Learning to Explain Cyberattacks: Insights from Random Forest and Decision Predicate Graphs

Eron Ponce Pereira[1], Azin Moradbeikie[2], Bruno Bogaz Zarpelão[1] and Sylvio Barbon Junior[2,*]

[1]Computer Science Department, State University of Londrina, Londrina-PR, Brazil
[2]Department of Engineering and Architecture, University of Trieste, Trieste, Italy

Abstract

Intrusion detection systems (IDS) must combine high accuracy with transparent decision-making to support network security operations. While post-hoc explanation techniques like SHAP and LIME highlight feature importance, they often suffer from biases in imbalanced datasets, especially failing to explain rare attack classes. To address this, we introduce a novel interpretability framework based on Decision Predicate Graphs (DPGs), which enhance the transparency of Random Forest-based IDS models. Each network flow is transformed into a DPG, where nodes represent decision predicates and edges encode logical and data dependencies, enabling a structured view of the model's reasoning. Experiments on the CIC-IoT-2023 benchmark dataset show that DPGs reveal class-specific decision boundaries derived from the Random Forest model, which are both interpretable and empirically selective. Predicates with high structural centrality also align with semantically meaningful thresholds—such as low ACK counts and high packet rates—that separate DoS/DDoS from benign traffic.

Keywords

Machine Learning, Cybersecurity, Explainable AI, DDoS, DoS, Reconnaissance, Spoofing

1. Introduction

Traditional classifiers, such as Random Forest (RF), are widely adopted in the cybersecurity field due to their robustness and predictive accuracy, particularly in high-dimensional feature spaces. To be explored by Random Forest models, raw network flow records are transformed into a structured dataset composed of features such as total packet size, transmission rate, header length, inter-arrival time, and protocol indicators, which can be used for classification. However, these models are inherently opaque and provide limited interpretability regarding how specific decisions are made, especially in the context of evolving attack strategies. In other words, security administrators and analysts must understand *why* a particular flow or event was flagged as malicious or safe. This understanding is crucial not only for validating alerts (false positives and false negatives) but also for understanding classification boundaries, feature relevance, or behavioural patterns. For example, it is important to identify which specific patterns, such as unusually high packet rates or abnormal TTL values, led the model to classify certain network flows as DoS/DDoS or Reconnaissance attacks. A growing body of research in Explainable Artificial Intelligence (XAI) seeks to address this limitation [1]. Notably, post-hoc methods such as SHAP (SHapley Additive exPlanations) [2] and LIME (Local Interpretable Model-agnostic Explanations) [3] have gained traction across domains due to their model-agnostic nature and ability to assign local importance scores to input features.

Despite their popularity, these XAI methods are not immune to limitations, particularly in imbalanced data settings common to cybersecurity applications. SHAP, for instance, relies on background data distributions to compute marginal contributions of features. Similarly, LIME approximates model behaviour through locally linear surrogate models constructed from perturbed samples in the feature space.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
eron.ponce.pereira@uel.br (E. P. Pereira); azin.moradbeikie@dia.units.it (A. Moradbeikie); brunozarpelao@uel.br (B. B. Zarpelão); sylvio.barbonjunior@units.it (S. Barbon Junior)
0000-0001-9172-3578 (B. B. Zarpelão); 0000-0002-4988-0702 (S. Barbon Junior)

To address these shortcomings, we propose a novel pipeline that integrates structural information derived from *Decision Predicate Graphs* (DPGs) [4] into both the classification and explanation phases. In this approach, each instance is represented as a graph of logical predicates extracted from its feature-based decision logic. Topological metrics such as *Local Reaching Centrality* (LRC) and detected nodes communities highlight structurally influential decisions and allow to identify functional subgraphs that correlate with specific attack behaviors. This graph-enriched representation not only facilitates improved detection but also enables interpretable explanations that account for both global structural patterns and localized anomaly signals, including those associated with minority class instances.

DPG-based approach yields class-specific, human-readable rules that preserve the structure of the underlying decision logic. These constraints serve as global approximations of decision boundaries and are especially useful in (i) Validating alerts: network analysts can trace alerts back to feature bounds that match known behavioral signatures, (ii) Rule derivation: the intervals can be translated into high-precision detection rules or policy signatures. (iii) Minority class resolution: classes with few training instances (e.g., *Spoofing*) were still represented with stable bounds, highlighting the robustness of the method to class imbalance.

