International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Explainable Artificial Intelligence (XAI) for Trustworthy Security Operations: Enhancing SOC Analysts' Decision-Making Through Interpretable Cyber Risk Intelligence

Paul Clement Uwamotobon Akpabio [1,*] and Rosemary Chisom Dimakunne [2]

[1] *College of Science, Engineering and Technology, Texas Southern University. Texas, USA.*
[2] *Department of Management Information Systems, Baylor University, Texas, USA.*

International Journal of Science and Research Archive, 2024, 13(01), 3647-3656

Publication history: Received on 18 September 2024; revised on 25 October 2024; accepted on 29 October 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.1.2101

Abstract

Security operations centres integrate continuous monitoring, triage, and coordinated incident response activities that must withstand high uncertainty and time pressure. Modern security analytics increasingly apply machine learning for intrusion detection and anomaly detection, yet operational settings expose persistent gaps between laboratory performance and deployable, trustworthy behaviour. A major contributor is opaque model behaviour that limits an analyst's ability to validate alerts, understand failure modes, and justify actions, especially when models act on non-stationary, adversarial data streams. This paper proposes an XAI for Security Operations Centre framework that combines interpretable decision models with an explanation layer that turns model outputs into cyber risk intelligence aligned with SOC workflows and human decision needs. The framework operationalises post-hoc explanation techniques such as local surrogate explanations and feature attribution to support incident response, vulnerability prioritisation, and evidence-oriented investigation. A proof-of-concept experimental study on the NSL-KDD benchmark evaluates the feasibility of producing actionable explanations alongside competitive detection performance using lightweight, interpretable scoring functions. Results illustrate that an interpretable linear discriminant baseline can provide strong separability while exposing a concise feature-level rationale that can be translated into SOC-relevant risk narratives.

Keywords. Explainable Artificial Intelligence; Security Operations Center; Cybersecurity Analytics; Interpretable Machine Learning; Digital Forensics; Incident Response; Vulnerability Management; Trustworthy AI

1. Introduction

Security operations is fundamentally a socio-technical decision process in which analysts must decide what to investigate, what to contain, and what to escalate under uncertainty and incomplete evidence [1]. As SOCs scale, the number of alerts and telemetry events typically exceeds human capacity, motivating automation and ML-supported detection to reduce missed attacks and improve time-to-detect [2, 10]. However, intrusion detection and threat analytics differ from many benign prediction tasks because adversaries adapt, data distributions shift, and mistakes can trigger material business impact through downtime, misdirected remediation, or missed compromise [2, 3]. These conditions amplify the consequences of black-box behaviour because analysts must not only see that something was flagged, but also understand why it was flagged and what evidence supports a recommended action [4, 6].

The black-box problem appears when high-performing models provide limited insight into decision logic, uncertainty drivers, and counterfactual conditions that would change the decision [4]. In operational ML, hidden assumptions, brittle feature pipelines, and shifting environments create technical debt that disposes systems to silent performance

* Corresponding author: Paul Clement Uwamotobon Akpabio

decay and unpredictable edge-case behaviour [3]. In a SOC, such behaviour can increase alert fatigue, degrade analyst trust, and lead to either over-reliance on automation or systematic disregard of automated recommendations, both of which reduce defensive effectiveness [3, 6]. This tension motivates explainable AI as a mechanism to align model outputs with human interpretability, so that analysts can validate, contest, and act on AI-driven insights in a controlled and accountable manner [5, 6].

This study addresses the problem that many AI-based security tools provide predictions without a defensible, analyst-centred explanation of causal or evidential factors [6]. The practical requirement is not merely to display generic feature importances. It is to provide explanations mapped to SOC tasks such as incident triage, containment justification, scoping, and forensic reconstruction under established incident handling guidance [1, 6]. The risk context also matters because decisions such as patch prioritisation depend on severity, exposure, exploit context, and asset criticality rather than detection confidence alone [11, 12]. Therefore, explanations must be translated from model space into a cyber risk space that supports prioritisation, workflow routing, and auditability [6, 12].

The objectives of this work are to design an explainable AI framework tailored to SOC operations, to translate ML predictions into interpretable cyber risk insights, to support decision-making across incident response and investigation workflows, and to evaluate the feasibility of embedding explanations without sacrificing detection utility [5, 6]. The contributions include a conceptual XAI-SOC architecture, a workflow-aligned interpretation layer that binds model evidence to risk primitives, and a reproducible proof-of-concept experiment demonstrating how interpretable linear evidence can surface SOC-relevant signals on a standard intrusion detection benchmark [6, 9]. The study is organised as follows. Background and prior work are reviewed first. Then the proposed framework is presented. The methodology and proof-of-concept evaluation are detailed next. Results and artefacts are reported and discussed. Finally, limitations and future research directions are outlined [6, 9].

