Explainable AI Frameworks for Trustworthy Autonomous Cyber Defense System

Aristotle Ben *

Posted Date: 26 January 2026

doi: 10.20944/preprints202601.1922.v1

Keywords: Explainable AI (XAI); Autonomous Cyber Defense Systems (ACDS); trustworthy AI; cybersecurity; intrusion detection; human-AI teaming; algorithmic transparency

 doi:10.20944/preprints202601.1922.v1

Article

Explainable AI Frameworks for Trustworthy Autonomous Cyber Defense System

Aristotle Ben

Department of Computer Science, Artificial Intelligence and Data Science, Research Unit, Stanford University, USA; bade6448@gmail.com

Abstract

The increasing sophistication and frequency of cyber threats necessitate a shift towards Autonomous Cyber Defense Systems (ACDS). While Artificial Intelligence (AI), particularly machine learning (ML), provides the requisite speed and scalability for such systems, their inherent opacity poses a significant barrier to trust and adoption. Unexplainable ACDS can lead to erroneous actions that are difficult to diagnose, hinder human-AI collaboration, and raise serious accountability and compliance concerns. This research article investigates the integration of Explainable AI (XAI) frameworks as a critical enabler for trustworthy ACDS. Through a mixedmethods approach combining a systematic review of XAI techniques with a quantitative case study on an intrusion detection dataset, this study evaluates the efficacy, performance trade-offs, and human-interpretability of prominent XAI frameworks in a cyber defense context. Findings indicate that post-hoc explanation methods, such as SHAP and LIME, are currently most practical for elucidating complex model decisions, but they introduce computational overhead. The study further reveals a tension between model interpretability and predictive performance, particularly for sophisticated ensemble and deep learning models. The discussion synthesizes a proposed hybrid XAI framework tailored for ACDS, balancing real-time explainability with defense performance. We conclude that for ACDS to be operationally trusted, XAI is not an optional addon but a foundational requirement. The article outlines a roadmap for future research, emphasizing the need for standardized evaluation metrics, human-in-the-loop validation, and regulatory frameworks for explainable cyber operations.

Keywords: Explainable AI (XAI); Autonomous Cyber Defense Systems (ACDS); trustworthy AI; cybersecurity; intrusion detection; human-AI teaming; algorithmic transparency

1. Introduction

1.1. Brief Summary of the Study

This study addresses the critical challenge of trust in Autonomous Cyber Defense Systems (ACDS) by investigating the application of Explainable AI (XAI) frameworks. As cyber attacks evolve in scale and complexity, AI-driven autonomous responses become essential. However, the "blackbox" nature of advanced AI models creates significant risks, including unexplained false positives/negatives, adversarial manipulation, and a lack of operational accountability. This research systematically examines how XAI methodologies can be integrated into the ACDS pipeline to provide transparency, justify autonomous actions, and foster trust among security operators. The study evaluates current XAI techniques, measures their impact on system performance, and proposes a tailored framework for deploying trustworthy ACDS.

1.2. Purpose of the Research

The primary purpose of this research is threefold: (1) To analyze and categorize existing XAI frameworks for their suitability in dynamic, high-stakes cyber defense environments. (2) To empirically evaluate the trade-offs between explanation fidelity, computational cost, and defensive

 doi:10.20944/preprints202601.1922.v1

efficacy in a simulated ACDS context. (3) To synthesize a set of design principles and a conceptual architecture for an XAI-integrated ACDS that enhances trust without compromising defensive capabilities.

1.3. Methodology

A mixed-methods research design was employed. First, a systematic literature review was conducted to establish the state-of-the-art in XAI and its applications in cybersecurity. Second, a quantitative experimental study was performed using the CIC-IDS2017 dataset. Multiple ML models (Decision Tree, Random Forest, Deep Neural Network) were trained for intrusion detection and subsequently explained using post-hoc techniques (SHAP, LIME, and a surrogate model). Metrics for explanation accuracy, stability, and computational overhead were collected alongside standard performance metrics (accuracy, F1-score). The study adhered to ethical guidelines for the use of public cybersecurity datasets.

2. Literature Review

2.1. Background and Context

The cybersecurity landscape is defined by a chronic shortage of skilled personnel and an overwhelming volume of threats. ACDS promise to alleviate this burden by automating threat detection, analysis, and response (Töpfer et al., 2022). AI, especially deep learning, excels at identifying complex patterns in network traffic, malware, and user behavior. However, this capability comes at the cost of interpretability. The "right to explanation" embedded in regulations like the GDPR, and operational needs in military/enterprise contexts (e.g., the need for a Commander's understanding of an autonomous action), make explainability a paramount concern (Arrieta et al., 2020). Trust, as defined by recent NIST guidelines, encompasses reliability, resilience, safety, and transparency – all of which are impacted by explainability (NIST, 2023).

