CENTRAL ASIAN JOURNAL OF INNOVATIONS ON TOURISM MANAGEMENT AND FINANCE

https://cajitmf.centralasianstudies.org/index.php/CAJITMF

Volume: 07 Issue: 03 | Juny 2026 ISSN: 2660-454X

Article

Per-Attack-Type Evidence Aggregation for Interpreting Multi-Agent Ddos Detection

Bekov Sanjar Nigmandjanovich *[1]

1. Independent Researcher at Tashkent International University
* Correspondence: sanjar.bekov@gmail.com

Abstract: Multi-agent Distributed Denial-of-Service (DDoS) detection decomposes the decision-making process into a supervised classification agent, an anomaly detection agent, a normal-behavior or baseline-deviation agent, and a transparent rule-based evidence agent. The outputs of these agents are subsequently integrated by an evidence-fusion agent to yield a unified risk score and discrete risk level. While this architectural approach enhances modularity and interpretability, it poses a pivotal evaluation challenge: for each attack category, which detector provides the primary discriminative signal, and how does evidence fusion reconcile partial or conflicting evidence? This paper introduces a per-attack-type evidence aggregation methodology, accompanied by an analyst console visualization. For each analysis window, five normalized signals are retained: classification probability, anomaly score, baseline-deviation score, rule-evidence score, and fused risk. Records are grouped by scenario, and empirical means, dispersion statistics, peak risk, and maximum ordinal risk levels are computed for SYN, UDP, HTTP, amplification, and benign reference scenarios. The resulting visualization elucidates detector complementarity, disagreement, and the consistency of evidence fusion across the attack taxonomy. This methodology is explicitly diagnostic and not intended as a replacement for accuracy-based evaluation; it is used in conjunction with per-class precision, recall, F1 score, confusion matrices, cross-dataset validation, and agent ablation experiments. The principal contribution is an analyst-centered methodology for interpreting multi-agent DDoS detection outcomes, while rigorously controlling for alert-selection bias, target leakage, imbalanced sample counts, and intra-class variability.

Keywords: DDoS detection; multi-agent systems; evidence aggregation; evidence fusion; analyst console; attack taxonomy; explainable machine learning; ablation analysis.

Citation: Nigmandjanovich B. S Per-Attack-Type Evidence Aggregation for Interpreting Multi-Agent Ddos Detection. . Central Asian Journal of Innovations on Tourism Management and Finance 2026, 7(3), 406-420.

Received: 20th Apr 2026
Revised: 30th Apr 2026
Accepted: 18th May 2026
Published: 09th Jun 2026

Introduction

Distributed Denial-of-Service (DDoS) attacks present a significant threat to the availability of Internet services, cloud platforms, financial applications, and Internet of Things infrastructures. The manifestations of these attacks are highly heterogeneous: certain attacks generate elevated packet or byte volumes, others exploit vulnerabilities in TCP state management, and application-layer attacks may closely mimic legitimate requests. Consequently, a detection system predicated on a single feature family or learning algorithm may exhibit strong performance for specific attack categories while failing to generalize to others. This inherent variability underscores the rationale for adopting hybrid and multi-agent detection architectures, wherein statistically diverse detectors contribute distinct forms of evidence to a unified decision-making process.[1]

Within the proposed architecture, the detection process is partitioned among four autonomous agents. The supervised classification agent estimates the probability of

known attack classes based on previously labeled data. The anomaly detection agent evaluates whether the current traffic window deviates significantly from its learned normal representation. The normal-behavior agent quantifies deviations from service-, host-, or time-specific baselines. The rule-evidence agent implements transparent protocol and traffic rules, such as detecting abnormal SYN-to-ACK ratios or unusual concentrations of requests targeting a single destination. Subsequently, the evidence-fusion agent integrates the four normalized outputs to produce a unified risk score and a discrete operational risk level.

This decomposition enhances modularity, facilitates independent model replacement, improves fault isolation, and augments interpretability. However, it also introduces a novel evaluation challenge. A high fused risk score does not inherently indicate whether the decision was predominantly influenced by classification, anomaly detection, contextual deviation, or explicit rule-based evidence. The aggregate detector may appear accurate even when an individual agent exhibits persistent overconfidence, inactivity, redundancy, or reliance on dataset-specific artifacts. Conversely, a low final risk score may obscure substantive disagreement among the contributing agents. Understanding these inter-agent dynamics is imperative, particularly when evidence fusion is posited as a means of mitigating biases in single-model detection systems.[2]

Traditional chronological dashboards depict scores as a function of time. Although such visualizations are instrumental for incident reconstruction-illuminating the timing of evidentiary changes and the system's response latency-a temporal framework is not optimized for comparing the characteristic responses of detection agents across diverse attack classes. These dashboards may include repeated analysis windows from a single scenario, varying attack durations, and benign intervals that dominate the representation. Accordingly, there is a clear need for a complementary analytical perspective in which the horizontal axis is defined by attack categories rather than by temporal progression.

