Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Ethical and Transparent AI in Cyber Defense: Design and Development of an Explainable AI Dashboard for Network Intrusion Detection: A Comparative Study of SHAP, LIME, and ELI5

H.C.S. Peiris
Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

M. A. Thashini Kamalka
Faculty of Business Management
International College of Business and Technology
Colombo, Sri Lanka

Abstract **- Security Operations Centers (SOCs) increasingly rely on machine-learning-based intrusion detection systems (IDSs) to identify malicious network activity. However, high-performing models such as Random Forest (RF) and XGBoost (XGB) often operate as black boxes, making their predictions difficult for analysts to interpret during incident triage. This study presents the design and evaluation of an explainable artificial intelligence (XAI) plugin prototype that integrates IDS predictions with human-interpretable explanations using SHAP, LIME, and ELI5. The proposed framework evaluates explanation methods across stability, fidelity, runtime efficiency, and cross-method agreement. RF and XGB models are trained and evaluated on the CICIDS2017 dataset using five repeated experimental runs to improve statistical reliability. In addition to quantitative evaluation, an interactive Streamlit-based dashboard is developed to operationalize explanations for analyst workflows. Results show that explanation techniques can provide stable and informative attributions while maintaining latency suitable for SOC decision support. The framework demonstrates how explainability can be integrated as a deployable IDS component to support analyst trust, improve transparency, and strengthen cybersecurity operations.**

Keywords - Intrusion detection; explainable artificial intelligence; SHAP; LIME; ELI5; cybersecurity operations

I. INTRODUCTION

The rapid evolution of cyber threats has increased the need for machine-learning (ML) techniques in intrusion detection systems (IDSs). Traditional signature-based detection mechanisms struggle to identify zero-day exploits, polymorphic malware, and sophisticated lateral movement behaviors. Consequently, ensemble learning algorithms such as Random Forest (RF) and gradient-boosted decision trees have become high-performing alternatives for anomaly and misuse detection. However, while these models often provide strong predictive performance, they introduce a critical limitation: interpretability.

In operational Security Operations Centers (SOCs), IDS models do not function in isolation. Automated alerts must be evaluated, validated, and often justified before incident escalation. Analysts must determine whether a flagged event represents a legitimate attack or a false positive. When a model produces a classification without explanation, analysts must

either trust opaque outputs or manually reinvestigate traffic features, increasing cognitive workload, delaying response, and contributing to alert fatigue.

The need for explainable artificial intelligence (XAI) in cybersecurity is therefore operational as well as academic. Governance frameworks such as the OECD AI Principles and the General Data Protection Regulation (GDPR) motivate transparency and accountability in automated decision-making [22, 23]. In cybersecurity contexts, especially within enterprise and critical-infrastructure environments, black-box IDS decisions can be difficult to justify during triage, post-incident review, and stakeholder communication. In this study, these frameworks are used as motivation for transparency; the evaluation focuses on technical and operational XAI properties rather than legal compliance.

Although several studies have explored the integration of XAI into IDS pipelines, many focus primarily on qualitative visualization rather than quantitative evaluation. Common limitations include single-run experiments, lack of confidence intervals or hypothesis testing, limited stability and fidelity analysis, minimal runtime benchmarking, and the absence of deployment-oriented interfaces for analysts. As a result, the robustness, reproducibility, and operational feasibility of explanation outputs remain insufficiently validated.

This study addresses these limitations by presenting a statistically grounded and operationally deployable explainable IDS framework. The approach integrates RF and XGB models trained on the CICIDS2017 dataset. Rather than reporting single-run accuracy, the framework uses five repeated runs with random seeds 42-46. Performance differences are assessed using paired t-tests and 95% confidence intervals.

Beyond classification performance, the study performs a structured quantitative evaluation of SHAP, LIME, and ELI5. The evaluation considers explanation stability, fidelity through top-k feature removal, runtime efficiency, and cross-method ranking agreement using Spearman correlation and Jaccard overlap. This multi-dimensional evaluation moves beyond

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

visual explanation toward measurable interpretability robustness.

