedents that prompted an alert. Instead of making use of a predicted class label only, analysts can get to see what actual behavioural indicators have led to the detection result. This supports directly on the rapid triage and allows analysts to assess whether an alert means a significant suspicious behaviour or unsubstantiated operational behaviour. The behaviour-centric features are also used, which makes the generated explanations interpretable and consistent with the common practices when conducting a SOC investigation.

Figure X demonstrates an instance-level breakup of why a single alert was detected with the most influential behavioural features to the decision to do so being shown with the relative contribution. The characteristics associated with the abnormal frequency of connection and the anomalous volume of traffic are presented as the most important features, which means that the alert is largely influenced by the anomaly in communication behaviour, but not independent protocol-level characteristics. This example shows how the suggested framework allows analysts to find the behavioural factors behind an alert as quickly as possible and make quick investigations and validation in threats hunting. The numerical figures in this figure depict the SHAP values which were calculated using the predictions of the model during the training of the baseline on the evaluation dataset. Every SHAP value shows the effect of a given feature to the model output of a particular instance. The SHAP value (change in the model prediction) is presented on the horizontal axis, and the bar plot to the left indicates the average value of the absolute SHAP value of the features to illustrate the overall importance of the features. Every instance of the plot is one data point and the colour is used to show the priority of the feature value (low to high)

.

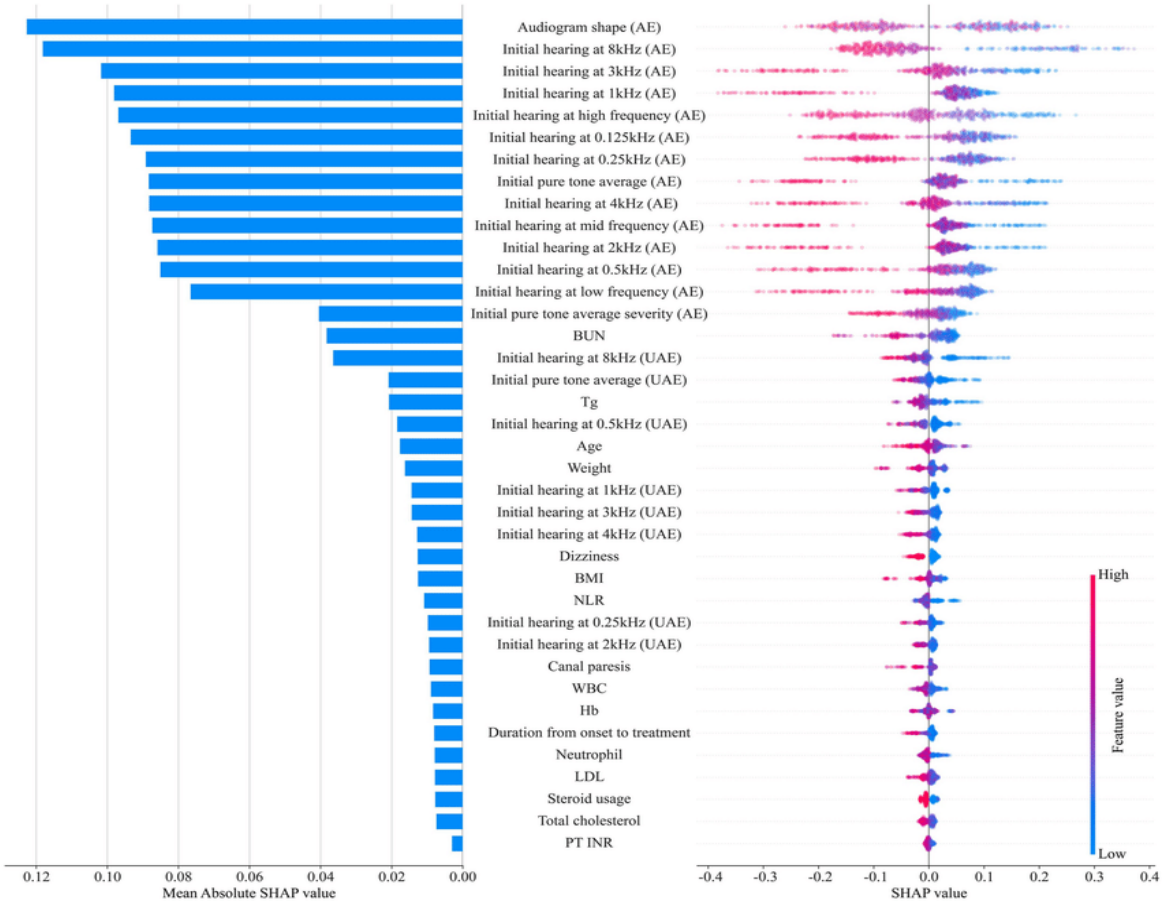

Figure X. Illustrative SHAP-based instance-level explanation for a single security alert.

5.3 Behavioural Interpretation for Threat Hunting

The influential features which were determined by instance-level explanations were further elaborated in the higher level behavioural patterns. The high frequency of short-period connection and high flow count may be an indication of scanning and reconnaissance behaviour, and the long duration of the session and unusual data volume profile may indicate command-and-control communication or data exfiltration efforts. These interpretations of behaviour can enable the analysts to go beyond individual alerts and develop investigative hypotheses. As a result, explainability layer allows proactive threat-hunting whereby analysts may search for related activities of similar behavioural signatures across the monitored environment.

5.4 Explained Behaviour to Known Attack Patterns Mapping

In order to determine the consistency of the generated explanations with the established adversarial behaviour, the behavioural indicators of the explainability results were analysed concerning the MITRE ATT&CK framework. The output of the explanation is highly aligned with the typical types of attacks, including reconnaissance, accessing of credentials and subsequent movement. This consistency shows that the produced explanations are related to operationally significant behaviours as opposed to statistical artefacts, which can be used to operationalise structured investigation and reporting in SOC settings. To further demonstrate explainability outputs in aiding analyst reasoning, Table X provides a summary of the typical mappings between behavioural features, analyst interpretation and threat-hunting meaning.