This paper presents a per-attack-type evidence-aggregation methodology for interpreting multi-agent Distributed Denial-of-Service (DDoS) detection outcomes. For each analysis window, the system records the four constituent evidence scores, the fused risk score, unique identifiers, timestamps, model versions, and relevant evaluation metadata. The records are subsequently organized by scenario, and summary statistics are computed for each detector. The resulting evidence profiles enable analysts to systematically compare the responses of classification, anomaly detection, baseline deviation, and rule-based agents to SYN floods, UDP floods, HTTP floods, amplification variants, and benign reference traffic.[3]

The proposed visualization is not intended to independently establish detection accuracy. Aggregated curves characterize the central tendency of stored model outputs rather than the prevalence of correct classifications. Accuracy assessment remains contingent on ground-truth labels, confusion matrices, precision, recall, F1 scores, calibration, and false-positive and false-negative rates. Instead, the visual evidence profiles are designed to facilitate the diagnosis of detector complementarity, inter-agent disagreement, and fusion consistency, and to inform the design of formal ablation experiments.

This study investigates five primary research questions: (1) How do individual agent responses vary across distinct attack categories? (2) Which agents exert dominant influence on the evidentiary landscape for various DDoS families? (3) In what manner does evidence fusion reconcile inter-agent agreement and disagreement? (4) Can per-attack evidence profiles expose model bias or target leakage? and (5) Are visually apparent agent contributions substantiated by controlled agent-ablation experiments? The principal contribution is an analyst-centered methodology that integrates stored evidence, scenario-level aggregation, visualization, and formal evaluation.[4]

Literature Review.

Research on Distributed Denial-of-Service (DDoS) attacks has consistently highlighted the heterogeneity inherent in both attack and defense taxonomies. Mirkovic and Reiher systematically classified DDoS attacks and defense mechanisms along dimensions such as source distribution, exploited vulnerabilities, attack rate, and defense location. Zargar, Joshi, and Tipper provided a comprehensive review of prevention, detection, source identification, and response strategies. These taxonomies suggest that a detection system should not be assessed solely as a global binary classifier; rather, the performance of its constituent components should be evaluated across protocol, volumetric, reflective, and application-layer categories.

Machine-Learning-Based Detection[5]

Machine-learning-based intrusion detection introduces additional diversity, as different models encapsulate distinct statistical properties. Supervised classifiers are designed to learn mappings from feature vectors to categorical labels and are generally effective when the distributions underlying training and deployment are sufficiently aligned. Murphy conceptualizes classification as the process of learning a mapping from inputs to categorical outputs, underscoring the importance of model selection, mitigating overfitting, and quantifying uncertainty. In the context of DDoS detection, classification probabilities can provide robust evidence for known attack families; however, these probabilities may be poorly calibrated when applied to data outside the training distribution.

Anomaly Detection & Normal Behavior[6]

Anomaly detection is employed to address previously unseen or weakly labeled behaviors by modeling normality or data density. The output generated by anomaly detection fundamentally differs from class probability:

- A traffic window exhibiting a high degree of anomaly is not inherently malicious.
- A familiar attack may be classified with high confidence despite lacking global unusualness.

A distinct normal-behavior agent contributes contextual information by comparing current traffic with expected patterns for specific services, destinations, or time periods. This distinction is particularly salient in scenarios involving flash crowds, scheduled batch processing, software updates, and other legitimate workload fluctuations.

Rule-Based Systems and Evidence Fusion[7]

Transparent rule-based systems retain significant value owing to their capacity to encode protocol-specific knowledge and generate explanations readily interpretable by system operators. Such rules can identify indicators such as:

- Elevated SYN-to-ACK ratios
- Persistent UDP activity
- A large number of sources targeting a single service

The principal limitations of rule-based approaches are their restricted coverage and brittleness when confronted with novel conditions. Evidence fusion seeks to integrate the strengths of classification, anomaly detection, baseline modeling, and rule-based reasoning, thereby preventing any single method from unilaterally determining the final verdict.

Agent-Oriented Intrusion Detection[8]

Agent-oriented intrusion detection constitutes a natural implementation paradigm for this decomposition. Shoham introduced agent-oriented programming as a conceptual framework for computational entities endowed with explicit mental-state representations, while Wooldridge and Jennings defined intelligent agents in terms of autonomy, reactivity, proactiveness, and social ability.

The AAFID architecture distributed intrusion detection functions across autonomous agents, and subsequent research underscored the importance of modularity and localized processing. More recent studies have applied intelligent agents and automated feature selection techniques directly to DDoS detection.

