Feature-Level Interpretability of a Hybrid Ensemble for Multi-Category Cyber Threat Classification: A SHAP-Based Analysis of the ADCTI-AR Framework

Dr. Sesha Bhargavi Velagaleti[1], Postdoctoral Researcher[1], Lincoln University College[1],
pdf.seshabhargavi@lincoln.edu.my[1]
Dr. Upendra Kumar[2], Institute of Engineering and Technology[2], Lucknow, India[2]
Adjunct Research Faculty, Lincoln University College, Malaysia

ABSTRACT
High detection accuracy alone does not make a cyber-threat intelligence (CTI) system operationally trustworthy: Security Operations Center (SOC) analysts must also understand why a given event was classified as malicious before acting on automated recommendations. Stage II of the ADCTI-AR (AI-Driven Cyber Threat Intelligence and Adaptive Response) research programme reported a hybrid ensemble—combining Deep Neural Networks, Random Forests, Long Short-Term Memory networks, and autoencoder-based anomaly detection—that achieved 99.31% detection accuracy and 0.43% false positive rate across a 92-dimensional, three-category feature space (statistical, behavioral, and contextual). That work included a single, brief SHAP-based observation that behavioral features dominate Advanced Persistent Threat (APT) classification while statistical volume features dominate volumetric Denial-of-Service (DoS) detection. This paper isolates, extends, and rigorously develops that observation into a dedicated feature-level interpretability analysis. We examine SHAP attributions and Random Forest impurity-based importance side by side across attack categories, analyze their convergence as a cross-validation of interpretability itself, and derive concrete implications for feature engineering priorities and SOC analyst triage workflows. This analysis constitutes a distinct methodological contribution—an interpretability study in its own right—rather than a restatement of the detection-accuracy results already reported in Stage II.

Keywords: Explainable AI, SHAP, Feature Importance, Random Forest, Interpretability, Cyber Threat Intelligence, ADCTI-AR, Intrusion Detection, Feature Engineering, APT Detection.

INTRODUCTION
The transition of AI/ML-based cyber threat intelligence from research prototypes to operational Security Operations Center (SOC) tools depends not only on raw detection performance but on the interpretability of that performance to the human analysts who must act on it [1]. Opaque classifications—however accurate—slow analyst triage and erode trust in automated recommendations [2]. SHapley Additive exPlanations (SHAP) [3] and tree-based impurity importance have become the two dominant approaches to post-hoc interpretability in ML-based intrusion detection [4], yet these two families of explanation are rarely examined together for the degree to which they agree or diverge on a shared model.

Stage II of the ADCTI-AR research programme developed a hybrid ensemble—Deep Neural Network (DNN), Random Forest (RF), Long Short-Term Memory (LSTM) network, and autoencoder—trained on a 92-dimensional feature space spanning statistical, behavioral, and contextual feature families, achieving 99.31% accuracy and 0.43% false positive rate on a curated dataset of 3.85 million labeled security events. That work reported, in a single passage, that SHAP attributions identified behavioral features (notably destination port entropy) as dominant for APT-category classification, while statistical volume features dominated volumetric DoS detection. This finding was noted but not developed: no prior stage of this research programme has systematically compared SHAP and RF-native feature importance across attack categories, nor examined what this convergence or divergence implies for feature engineering and analyst workflow design.

This paper addresses that gap directly. The contributions are: (i) a structured, category-by-category comparison of SHAP and Random Forest feature importance across the three engineered feature families; (ii) an analysis of the degree of agreement between these two independent interpretability

methods, used as a form of cross-validation of the underlying explanations; (iii) derived, data-grounded recommendations for feature engineering prioritization and SOC analyst alert-triage design; and (iv) an explicit account of this paper's methodological boundary—an analysis of existing model behavior rather than a new human-subjects evaluation, with the latter identified as a distinct direction for future work.

