A Framework for Network Traffic–Based DDoS Attack Detection and Explanation

Roberta Viola[1,2], **Michele Nogueira**[2], **Adriano Veloso**[1,2]

[1]Instituto Kunumi
Belo Horizonte, MG – Brazil

[2]Department of Computer Science – UFMG
Belo Horizonte, MG – Brazil

roberta@kunumi.com, michele@dcc.ufmg.br, adriano@kunumi.com

Abstract. *This paper presents TRACE-NET, an auditable and governed framework for network traffic-based DDoS attack detection and explanation. TRACE-NET combines a traffic classifier with feature attribution to generate instance-level explanations grounded in observable flow behavior. A deterministic jury score summarizes the epistemic support of each detection by jointly assessing model confidence and the structure of characteristic attributions, penalizing weak, ambiguous, or single-dominant evidence independently of ground truth. A Large Language Model (LLM) is used strictly as a post-hoc translation layer. Experiments on the CICDDoS2019 dataset show that TRACE-NET reduces explanation risk inflation from 88% to 32%,by anchoring explanations to auditable reliability signals, aligning operational security requirements with emerging regulatory demands for transparent, accountable, and risk-aware AI systems in network security.*

1. Introduction

Distributed Denial of Service (DDoS) attacks remain among the most disruptive threats to modern networks, overwhelming services with malicious traffic, and preventing legitimate access [Wang et al. 2025]. The number of DDoS attacks has increased 40% year-over-year, averaging 3,780 attacks per hour globally. These attacks increasingly target critical sectors. AI companies, for instance, experienced a 347% month-over-month increase in observed DDoS attack traffic volume in September 2025, while the mining and automotive industries have faced increased activity amid geopolitical tensions [Cloudflare 2025]. Beyond service outages, large-scale incidents can impact critical infrastructure, public services, and economic stability. According to the last IBM report [IBM Security and Ponemon Institute 2025], the global average total cost of a data breach reached USD 4.44 million, reflecting the substantial financial impact that disruptive cyber incidents can impose on organizations worldwide. As attack strategies continue to evolve in volume, coordination, and protocol diversity, detection systems are increasingly required not only to be accurate, but also to be trustworthy, interpretable, explainable and auditable in operational settings [Zhou et al. 2023].

From a networking perspective, DDoS attack detection is typically formulated as a supervised classification problem over packet or flow-level traffic representations [Abiramasundari and Ramaswamy 2025, Liu et al. 2023]. Features such as packet

rates, burstiness, protocol usage, temporal concentration, and connection asymmetries are extracted from packet traces or network flow records and used as inputs to machine learning or deep learning models [Hill et al. 2025]. While these models can achieve strong detection performance, their decisions are often opaque, limiting an analyst ability to assess the reliability of individual detections [Wei et al. 2023].

This opacity raises both practical and governance challenges [Batool et al. 2024]. When a machine learning model flags a network connection as malicious, security analysts often receive little more than a risk score or binary label. Without understanding which features drove the decision, whether it was an unusual port number, a suspicious domain pattern, or an anomalous sequence of API calls, analysts cannot effectively triage alerts, validate findings, or refine detection rules. This creates a bottleneck: high-confidence false positives consume investigation time, while subtle true positives may be dismissed due to lack of supporting evidence. For example, a Random Forest model might flag a login attempt based on a combination of timestamp, geolocation, and user-agent string, but if the model merely outputs "anomaly score: 0.87", the analyst has no basis to distinguish a legitimate employee traveling abroad from an attack.

Security analysts must determine whether alerts are well-supported or based on fragile evidence, while organizations face increasing regulatory and operational demands for transparency, accountability, and risk-aware automation. In Brazil, this pressure for transparency takes on specific characteristics. The General Data Protection Law (Lei Geral de Proteção de Dados, LGPD), in force since 2020, establishes the right to review automated decisions that affect the interests of data subjects. Although the LGPD does not explicitly mandate technical explainability in all cybersecurity contexts, Brazilian organizations face increasing scrutiny from the National Data Protection Authority (ANPD) and internal audits regarding how automated systems make decisions that may result in access blocking, account suspension, or information sharing with law enforcement authorities. In addition, regulated sectors, such as financial institutions under the supervision of the Central Bank of Brazil and operators of critical infrastructure, are required to demonstrate robust governance over their detection tools, including auditability and the ability to justify alerts in incidents.