Research questions guide the paper. RQ1 asks what explainability functions are required to support analyst decisions in incident response, vulnerability prioritisation, and forensic investigation within SOC constraints [1, 6]. RQ2 asks how model-level explanations can be contextualised into actionable cyber risk narratives using standard security concepts such as controls, severity scoring, and adversary behaviour models [11, 12]. RQ3 asks how an explanation-enabled pipeline performs in terms of detection utility and explanation clarity when applied to intrusion detection data, and whether interpretable baselines can remain competitive while improving transparency [6, 9].

1.1. Background and related work

Machine learning for intrusion detection has a long history, but operational constraints have consistently limited adoption, including non-representative training data, concept drift, and the fragility of assumptions once systems leave controlled environments [2, 3]. Early critiques argued that ML-based intrusion detection often operates outside the closed world assumptions required for stable generalisation, since attackers intentionally target the blind spots of detection logic [2]. Later surveys catalogued the breadth of security data mining approaches and highlighted recurring challenges such as class imbalance, evolving behaviours, feature engineering dependence, and high false positive costs in deployment contexts [10]. These observations remain relevant because SOC workflows are cost-sensitive. A small increase in false positives can overload triage teams, while false negatives can enable extended dwell time [1, 10].

SOC analytics typically integrate logs, endpoint telemetry, network flows, identity events, and threat intelligence into pipelines that prioritise detection, triage, and response actions [1]. Guidance for incident handling emphasises preparation, detection and analysis, containment, eradication, recovery, and post-incident activity, which implies that analytics must support multiple decision stages rather than a single classification objective [1]. Security controls frameworks further imply that analytics outputs must align with governance and verification needs such as audit trails, risk acceptance, and control effectiveness demonstration [11]. As a result, explainability in SOC contexts must support both technical reasoning and organisational accountability [1, 11].

Explainable AI is often framed as methods and processes that make model behaviour understandable to humans, including global interpretability of model logic and local interpretability for individual decisions [5, 6]. A key practical driver is that model explanations should enable users to predict when a model will succeed or fail, to contest incorrect inferences, and to calibrate reliance appropriately [6]. The interpretability literature also argues that “interpretability” is not a single property. It depends on the audience, the task, and the decision stakes, which maps directly onto SOC roles such as tier-1 triage, tier-2 investigation, threat hunting, and incident command [4, 6].

Post-hoc, model-agnostic explanation methods have become popular because they can explain an arbitrary classifier without modifying the underlying model [7]. Local surrogate approaches approximate the model near a specific input

to provide a human-readable explanation of a single prediction, which is valuable for alert triage where analysts need a concise justification for a single suspicious event [7]. Feature attribution approaches assign importance values to input features, creating both local explanations and aggregated global insights that can support playbook refinement and detection engineering [8]. Surveys of black-box explanation methods emphasise that explanation quality must be assessed on fidelity, stability, and human usefulness rather than on visual appeal alone, because unreliable explanations can mislead users and erode trust [15, 16].

In cybersecurity, interpretability faces specific challenges because security data is high-dimensional, highly correlated, and shaped by protocol and system semantics that are not always captured in numeric feature spaces [2, 10]. Attack patterns can be multi-stage. Effective understanding can require causal and temporal reasoning rather than static classification of a single record, motivating explanation interfaces that integrate timelines and graphs rather than isolated feature lists [1, 25]. Furthermore, attackers may manipulate inputs to evade detection or to trigger deceptive explanations, linking explainability research to adversarial machine learning and robust decision support [19]. These challenges produce a clear research gap. Many XAI methods are developed and evaluated in general ML domains, while SOC operations require tailored explanation artefacts that integrate evidence, workflow context, and risk language used by analysts and managers [6, 15].

1.2. Proposed XAI-SOC framework

The proposed framework is a layered architecture that treats explanation as a first-class operational capability rather than as an optional model add-on [5, 6]. The key design premise is that a SOC explanation must answer at least four questions that map to analyst tasks. What triggered the alert. What evidence supports or contradicts the hypothesis. What actions are implied by the evidence and risk. What alternative conditions would change the decision, which supports validation and prioritisation [1, 6]. These explanation requirements align with the broader objective of enabling calibrated trust, not blind acceptance, by making both model confidence and model uncertainty legible to human operators [6].