The study also develops an interactive dual-model dashboard designed for SOC workflows. The dashboard allows analysts to compare RF and XGB predictions, inspect explanation outputs from multiple XAI methods, and observe model-agreement indicators in real time. In this article, the term plugin refers to a modular explainability layer that can be attached to an IDS pipeline to generate human-interpretable explanations. Due to limited access to live SOC datasets and cloud integration, the plugin is implemented and validated as a standalone Streamlit dashboard prototype.

The main contributions of this paper are: (1) a statistically validated comparison of RF and XGB on CICIDS2017 using repeated experimentation; (2) quantitative evaluation of SHAP, LIME, and ELI5 across stability, fidelity, runtime, and agreement; (3) per-attack and attribution analysis to identify feature-bias patterns; and (4) development of an explainable dual-model dashboard for operational cybersecurity decision support, positioning explainability as a measurable and deployable IDS component.

II. RELATED WORK

A. Explainable artificial intelligence foundations

XAI has emerged as a major research direction for addressing the opacity of complex ML models. SHAP, introduced by Lundberg and Lee [8], provides additive feature attribution grounded in Shapley values from cooperative game theory. It offers desirable properties such as local accuracy, missingness, and consistency, making it a theoretically grounded method for model interpretation.

LIME, proposed by Ribeiro et al. [9], explains black-box model behavior by training locally faithful surrogate models around individual predictions. While LIME is model-agnostic and flexible, it introduces stochasticity through neighborhood sampling, which may affect explanation stability.

ELI5 provides model-inspection utilities including permutation-based feature importance, which can support global interpretability by estimating the effect of feature perturbation on model performance [11]. Broader surveys by Adadi and Berrada [12] and Guidotti et al. [13] emphasize that explanation quality should be assessed beyond visualization, including stability, faithfulness, and usability. Rai [10] and Tjoa and Guan [14] similarly highlight the importance of moving from black-box prediction toward transparent and accountable decision support.

B. Machine learning in intrusion detection

The CICIDS2017 dataset introduced by Sharafaldin et al. [6] has become a common benchmark for evaluating ML-based IDS systems because it contains realistic traffic generation and diverse attack categories. Ensemble methods such as RF and gradient-boosted trees consistently show strong performance on tabular network-flow features.

Ahmed et al. [4] integrates SHAP-based feature selection into hybrid boosting frameworks and report improved predictive performance. However, their work primarily focuses on prediction accuracy and does not include repeated statistical testing across runs. Wang et al. [17] present an explainable ML framework for IDS, while more recent studies explore XAI-supported threat detection and algorithm comparison in cybersecurity contexts [20, 21]. These works demonstrate growing interest in XAI-enabled IDS but also reveal the need for more systematic evaluation of explanation reliability.

Gaspar et al. [3] compare SHAP and LIME explanations for intrusion detection, but their work does not evaluate explanation stability, computational overhead, or operational deployment. Arreche et al. [7] propose an XAI-enhanced IDS framework, while Sharma et al. [19] investigates explainable deep learning for IoT intrusion detection. These studies are valuable, but many rely on limited or single-trial evaluation and provide insufficient evidence on reproducibility and deployment feasibility.

C. Stability and fidelity in XAI

Explanation stability refers to the consistency of feature attributions under repeated evaluations or minor perturbations. Unstable explanations can undermine analyst trust because the same or similar network event may appear to be justified by different features. SHAP is deterministic for many tree-based settings, while LIME can vary because of random sampling. Few IDS studies explicitly measure rank stability using correlation metrics.

Fidelity, or faithfulness, measures whether explanations reflect the actual decision behavior of the model. A common approach is perturbation analysis, in which highly ranked features are removed or modified, and the model response is observed. If removing the top-ranked features does not affect the prediction, the explanation may not reflect true model reasoning. This issue is especially important for IDS contexts, where explanation outputs may be used to support escalation and response decisions.