BACKGROUND AND RELATED WORK

SHAP-based explanation has become the standard approach for post-hoc interpretability of deep learning-based intrusion detection systems, owing to its grounding in cooperative game theory and its consistency guarantees [3]. Tree ensemble models such as Random Forest offer a complementary, model-native form of interpretability through mean decrease in impurity or permutation importance [5]. Prior literature on ML-based intrusion detection has applied SHAP or RF importance individually to explain classifier behavior [6], [7], but comparative studies examining whether two independent interpretability methods converge on the same underlying feature signal—as opposed to reporting divergent or method-specific artifacts—remain limited in the CTI literature specifically. This gap motivates the comparative design adopted here: where SHAP and RF importance agree, the resulting explanation is less likely to be a methodological artifact of either technique and more likely to reflect genuine discriminative structure in the data.

This work builds directly on the feature taxonomy, dataset, and model architecture reported in Stage II of this research programme [8], and the single-paragraph SHAP observation in that work is the starting point—not the endpoint—of the analysis developed here.

DATA AND MODEL RECAP

The underlying dataset and models are as reported in Stage II [8] and are summarized here for completeness. The curated dataset integrates NSL-KDD, CICIDS-2017, UNSW-NB15, and MITRE ATT&CK evaluation data, comprising 3,857,422 labeled instances across 17 threat categories plus a normal traffic class, with Conditional GAN-based augmentation addressing zero-day class imbalance. The hybrid ensemble combines DNN, Random Forest (500 trees, max depth 30), and LSTM (three bidirectional layers, 128 units) posterior probabilities with autoencoder reconstruction error via an XGBoost meta-learner, achieving 99.31% accuracy, 98.89% precision, 99.14% recall, 0.43% false positive rate, and AUC-ROC of 0.9984 on benchmark evaluation.

FEATURE TAXONOMY AND SELECTION

Ninety-two candidate features were organized into three families. Statistical Network Features (38 features) capture quantitative flow and packet characteristics, including inter-arrival time statistics, volume distributions, packet ratios, flow duration, and protocol header entropy. Behavioral Features (31 features) encode temporal activity structure: per-source-IP connection frequency over sliding windows, protocol mixture indices, unique destination port counts, and inter-event burstiness coefficients. Contextual Features (23 features) incorporate domain knowledge such as asset criticality scores, IP reputation scores, temporal encodings, and CVE-associated CVSS scores.

Feature selection proceeded in two stages: Mutual Information (MI) scoring eliminated 18 features falling below a threshold of $\tau = 0.01$ nats, and Recursive Feature Elimination with Cross-Validation (RFECV), applied to the remaining 74 features plus 36 derived interaction features, yielded the final 92-feature matrix accounting for 99.7% of classification variance on validation data. This two-stage selection process is itself a form of interpretability signal, since features surviving both statistical filtering and model-driven elimination are, by construction, those with the most robust discriminative contribution.

COMPARATIVE FEATURE ATTRIBUTION ANALYSIS

Two independent interpretability signals were compared across attack categories: SHAP attributions computed on the DNN component of the ensemble, and native impurity-based feature importance from the Random Forest component. For APT-category classification, both methods concur that

behavioral features dominate: destination port entropy and connection burst frequency register as the highest-importance features under both SHAP and RF impurity scoring. This convergence is consistent with the underlying threat semantics—APT campaigns are characterized by low-and-slow reconnaissance and lateral movement, patterns that manifest as temporal and behavioral irregularities rather than volumetric anomalies, and both interpretability methods independently recover this structure from the trained models.

For volumetric DoS detection, both methods instead identify statistical network features—packet volume distributions and protocol header entropy—as dominant, again consistent with the threat semantics of DoS attacks as high-volume, low-behavioral-complexity events. The agreement between a game-theoretic attribution method (SHAP) and a structurally distinct, model-native importance measure (RF impurity) on both attack categories strengthens confidence that these findings reflect genuine discriminative structure rather than an artifact specific to either interpretability technique.

Contextual features, by contrast, show comparatively lower and more evenly distributed importance scores across both methods for either attack category, suggesting their primary operational value lies less in driving the initial classification decision and more in downstream risk-prioritization—consistent with their design purpose in the ADCTI-AR pipeline as inputs to the Threat Intelligence Generation module's risk-based ranking rather than to the classification layer itself.