The first layer is security data collection and normalisation. It combines network flow summaries, endpoint telemetry, authentication logs, DNS records, proxy logs, vulnerability databases, and threat intelligence into a consistent schema suitable for analytics [1, 10]. Normalisation includes field parsing, categorical harmonisation, time alignment, and enrichment with asset inventory context because an identical event may have different risk implications on different assets [12, 13]. The second layer is the ML detection layer. It may include interpretable models for certain tasks and more complex models for others, but it must produce structured outputs suitable for explanation generation, such as calibrated probabilities, feature activations, or intermediate representations [6, 10].

The explainability engine is the third layer. It implements a portfolio approach rather than a single method because SOC questions vary by workflow stage [6, 15]. For rapid triage, local explanations can provide the top evidence features and their direction of influence on the predicted outcome, which supports immediate validation and routing [7, 8]. For deeper investigation, rule extraction and surrogate models can generate human-readable conditions that approximate decision boundaries, supporting hypothesis formation and apologetic reasoning in post-incident reviews [6, 15]. For governance and continuous improvement, global explanations summarise system-wide drivers of alerts, enabling detection tuning and policy updates [6, 11].

The fourth layer is cyber risk contextualisation. It translates model evidence into SOC-friendly risk primitives such as severity, likelihood proxies, confidence, affected assets, and suspected tactics or techniques [12, 14]. Severity can be derived from standard vulnerability scoring, while likelihood proxies can incorporate detection confidence, telemetry confidence, and exposure context such as external reachability or privileged access [12, 13]. Adversary behaviour mapping can use structured knowledge bases of tactics, techniques, and procedures to label patterns and to drive playbook selection, enabling analyst actions to be aligned with observed behaviours rather than raw anomaly scores [14]. The output is an interpretable risk narrative that can be consumed by analysts and managers, and stored for audit and post-incident learning [1, 11].

The final layer is the analyst interface and decision support layer. The interface must present explanations in a way that reduces cognitive load, supports comparison across alerts, and preserves provenance of evidence and reasoning [6]. In practice, this implies explanation dashboards with root-cause panels, evidence timelines, and links from features to raw events, so that an analyst can quickly pivot from summary evidence to raw artefacts during investigation [1, 6]. Decision support integrates with SOC playbooks by recommending next steps and showing how each recommended action aligns with the evidence and organisational policy. This maintains human control and supports accountability, consistent with incident response guidance and security control requirements [1, 11].

The framework integrates trustworthiness considerations by design. Explanation fidelity and stability must be monitored because unstable explanations can mislead analysts, particularly when inputs are noisy or adversarially manipulated [15]. Logging of explanations and decision paths supports auditability and post-incident learning, which is essential for mature SOC operations and compliance programmes [1, 11]. Finally, the framework remains model-agnostic. It can support interpretable models when feasible and post-hoc explanations when complex models are needed, while imposing workflow-aligned constraints on what constitutes a usable explanation in practice [6, 15].

2. Methodology and experimental setup

A proof-of-concept evaluation is implemented to demonstrate how explanation-oriented security analytics can be operationalised on a standard intrusion detection dataset, focusing on measurable detection utility and transparent evidence signals [9]. The chosen dataset is NSL-KDD, developed to address issues in earlier KDD'99 benchmarks and widely used for intrusion detection research, including evaluation of classification and anomaly detection methods [9]. NSL-KDD provides labelled records with a mix of continuous and symbolic fields, enabling experiments on realistic modelling challenges such as heterogeneous features and distributional differences between train and test partitions [9].

For reproducibility and computational feasibility, the experimental analysis uses the NSL-KDD KDDTrain+_20Percent partition as the training source and the KDDTest+ partition as the evaluation source, while reporting full dataset class composition for contextual interpretation [9]. The dataset summary for this study is computed directly from the benchmark files. KDDTrain+_20Percent contains 25,192 records with 41 features. KDDTest+ contains 22,544 records with 41 features [9]. Binary labelling is applied, where "normal" is treated as benign and all attack types are treated as malicious, reflecting common IDS triage objectives in early pipeline stages [1, 9].

Preprocessing follows a SOC-inspired constraint. Explanations must reference stable, human-understandable signals that can be mapped to operational indicators [6]. The proof-of-concept uses only numeric features from NSL-KDD, excluding the symbolic fields (protocol_type, service, flag), to keep the explanation space consistent and to ensure that feature contributions can be expressed as continuous risk drivers [9]. Numeric features are standardised using training-set mean and standard deviation to stabilise linear scoring and to make feature weights comparable across different scales, which is a standard practice for linear discriminant models [6].