D. Operational deployment gaps and positioning of this work

Prior work on explainable intrusion detection can be grouped into three main areas: studies that apply SHAP for interpretability or feature selection, studies that use LIME for local post-hoc explanations, and broader XAI-IDS frameworks that discuss explainability conceptually but do not extend to operational deployment. SHAP-based IDS studies often emphasize feature attribution and model transparency, while LIME-based studies primarily focus on instance-level interpretability. However, most existing work evaluates explanations qualitatively or through single-run experiments, with limited attention to stability, fidelity, runtime overhead, or analyst-facing interfaces.

Representative studies demonstrate the value of explainability in intrusion detection but also reveal clear methodological gaps. Lundberg and Lee [8] and Ribeiro et al. [9] established the foundations of SHAP and LIME, respectively, but their work is not specific to IDS deployment. Ahmed et al. [4] used SHAP-supported feature selection to improve IDS prediction, although the emphasis remained primarily on predictive performance rather than explanation stability, fidelity, and runtime. Gaspar et al. [3] compared SHAP and LIME for intrusion detection without repeated statistical validation or an operational analyst interface, while Arreche et al. [7] proposed an XAI-enhanced IDS framework with limited quantitative runtime benchmarking.

More recent studies, including Sharma et al. [19], Mohale and Obagbuwa [2], and Hermosilla et al. [1], further demonstrate the relevance of explainability to cybersecurity, but they do not combine repeated model validation, multi-dimensional XAI assessment, cross-method agreement

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

analysis, and an analyst-facing dashboard in a single framework. The present study addresses this gap by evaluating SHAP, LIME, and ELI5 across stability, fidelity, runtime, and agreement, and by operationalizing the resulting explanations through a dual-model Streamlit prototype for SOC-oriented decision support.

III. METHODOLOGY

A. Dataset and preprocessing

The CICIDS2017 dataset [6] was selected because it provides realistic traffic simulation, diverse attack categories, and widespread adoption in intrusion-detection research. The dataset includes benign traffic and multiple attack types, including DDoS, PortScan, Botnet, Web attacks, brute-force attacks, and infiltration. It contains flow-based network features extracted using CICFlowMeter, including packet counts, flow duration, byte statistics, and header-based attributes.

Duplicate records were removed to reduce potential data leakage. Only numeric features were retained, and non-numeric categorical identifiers were excluded to ensure compatibility with tree-based ensemble methods. Missing and infinite values were handled through standard cleaning procedures to maintain model stability.

Binary classification was implemented by grouping all attack categories under a single Attack label and preserving Benign as the negative class. While multi-class classification provides fine-grained attack differentiation, binary classification reflects a common operational SOC task: distinguishing malicious from benign traffic. Labels were encoded as Normal = 0 and Attack = 1.

B. Data splitting and class imbalance handling

A fixed 70/15/15 train-validation-test split was used with stratified sampling to preserve class proportions across subsets. The training set was used for model fitting, the validation set for hyperparameter confirmation, and the test set for final evaluation. Cross-validation was not used because the objective was a reproducible repeated-run comparison under fixed split conditions. Robustness was instead evaluated through multiple seeded runs.

CICIDS2017 exhibits imbalance across attack types and between benign and malicious traffic. Rather than applying oversampling or synthetic methods such as SMOTE, which may distort real-world traffic distributions, class imbalance was addressed using algorithm-level weighting. RF used inverse-frequency class weights, while XGB used the scale_pos_weight parameter. This approach preserves the original traffic distribution while reducing bias toward majority classes.

C. Model configuration

RF was selected for its robustness, relative interpretability, and strong performance on tabular IDS data. Hyperparameters were manually validated using validation curves across n_estimators ranging from 50 to 250. The final RF configuration used n_estimators = 200, max_depth = 12, min_samples_split = 20, min_samples_leaf = 10, max_features = 'sqrt', and max_samples = 0.7. The final XGB configuration used n_estimators = 200, max_depth = 8, learning_rate = 0.08, subsample = 0.7, colsample_bytree = 0.7, gamma = 0.3, reg_alpha = 0.6, and reg_lambda = 6.

D. Experimental protocol and statistical testing

To improve reproducibility and statistical validity, each model was trained and evaluated five times using random seeds 42-46. For each run, accuracy and per-attack detection behavior were recorded and aggregated. Mean and standard deviation were computed across runs.