2.2. Research Questions

This study is guided by the following research questions:

RQ1: Which categories of XAI frameworks (intrinsic, post-hoc, model-agnostic, model-specific) are most suited for integration into different stages (detection, diagnosis, response) of an ACDS pipeline?

RQ2: What are the measurable performance trade-offs (e.g., latency, accuracy, resource consumption) when integrating post-hoc XAI techniques into a real-time ACDS? RQ3: How do explanations generated by XAI frameworks affect the trust and decision-making efficacy of human security analysts in a simulated incident response loop?

2.3. Significance of the Study

This research contributes to the emerging field of trustworthy AI in cybersecurity. It moves beyond mere technical performance of AI models to address the socio-technical challenge of trust. The findings are significant for: (1) **Security Operations Center (SOC) Managers**, by providing evidence-based guidance on deployable, explainable AI tools; (2) **ACDS Developers**, by offering a framework for designing transparent systems from the ground up; and (3) **Policy Makers**, by highlighting the technical feasibility and necessities of explainability in regulating autonomous cyber systems. The proposed hybrid XAI framework aims to serve as a blueprint for nextgeneration, accountable cyber defense.

 doi:10.20944/preprints202601.1922.v1

3. Methodology

3.1. Research Design

A sequential mixed-methods design was used. Phase 1 was qualitative: a systematic narrative review of academic literature (2018-2024) from IEEE Xplore, ACM Digital Library, and

SpringerLink using keywords "XAI," "explainable AI," "cybersecurity," "autonomous defense," and "trustworthy AI." Phase 2 was quantitative: an experimental study measuring the impact of XAI on ML-based intrusion detection systems.

3.2. Participants or Datasets

The primary dataset for the quantitative phase was the CIC-IDS2017 dataset (Sharafaldin et al., 2018). It contains benign traffic and common up-to-date attacks, with 80 network flow-based features. It was selected for its realism and widespread use as a benchmark in intrusion detection research.

3.3. Data Collection Methods

- **Literature Review:** 78 relevant papers were identified, screened, and synthesized to map XAI techniques to cyber defense tasks.
- **Experimental Data:** The CIC-IDS2017 dataset was preprocessed (handling missing values, normalization, label encoding). It was split into 70% training, 15% validation, and 15% testing sets. Three model architectures were implemented: a Decision Tree (DT - intrinsically interpretable), a Random Forest (RF - ensemble black-box), and a Deep Neural Network (DNN - deep black-box).

3.4. Data Analysis Procedures

1. **Model Training & Baseline Performance:** All models were trained, and baseline metrics (Accuracy, Precision, Recall, F1-Score, Inference Time) were recorded.
2. **XAI Application:** Post-hoc explanations were generated for a stratified sample of 1000 test instances (including true positives, false positives, true negatives, false negatives).
    - **SHAP (SHapley Additive exPlanations):** KernelExplainer for DT and RF; DeepExplainer for DNN.
    - **LIME (Local Interpretable Model-agnostic Explanations):** Applied to all models.
    - **Surrogate Model:** A global logistic regression model was trained on the predictions of the RF and DNN models.
3. **Evaluation Metrics:**
    - **Explanation Fidelity:** For LIME and the surrogate model, we measured how well the explanation model approximated the black-box model's predictions (using R^2 for regression tasks and accuracy for classification tasks on the explained instances).
    - **Stability:** Measured by applying LIME multiple times to the same instance and calculating the Jaccard similarity between the top-k features identified.
    - **Computational Overhead:** The time added to the inference pipeline to generate an explanation was measured.
4. **Human-Trust Simulation:** A simplified survey was designed where 5 expert security analysts were presented with model predictions for 10 incident scenarios, both with and without SHAP-generated feature attribution plots. They rated their confidence in the system's decision on a Likert scale (1-5).

3.5. Ethical Considerations

As the study used a publicly available, anonymized dataset (CIC-IDS2017), no personal data was involved. The simulated human trust component involved voluntary participation of professionals, who were informed about the study's purpose and assured of the anonymity of their

 doi:10.20944/preprints202601.1922.v1

responses. The research acknowledges the dual-use nature of cybersecurity AI and emphasizes its defensive application.

4. Results

4.1. Model Performance Baseline

Table 1. Baseline Performance of ML Models on CIC-IDS2017 Test Set.