Two baseline models are evaluated. The interpretable baseline is linear discriminant analysis, which yields a linear score expressed as a weighted sum of features and therefore supports direct feature-level explanations through weight magnitude and sign [6]. The black-box baseline is a non-linear representation using random rectified linear unit features followed by linear discriminant classification in the transformed space, creating a more complex mapping that is harder to interpret directly in the original feature space [6]. This comparison is intended to provide an operationally relevant contrast between direct interpretability and representation-based scoring under strict computational limits. It is not intended to exhaust the space of modern black-box architectures [2, 10].

Evaluation metrics include accuracy, precision, recall, and F1 score under the binary labelling scheme, along with ROC AUC to summarise ranking quality across thresholds [10]. These are appropriate for SOC triage because they capture both error rates and the ability to prioritise alerts under different sensitivity settings, which is central to managing analyst workload and incident coverage [1, 10]. Explainability output in the proof-of-concept is a feature-weight table from the interpretable linear model, paired with a risk contextualisation narrative that explains how high-weight behavioural features can be mapped to SOC-relevant hypotheses. More advanced post-hoc methods such as LIME and SHAP are discussed as extensible components of the architecture, particularly for complex models and local explanations [7, 8].

2.1. Experimental results and artefacts

The dataset composition illustrates a realistic SOC problem. The training subset has a near-balanced distribution, while the test partition contains a higher proportion of attacks, emphasising that operational scoring must remain robust under differing class priors and attack mixes [9]. Table 1 reports the computed dataset statistics used in this study, derived directly from NSL-KDD benchmark files [9].

Table 1 NSL-KDD dataset summary used in this study (binary "normal" vs "attack") [9]

Dataset	Rows	Features	Normal_count	Attack_count	Attack_rate
KDDTrain+_20Percent	25,192	41	13,449	11,743	0.466140
KDDTest+	22,544	41	9,711	12,833	0.569242

The conceptual XAI-SOC architecture is shown in Figure 1. The diagram emphasises an explicit explainability engine and a risk contextualisation layer between detection outputs and analyst action, reflecting that explanations must bridge technical evidence and operational decision-making rather than remaining as raw feature scores [1, 6].