In this context, explanation quality becomes a first-class concern, distinct from detection accuracy itself. By explanation, this work refers to interpretive artifacts that make machine learning detections understandable to humans. These include for instance feature attribution scores indicating influential inputs, rule-based approximations of model behavior, and counterfactual descriptions of minimal changes that would alter a detection. Importantly, explanation quality is not necessarily aligned with classification accuracy: highly accurate models may still produce inconsistent or misleading explanations, while less accurate models can yield more reliable and actionable justifications.

Recent works have explored feature attribution methods and large language models (LLMs) to generate human-readable explanations for intrusion detection decisions [Ziems et al. 2023, Wang et al. 2024, Ali and Kostakos 2023]. However, most existing approaches implicitly assume that accurate detections yield reliable explanations, conflating model decision-making with explanation generation. This assumption is flawed: correct detections feature or communicated through overconfident or misleading language [Agarwal et al. 2024, Ghafouri et al. 2025]. Conversely, explanations may ap-

pear plausible even when underlying evidence is fragile, resulting in *risk inflation*, where natural-language outputs exaggerate certainty or severity beyond what is warranted by the model internal evidence [Epstein et al. 2025].

Hence, this work introduces TRACE-NET, a model-agnostic audit framework instantiated in the context of DDoS detection but defined independently of any specific attack type or learning algorithm. The framework is specified in terms of information flow, allowing classifiers, attribution methods, and language models to be substituted without altering the core audit logic. While DDoS traffic motivates the study due to its operational relevance and well-defined flow-level features, the approach applies to any network security setting where probabilistic detections and feature attributions are available. At its core, TRACE-NET employs a deterministic jury, an heuristic that evaluates the structural support of an attack detection. The premise of the jury score is to decouple decision-making from justification. Rather than assessing whether an attack detection is correct, it evaluates whether the decision is epistemically well supported by the available network traffic evidence. In operational settings, operators must judge how much trust to place in a model's output before acting. The jury score formalizes this requirement by inspecting the internal support structure of a detection using only deterministic, model-derived signals such as confidence and attribution patterns.

TRACE-NET is evaluated using a public network traffic trace-based dataset under an audit setting. A supervised classifier is combined with feature attributions to generate instance-level explanations, which are then assessed by the deterministic jury. The evaluation analyzes jury score distributions, epistemic flags, and their relationship with model confidence, as well as the alignment between deterministic audit signals and LLM-generated explanations. The results show that the jury provides fine-grained reliability assessment beyond detection probability and significantly reduces explanation risk inflation when compared to an LLM-only baseline.

The paper is organized as follows. Section 2 reviews related work on DDoS detection and explainable network security. Section 3 describes the TRACE-NET architecture and the deterministic jury mechanism. Section 4 presents the experimental setup and evaluation methodology. Section 5 discusses the results and analysis. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Related Work

Research on DDoS detection spans classical machine learning, deep learning, and more recent LLM-based approaches. Early supervised learning methods showed that traditional ML models can effectively distinguish attack traffic from benign flows using public datasets [Abiramasundari and Ramaswamy 2025]. Subsequent work incorporated explainability mechanisms to improve transparency and trust in intrusion detection systems, with explainers such as SHAP gaining prominence [Wali and Khan 2021]. More recent studies investigated the use of large language models for network traffic analysis and DDoS attack detection, with LLMs employed directly as attack detectors or classifiers [Li et al. 2024]. Other work explored fine-tuning LLMs for DDoS attack detection, reporting high accuracy on public datasets while also generating natural-language explanations [Guastalla et al. 2024]. In addition, hybrid architectures that combine lightweight machine learning detectors with LLM-based reasoning have been proposed to produce

real-time, human-readable threat assessments [Jamshidi et al. 2025].

Several frameworks position LLMs as active components in attack detection or mitigation. There are examples using an LLM to interpret traffic patterns and recommend attack mitigation actions, effectively assigning the model a control and reasoning role [Ziems et al. 2023, Wang et al. 2024]. Similarly, other approaches employed an LLM as a central reasoning engine to correlate alerts, summarize logs, and assist analysts across heterogeneous security signals [Ali and Kostakos 2023]. While these approaches demonstrate the operational potential of LLMs, they tightly couple attack detection, reasoning, and explanation within the same component.