Attack Category	Dominant Feature Family (SHAP)	Dominant Feature Family (RF Importance)
APT / lateral movement	Behavioral (destination port entropy, connection burst frequency)	Behavioral (same top features)
Volumetric DoS	Statistical (packet volume, protocol header entropy)	Statistical (same top features)
Risk prioritization (downstream)	Contextual (diffuse, lower magnitude)	Contextual (diffuse, lower magnitude)

TABLE I. CONVERGENCE OF SHAP AND RANDOM FOREST FEATURE ATTRIBUTIONS ACROSS ATTACK CATEGORIES

DISCUSSION

Three implications follow from this analysis. First, the cross-method convergence observed here provides a stronger interpretability guarantee than either method alone: an explanation that survives scrutiny under two structurally independent attribution techniques is more defensible to SOC analysts and auditors than a single-method explanation, and this convergence-based validation approach could be adopted as a standard interpretability check for future CTI model deployments. Second, the category-specific dominance patterns—behavioral features for APT, statistical features for DoS—offer a concrete, data-grounded prioritization for future feature engineering effort: incremental engineering investment is likely to yield the greatest return by refining behavioral feature granularity for APT-relevant slow-attack detection, an area the Stage II feature set addresses with only 31 of 92 total features. Third, the comparatively diffuse contextual feature importance suggests that these features are best understood, and best presented to analysts, as risk-prioritization inputs rather than classification drivers—an architectural distinction with direct implications for how the TABAP platform's alert dashboard should surface explanation information to SOC staff.

These findings should be read as an analysis of existing trained-model behavior on the Stage II dataset, not as new evidence about analyst decision-making in practice. The Stage III observation that SHAP-based explanation improved analyst decision-making by 34% during deployment remains an operationally important but separately-evaluated claim; a dedicated, controlled human-in-the-loop study—of the kind outlined but not completed here—is identified as the natural next stage and is left as explicit future work rather than folded into this paper's claims.

LIMITATIONS

This analysis is conducted on the same benchmark-derived dataset used in Stage II and does not include independent validation on live production traffic beyond the deployment reported in Stage III. SHAP and RF importance, while structurally independent, both operate on the same trained model instances; convergence between them is evidence against method-specific artifacts but is not equivalent to ground-truth validation of feature relevance from domain experts or from a controlled analyst study. Reproducing this comparison across additional model retrainings and additional benchmark datasets would further strengthen the robustness claims made here.

CONCLUSION

By examining SHAP and Random Forest importance side by side across attack categories, it is observed that both methods converge on behavioral features as dominant for APT detection and statistical features as dominant for volumetric DoS detection, and that contextual features play a comparatively diffuse, risk-prioritization role rather than a classification-driving one. These findings yield concrete guidance for feature engineering prioritization and for the design of analyst-facing explanation interfaces, while explicitly identifying a controlled human-in-the-loop evaluation of analyst decision-making as a distinct and necessary next contribution to this research.

REFERENCES

[1] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in Proc. IEEE Symp. Security and Privacy, Oakland, CA, 2010, pp. 305–316.
[2] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "On the Effectiveness of Machine and Deep Learning for Cyber Security," in Proc. 10th Int. Conf. Cyber Conflict (CyCon), Tallinn, Estonia, 2018, pp. 371–390.
[3] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.
[4] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," J. Inf. Security and Applications, vol. 50, p. 102419, 2020.
[5] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[6] T. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection," IEEE Trans. Emerging Topics Comput. Intell., vol. 2, no. 1, pp. 41–50, Feb. 2018.
[7] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of Intrusion Detection Systems: Techniques, Datasets, and Challenges," Cybersecurity, vol. 2, no. 1, pp. 1–22, 2019.
[8] S. B. Velagaleti, "Data Engineering and Feature Design for AI/ML-Driven Cyber Threat Intelligence," GNITS Research Programme, Hyderabad, India, 2025.