To determine whether the observed difference in performance between RF and XGB was statistically significant, a paired t-test was conducted across the five repeated runs using the same random seeds. The null hypothesis stated that there was no significant difference between the mean accuracies of the two models, while the alternative hypothesis stated that a significant difference existed. The paired design ensures that both models are compared under matched experimental conditions.

E. Quantitative XAI evaluation

Three explanation methods were evaluated: SHAP for local additive attribution, LIME for local surrogate interpretation, and ELI5 for permutation-based global feature importance. The evaluation considered four dimensions: stability, fidelity, runtime efficiency, and cross-method agreement.

Stability was measured by repeating LIME explanations on a stratified subset of held-out test instances and computing Spearman rank correlation over the top-15 ranked features. Fidelity was evaluated through top-k feature removal, where the top-ranked features were ablated by setting them to 0 in standardized feature space, equivalent to replacing them with the training-set mean under StandardScaler. Predictions were then recomputed, and prediction flips, accuracy drops, and changes in attack probability were recorded.

Runtime benchmarking measured explanation latency on the same reproducible stratified subset. SHAP and LIME were evaluated as per-instance explainers, while ELI5 was evaluated as a one-off global computation. Agreement between methods was assessed using Spearman rank correlation over top-15 features and Jaccard overlap over top-10 feature sets.

IV. RESULTS AND QUANTITATIVE XAI EVALUATION

A. Classification performance

Across five repeated runs, both models achieved high classification accuracy. Table I summarizes the aggregated performance results.

TABLE I. CLASSIFICATION PERFORMANCE ACROSS REPEATED RUNS

Model	Mean accuracy (%)	Standard deviation (%)	95% confidence interval
Random Forest (RF)	98.85	0.07	[98.76, 98.93]
XGBoost (XGB)	98.10	0.03	[98.07, 98.13]

RF achieved a mean accuracy of 98.85% +/- 0.07%, while XGB achieved 98.10% +/- 0.03%. The paired t-test produced $t = 35.2$ with $p < 0.001$, indicating that the difference in mean accuracy was statistically significant. The null hypothesis was therefore rejected, confirming that RF achieved significantly higher mean accuracy than XGB under the evaluated setting. However, XGB showed lower variance across runs, indicating greater consistency in point estimates.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

B. Per-attack behaviour and attribution analysis

Feature attribution analysis of false-positive cases using SHAP, LIME, and ELI5 showed that flow duration, packet-length mean, and forward-packet statistics frequently contributed to benign flows being classified as attacks. This suggests that benign long-duration or high-volume flows can share statistical characteristics with attack patterns. From an SOC perspective, identifying these attribution patterns can support threshold refinement and reduce unnecessary investigation effort.

Although the classifier is binary, attack-family detection was evaluated by grouping held-out test samples using the original CICIDS2017 labels and computing the proportion correctly predicted as Attack within each family. Results showed strong detection for high-volume families such as DDoS and DoS Hulk, while several minority families showed low or unstable detection because of extreme class imbalance and limited representation in the test split. Infiltration was particularly unreliable because of the very small number of test samples. These findings show why overall accuracy can remain high even when minority attack families are missing. Operational deployment should therefore incorporate per-family monitoring, false-negative analysis, threshold calibration, and targeted retraining for rare attack families.

C. Stability analysis

Explanation stability is important for analyst trust because inconsistent explanations can reduce confidence in automated alerts. SHAP was deterministic for the evaluated tree-based models, while ELI5 provided repeatable global feature rankings under a fixed dataset and random seed. Because LIME is stochastic due to neighborhood sampling, its stability was evaluated explicitly on a reproducible stratified subset of held-out test instances using five repeated runs per instance. Table II summarizes the LIME stability results.

TABLE II. LIME EXPLANATION STABILITY OVER REPEATED RUNS

Model	Mean Spearman	Median Spearman	Instances with rho < 0.5
Random Forest (RF)	0.7426	0.7656	0%
XGBoost (XGB)	0.9311	0.9361	0%

The results indicate stable LIME rankings for both models, with XGB showing stronger rank consistency than RF. No evaluated instances produced a Spearman correlation below 0.5, suggesting that LIME explanations remained sufficiently consistent under the controlled experimental design.