This work differs fundamentally in scope and design philosophy. The TRACE-NET framework deliberately restricts the LLM to a post-hoc interpretability and auditing role, while attack detection is performed exclusively by a conventional Random Forest classifier. This design leverages the lower computational cost, latency, and deployment complexity of classical traffic classifiers, which are better suited for operational network environments. SHAP provides model-grounded feature attributions, and the LLM is used only to translate deterministic signals into analyst-facing explanations and to audit the resulting narratives, without influencing detection outcomes. This separation enables the systematic identification of explanation-level failure modes that are often obscured in LLM-centric systems, such as overconfident language, unjustified causal claims, or omission of uncertainty. Recent work on AI accountability and algorithmic auditing highlights the need for independent evaluation of transparency, robustness, and auditability, particularly in regulated and safety-critical domains [BC 2024]. TRACE-NET aligns with these principles by introducing an explicit *LLM Translation Audit* that evaluates explanation quality independently of detection correctness, consistent with emerging frameworks for AI governance and auditability [Li and Goel 2025].

Although existing explainability approaches based on SHAP and LIME are widely used to provide post-hoc local and global interpretability [Gaspar et al. 2024], they do not explicitly audit the quality or reliability of the generated explanations. TRACE-NET addresses this gap by treating explanation quality as a measurable and auditable artifact, extending accountability beyond detection performance.

3. TRACE-NET: An Auditable Framework for Network Traffic–Based Attack Detection and Explanation

This section introduces TRACE-NET, a governed, modular framework, for auditable network traffic analysis. By governed, it is meant that explanations are constrained, auditable, and reproducible. The term *TRACE* reflects the framework emphasis on traceability: each detection, explanation, and associated risk can be systematically tracked, inspected, and audited across all stages of the decision process. TRACE-NET is designed to support machine learning–based detection while explicitly treating explanations as first-class artifacts subject to evaluation and governance. The framework operates on flow-level network features and remains agnostic to the specific learning algorithms.

As shown in Figure 1, TRACE-NET comprises four sequential stages. First, an attack detection model estimates attack likelihoods; any classifier able of producing confidence scores, such as Random Forest or XGBoost [Breiman 2001, Chen and Guestrin 2016], can be integrated without affecting the downstream logic. Sec-

ond, instance-level feature attributions are computed to explain each individual detection by decomposing the model output into feature-level contributions for a single network flow. Third, a deterministic jury evaluates epistemic risks, including weak evidence, feature dominance, and probability ambiguity. Finally, a constrained LLM generates concise, human-readable explanations grounded in the audited evidence.

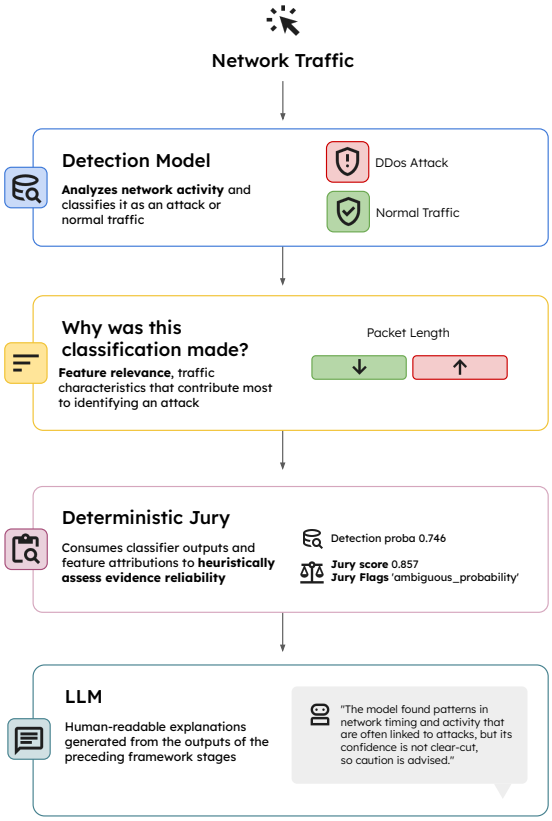

Figure 1. The framework emphasizes deterministic evidence interfaces

Overall, TRACE-NET provides a modular framework for governed and accountable network traffic analysis, in which explanations are treated as auditable artifacts rather than informal by-products of detection. By constraining human-facing explanations with deterministic evidence signals, the framework enables calibrated risk communication and independent assessment of explanation quality, even when detection accuracy cannot be verified. The next subsections describe each stage of the framework.

3.1. Stage 1: Attack Detection

The framework operates on flow-level network traffic observations, where each instance corresponds to an aggregated bidirectional flow defined over a fixed temporal window. The input feature set consists of statistical and behavioral descriptors commonly used in DDoS detection, including traffic volume metrics (e.g., packet and byte rates) and protocol-level distribution statistics. These features are selected due to their established relationship with volumetric and protocol-based DDoS behaviors, such as abnormal traffic concentration, amplification patterns, and sudden rate escalation.