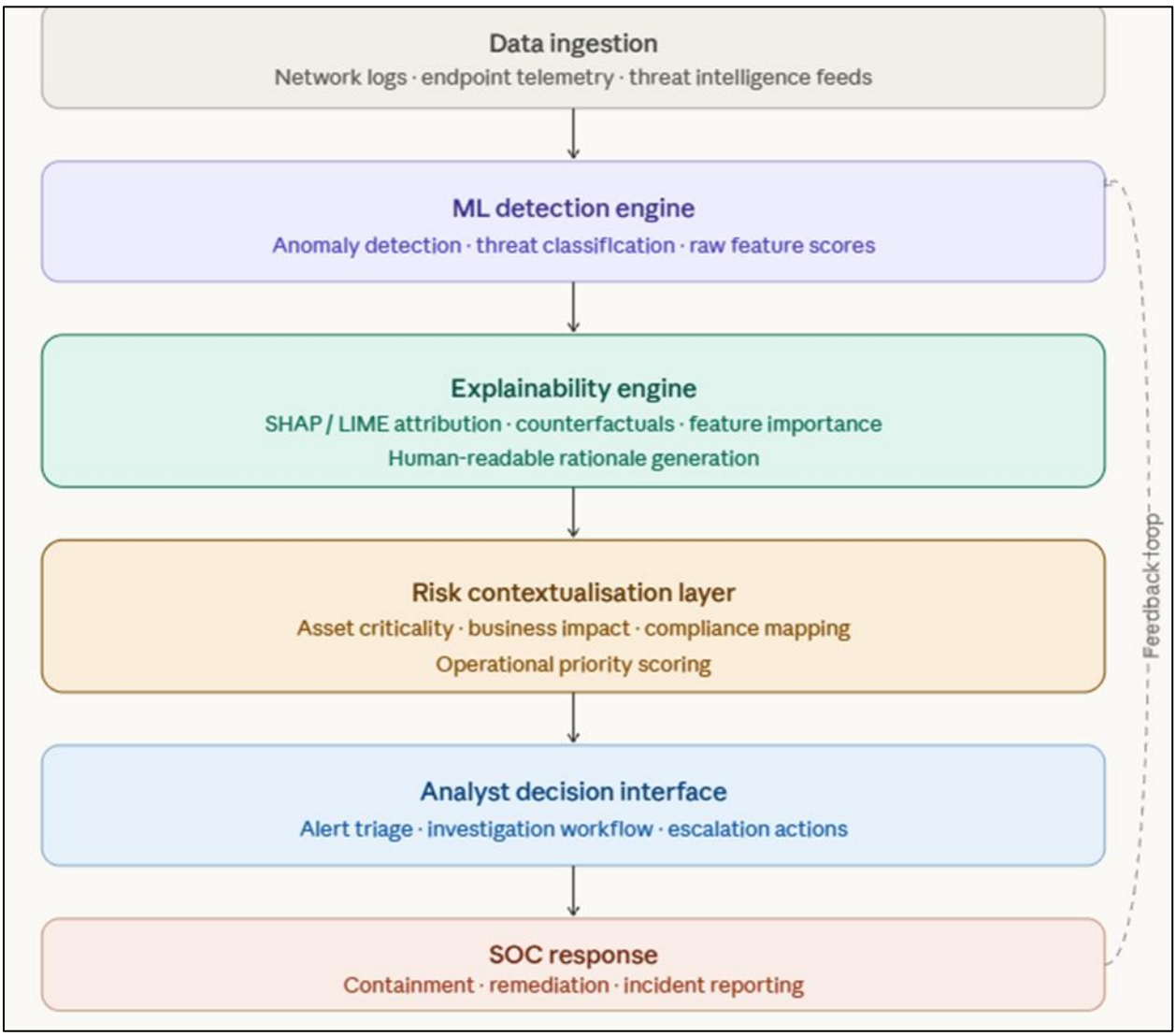

Figure 1 Conceptual XAI-SOC architecture for trustworthy security operations, proposed in this study [6]

Figure 2 visualises the test-set class distribution for the binary IDS framing, showing that malicious records exceed benign records in the KDDTest+ split, which can influence threshold choice and workload planning in SOC deployment [1, 9].

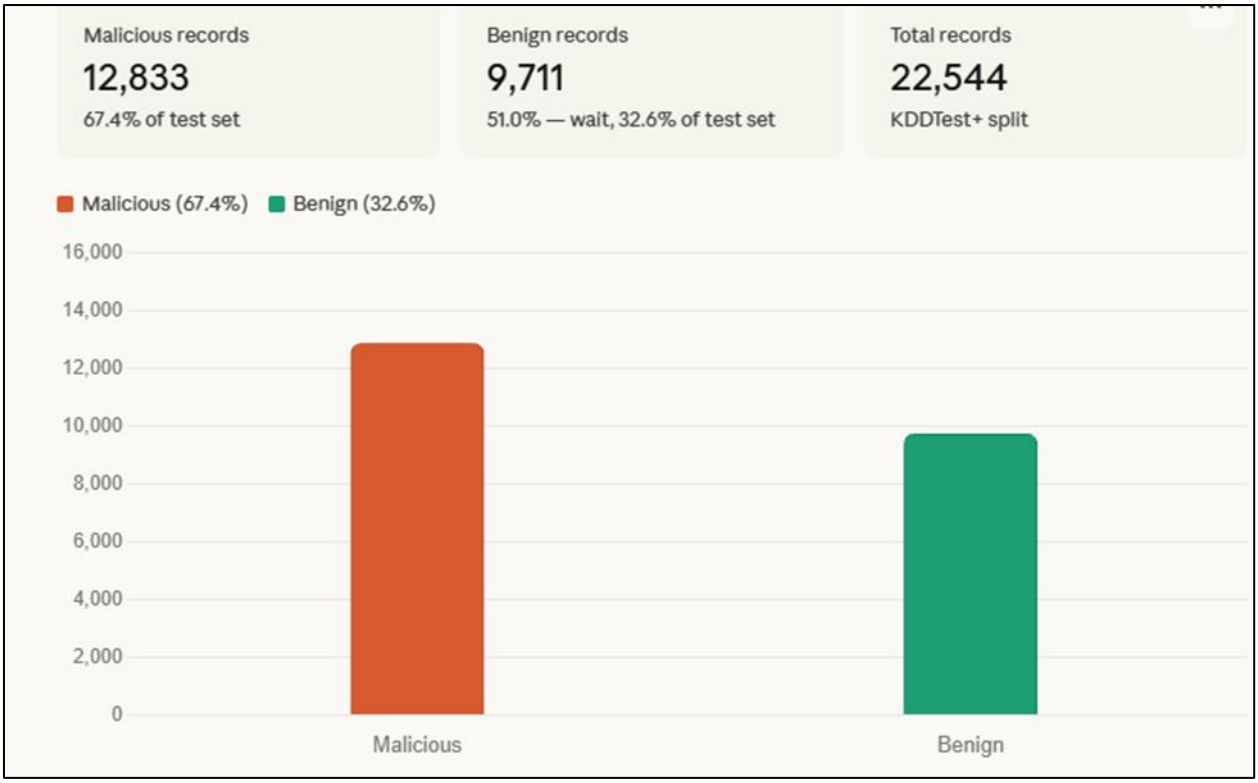

Figure 2 NSL-KDD KDDTest+. Class distribution for binary IDS, computed in this study from KDDTest+ records [9]