Explainability and Interpretation[9]

Most explainable machine learning methodologies concentrate on individual prediction interpretation or the assessment of global feature importance:

- LIME constructs a locally interpretable approximation centered on a specific prediction (Ribeiro, Singh, & Guestrin).
- SHAP provides additive feature attributions for a given model output.

While these techniques are valuable in the context of the classification agent, they do not directly elucidate interactions among heterogeneous agents whose outputs have distinct semantic interpretations. A per-attack evidence profile provides an alternative explanatory layer by revealing which decision components respond strongly or weakly to each attack category.

Evaluation Methodology and Data Integrity

Evaluation methodology is of paramount importance in the context of security data. Sommer and Paxson cautioned that exemplary benchmark performance does not necessarily equate to effective operational intrusion detection. Ring et al. conducted a comprehensive review of network-based datasets, underscoring the necessity of realism, diversity, and thorough documentation. Engelen, Rimmer, and Joosen demonstrated that choices regarding flow construction and labeling can significantly influence outcomes. Scenario identifiers, capture dates, IP addresses, and other experimental artifacts may inadvertently serve as shortcuts if incorporated into learning features.[10]

Resampling and Temporal Challenges

Cross-validation and class imbalance introduce additional methodological risks. Kohavi systematically evaluated resampling methods for model selection, while Varma and Simon demonstrated that using cross-validation for both hyperparameter tuning and final performance estimation can introduce optimistic bias. Saito and Rehmsmeier asserted that precision-recall analysis provides particularly valuable insights for imbalanced classification tasks.

Note: These findings are salient to the proposed analyst-centric view, as a visually compelling evidence profile cannot compensate for an invalid test partition or an alert repository that omits missed detections.

Concept drift introduces a temporal dimension to the analysis. Gama et al. provided a comprehensive survey of methodologies for adapting to evolving data distributions and predictive relationships. A chronological console visualization can identify the timing of baseline changes, whereas a per-attack-type perspective can determine whether such changes disproportionately affect specific attack categories or individual detectors. In combination, these perspectives facilitate the diagnosis of both temporal drift and category-specific model dependencies.

Identified Literature Gap

A distinct gap is evident in the extant literature. While prior studies address DDoS classifiers, anomaly detectors, agent-based frameworks, ensemble methodologies, and explainability, few offer a formal, analyst-oriented methodology for aggregating heterogeneous detector outputs by attack category.

The approach advanced herein addresses this deficiency by specifying a standardized evidence record, scenario-based grouping, summary statistical measures, ordering rules, interpretative guidelines, and safeguards against erroneous or misleading conclusions.[11]

Research Methodology.

This research adopts a design-science methodology, underpinned by controlled experimental procedures. The primary artifact comprises an analyst console visualization and an aggregation service, both of which operate on evidence generated by a multi-agent Distributed Denial-of-Service (DDoS) detection system. The evaluation is structured around two principal objectives: first, to assess whether the visualization accurately represents the underlying agent outputs; and second, to determine whether patterns observed within the visualization correspond to quantifiable detector contributions as established by formal performance evaluations.

The underlying pipeline comprises a Traffic Monitoring Agent, Feature Extraction Agent, Supervised Classification Agent, Anomaly Detection Agent, Normal Behavior Agent, Rule-Based Evidence Agent, Evidence Fusion Agent, Explainability Agent, Alert Store Agent, Control Gateway, and Bias Monitoring Agent. The initial four detection agents generate scalar scores normalized to the interval [12]. The Evidence Fusion Agent synthesizes these outputs into a fifth scalar score, which is subsequently mapped to an ordinal risk category. The Alert Store preserves the complete evidence vector, thereby ensuring that the final verdict remains fully auditable.

The primary labeled dataset utilized in this study is CIC-DDoS2019, owing to its inclusion of multiple DDoS families and flow-level records [13]. CSE-CIC-IDS2018, CAIDA DDoS 2007, and Bot-IoT serve as sources for external or transfer evaluation when feature compatibility permits. The console design remains agnostic to any particular dataset; any experimental dataset may be employed, provided that each scored window is associated with a correlation identifier, ground-truth or evaluation scenario, and all five evidence values.

Scenario labels may be assigned based on controlled experimental metadata or correlation identifiers, such as syn-flood, udp-flood, http-flood, dns-amplification, ntp-amplification, or benign-reference. These scenario labels serve solely as evaluation metadata and must not be incorporated into any feature vector or provided to any detection or fusion model. This strict separation is essential to prevent target leakage, in which the detector infers the experiment label rather than accurately learning network behavior.