A Random Forest–based probabilistic detection model is employed to estimate the likelihood of an ongoing DDoS attack. Instead of producing only a hard classification, the model outputs a probability estimate, together with the corresponding detection label, enabling downstream uncertainty-aware reasoning.

3.2. Stage 2: Instance-Level Evidence Extraction

Stage 2 receives both the predicted probability and the corresponding flow-level features from Stage 1 and applies a local feature attribution method based on SHAP (SHapley Additive exPlanations). SHAP is used to decompose each individual detection into per-feature contribution scores, quantifying how specific traffic characteristics support or oppose the attack hypothesis for that instance. This stage produces instance-level evidential representations that link the model's decision to concrete and observable traffic attributes, enabling explainability.

3.3. Stage 3: Deterministic Jury

In this stage, the framework evaluates the structural reliability of each detection by combining the model probability score with the extracted feature attributions. A deterministic jury mechanism applies a set of predefined heuristic rules designed to assess epistemic risks associated with the decision.

Specifically, penalties are assigned based on: (i) weak evidence, when the aggregate magnitude of supporting feature attributions is low; (ii) feature dominance, when the decision relies disproportionately on a single feature, indicating fragility; and (iii) probability ambiguity, when the predicted probability lies close to the decision threshold. These penalties are combined into a jury score that reflects the robustness of the detection under interpretable and uncertainty-aware criteria:

$$J = \mathrm{clip}(100 - P_{\text{weak}} - P_{\text{dom}} - P_{\text{amb}},\ 0,\ 100) \tag{1}$$

where J is the deterministic jury score obtained by subtracting penalties for weak evidence, feature dominance, and probability ambiguity from a base score of 100, clipped to the $[0, 100]$ range.

3.4. Stage 4: Governed Human-Readable Explanation

The final stage translates the deterministic outputs of the framework into analyst-facing explanations. The predicted label, probability score, feature attributions, jury score and epistemic risk flags are provided as structured input to a constrained large language model (GPT-4.1), which is restricted to explanation synthesis rather than inference. The LLM generates concise, human-readable explanations that faithfully reflect the underlying signals, suitable for operational analysis, auditing, and post-incident review.

4. Evaluation Methodology

This section details the evaluation of the framework in a concrete experimental setting. While the TRACE-NET is instantiated using public datasets and standard models, the evaluation scenario is intentionally generic. Any network traffic dataset with flow-level features and any probabilistic classifier could be used without modifying the framework. This section does not benchmark a specific detector, but it presents the evaluation of TRACE-NET operating end-to-end in a realistic network security context.

4.1. Data, Preprocessing and Training

In this evaluation, both the machine learning detector and the LLM-based auditing components use the public CICDDoS2019 dataset. It contains labeled network traffic flows categorized as *Attack* or *Benign*. The dataset provides flow-level statistics commonly used in network intrusion detection systems, such as packet sizes, durations, and protocol-related features. Because the original dataset contains approximately three times more attack samples than benign ones, benign traffic was intentionally over-sampled to better approximate real-world network traffic distributions, in which benign flows dominate. Oversampling was performed using SMOTE [Chawla et al. 2002], a synthetic over-sampling technique for addressing class imbalance.

Furthermore, the use of a single dataset is primarily driven by the limited availability of high-quality, publicly accessible DDoS datasets. Real-world network traffic containing verified DDoS attacks is rarely released due to privacy, security, and operational constraints, which makes reproducible research challenging. Consequently, publicly available datasets such as CICDDoS2019 have become widely adopted benchmarks in the literature. This is evidenced by prior works such as [Li et al. 2024, Abiramasundari and Ramaswamy 2025], enabling consistent and fair comparisons with existing approaches.

Random Forest (RF) is a well-established and widely adopted algorithm for DDoS attack detection [Abiramasundari and Ramaswamy 2025, Wali and Khan 2021, Choubisa et al. 2022, Markovic et al. 2022], offering strong performance while enabling interpretability. In this work, RF is deliberately chosen not as a novelty in detection, but as a stable and trustworthy backbone that allows us to focus on the downstream explanation, auditing, and transparency framework. The RF classifier is trained to detect whether a network flow corresponds to an attack, producing both class labels and probabilistic confidence scores. These outputs serve as inputs to the subsequent TRACE-NET stages.

The classification performance is evaluated using common metrics (precision, recall and F1-score), as summarized in Table 1. The detector exhibits high precision for attack traffic and perfect recall for benign flows, despite class imbalance. As near-perfect accuracy is common in DDoS attack detection on public datasets, the model capacity is deliberately constrained to avoid overfitting and to ensure that subsequent analysis focuses on explanation quality rather than raw detection performance.