2. Related Work

In IoT networks, Houda et al. [5] propose a deep learning approach that embeds domain-specific constraints to preserve interpretability even under resource limitations. While deep learning models have shown impressive performance in intrusion detection, graph-based representations have been seen as innovative tools for modeling complex relationships in network traffic. Pujol-Perich et al. [6] demonstrated how graph-based features can capture the structural patterns of network flows, improving detection performance compared to traditional feature vectors, though their focus remained on accuracy rather than interpretability. Arrighi et al. [4] further explored the role of community detection in revealing latent decision structures within ensemble models, enabling insights into class-specific feature interactions and highlighting complex classification regions beyond traditional feature-based analysis.

The gap between research advancements and the operational adoption of AI-based security systems has been consistently reported. Dietz et al. [7] argue for interpretable models that can earn the trust of security professionals, a conclusion reinforced by Mink et al. [8], who find that explainability is a key factor in the acceptance of machine learning tools in practical environments. Nascita et al. [9] provide a systematic overview of explainable AI methods for network intrusion detection, including SHAP and LIME. They note that popular post-hoc techniques like SHAP and LIME, while accessible, may not adequately capture the internal logic or structural dependencies within the model, particularly critical in network contexts where behavior is shaped by temporal and protocol-based correlations. Arrighi et al. [4] recently introduced the DPG as a novel, model-agnostic graph structure for interpreting tree ensemble models. The authors emphasize that DPG leverages graph concepts (e.g. centrality, community structure) to provide additional quantitative insights, complement visualizations, and expand the description of the model's decision space.

In the following, other studies have emphasized the importance of structural transparency in security-focused models besides accuracy. Houda et al. [5] proposed a hybrid framework that combines decision trees with deep learning to enhance both accuracy and explainability, emphasizing the role of structural representations in model transparency using *RuleFit*. In Liu et al. [10], the authors further highlight that tree-based models strike a favorable balance between performance and interpretability, recommending that explanation strategies be tailored to the domain-specific requirements of security practitioners. Complementing this, Kumar and Thing [11] stress the need for diverse explanation techniques, as each may reveal different facets of the model's decision-making process. Building on these insights, Patel et al. [12] advocate for concept-based explanations that align with practitioner mental models.

3. Material and Method

3.1. Dataset

We used the *Merged01.csv* file from the CIC-IoT-2023 dataset [13], a large-scale benchmark containing benign traffic and multiple attack types. To mitigate class imbalance, DoS/DDoS variants were limited to 2000 instances each, while other classes remained unchanged. Following Neto et al. [13], we grouped the original labels into four new categories: Benign, DoS/DDoS (including Mirai variants), Reconnaissance (e.g., scans, pings), and Spoofing (e.g., ARP/DNS spoofing). Web-based and brute-force attacks were excluded due to low representation and lack of distinct flow-level patterns. We adopted the same set of statistical and protocol-specific features used by Neto et al. [13] to encode each flow, and 20% of the data was reserved for testing.

3.2. Cyberattack Detection Model

As the classification engine, we trained a RF intrusion detection model using scikit-learn's RandomForestClassifier. The RF model was optimized using a Genetic Algorithm (GA) to maximize F1-score performance by tuning four key hyperparameters based on validation accuracy [14]: number of trees (`n_estimators=20`), tree depth (`max_depth=16`), minimum samples required to split a node (`min_samples_split=2`), and minimum samples required at a leaf node (`min_samples_leaf=2`). DPGs were constructed using the official implementation[1] and configured with `max_features='sqrt'`, `criterion='gini'`, `bootstrap=True`, `class_weight='balanced'`, and `oob_score=False`.

The RF best performance was obtained with 20 decision trees, each limited to a maximum depth of 16 to prevent overfitting. The minimum samples to split a node was set to 2, and minimum samples per leaf to 2. All models were trained on the training split and evaluated on the held-out test split, following an 80% training and 20% testing split. After training, the RF achieved approximately 90.91% accuracy on the test set, indicating that the ensemble of trees was able to capture the distinguishing characteristics of IoT attacks and normal traffic. Key features for the model included statistics such as number of packets in the window (Number), longest packet length in the window (Max), total sum of packet lengths in the window (Tot_sum), ack flag occurrences in the window (ack_count), and shortest packet length in the window (Min). Feature importance analysis highlighted these attributes as the most influential for distinguishing among the traffic categories.