The storage system should retain all scored windows during evaluation, rather than limiting storage to those windows that surpass the alert threshold. Restricting storage to alert-generating windows introduces selection bias, as undetected attacks and most benign windows are excluded from subsequent analyses. If operational constraints necessitate selective storage, the retained population and sampling policy must be thoroughly documented, and the resulting analytical perspective must be explicitly labeled as conditional upon alert selection.[14]

For each scenario, the aggregation service computes the count, arithmetic mean, median, standard deviation, interquartile range, maximum, and selected percentiles for each evidence signal. The mean provides a succinct profile, while measures of dispersion and sample size help prevent overinterpretation. Bootstrap confidence intervals may be estimated when independent windows are available. In instances where adjacent windows from the same flow exhibit strong correlation, a capture session should be conducted at the capture session episode level, rather than treating each window as an independent observation.

Scenarios are visualized on a standardized ordinate, scaled from 0 to 1 or equivalently from 0% to 100%. The default ordering is based on descending peak fused risk, thereby prioritizing categories eliciting the most pronounced detector responses. Given the susceptibility of maximum values to outliers, the console also supports ordering by mean, median, or 95th-percentile-fused risk. Each category is annotated with the highest observed discrete risk level and the corresponding sample count.[15]

Interpretation is facilitated by three complementary analytical approaches. First, vertical comparisons within a given category identify which agent outputs are relatively strong. Second, horizontal comparisons across categories assess whether a specific agent responds consistently or is selective to particular attack types. Third, disagreement statistics quantify the variability among the four contributing inputs. These descriptive analyses are subsequently augmented by agent-ablation experiments that evaluate whether a seemingly dominant agent substantively improves per-class performance.

The formal evaluation methodology involves comparing the complete fusion model with four leave-one-agent-out variants. For each attack category, the analysis reports precision, recall, F1 score, false positive rate, false negative rate, precision-recall area, and calibration metrics when probabilistic outputs are available. A detector's contribution is considered substantiated when excluding the corresponding agent results in reduced performance or degraded calibration, as demonstrated in reproducible, held-out evaluations. Visual prominence, in isolation, is not regarded as sufficient causal evidence.

The analyst-console implementation may leverage an event-driven microservice architecture. Detection agents publish versioned events, which the Alert Store Agent subsequently correlates using unique identifiers. The Control Gateway provides endpoints, such as GET /analytics/evidence/by-attack-type, to facilitate data retrieval. A web-based console visualizes the aggregated evidence series and enables analysts to toggle between chronological, per-attack-type, and bias-monitoring perspectives. All aggregation computations are fully reproducible from the persisted evidence records.[16]

Result and discussion.

System Model and Data Provenance.

For each analysis window w, the pipeline generates four component scores and a fused score. All values are normalized to the unit interval, thereby enabling their representation on a unified scale. It is important to note that normalization does not imply identical calibration across scores; rather, it solely establishes a common numerical range. Table 1. The signals are defined as follows:

Symbol	Evidence source	Meaning
c(w)	Classification Agent	Estimated probability or confidence that the window represents DDoS.
a(w)	Anomaly-Detection Agent	Degree to which the window differs from the anomaly model's learned distribution.
b(w)	Normal-Behaviour Agent	Deviation from the expected host, service, destination, or time-specific baseline.
r(w)	Rule-Evidence Agent	Normalized strength of transparent protocol and traffic rules fired for the window.
f(w)	Evidence-Fusion Agent	Combined risk score produced from c(w), a(w), b(w), and r(w).

Table 1. Evidence signals stored for every analysis window.

A general weighted fusion rule is written as:

$$f(w) = \omega c\, c(w) + \omega a\, a(w) + \omega b\, b(w) + \omega r\, r(w),$$
$$\text{where } \omega c, \omega a, \omega b, \omega r \geq 0 \text{ and } \omega c + \omega a + \omega b + \omega r = 1.$$

An initial implementation may assign weights of 0.40, 0.25, 0.20, and 0.15 to classification, anomaly detection, baseline deviation, and rule-based evidence, respectively. These values serve as design defaults rather than universally optimal weights and must be selected or validated without reference to the final test set. A learned meta-classifier may supplant the fixed weighted sum; however, the component outputs should continue to be recorded to support interpretability and auditability.[17]

Table 2. The fused score is mapped to a discrete risk level using operational thresholds:

Fused-risk interval	Risk level
$0.00 \leq f < 0.40$	NORMAL
$0.40 \leq f < 0.70$	SUSPICIOUS
$0.70 \leq f < 0.85$	PROBABLE_DDOS
$0.85 \leq f \leq 1.00$	HIGH_CONFIDENCE_DDOS

Table 2. Example mapping from continuous fused risk to an ordinal risk level.

Per-Attack-Type Evidence Aggregation Method.