D. Fidelity evaluation through top-k feature removal

Fidelity was evaluated by removing the three and five highest-ranked features identified by each explanation method and then measuring prediction flips, accuracy reduction, and change in attack probability. For RF at top-3 removal, SHAP produced 40% prediction flips and a 38.0% accuracy drop, with mean and median attack-probability reductions of 0.5669 and 0.6578. LIME produced 36% flips and a 34.0% accuracy drop, with corresponding probability reductions of 0.4898 and 0.5040, whereas ELI5 produced only 4% flips and a 4.0% accuracy drop, with reductions of 0.3265 and 0.4583.

When the top five RF features were removed, SHAP produced 47% prediction flips and a 45.0% accuracy drop, LIME produced 45% flips and a 43.0% drop, and ELI5 increased to 37% flips with a 35.0% drop. The mean attack-probability reductions were 0.5974 for SHAP, 0.5531 for LIME, and 0.4488 for ELI5. For XGB at top-3 removal, SHAP, LIME, and ELI5 produced prediction-flip rates of 26%, 28%, and 38%, respectively, with accuracy drops of 24.0%, 28.0%, and 38.0%. Their mean attack-probability reductions were 0.4343, 0.5041, and 0.7448.

At top-5 removal for XGB, each method produced a 38% prediction-flip rate; the accuracy drops were 32.0% for SHAP and 38.0% for both LIME and ELI5. Mean attack-probability reductions increased to 0.5566, 0.7267, and 0.7487, respectively. These findings indicate strong instance-level fidelity for SHAP and LIME on RF, while ELI5 became more influential as additional globally important features were removed. For XGB, all three methods showed substantial fidelity, with ELI5 producing the strongest probability-level effect. Because feature ablation may create combinations of flow characteristics that do not occur naturally, these values are interpreted as controlled model-sensitivity measures rather than realistic counterfactual network scenarios.

E. Runtime benchmarking

Operational deployment requires explanation latency compatible with SOC workflows. For RF, SHAP was the fastest per-instance method, with a mean runtime of 0.0057 s, a median of 0.0056 s, and a 95th-percentile runtime of 0.0072 s. LIME required a mean of 0.4771 s, a median of 0.4797 s, and a 95th-percentile runtime of 0.5491 s. ELI5 required 22.73 s as a one-off global computation.

For XGB, LIME achieved a mean per-instance runtime of 0.1697 s, while SHAP required 0.2897 s. ELI5 completed its global computation in 0.49 s. Therefore, SHAP is highly suitable for interactive RF explanations, whereas LIME was the fast local explainer for the evaluated XGB implementation. ELI5 is better suited to periodic global interpretation and reporting than to high-frequency per-alert reasoning because it provides model-level importance rather than an explanation of an individual prediction.

F. Agreement between explanation methods

Agreement between explanation methods was evaluated using Spearman rank correlation over the top-15 ranked features and Jaccard overlap over the top-10 feature sets. For RF, SHAP and LIME showed the strongest alignment, with mean and median Spearman correlations of 0.4060 and 0.4144 and mean and median Jaccard overlaps of 0.3437 and 0.3333. SHAP and ELI5 showed lower agreement, with Spearman values of 0.0813 and 0.0800 and Jaccard values of 0.2119 and 0.2500. The RF LIME-ELI5 comparison produced a mean Spearman correlation of 0.1385 and a mean Jaccard overlap of 0.1943.

For XGB, SHAP and LIME achieved a mean Spearman correlation of 0.2870 and a mean Jaccard overlap of 0.4024. SHAP and ELI5 showed a mean Jaccard overlap of 0.3283, while LIME and ELI5 produced the strongest XGB feature-set overlap, with a mean Jaccard value of 0.4632 and a median of 0.4286; their mean Spearman correlation was 0.2505. The incomplete convergence is expected because SHAP and LIME explain individual predictions, whereas ELI5 provides global permutation importance. Agreement involving ELI5 should therefore be interpreted as cross-level consistency rather than direct equivalence. Overall, the methods provide complementary views of model reasoning rather than identical feature rankings.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

V. EXPLAINABLE DASHBOARD ARCHITECTURE AND OPERATIONAL INTEGRATION

While quantitative evaluation of models and explanation methods is critical, real-world deployment also requires operational usability. Many IDS-XAI studies conclude with offline feature-importance plots, leaving a gap between experimentation and deployment. This study addresses that gap through a dual-model explainable dashboard designed for SOC-oriented workflows.