3.3. Decision Predicate Graph

We employed the DPG [4] as a post-hoc explainability tool to capture the structural logic of a RF classifier. The resulting graph structure (i.e., DPG) allows us to identify class-specific constraints (*class bounds*) and measure each predicate's influence using LRC. High-LRC predicates tend to dominate early or frequent decision paths, supporting global interpretability. These bounds offer human-readable summaries of each class and enable controlled semantic manipulation of input flows.

With the DPG constructed from the optimized RF model, we computed the LRC for each node to identify structurally influential predicates. High-LRC nodes acted as key decision points, influencing numerous paths across the decision trees. In contrast, predicates with lower LRC values played more specialized roles deeper in the decision structure. This use of centrality metrics to identify influential nodes builds on the work of Arrighi et al. [4].

In addition to centrality analysis, we applied community detection to the DPG to uncover groups of predicates that frequently co-occur within subgraphs of the decision forest. This graph-based perspective enables the identification of semantically related decision patterns and supports modular interpretation of the model.

[1] https://github.com/LeonardoArrighi/DPG

4. Results and Discussion

RF model, optimized through a genetic algorithm, demonstrated strong classification performance across four traffic classes: Benign, DoS/DDoS, Reconnaissance, and Spoofing. The model achieved an overall accuracy of 90.91%, with a macro-averaged F1-score of 0.87, indicating balanced performance across all classes, including minority attack types. Per-class F1-scores were particularly high for DoS/DDoS (0.9998) and Spoofing (0.8800), while Reconnaissance (0.7638) and Benign (0.8366) also showed consistent results. The balanced class weighting proved effective in compensating for the skewed distribution of samples, particularly where DoS attacks were overrepresented. These results validate the suitability of the RF not only as a high-performing classifier but also as a reliable surrogate model for subsequent interpretability through Decision Predicate Graph analysis.

To validate the relationship between the structural importance of nodes in the DPG (measured by LRC) and their importance in the original RF model, we conducted a correlation analysis between LRC rankings and RF feature importance rankings. The results revealed strong correlations across multiple statistical measures: Spearman ($\rho = 0.89, p < 0.001$), Kendall Tau ($\tau = 0.71, p < 0.001$), and Pearson ($r = 0.90, p < 0.001$). These statistically significant correlations confirm that the DPG effectively preserves the decision structure of the original RF model, with structurally central nodes in the graph corresponding to features that the model relies on most heavily for classification decisions.

We derived interpretable class bounds for key features across four classes: *Benign*, *Spoofing*, *DoS/DDoS*, and *Reconnaissance*. These bounds act as global approximations of the classification logic and allow for clear behavioral differentiation between traffic types.

Table 1 presents the top-ranked constraints based on their LRC. Higher LRC values indicate predicates that reside in topologically central regions of the DPG and exert significant influence over downstream decision logic. The highest-ranked constraint, `ack_count` $\leq$ `1.5`, with an LRC of 2.7062, suggests that this predicate is a highly influential structural component. This threshold is aligned with detection of suspicious traffic flows, such as incomplete TCP handshakes or scanning activity, where the number of acknowledgment packets is unusually low.

Table 1
Top Constraints Ranked by Local Reaching Centrality

LRC Value	Constraint	LRC Value	Constraint
2.7062	ack_count $\leq$ 1.5	2.0045	ack_flag_number $\leq$ 0.8
2.4273	UDP $\leq$ 0.58	1.9963	Header_Length $\leq$ 21.18
2.3112	Variance $\leq$ 398984.47	1.9891	Header_Length $\leq$ 20.38
2.2248	Time_To_Live $>$ 61.37	1.8671	Rate $\leq$ 154145.7
2.1315	Rate $>$ 1476.74	1.8485	Tot sum $>$ 5948.5
2.1249	Header_Length $\leq$ 21.58	1.8390	Time_To_Live $\leq$ 72.38
2.0795	Tot sum $>$ 5999.0	1.8212	Std $\leq$ 145.5
2.0462	Time_To_Live $>$ 57.92	1.6943	Time_To_Live $\leq$ 89.91
2.0076	Tot size $\leq$ 957.13	1.6686	Rate $>$ 1482.72

Notably, certain features appear multiple times with slightly different thresholds (e.g., `Rate` $>$ `1476.74` and `Rate` $\leq$ `154145.7`), underscoring the importance of capturing feature ranges rather than binary decisions. Similarly, the repeated presence of `Time_To_Live` across several entries (both upper and lower bounds) reflects its non-linear, class-dependent behavior in the model.