Let W denote the set of scored windows available to the console and S the set of represented attack scenarios. Each window w in W is associated with an evaluation label s(w) in S. The scenario-specific partition is defined as:

$$W\_s = \{ w \text{ in } W : s(w) = s \}.$$

The number of windows associated with scenario s is $n\_s = |W\_s|$. For every evidence signal x in {c, a, b, r, f}, the primary plotted statistic is the empirical scenario mean:

$$mean\_x(s) = (1 / |W\_s|) \sum_{w \text{ in } W\_s} x(w), \quad for\ |W\_s| > 0.$$

The resulting evidence profile for scenario s is:

$$E\_s = [mean\_c(s), mean\_a(s), mean\_b(s), mean\_r(s), mean\_f(s)].$$

The five profile components are depicted as overlaid curves or grouped markers on a unified vertical scale. The fused-risk series is visually accentuated as the final verdict, while the four input series remain accessible to facilitate examination of the relationships between individual inputs and the overall verdict. The categorical axis supplants the temporal axis, thereby enabling direct comparison of the characteristic responses to each evaluated attack family.

Relying solely on mean values is inadequate when category sizes vary, or score distributions are skewed. Accordingly, each category should report the sample size (n_s) and provide access to additional descriptive statistics as needed. The median offers robustness to isolated outliers, while the standard deviation and the interquartile range characterize within-category variability. Confidence intervals quantify the uncertainty associated with the estimated mean. In the context of extended captures divided into overlapping windows, statistical dependence must be accounted for, as a nominally large n_s may correspond to only a limited number of independent episodes.[18]

The default category order is based on peak fused risk:

$$peak\_f(s) = \max_{w \text{ in } W\_s} f(w).$$

Categories are organized in descending order of peak_f(s), thereby presenting scenarios with the highest model-assigned risk at the forefront. It is crucial to emphasize that model-assigned risk, as represented by peak_f(s), does not constitute an objective measure of real-world attack severity; rather, it reflects the maximum response elicited by the implemented detector and may be susceptible to outliers or calibration inaccuracies. Consequently, the console should facilitate alternative ordering options, such as sorting by mean, median, or 95th-percentile fused risk.

Each category receives a worst observed risk annotation:

$$L(s) = \max_{w \text{ in } W\_s} riskLevel(w),$$

where the maximum follows the ordinal relationship NORMAL < SUSPICIOUS < PROBABLE_DDOS < HIGH_CONFIDENCE_DDOS. This categorical label complements the continuous mean profile and allows rapid operational scanning without discarding the underlying scores.

Two supplementary diagnostic quantities strengthen interpretation. The first is an agent-dominance proportion for input j and scenario s:

$$D\_j(s) = mean\_j(s) / (mean\_c(s) + mean\_a(s) + mean\_b(s) + mean\_r(s) + epsilon),$$

where epsilon is a small constant used only to avoid division by zero. D_j(s) describes the share of average input evidence associated with one agent; it is descriptive and should not be interpreted as causal contribution. The second is a disagreement index:

$$G(s) = (1 / |W\_s|) \, sum\_\{w \, in \, W\_s\} \, SD(c(w), a(w), b(w), r(w)),$$

where SD is the standard deviation among the four input scores for one window. A high G(s) indicates that the agents frequently disagree for scenario s. Such disagreement can reveal novelty, model overconfidence, baseline instability, or rule coverage gaps and should trigger deeper inspection.

The aggregation service returns both the displayed values and their provenance. A representative response object contains scenario, sampleCount, meanClassification, meanAnomaly, meanBaselineDeviation, meanRuleEvidence, meanFusedRisk, medianFusedRisk, standard deviations, peakFusedRisk, percentile95FusedRisk, worstRiskLevel, and model-version identifiers. This prevents the chart from becoming an untraceable visual summary.[19]

Analyst Console Design and Interpretation.

The analyst console encompasses three complementary analytical perspectives. The chronological view visualizes evidence as a function of timestamp, thereby facilitating incident reconstruction, response latency analysis, and concept drift investigation. The per-attack-type perspective organizes evidence by scenario, enabling comparative analysis across the attack taxonomy. The bias and performance perspective presents confusion matrices, per-class performance metrics, calibration assessments, and agent-ablation results. It is important to note that no single perspective is sufficient in isolation.

The per-attack-type view is interpreted through vertical comparison. For a given attack category, the relative magnitudes of mean_c(s), mean_a(s), mean_b(s), and mean_r(s) reveal which detectors exhibit strong responses. The value of mean_f(s) reflects the manner in which the fusion stage integrates these individual signals. For example, if both classification and rule-based evidence are elevated in response to a SYN flood, while anomaly evidence remains moderate, this pattern may indicate a familiar signature-based attack. Conversely, if anomaly and baseline-deviation signals are pronounced but classification evidence is low, the scenario may represent a novel, distributionally shifted, or poorly represented instance in the labeled training data.