The implemented dashboard follows a layered architecture. The input layer ingests structured network-flow features and applies preprocessing consistent with training conditions. The model-inference layer runs RF and XGB in parallel, supporting cross-model comparison, disagreement detection, and confidence contrast. The explanation-engine layer generates SHAP, LIME, and ELI5 explanations. The consensus and visualization layer presents prediction labels, probabilities, model-agreement indicators, and top contributing features to analysts.

Model agreement is used as an operational risk indicator. When RF and XGB agree on a classification with high confidence, prediction reliability is strengthened. Conversely, disagreement may indicate borderline cases or feature-distribution anomalies that require closer analyst review. This agreement-based risk signal provides a second validation layer beyond single-model confidence scores.

The moderate correlations observed between explanation methods suggest that no single method fully captures model reasoning. The dashboard therefore supports explanation complementarity by presenting SHAP local attributions, LIME local surrogate reasoning, and ELI5 global importance. This triangulation allows analysts to compare reasoning patterns, identify inconsistencies, and avoid blind reliance on a single explanation method.

The dashboard was functionally evaluated using held-out test instances to confirm consistent preprocessing, dual-model inference, explanation generation, model-agreement indicators, and visualization. This evaluation demonstrates prototype feasibility but does not constitute a formal user study.

VI. ETHICAL AND GOVERNANCE IMPLICATIONS

The integration of explainability into IDS design extends beyond usability; it also supports transparency, accountability, and responsible deployment. Although IDS outputs do not directly determine legal outcomes, they can influence incident response, escalation decisions, and audit documentation. An explainable IDS can therefore provide justifiable decision pathways, traceable feature influence, and auditable alert reasoning.

False-positive attribution analysis revealed concentration around features such as flow duration and packet statistics. Such concentration may reflect structural similarities between benign high-volume flows and attack traffic or model sensitivity to traffic bursts. Without explanation mechanisms, these patterns may remain hidden. Feature-level reasoning therefore enables bias auditing and model refinement.

Trust calibration is also essential. Blind trust in high-accuracy models can lead to overconfidence, while unstable explanations can undermine analyst confidence. By evaluating stability, fidelity, runtime, and agreement, this study calibrates trust in explanation outputs rather than assuming that all visual explanations are reliable.

VII. THREATS TO VALIDITY AND LIMITATIONS

The study was designed to improve methodological rigor through repeated experimentation, statistical testing, and multiple explanation-quality metrics. Nevertheless, several boundaries should be considered when interpreting the results.

First, evaluation was limited to the CICIDS2017 dataset. Although this dataset is widely used in IDS research, results may vary on newer datasets, enterprise traffic, encrypted flows, or cloud-native telemetry. Second, the study focused on RF and XGB because of their strong performance on tabular network-flow data. Findings may not directly generalize deep learning or sequence-based IDS architectures.

Third, the dashboard was validated as a standalone Streamlit prototype rather than as a fully integrated production SOC plugin. This was appropriate for demonstrating operational feasibility, but future work should evaluate integration with live alert pipelines, SIEM tooling, and analyst case-management systems. The dashboard evaluation focused on functional prototype validation rather than a formal user study.

VIII. CONCLUSION

This study presented a statistically validated and operationally oriented explainable intrusion-detection framework integrating RF and XGB models with SHAP, LIME, and ELI5 explanation methods. Unlike prior IDS-XAI research that emphasizes qualitative visualization, this work combined five-run statistical validation, confidence-interval reporting, quantitative stability measurement, top-k feature-removal fidelity testing, runtime benchmarking, cross-method agreement analysis, and dashboard-based operational integration.