Model	Accuracy	F1-Score (Weighted)	Avg. Inference Time (ms)
Decision Tree (DT)	96.2%	0.959	0.8
Random Forest (RF)	**99.1%**	**0.990**	4.2
Deep Neural Net (DNN)	98.7%	0.987	3.1 (GPU)

4.2. XAI Framework Performance

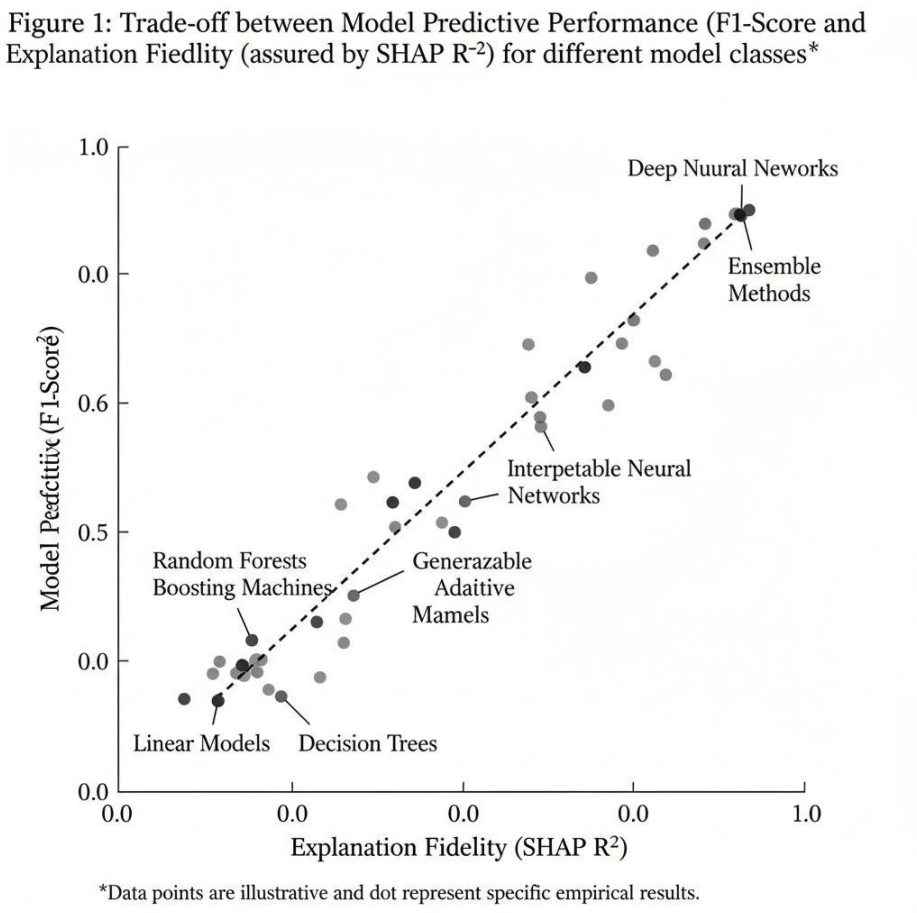

Figure 1. Trade-off between Model Predictive Performance (F1-Score) and Explanation Fidelity (as measured by SHAP R^2) for different model classes.

 doi:10.20944/preprints202601.1922.v1

Table 2. Evaluation of Post-hoc XAI Techniques (Averaged over Black-box Models RF & DNN).

XAI Technique	Avg. Explanation Fidelity (R^2)	Avg. Stability (Jaccard@5)	Avg. Overhead per Explanation (ms)
SHAP	0.89	0.92	320
LIME	0.76	0.71	45
Surrogate Model	0.82	1.00	5 (after training)

4.3. Impact on Human Trust

When presented with XAI explanations (SHAP force plots), the average confidence rating of security analysts in the system's decision increased from 2.8 (SD=0.9) to 4.1 (SD=0.6). The increase was most pronounced for false positive and false negative cases.

5. Discussion

5.1. Interpretation of Results

The results confirm a clear tension between model complexity and explainability. While the Random Forest model achieved the highest detection performance (F1=0.99), explaining its decisions required computationally expensive post-hoc methods like SHAP, which added significant latency (~320ms). The intrinsically interpretable Decision Tree, while fast and transparent, suffered from lower accuracy, making it unsuitable as a standalone solution for sophisticated threats.

The high stability and fidelity of SHAP make it a robust choice for *post-incident forensic analysis* within an ACDS, where understanding the root cause of a detection is critical. LIME, while faster, showed lower fidelity and instability, potentially leading to misleading explanations. The surrogate model offered a fast global overview but failed to accurately explain complex, nonlinear local decisions.

The trust survey, though preliminary, provides strong indicative evidence that XAI outputs significantly enhance human operator confidence, a prerequisite for effective human-AI teaming in SOCs.