The same view is interpreted horizontally to discern agent-specific behavioral patterns. A classification curve that remains elevated across all categories, including benign reference traffic, may indicate poor calibration or target leakage. An anomaly curve that exhibits consistently high values across scenarios may suggest instability in the normality model. A rule-based evidence curve that remains near zero across most known attack types may indicate inadequate rule coverage. Substantial fluctuations in the baseline curve across different capture days may indicate operational drift rather than attack-specific evidence.

Benign reference traffic holds particular significance in the evaluation process. An appropriately designed system should generally suppress all four input signals and the fused verdict in representative benign scenarios. However, the benign traffic category must encompass realistic high-load and burst conditions; reliance on a quiet-only reference class renders the detection task artificially straightforward. Accordingly, flash-crowd, backup, update, and scheduled processing scenarios should be incorporated as challenging benign cases.

Disagreement among agents does not inherently constitute an error. The agents are intentionally heterogeneous, each designed to detect distinct facets of the same event. The primary objective of evidence fusion is to integrate complementary sources of information. Nevertheless, instances of persistent disagreement should be explicable. The console should enable analysts to select a specific category, examine its window-level distribution, access representative evidence records, and review feature attributions and activated rules.

A high fused curve above all inputs may be impossible under a convex weighted sum and can reveal a data or visualization defect. A fused curve far below several consistently high inputs may be mathematically valid when their weights are small, but it may also indicate inappropriate weight selection. Thus, the view serves as both a software validation instrument and a machine-learning diagnostic.

Color and ordering selections must be designed to prevent misleading emphasis. While fused risk may be presented as the primary series, component signals should remain clearly distinguishable, without suggesting that any particular color connotes correctness. Each category should display sample counts, and tooltips should reveal measures of central tendency and dispersion. To ensure accessibility, alternative encodings-such as line style, marker shape, or direct labeling-should be employed in addition to color.

The console should facilitate filtering by dataset, capture day, destination, model version, and time interval. In the absence of such filters, aggregating data across heterogeneous deployments may lead to Simpson's paradox, in which trends observed at the aggregate level are reversed within specific sub-environments. Version-aware filtering is especially critical after model retraining, as evidence scores from models with different calibration may not be directly comparable.

Prototype Analyst-Console Output.

Figures 1(a) and 1(b) illustrate two complementary perspectives derived from the prototype analyst console. The first visualization aggregates evidence by attack category, while the second preserves the chronological sequence of recent detections. The panels were captured from live or replayed streams at proximate, though not identical, time points; thus, their respective counters differ marginally and should not be interpreted as a perfectly synchronized statistical sample.

Figure 1(a). Per-attack-type evidence aggregation. The plot summarises mean agent scores across 15,079 detected windows; all displayed categories, including the benign reference, are annotated as Suspicious.

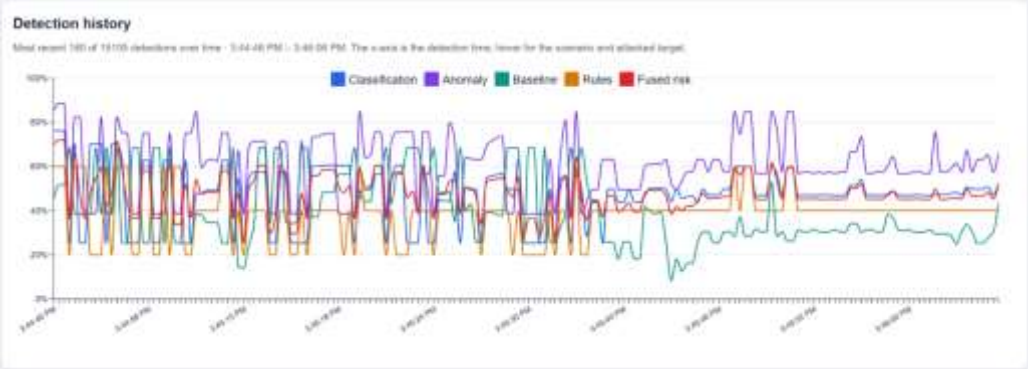

Figure 1(b). Chronological evidence view. The panel shows the most recent 180 of 15,105 detections over approximately 80 seconds and exposes short-lived score changes that category means suppress.

The category-level visualization reveals that anomaly evidence is the strongest signal across all scenarios, whereas baseline evidence consistently has the lowest mean values. The fused risk score closely tracks the classification series and remains within the Suspicious band across all categories. Notably, the benign reference category is also labeled "Suspicious". This outcome does not attest to detection accuracy; rather, it highlights a calibration or threshold-separation issue that warrants further investigation using comprehensive ground truth, challenging benign traffic scenarios, probability calibration, and agent-ablation experiments. Accordingly, the current screenshot serves as diagnostic evidence of the proposed visualization's utility, rather than as proof of effective discrimination between attacks.