Table 1. Classification performance on the CICDDoS2019 test set

Class	Precision	Recall	F1-score
Attack	1.00	0.97	0.98
Benign	0.91	1.00	0.95
Accuracy			0.97
Macro Avg	0.95	0.98	0.96
Weighted Avg	0.98	0.97	0.97

4.2. SHAP-Based Attribution of Network Traffic Features

SHAP (SHapley Additive exPlanations) [Lundberg and Lee 2017] interprets individual attack detections, quantifying how each network traffic feature influences the model out-

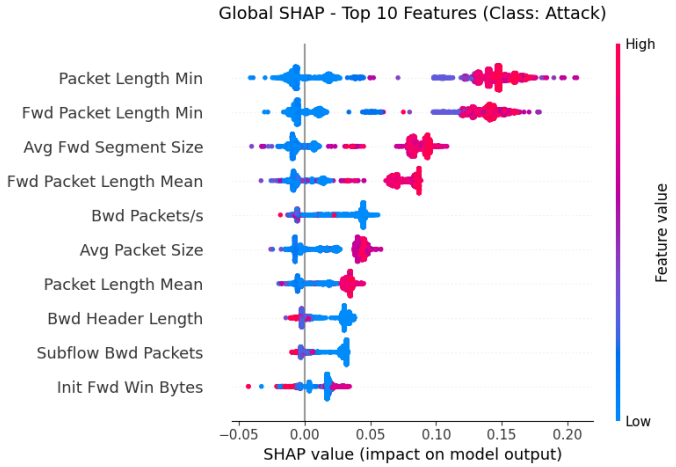

Figure 2. SHAP Top 10 Features - Attack Class

put. For a given network flow, SHAP estimates the contribution of each feature by comparing the detected attack probability under different combinations of observed and absent features. Resulting attributions form an additive decomposition of the detection, indicating if specific traffic characteristics increase or decrease the likelihood of an attack.

In order to explain model decisions, SHAP values were computed for the *attack* class, which represent feature-level contributions to the attack probability. The analysis is conducted at both the global level, to identify feature importance across the dataset, and the local level, to explain individual detections and support downstream auditing. For evaluating the proposed framework, SHAP is applied to two complementary subsets. A visualization subset of 200 randomly sampled flows is used to generate global feature importance, as shown in Figure 2. In parallel, an audit subset of 100 flows is stratified by preliminary jury scores (25 per quartile), ensuring balanced coverage of low, medium, and high-confidence detections for explanation auditing.

SHAP attributions are sensitive to factors such as feature collinearity and background distributions and are therefore not treated as causal or absolute measures of importance. TRACE-NET uses only coarse distributional properties of the attribution structure, such as total magnitude and dominance, as relative, model-internal proxies for evidential support. Jury penalties are calibrated using population-level statistics from the same model and dataset, and the audit logic remains agnostic to the specific attribution method.

4.3. Deterministic Jury

TRACE-NET introduces an deterministic auditing layer that evaluates machine learning decisions without access to ground truth, a common constraint in operational network environments. Rather than assessing detection correctness, this layer examines if a decision is supported by sufficient, consistent, and non-fragile evidence derived from the model and the observed traffic features. Since a language model cannot identify classification errors without explicit contradiction signals, TRACE-NET does not rely on an LLM to judge decision validity. Instead, a deterministic jury heuristically evaluates internal consistency indicators, including model confidence and SHAP-based feature attributions, to flag decisions that may be unreliable or high-risk. Crucially, the jury score does not determine whether the traffic is truly malicious; it assesses whether the detection is epis-

temically well-supported by network traffic evidence. This distinction enables risk-aware decision auditing even when ground truth labels are delayed or unavailable.

The jury assigns a continuous reliability score initialized at 100 and reduced by three bounded penalty terms reflecting distinct risks. Each penalty is a monotonic, proportional function calibrated using population-level statistics computed on the evaluation data. Weak evidence is penalized when the total magnitude of SHAP contributions falls below the median attribution strength observed in the population, with larger deviations incurring higher penalties. Single-feature dominance is penalized when the ratio between the largest absolute SHAP value and the total attribution magnitude exceeds the population median dominance ratio, reflecting over-reliance on a single traffic feature. Ambiguous probability is penalized when the detected probability lies within the inter-quartile range of detected probabilities, with penalties increasing toward the center of this region. Each penalty is bounded to prevent dominance by a single factor, and the final score is clipped to the range $[0, 100]$. In addition to the continuous score, coarse epistemic flags are raised when penalty terms exceed predefined ratios relative to their calibrated baselines. These flags serve as qualitative indicators of risk, while the numeric score captures fine-grained degradation in epistemic support. The resulting score and flags are immutable and fully determined by model outputs, ensuring reproducibility and auditability.