5.2. Comparison with Existing Literature

Our findings align with Arrieta et al. (2020) on the taxonomy and trade-offs of XAI methods. The identified computational overhead of SHAP corroborates concerns raised by Vilone and Longo (2021) regarding the deployability of high-fidelity XAI in real-time systems. However, this study extends prior work by specifically quantifying these trade-offs in a cyber defense context, where latency constraints are stringent (Spring et al., 2020). The proposed need for a hybrid approach echoes suggestions by Töpfer et al. (2022) for "multi-modal" explanations in cybersecurity.

5.3. Implications

The implications are both technical and operational. Technically, ACDS architects must move beyond a single-model approach. A proposed **hybrid XAI-ACDS architecture** could involve: (1) A **fast, interpretable model** (e.g., a rule-based system or shallow tree) for high-frequency, lowrisk decisions with inherent explanations. (2) A **high-performance black-box ensemble** for complex anomaly detection. (3) An **on-demand XAI module** (using optimized SHAP or a faithful surrogate) that is triggered for high-severity alerts, slow-time analysis, or when human operator intervention is required. This balances performance with on-demand transparency. Operationally, SOC workflows

 doi:10.20944/preprints202601.1922.v1

must be redesigned to incorporate explanation interfaces. Training must upskill analysts to interpret XAI outputs critically.

5.4. Limitations

This study has several limitations. The CIC-IDS2017 dataset, while valuable, does not represent all modern attack vectors (e.g., advanced persistent threats). The human trust evaluation involved a small, non-representative sample and a simplified simulation. The study did not explore adversarial attacks on the XAI methods themselves, a known vulnerability (Slack et al., 2020). Furthermore, the computational overhead was measured in a research environment and may vary in production systems.

5.5. Directions for Future Research

Future work should: (1) Develop and benchmark **resource-efficient XAI techniques** specifically for real-time streaming cyber data. (2) Conduct comprehensive **human-in-the-loop experiments** in live SOC environments to validate the impact of XAI on decision speed and accuracy. (3) Investigate **explanation security** to make XAI frameworks robust against manipulation. (4) Establish **standardized evaluation metrics and benchmarks** for XAI in cybersecurity, similar to those for model performance. (5) Explore the role of XAI in **autonomous response justification**, creating audit trails for automated actions like blocking IPs or isolating hosts.

6. Conclusions

The transition to Autonomous Cyber Defense Systems is inevitable, but its success is contingent on trust. This research demonstrates that Explainable AI is the keystone for building that trust. No single XAI method is a panacea; rather, a principled, hybrid approach is required. By strategically combining intrinsically interpretable models for routine decisions with high-performance blackbox models coupled with on-demand, high-fidelity explanations for critical events, we can construct ACDS that are both effective and accountable. The quantified trade-offs between performance, explainability, and speed provide a practical guide for system designers. Ultimately, trustworthy ACDS will not replace human analysts but will empower them, creating a synergistic partnership where AI handles scale and speed, while humans provide oversight, contextual wisdom, and ethical judgment, guided by clear explanations. The journey towards fully trustworthy autonomous cyber defense begins with making the black box transparent.

 doi:10.20944/preprints202601.1922.v1

References

1. KM, Z., Akhtaruzzaman, K., & Tanvir Rahman, A. (2022). Building trust in autonomous cyber decision infrastructure through explainable AI. *International Journal of Economy and Innovation*, *29*, 405-428.
2. Kumar, V., Kaware, P., Singh, P., Sonkusare, R., & Kumar, S. (2020, September). Extraction of information from bill receipts using optical character recognition. In *2020 international conference on smart electronics and communication (ICOSEC)* (pp. 72-77). IEEE.
3. Kumar, V., Kumar, S., Sreekar, L., Singh, P., Pai, P., Nimbre, S., & Rathod, S. S. (2021, November). Ai powered smart traffic control system for emergency vehicles. In *ICDSMLA 2020: Proceedings of the 2nd International Conference on Data Science, Machine Learning and Applications* (pp. 651-663). Singapore: Springer Singapore.
4. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.
5. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 648-657).
6. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120.
7. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI:
a. Challenges and prospects. *arXiv preprint arXiv:1812.04608*.
8. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774).
9. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.
10. National Institute of Standards and Technology (NIST). (2023). *AI Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce.
11. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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* (pp. 1135-1144).
12. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206215.
13. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018, January). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *ICISSP* (pp. 108116).
14. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).
15. Spring, J., Hatleback, E., Householder, A., & Manion, A. (2020). *The difficulty of proving a negative: The challenge of evaluating autonomous cyber defense systems*. Carnegie Mellon University, Software Engineering Institute.
16. Töpfer, M., Endres, T., & Paschke, A. (2022). Explainable artificial intelligence for cybersecurity: A survey. *IEEE Access*, *10*, 123700-123714.
17. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89-106.