The chronological panel provides a complementary temporal perspective. During the initial portion of the interval, scores display significant fluctuations. Subsequently, the classification, rule-based, and fused risk signals attain greater stability, whereas the anomaly detection agent continues to produce intermittent peaks and the baseline agent shifts to a lower operational range. These temporal dynamics are largely concealed when examining only scenario-level mean values. Taken together, the two panels highlight the necessity of integrating category-based aggregation with temporal analysis: the former reveals systematic agent-to-scenario relationships, while the latter exposes transient disagreement, concept drift, and evolving operational regimes.

The discrepancy between the 15,079-window count in Figure 2(a) and the 15,105-detection count in Figure 2(b) further underscores the importance of provenance tracking. Production consoles should annotate every exported figure with a snapshot identifier, query timestamp, model versions, filter set, and retention mode. Absent such metadata, two views captured only seconds apart may be erroneously interpreted as representing identical populations.

Evaluation Protocol.

The evaluation process is partitioned into descriptive verification, predictive evaluation, and causal contribution analysis. Descriptive verification assesses whether console aggregates correspond to independently recomputed statistics derived from the evidence store. Unit tests employ small, deterministic examples, whereas integration tests validate the integrity of correlation identifiers, the handling of missing agents, normalization procedures, risk-level mapping, and model version propagation.

Predictive evaluation is conducted using an untouched, held-out test set, with results reported by attack type. Essential evaluation metrics include precision, recall, F1 score, macro-F1, false positive rate, false negative rate, confusion matrices, precision-recall area, and calibration error. Accuracy is reported only as a secondary metric, given that class

imbalance can render it misleading. Additionally, the evaluation should quantify latency and throughput to ensure that interpretability features do not compromise operational feasibility.

Agent-ablation analysis involves evaluating the complete detection system alongside four leave-one-agent-out variants that exclude classification, anomaly detection, baseline deviation, or rule-based evidence. Fusion weights must be appropriately renormalized or retrained within the validation procedure. For a given metric M and agent j, the contribution in scenario s can be summarized as follows:

$$Delta\_M\_j(s) = M\_full(s) - M\_without\_j(s).$$

A positive Delta_M_j(s) indicates that removing agent j reduces the chosen performance measure. For error measures, the sign is interpreted accordingly. Confidence intervals across outer validation folds or independent capture sessions should be reported because small differences can arise from sampling variation.

The interpretation of evidence profiles is directly compared with ablation results. If the rule-based agent exhibits the strongest mean evidence for SYN flood scenarios and its removal results in a significant reduction in SYN-flood recall, the observed visual dominance is corroborated by empirical performance data. Conversely, if an agent appears visually dominant but its removal does not affect performance, it may be redundant or highly correlated with another detector. If an agent demonstrates moderate evidence, yet its exclusion leads to a substantial decline in performance, it may provide complementary information that is proximate to the decision boundary.

Cross-dataset evaluation assesses the stability of evidence profiles and agent contributions when applied to datasets beyond the principal training set. For instance, a model trained on CIC-DDoS2019 may be evaluated on compatible scenarios from an alternative dataset following feature alignment. The analysis should report both the transfer of performance metrics and any shifts in evidence profiles. Substantial changes in profile characteristics may signal environment-specific dependencies, even when aggregate accuracy remains within acceptable bounds.

Threshold sensitivity is evaluated by conducting assessments across a continuum of fused-risk thresholds. The per-attack evidence profile remains independent of threshold selection only when all scored windows are retained. In contrast, an alert-only storage policy alters the composition of retained windows as the threshold varies, thereby affecting the calculated means. This interaction should be explicitly demonstrated to ensure that analysts comprehend the implications of the selection mechanism.

Finally, the evaluation encompasses challenging benign scenarios, temporally based data splits, and concept drift assessments. Analyst feedback is documented subsequent to the initial test and is not utilized to modify the held-out results. Confirmed cases may inform subsequent online adaptation studies; however, the original evaluation is preserved in its entirety to ensure reproducibility.

The proposed visualization establishes a connection between architectural interpretability and statistical evaluation. While agent-oriented design facilitates the availability of component outputs, mere accessibility does not guarantee their interpretability. Scenario-level aggregation organizes extensive event streams into structured profiles, enabling direct comparison with the Distributed Denial-of-Service (DDoS) taxonomy and operational benchmarks.

The principal utility of the proposed visualization is diagnostic rather than promotional. While a chart displaying high fused risk for attack scenarios and low risk for benign traffic may offer reassurance, the most valuable insights often arise from observed inconsistencies: for instance, one detector responding to all scenarios, a benign category exhibiting elevated anomaly evidence, a fusion score misaligned with its configured

weights, or an attack category supported solely by a single brittle rule. Such patterns serve to highlight areas warranting formal investigation.