4.4. LLM Experimental Setup

All experiments involving large language models use a GPT 4.1. LLM is accessed in zero-shot mode using a structured prompt template that includes only deterministic, model-derived artifacts. Decoding parameters are fixed across all experiments, including a fixed maximum token budget. The same prompt template and decoding configuration are used for both the baseline LLM condition and the proposed governed audit framework, ensuring direct comparability. TRACE-NET employs the large language model as a controlled interpretation layer that operates on model-derived artifacts.

The LLM performs a single, constrained function: generating a human-readable explanation grounded in deterministic evidence. For each audited detection, it translates the most influential SHAP traffic features (renaming them into analyst-friendly terms) while contextualizing the detection using immutable inputs, including the predicted label, model confidence, jury score, and epistemic flags. Based on these signals, the LLM explains why the detection received its jury score, converts technical risk flags into intuitive categories, assigns an operational risk level (low, medium, or high), and produces a concise, non-mathematical explanation. The output is a structured JSON object containing *risk_level* and *human_explanation*. Crucially, the LLM does not inspect raw traffic, re-evaluate the classifier, or infer causality, it only interprets and communicates the deterministic jury assessment, ensuring that human-facing explanations remain aligned with the underlying signals.

5. Results

TRACE-NET is evaluated by examining the interaction between model confidence, SHAP-based evidence, deterministic jury scores, and LLM-generated explanations to characterize epistemic decision reliability. Figure 3 shows 50 representative attack detections using a SHAP decision plot. The trajectories indicate that attack predictions are

generally supported by multiple reinforcing traffic characteristics, rather than isolated features, reflecting structured evidence across samples. The observed variability across paths highlights heterogeneity in attack traffic and motivates instance-level evidence assessment beyond aggregate confidence scores.

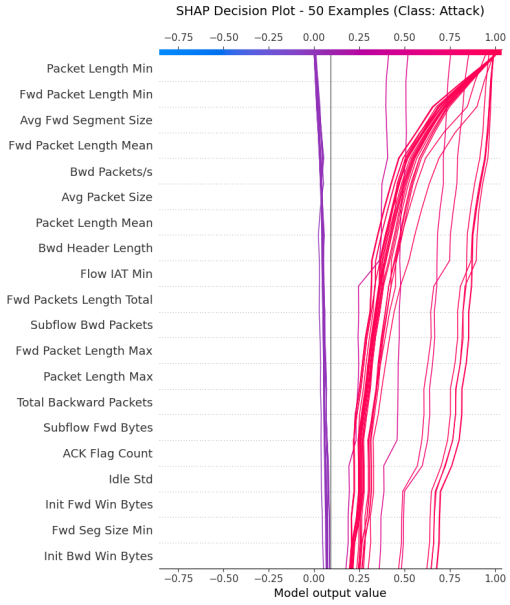

Figure 3. SHAP decision plot illustrating cumulative feature contributions for representative attack detections.

Figure 4 relates predicted attack probabilities to jury scores over 100 audited detections. Although higher probabilities tend to yield higher jury scores, the relationship is non-deterministic. Detections with similar confidence values may receive different scores, reflecting differences in strength, feature dominance, and probability ambiguity. Low jury scores are predominantly associated with *weak evidence* flags, while higher scores are mainly affected by *ambiguous probability* flags. No weak-evidence cases appear among the highest-score detections, confirming that higher jury scores correspond to stronger and more distributed SHAP support. Importantly, the jury score captures evidential robustness rather than correctness, enabling epistemic risk assessment independently of detection accuracy.

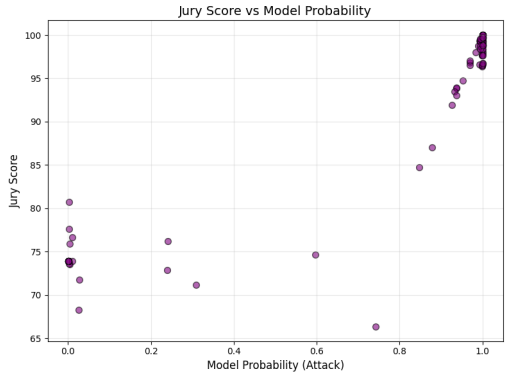

Figure 4. Jury score as a function of predicted attack probability

Qualitative inspection shows that LLM-generated explanations generally follow the epistemic signals produced by the jury. Strong evidence leads to confident but calibrated explanations, whereas weak-evidence cases result in more cautious language. Overall, explanation risk is governed by the structure of supporting evidence rather than by model confidence alone. While the previous analysis demonstrates that explanation quality closely follows the epistemic signals produced by the deterministic jury, it does not isolate the contribution of each evidential component to this behavior. To assess how individual sources of epistemic risk influence the jury score and its downstream effects, we next conduct an ablation study that systematically removes each penalty term while preserving the original calibration baselines.