Table X. Mapping of explainability outputs to analyst-oriented threat-hunting interpretation

Example behavioural feature highlighted by explainability	Analyst interpretation	Threat-hunting meaning	Related ATT&CK tactic
High connection count in a short time window	Repeated probing activity	Reconnaissance/ network scanning	Reconnaissance
Abnormally long session duration	Persistent remote connection	Possible command-and-control communication	Command and control
Unusual traffic volume or burst transfer	Data transfer anomaly	Potential data exfiltration	Exfiltration
Irregular or uncommon protocol usage	Non-standard communication pattern	Suspicious lateral or remote activity	Lateral Movement

5.5 Cross-Dataset Behavioural Consistency

Multiple public datasets resembling SIEM were analyzed using the explainability-driven analysis. Although there were variations in the process of traffic generation and data distribution, the explainability layer consistently indicated the similarity of behavioural features. This behavioural consistency point suggests that the proposed framework is not tied closely to a particular dataset and may be used in a heterogeneous monitoring environment. The generalizability of the explanatory patterns is a factor that supports the high degree of usefulness of the proposed explainable threat-hunting methodology.

5.6 Discussion Discussion and Practical Implications.

The qualitative analysis shows that the incorporation of post-hoc explainability with the current detection models can significantly enhance the informational interpretation and practical usefulness of security alerts. The suggested framework will allow analysts to comprehend the rationale of behaviours which lead to alerts, justify the results of detection through interpretable indicators and match suspicious behaviours with known adversarial behaviours. In contrast with the traditional detection-based methodologies, the suggested methodology directly facilitates the reasoning of analysts, planning of investigations and threat-hunting processes. Moreover, the model-agnostic nature of the explainability layer makes it possible to combine the framework with the existing SIEM and cloud security systems without the need to make changes to deployed detection engines. In general, the findings show that explainability is a significant factor in converting the output of automated detection into actionable and analyst friendly threat-hunting intelligence.

5.7 Limitations and Future Work

The current paper is dedicated to a qualitative assessment of the explainability and threat-hunting support that is oriented at analysts. These quantitative evaluations of the performance of detection (in a large scale) and large-scale empirical validation of its work with operational SOC data will be included in further work.

6. Conclusion

The work demonstrated a clarifiable AI-inspired model of cyber threat hunting in SIEM and cloud security systems, where the emphasis is placed on facilitating the investigation approach and behavioural meaning of security alerts to the needs of analysts. The proposed methodology, in contrast to traditional detection-focused designs, introduces post-hoc explainability in a combination with baseline detection models that allow analyzing suspicious activities in a transparent and understandable way. The qualitative analysis reveals that the suggested framework enables valuable interpretation of behaviour signals, helps map the alerts onto the known attack patterns, and allows proactive threat-hunting processes to security analysts. The model-agnostic form also permits the framework to be combined with the already present detection engines without altering architecture.

References

[1] L. N. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[2] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu and C. Wang, "Machine Learning and Deep Learning Methods for Cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018.

[3] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2010.

[4] M. T. Ribeiro, S. Singh and C. Guestrin, "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.

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[6] R. Marcinkevics and J. E. Vogt, "Interpretable and Explainable Machine Learning for Security and Privacy," *IEEE Security & Privacy*, vol. 19, no. 5, pp. 55–63, 2021.

[7] A. Sarker, R. K. Karim, H. Kayes, S. Badsha, H. Alqahtani and P. Watters, "Cybersecurity Data Science: An Overview from Machine Learning Perspective," *Journal of Big Data*, vol. 7, no. 41, 2020.

[8] V. Chandola, A. Banerjee and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, 2009.

[9] A. Patcha and J.-M. Park, "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[10] H.-J. Liao, C.-H. Lin, Y.-C. Lin and K.-Y. Tung, "Intrusion Detection System: A Comprehensive Review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16–24, 2013.

[11] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," *Military Communications and Information Systems Conference*, 2015.

[12] J. Zhang, M. Zulkernine and A. Haque, "Random-Forest-Based Network Intrusion Detection Systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 38, no. 5, pp. 649–659, 2008.

[13] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," *ICISSP*, 2018.

[14] M. Ring, S. Wunderlich, D. Grüdl, D. Landes and A. Hotho, "A Survey of Network-based Intrusion Detection Data Sets," *Computers & Security*, vol. 86, pp. 147–167, 2019.

[15] G. Apruzzese, M. Colajanni, L. Ferretti, M. Marchetti and M. Mori, "On the Effectiveness of Machine and Deep Learning for Cyber Security," *IEEE International Conference on Cyber Conflict (CyCon)*, 2018.

[16] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[17] W. Samek, G. Montavon, A. Vedaldi, L. Hansen and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 2019.

[18] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence for Cybersecurity," *IEEE International Conference on Big Data*, 2020.

[19] A. Warnecke, D. Arp, C. Wressnegger and K. Rieck,
"Evaluating Explanation Methods for Deep Learning in Security," *IEEE European Symposium on Security and Privacy Workshops*, 2020.

[20] M. Hind et al., "TED: Teaching AI to Explain Its Decisions," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[21] E. Begoli, T. Bhattacharya and D. Kusnezov, "The Need for Uncertainty Quantification and Explainable AI for ML in Cybersecurity," *IEEE Security & Privacy Workshops*, 2019.

[22] R. M. Strom et al., "MITRE ATT&CK: Design and Philosophy," MITRE Technical Report, 2018.

[23] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.