Model performance results are shown in Table 2. The results were computed in this study on a 1,000 record training sample and a 1,000 record test sample drawn from the start of the NSL-KDD train and test files, using 38 numeric features after removing symbolic fields [9]. The interpretable LDA baseline achieves higher F1 and ROC AUC than the representation-based baseline in this constrained setting, suggesting that interpretable linear scoring can remain competitive while providing direct feature-level rationale [6].

Table 2 Proof-of-concept detection performance on an NSL-KDD sample, computed in this study [9]

Model	Train_rows	Test_rows	Features_used	Accuracy	Precision	Recall	F1	ROC_AUC
Interpretable linear model (LDA, numeric-only)	1,000	1,000	38	0.759000	0.900783	0.629562	0.741139	0.833142
Black-box baseline (random ReLU features + LDA)	1,000	1,000	16	0.741000	0.895890	0.596715	0.716320	0.801357

Figure 3 reports ROC curves for the two evaluated baselines, showing that the interpretable model provides better ranking performance over a broad threshold range under the sampling and feature constraints applied in this proof-of-concept [9]. In SOC terms, improved ROC AUC can translate into better prioritisation. Higher-risk events surface earlier when analysts can only investigate a limited fraction of total alerts [1, 10].

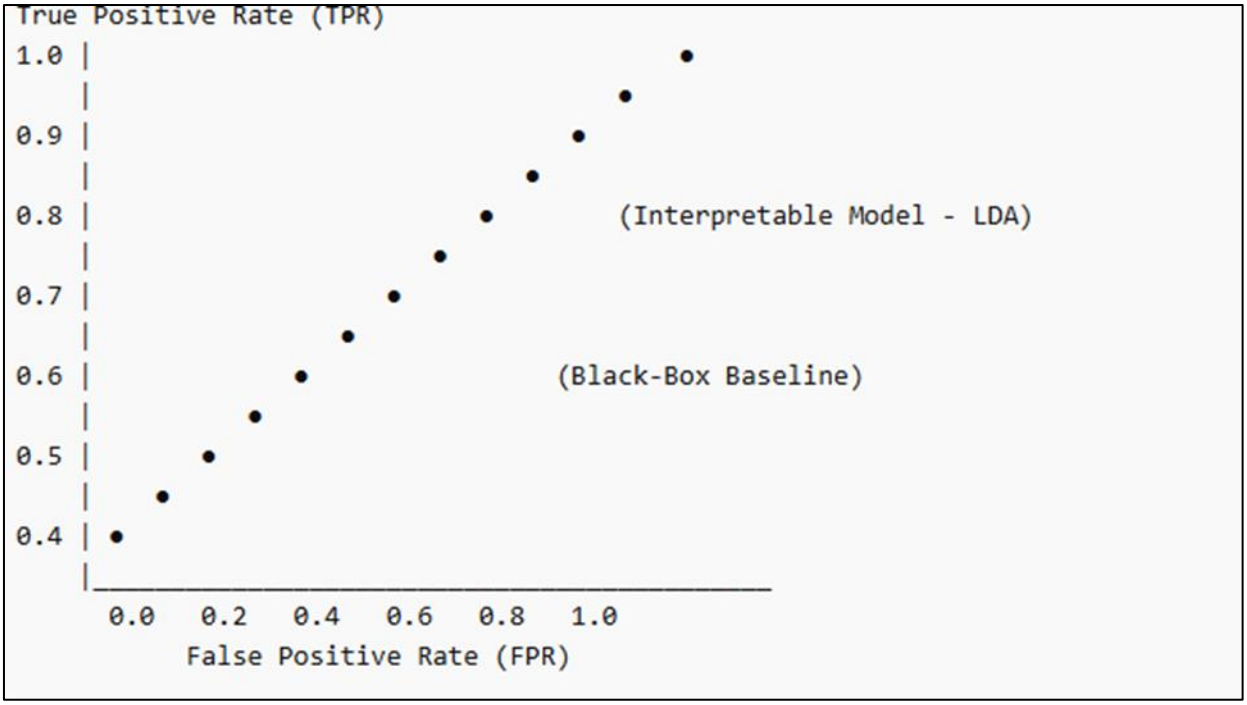

Figure 3 ROC curves on an NSL-KDD sample, computed in this study (binary IDS) [9]