A brief coverage analysis of the DPG predicates reveals very strong class-specific selectivity. For instance, the predicate ack_count $\leq$ 1.5 is triggered by 74% of DoS/DDoS samples but only about 1–2% of benign samples, reflecting that benign IoT traffic in the monitored network is largely TCP-based (with multiple ACKs per connection). This example highlights an interpretable decision boundary: in the monitored network, flows with essentially no ACK packets are likely malicious. Similar predicate thresholds (e.g. low SYN or high RST counts) exhibit analogous skewed coverage by class. These simple rules not only clarify model behavior (e.g. "few ACKs implies DoS-like traffic") but can be directly

translated into alert conditions or understandable limits. At the same time, the uneven coverage points out potential dataset biases. For example, if attack flows were collected with larger windows than benign flows, predicates on packet counts or durations could reflect that artifact. In practice, such coverage statistics can validate alerts (by confirming high precision of certain rules), set human-readable thresholds, and flag training biases (e.g. identifying predicates that exploit uneven window lengths rather than intrinsic malicious patterns).

Several other top-ranked constraints are related to protocol and transport layer features, such as $\text{UDP} \leq 0.58$, $\text{ack\_flag\_number} \leq 0.8$, $\text{Header\_Length} \leq 21.58$, and $\text{Min} \leq 641.0$, which capture deviations in protocol usage or header structuring—often indicative of spoofed or malformed traffic. Constraints reflecting traffic volume and packet length, including $\text{Tot sum} > 5999.0$, $\text{Tot size} \leq 957.13$, $\text{Variance} \leq 398984.47$, $\text{Rate} > 1476.74$, and $\text{Rate} \leq 154145.7$, also appear frequently and suggest that these characteristics—whether excessive or unusually limited—serve as effective discriminators for anomalous flows. Lastly, statistical distribution metrics such as $\text{Std} \leq 145.5$ and $\text{Variance} \leq 398984.47$, which represent the standard deviation and variance of packet length within a window, emphasize the importance of intra-flow variability in shaping the model's decision logic.

Applying community detection techniques to the DPGs revealed meaningful structural groupings of predicates that often correspond to specific traffic behaviors. The largest community, Community 1, contained 331 predicates and was dominated by high-volume and rate-based thresholds, with 22 predicates related to rate like `Rate > 3475.54`, which are indicative of DoS/DDoS behavior. Community 3, with 103 predicates, featured more nuanced indicators like `psh_flag_number <= 0.16` and `ack_flag_number <= 0.95`, commonly associated with reconnaissance or benign traffic. The features `ack_flag_number` and `psh_flag_number` represent the ratio of packets with the ACK and PSH flags, respectively, set to true in the packet window. For instance, `ack_flag_number = 0.95` means that 95% of the packets in the window have the ACK flag set to true. Community 2, comprising 24 predicates, identifies conditions involving `SSH` and `Tot size`. A detailed analysis revealed that non-zero SSH values occur in approximately 2.5% of benign samples and also in reconnaissance-labeled windows. Manual inspection of PCAP files corresponding to benign traffic showed that some of the SSH flows originate from a background IoT device communicating with external servers. Although this community predominantly reflects benign and reconnaissance patterns, it can also include certain DDoS traffic (approximately 0.41%), since SSH was explicitly documented as employed in attack scenarios within the original dataset description. Community 4, although smaller, exhibits characteristic like the *Rate>161.27* above 75th percentile of the whole data set and constraints such as `AVG <= 62.65`, `Number <= 12.5`, and `Min <= 63.0` suggest that the flows are short, composed of small packets compared to the overall dataset.

We observed that DoS/DDoS attacks are typically characterized by predicates that impose high thresholds for volume metrics, such as `Tot_sum > 5999` and `Rate > 1655.11`. These predicates tend to exhibit high LRC values, indicating their significant structural influence in the decision graph.