5.1. Ablation Study

In order to evaluate the structural contribution of each deterministic component of the jury heuristic, we perform an ablation study in which individual penalty terms are removed while all calibration baselines are kept unchanged. Each penalty term encodes a specific epistemic risk and reduces the jury score when such risk is detected. The analysis focuses on the sensitivity and expressiveness of the jury score, rather than on classification correctness. Removing the SHAP-based evidence strength component sharply compresses the score distribution, indicating that attribution magnitude is the primary driver of differentiation between well-supported and weakly grounded decisions. In contrast, removing the feature dominance or probability ambiguity components produces only minor variance changes, confirming their role as complementary safeguards that penalize structurally risky decisions without destabilizing the overall scoring behavior. Overall, the jury score responds to the internal structure of evidential support rather than to the predicted label, enabling the identification of epistemic fragility independently of correctness.

5.2. Effect of Jury Scores on LLM Risk Expression and Explanatory Behavior

In order to assess how epistemic grounding affects risk communication, Table 2 compares a purely LLM-driven explanation baseline with the governed TRACE-NET framework, in which the LLM is constrained by deterministic jury signals. Without structural reliability cues, the LLM-only baseline assigns high risk to most cases (88%) and frequently uses uncertainty-related language (87%), reflecting systematic risk inflation weakly tied to evidential support. In contrast, the governed framework issues fewer high-risk assessments (32%) while still acknowledging uncertainty when appropriate (65%), showing that anchoring explanations to jury signals yields more calibrated and evidence-aware risk communication, independently of detection accuracy.

Table 2. Comparison of risk expression between the LLM baseline and the governed framework

Framework	High-risk label (%)	Uncertainty (%)	Assertivity (%)
LLM baseline (no jury)	88	87	87
TRACE-NET framework	32	65	57

Table 3 compares selected examples from the LLM baseline and the governed framework, showing that the proposed approach produces more precise and interpretable

explanations. For high-risk cases, the framework explicitly focuses on unusual network packet characteristics, whereas the baseline relies on model confidence and generic feature contributions. For low-risk cases, the framework highlights weak evidence and encourages cautious interpretation, while the baseline assumes benign behavior with minimal nuance. The framework does not only correctly identify risk but it also communicates relative certainty and evidence relevance, providing an actionable explanation.

Table 3. Selected examples of human-redable explanations for high and low risk from the LLM baseline and TRACE-NET governed framework

System	Risk Level	Human Explanation
Framework	High	The system is almost certain this is an attack because the size and structure of the network packets are very unusual compared to normal activity.
Framework	Low	Although the features point toward benign activity, the evidence used by the model is weak, so while the result is likely safe, it should be interpreted with some caution.
Baseline	High	The model detects 'Attack' with absolute confidence (probability 1.0000), and the decision is based on multiple top contributing features. Such high certainty about an attack typically indicates a high-risk scenario. Even though feature details are limited, the combination of strong detection and relevant network packet measurements suggests a high perceived risk. There is some uncertainty due to lack of contextual information, but the model confidence pushes the perceived risk to high.
Baseline	Low	The model detects 'Benign' with a very low probability for the alternative (less than 1%). The top features all have small negative SHAP values, supporting the benign classification. No evidence in the provided information suggests significant risk or uncertainty.

6. Conclusion

This work advances network security research by shifting the focus from detection accuracy alone to the auditability and interpretability of network traffic–based decisions. TRACE-NET enables operators to assess not only what a model detects, but whether a detection is epistemically well supported. Central to this design, the LLM is strictly constrained to contextualizing deterministic evidence, ensuring alignment between human-facing explanations and underlying model signals. The explanation audit is intentionally performed on a small, jury-stratified subset to support controlled inspection of epistemic reliability rather than population-level performance estimation. TRACE-NET requires only: (i) flow-level network traffic features, (ii) a probabilistic classifier, and (iii) feature attributions. Consequently, the methodology can be applied unchanged to other network intrusion datasets, as well as to alternative classifiers.