The results show that explanation methods can provide stable and decision-relevant feature attributions while maintaining latency compatible with analyst workflows. The dual-model dashboard further demonstrates how explanation outputs can be translated into a deployable decision-support layer for SOC environments. By combining statistical rigor, quantitative interpretability assessment, and operational dashboard design, this framework advances explainable and responsible AI integration within cybersecurity operations. Future work should extend evaluation to additional datasets, conduct larger user-centered trust studies, and explore adaptive explanation frameworks for evolving threat landscapes.

REFERENCES

[1] P. Hermosilla, S. Berrios, and H. Allende-Cid, “Explainable AI for forensic analysis: a comparative study of SHAP and LIME in intrusion detection models,” Applied Sciences, vol. 15, no. 13, p. 7329, 2025, doi: 10.3390/app15137329.

[2] V. Z. Mohale and I. C. Obagbuwa, “Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability,” Frontiers in Computer Science, vol. 7, Art. no. 1520741, 2025, doi: 10.3389/fcomp.2025.1520741.

[3] D. Gaspar, P. Silva, and C. Silva, “Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron,” IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3368377.

[4] U. Ahmed et al., “Hybrid bagging and boosting with SHAP-based feature selection for enhanced predictive modeling in intrusion detection systems,” Scientific Reports, vol. 14, Art. no. 30532, 2024, doi: 10.1038/s41598-024-81151-1.

[5] M. Tavallaee, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” in Proc. IEEE Symp. Computational Intelligence for Security and Defense Applications, Ottawa, Canada, 2009, doi: 10.1109/CISDA.2009.5356528.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

[6] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "A detailed analysis of the CICIDS2017 data set," in Information Systems Security and Privacy, vol. 977, Cham, Switzerland: Springer, 2018, pp. 172–188, doi: 10.1007/978-3-030-25109-3_9.

[7] O. Arreche, T. Guntur, and M. Abdallah, "XAI-IDS: toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems," Applied Sciences, vol. 14, no. 10, p. 4170, 2024, doi: 10.3390/app14104170.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," arXiv:1705.07874, 2017, doi: 10.48550/arXiv.1705.07874.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[10] A. Rai, "Explainable AI: from black box to glass box," Journal of the Academy of Marketing Science, vol. 48, pp. 137–141, 2020, doi: 10.1007/s11747-019-00710-5.

[11] ELI5 Documentation, "ELI5: Explain Like I am 5," accessed Mar. 3, 2026.

[12] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[13] R. Guidotti et al., "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, Art. no. 93, 2018, doi: 10.1145/3236009.

[14] M. Tjoa and C. Guan, "A survey on explainable artificial intelligence: toward medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.

[15] A. Shapira, G. Blomberg, and M. Last, "Feature-based and example-based explanation analysis in XGBoost and SHAP for fraud detection," IEEE Access, vol. 9, pp. 145714–145726, 2021, doi: 10.1109/ACCESS.2021.3052482.

[16] G. Ke et al., "LightGBM: a highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[17] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," IEEE Access, vol. 8, pp. 73127–73141, 2020, doi: 10.1109/ACCESS.2020.2988359.

[18] SHAP Documentation, "SHAP: SHapley Additive exPlanations," accessed Mar. 3, 2026.

[19] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Explainable artificial intelligence for intrusion detection in IoT networks: a deep learning-based approach," Expert Systems with Applications, vol. 238, Art. no. 121751, 2024, doi: 10.1016/j.eswa.2023.121751.

[20] S. J. Mohammed and B. M. Nema, "Threat detection based on explainable AI and hybrid learning," Mesopotamian Journal of Cybersecurity, vol. 5, no. 2, 2025, doi: 10.58496/MJCS/2025/029.

[21] D. Satyanarayana and E. Saikiran, "Intrusion detection system in explainable artificial intelligence by using different algorithms," in Proc. Int. Conf. Distributed Computing and Optimization Techniques, 2024, pp. 1–4, doi: 10.1109/ICDCOT61034.2024.10515463.

[22] OECD, "Recommendation of the Council on Artificial Intelligence," OECD/LEGAL/0449, 2019.

[23] European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Official Journal of the European Union, 2016.