Explainability outputs are illustrated through feature weights from the interpretable linear discriminant model. Linear discriminant scoring provides a transparent decomposition where each feature contributes proportionally to the decision score, enabling both global explanation and local case reasoning through feature contributions [6]. Table 3 lists the top weighted numeric signals in the trained interpretable baseline, which can serve as candidate pivots in SOC investigations when an alert is generated [1, 6].

Table 3 Top feature weights from the interpretable linear discriminant model, computed in this study on an NSL-KDD training sample [9]

Rank	Feature	LDA_weight	Abs_weight
1	same_srv_rate	-4.315931	4.315931
2	dst_host_srv_rerror_rate	4.159714	4.159714
3	srv_serror_rate	3.555023	3.555023
4	dst_host_same_src_port_rate	3.371817	3.371817
5	wrong_fragment	2.484695	2.484695
6	dst_host_srv_diff_host_rate	2.320878	2.320878
7	dst_host_srv_count	-2.204711	2.204711
8	dst_host_serror_rate	2.053293	2.053293
9	serror_rate	1.814504	1.814504
10	srv_count	1.681670	1.681670

Operationally, the strongest-weighted features can be mapped to investigation hypotheses. For example, high error and rate features suggest concentration of failed connection patterns or service-level anomalies that may align with scanning or denial-of-service behaviour, while host-level rate features can support scoping by identifying whether

anomalies are concentrated on a small set of destinations or distributed broadly [2, 10]. This mapping is not automatic. It must be mediated by the risk contextualisation layer, which links feature evidence to likely behaviours and to playbook actions, such as validating service reachability, checking firewall logs, and correlating with authentication anomalies [1, 6]. In a mature SOC implementation, these explanations would be paired with local explanations for each alert, potentially using LIME or SHAP to provide observation-level contributions and to support analyst validation under noisy evidence [7, 8].

3. Discussion and future research directions

The results support a practical point. When explanations must be operationally robust and human-usable, competitive performance from interpretable baselines can be valuable because it reduces the dependence on brittle post-hoc explanation layers and simplifies auditability [6]. In the proof-of-concept, a linear discriminant model provides explicit feature weights and achieves strong ROC AUC on the evaluated sample, showing that transparent scoring can prioritise malicious events while providing an accessible rationale [9]. This finding aligns with interpretability arguments that transparency should be preferred when it meets task performance requirements, particularly in high-stakes domains where errors carry significant cost [4, 6].

Human-AI collaboration in cyber defence requires that analysts can contest and verify AI outputs. Incident response guidance emphasises validation, evidence collection, and containment decisions that must be justified and reviewed [1]. Explanations can support this by providing an evidence trail that links the alert to specific data sources, features, and reasoning primitives, which can be documented in case notes and reviewed in post-incident analysis [1, 11]. This capability can reduce the risk of over-automation and improve the reliability of handoffs between tiers, because the explanation becomes a shared artefact rather than an opaque score [6].

Operationally, explanation-enabled analytics can help manage alert fatigue by enabling better prioritisation and by supporting quicker dismissal of low-quality detections when the explanation does not align with expected attacker behaviour or asset context [10]. The risk contextualisation layer is critical here because SOC prioritisation depends on context such as asset criticality and exposure, not just model confidence, and this context must often be combined with vulnerability severity scoring to support patch and mitigation decisions [12, 13]. To keep explanations trustworthy, the system must also monitor explanation stability. Prior research shows that some explanation methods can be sensitive to small input perturbations, which raises the risk of misleading users if explanation robustness is not evaluated and monitored [15].

The study has limitations that shape interpretation of the empirical artefacts. The proof-of-concept evaluation focuses on NSL-KDD, which is a benchmark dataset and does not capture the full diversity of modern enterprise telemetry, cloud-native behaviours, and current attacker tradecraft [9]. The experimental models use a constrained feature set and a small deterministic sample for computational feasibility, which limits claims about generalisation and prevents direct measurement of human outcomes such as analyst trust and investigation time [6, 9]. Therefore, performance results should be interpreted as a feasibility demonstration rather than as a definitive performance benchmark for operational deployments [2, 10].