While this work focuses on methodological validation, we also acknowledge the importance of evaluating latency and real-time performance in operational settings. Addressing these aspects remains an important direction for future work, particularly in realistic deployment scenarios once appropriate data and infrastructure become available. Overall, TRACE-NET provides an auditable and governed framework that aligns operational security needs with emerging regulatory demands for transparent, accountable, and risk-aware AI systems in network security.

References

Abiramasundari, S. and Ramaswamy, V. (2025). Distributed denial-of-service (ddos) attack detection using supervised machine learning algorithms. *Scientific Reports*, 15.

Agarwal, C., Tanneru, S. H., and Lakkaraju, H. (2024). Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models.

Ali, T. and Kostakos, P. (2023). Huntgpt: Integrating machine learning-based anomaly detection and explainable ai with large language models (llms).

Batool, A., Zowghi, D., and Bano, M. (2024). Ai governance: A systematic literature review.

BC, C. (2024). Transparency and accountability in ai systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. ACM.

Choubisa, M., Doshi, R., Khatri, N., and Hiran, K. (2022). A simple and robust approach of random forest for intrusion detection system in cyber security.

Cloudflare (2025). DDoS Threat Report 2025 Q3. `https://blog.cloudflare.com/ddos-threat-report-2025-q3`. Accessed: 2026-01-14.

Epstein, E. L., Winnicki, J., Sornwanee, T., and Dwaraknath, R. (2025). Llms are overconfident: Evaluating confidence interval calibration with fermieval.

Gaspar, D., Silva, P., and Silva, C. (2024). Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron. *IEEE Access*, 12:30164–30175.

Ghafouri, B., Mohammadzadeh, S., Zhou, J., Nair, P., Tian, J.-J., Tsujimura, H., Goel, M., Krishna, S., Rabbany, R., Godbout, J.-F., and Pelrine, K. (2025). Epistemic integrity in large language models.

Guastalla, M., Li, Y., Hekmati, A., and Krishnamachari, B. (2024). Application of large language models to ddos attack detection. In *Smart Systems and IoT: Innovations and Analytics for a Sustainable World*, pages 65–78. Springer.

Hill, W., Mason, J., Aldrich, B., Acquaah, Y., and Roy, K. (2025). Enhancing ddos detection in software-defined networking: A machine learning and deep learning approach.

IBM Security and Ponemon Institute (2025). Cost of a data breach report 2025. Technical report, IBM. Annual industry report on data breach costs and trends.

Jamshidi, S., Nikanjam, A., Shahabi, N., Nafi, K., Khomh, F., Keivanpour, S., and Herrero, R. (2025). Think fast: Real-time iot intrusion reasoning using ids and llms at the edge gateway.

Li, C. and Goel, S. (2025). Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems. *International Journal of Accounting Information Systems*.

Li, Q., Zhang, Y., Jia, Z., Hu, Y., Zhang, L., Zhang, J., Xu, Y., Cui, Y., Guo, Z., and Zhang, X. (2024). Dollm: How large language models understanding network flow data to detect carpet bombing ddos.

Liu, Z., Wang, Y., Feng, F., Li, Z., and Shan, Y. (2023). A ddos detection method based on feature engineering and machine learning in software-defined networks. *Sensors*, 23.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Markovic, T., Leon, M., Buffoni, D., and Punnekkat, S. (2022). Random forest based on federated learning for intrusion detection.

Wali, S. and Khan, I. (2021). Explainable ai and random forest based reliable intrusion detection system.

Wang, J., Yu, L., Lui, J. C. S., and Luo, X. (2025). Modern ddos threats and countermeasures: Insights into emerging attacks and detection strategies.

Wang, T., Xie, X., Zhang, L., Wang, C., Zhang, L., and Cui, Y. (2024). Shieldgpt: An llm-based framework for ddos mitigation. In *APNet 2024: The 8th Asia-Pacific Workshop on Networking*, pages 108–114.

Wei, Y., Jang-Jaccard, J., Singh, A., Sabrina, F., and Camtepe, S. (2023). Classification and explanation of distributed denial-of-service (ddos) attack detection using machine learning and shapley additive explanation (shap) methods.

Zhou, Q., Li, R., Xu, L., Nallanathan, A., Yang, J., and Fu, A. (2023). Towards interpretable machine-learning-based ddos detection. *SN Comput. Sci.*, 5(1).

Ziems, N., Liu, G., Flanagan, J., and Jiang, M. (2023). Explaining tree model decisions in natural language for network intrusion detection.