Class scoring charts for each community can visually confirm the inferred dominant class and reveal possible overlaps or ambiguities between classes. This visual approach to interpretation addresses the needs identified by Mink et al. [8], who found that security practitioners value explanations that align with their mental models and domain knowledge.

5. Conclusion

This study demonstrated the effectiveness of the Decision Predicate Graph (DPG) as an interpretable framework for analyzing flow-based network traffic classification. By extracting class-specific bounds, we revealed feature intervals that distinguish between Benign, DoS/DDoS, Spoofing, and Reconnaissance traffic. The analysis showed that DoS/DDoS is the most structurally separable class, with consistently distinct bounds across volume- and rate-related features. In contrast, Spoofing and Reconnaissance often overlapped with Benign traffic, indicating greater ambiguity. Furthermore, LRC identified the most structurally influential constraints—such as ack_count, Rate, and Header_Length—highlighting their

central role in the model's decision logic. Together, these results validate the use of DPG to enhance transparency in classification models and support the extraction of robust, class-specific behavioral signatures.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase, and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. URL: http://dx.doi.org/10.1016/j.inffus.2024.102301. doi:10.1016/j.inffus.2024.102301.

[2] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, 2016. URL: https://arxiv.org/abs/1602.04938. arXiv:1602.04938.

[4] L. Arrighi, L. Pennella, G. M. Tavares, S. B. Junior, Decision predicate graphs: Enhancing interpretability in tree ensembles, 2024. URL: https://arxiv.org/abs/2404.02942. arXiv:2404.02942.

[5] Z. A. E. Houda, B. Brik, L. Khoukhi, "why should i trust your ids?": An explainable deep learning framework for intrusion detection systems in internet of things networks, IEEE Open Journal of the Communications Society 3 (2022) 1164–1176. doi:10.1109/OJCOMS.2022.3188750.

[6] D. Pujol-Perich, J. Suárez-Varela, A. Cabellos-Aparicio, P. Barlet-Ros, Unveiling the potential of graph neural networks for robust intrusion detection, 2021. URL: https://arxiv.org/abs/2107.14756. arXiv:2107.14756.

[7] K. Dietz, M. Mühlhauser, J. Kögel, S. Schwinger, M. Sichermann, M. Seufert, D. Herrmann, T. Hoßfeld, The missing link in network intrusion detection: Taking ai/ml research efforts to users, IEEE Access 12 (2024) 79815–79837. doi:10.1109/ACCESS.2024.3406939.

[8] J. Mink, H. Benkraouda, L. Yang, A. Ciptadi, A. Ahmadzadeh, D. Votipka, G. Wang, Everybody's got ml, tell me what else you have: Practitioners' perception of ml-based security tools and explanations, in: 2023 IEEE Symposium on Security and Privacy (SP), 2023, pp. 2068–2085. doi:10.1109/SP46215.2023.10179321.

[9] A. Nascita, G. Aceto, D. Ciuonzo, A. Montieri, V. Persico, A. Pescapé, A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection, IEEE Communications Surveys Tutorials (2024) 1–1. doi:10.1109/COMST.2024.3504955.

[10] N. Liu, D. Shin, X. Chu, J. Wu, M. Xu, Explainable security: A systematic review of approaches, applications and challenges, ACM Computing Surveys 55 (2022) 1–34.

[11] A. Kumar, V. L. L. Thing, Evaluating the explainability of state-of-the-art deep learning-based network intrusion detection systems, 2025. URL: https://arxiv.org/abs/2408.14040. arXiv:2408.14040.

[12] S. Patel, D. Han, N. Narodystka, S. A. Jyothi, Toward trustworthy learning-enabled systems with concept-based explanations, in: Proceedings of the 23rd ACM Workshop on Hot Topics in Networks, 2024, pp. 60–67.

[13] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, A. A. Ghorbani, Ciciot2023: A real-

time dataset and benchmark for large-scale attacks in iot environment, Sensors 23 (2023). URL: https://www.mdpi.com/1424-8220/23/13/5941. doi:10.3390/s23135941.

[14] D. E. Goldberg, J. H. Holland, Genetic algorithms and machine learning, Machine Learning 3 (1988) 95–99. doi:10.1023/A:1022602019183.