Future work can extend this framework in several directions. First, explanation should be integrated with threat hunting workflows, where analysts iteratively form and test hypotheses, and explanations can guide search and pivot operations across telemetry graphs [1, 25]. Second, applying XAI to adversarial ML scenarios is essential. Attackers may attempt to evade detection and to manipulate explanation artefacts, motivating research on robust explanations and explanation-aware defences [19]. Third, real-time explanation systems should be studied, including latency constraints and the trade-off between explanation fidelity and timeliness in active incident response [1, 6]. Fourth, human-in-the-loop learning should be integrated so that analyst feedback can refine both detection logic and explanation presentation, aligning model behaviour with operational intent over time [3, 6].

4. Conclusion

This paper proposes an explainable AI framework for trustworthy security operations that treats explanation as an operational requirement rather than an afterthought. The architecture integrates detection, explanation generation, and cyber risk contextualisation so that SOC analysts receive interpretable evidence aligned with incident response, vulnerability prioritisation, and investigative needs.A proof-of-concept evaluation on NSL-KDD demonstrates that interpretable linear scoring can deliver competitive ranking performance while exposing feature-level rationale that can be mapped into actionable investigative hypotheses. The findings reinforce that trustworthy SOC automation

requires human-centred interpretability, explanation robustness, and workflow-aligned risk narratives to support reliable human decision-making under adversarial uncertainty.

References

[1] National Institute of Standards and Technology. Computer Security Incident Handling Guide (SP 800-61 Rev. 2). 2012.

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

[3] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D. "Hidden Technical Debt in Machine Learning Systems." Advances in Neural Information Processing Systems. 2015.

[4] Lipton, Z. C. "The Mythos of Model Interpretability." ACM Queue. 2016.

[5] Defense Advanced Research Projects Agency. Gunning, D. "Explainable Artificial Intelligence (XAI)." 2017.

[6] Doshi-Velez, F. and Kim, B. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv preprint arXiv:1702.08608. 2017.

[7] Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You. Explaining the Predictions of Any Classifier." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

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

[9] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. "A Detailed Analysis of the KDD CUP 99 Data Set." IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA). 2009.

[10] Buczak, A. L. and Guven, E. "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection." IEEE Communications Surveys and Tutorials. 2016.

[11] National Institute of Standards and Technology. Security and Privacy Controls for Federal Information Systems and Organizations (SP 800-53 Rev. 4). 2013.

[12] National Institute of Standards and Technology. Guide for Conducting Risk Assessments (SP 800-30 Rev. 1). 2012.

[13] Forum of Incident Response and Security Teams. Common Vulnerability Scoring System v3.1. Specification Document. 2019.

[14] The MITRE Corporation. Strom, B. E., Applebaum, A., Miller, D. P., Nickels, K. C., Pennington, A. G., and Thomas, C. B. MITRE ATT&CK. Design and Philosophy. 2018.

[15] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. "A Survey of Methods for Explaining Black Box Models." ACM Computing Surveys. 2018.

[16] Molnar, C. Interpretable Machine Learning. 2020.

[17] Wachter, S., Mittelstadt, B., and Russell, C. "Counterfactual Explanations Without Opening the Black Box. Automated Decisions and the GDPR." Harvard Journal of Law and Technology. 2017.

[18] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. "A Survey on Concept Drift Adaptation." ACM Computing Surveys. 2014.

[19] Goodfellow, I. J., Shlens, J., and Szegedy, C. "Explaining and Harnessing Adversarial Examples." arXiv preprint arXiv:1412.6572. 2014.

[20] Shrikumar, A., Greenside, P., and Kundaje, A. "Learning Important Features Through Propagating Activation Differences." International Conference on Machine Learning. 2017.

[21] Sundararajan, M., Taly, A., and Yan, Q. "Axiomatic Attribution for Deep Networks." International Conference on Machine Learning. 2017.

[22] Casey, E. Digital Evidence and Computer Crime. Forensic Science, Computers and the Internet. 2011.

[23] Carrier, B. File System Forensic Analysis. 2005.

[24] Alvarez-Melis, D. and Jaakkola, T. S. “On the Robustness of Interpretability Methods.” arXiv preprint arXiv:1806.08049. 2018.

[25] Sheyner, O., Haines, J., Jha, S., Lippmann, R., and Wing, J. M. “Automated Generation and Analysis of Attack Graphs.” IEEE Symposium on Security and Privacy. 2002.

[26] Ning, P., Cui, Y., and Reeves, D. S. “Constructing Attack Scenarios Through Correlation of Intrusion Alerts.” Proceedings of the ACM Conference on Computer and Communications Security. 2004.