The visualization additionally facilitates bias analysis. Dataset bias may be indicated when classification evidence is elevated exclusively for scenarios drawn from the training collection. Feature bias may manifest when a volume-sensitive agent exerts dominant influence over both attack and flash-crowd scenarios. Threshold bias may be indicated by abrupt category transitions due to minor score fluctuations. Environment bias may become apparent when the same attack class exhibits divergent evidence profiles across different network environments. While these visual indicators do not constitute definitive proof of bias, they render underlying dependencies both observable and empirically testable.

Agent complementarity constitutes the central hypothesis underpinning multi-agent fusion. Protocol-specific flood attacks are expected to elicit strong classification and rule-based evidence when their signatures are present in the training data and in the encoded rules. Volumetric and amplification scenarios may yield more pronounced responses in the anomaly and baseline. Application-layer attacks are likely to generate moderate, distributed evidence across multiple agents. It is imperative to regard these expectations as testable hypotheses; the console is designed to elucidate empirical patterns of evidence, rather than to impose a predetermined interpretive framework.

The aggregation methodology is well-suited to a Spring-based microservice implementation. Detection agents publish events via Kafka or an alternative message broker; an alert-store service associates events using correlation identifiers; a control gateway provides read-only analytics; and the analyst console retrieves scenario-level summaries. This architecture ensures independent deployment and enables the replacement of individual agents without necessitating modifications to the aggregation contract. However, it also introduces engineering responsibilities, including schema versioning, maintenance of eventual consistency, management of missing events, and mitigation of duplicate message delivery.

The treatment of missing evidence necessitates explicit consideration. An agent failure should not result in the silent assignment of a zero value, as zero constitutes a meaningful score within the context of detection. Each record should include a missing-status flag, and the fusion stage should implement a documented fallback policy. The analyst interface should present incomplete records separately or restrict statistical calculations to instances where the requisite signals are present. Failure to do so may lead to the misinterpretation of operational failures as indicators of low attack evidence.

The proposed evidence profile does not supplant feature-level explanation. It elucidates which detector generated a response, whereas LIME, SHAP, and rule-based traces provide insight into the underlying rationale for that detector's response. A comprehensive analyst workflow progresses from l profiling to window-level evidence examination to feature attribution and ultimately to the inspection of raw, authorized telemetry. This layered explanatory approach offers greater utility than reliance on a singular global dashboard score.

The proposed methodology also confers significant value in the context of model governance. Evidence records are annotated with model version identifiers, thereby facilitating before-and-after comparisons following model retraining. For example, a newly deployed classifier may enhance macro-F1 performance while simultaneously increasing benign classification evidence in an unanticipated manner. The per-attack evidence profile renders such trade-offs transparent and establishes an audit trail to support model approval, rollback, and threshold adjustment processes.

Limitations and Threats to Validity.

The principal limitation is selection bias. If the storage system retains only alert-generating windows, the analysis omits undetected attacks and the majority of benign

windows. Consequently, the calculated means reflect only surfaced alerts, rather than the full operational behavior of the detector. Evaluation mode should therefore retain all scored windows or, at a minimum, a rigorously documented representative sample.

Arithmetic means may obscure multimodal or highly skewed distributions. Two attack categories may exhibit identical mean values, despite one displaying stable evidence and the other oscillating between extremely low and high scores. Accordingly, medians, interquartile ranges, distribution plots, and confidence intervals should be reported alongside the primary curves to provide a more comprehensive depiction of the underlying data distributions.

Analysis windows may not exhibit statistical independence. Overlapping windows, flows originating from the same botnet episode, or repeated requests within a single capture session can result in overly narrow uncertainty estimates. Consequently, resampling and significance testing procedures should employ the episode, capture session, or day as the independent observational unit where appropriate.

Ordering based on maximum fused risk is highly susceptible to outliers and should not be construed as an objective indicator of attack severity. Rather, it reflects the most pronounced model response observed within the retained dataset. To ensure balanced interpretation, alternative ordering criteria and explicit reporting of sample counts are essential.

Scenario labels derived from correlation identifiers may be erroneous or inconsistent. These labels must be rigorously validated against experimental metadata and completely segregated from all learning inputs. Failure to maintain this separation results in target leakage, thereby invalidating both performance and interpretability claims.

The relative height of an evidence curve does not constitute evidence of causal contribution. Component scores may be correlated, and an agent that appears visually dominant could, in practice, be operationally redundant. Claims regarding agent contribution should be substantiated through ablation studies, counterfactual masking, and independent evaluation on held-out data.

Normalization to the interval [0,1] does not ensure comparable calibration across distinct agent outputs. An anomaly score of 0.8 and a classification probability of 0.8 possess fundamentally different semantic interpretations. Accordingly, fusion and visualization processes must preserve these distinctions, and probability calibration should be assessed independently for each probabilistic agent.

The proposed paper adopts a methodological focus. It delineates the data model, aggregation strategy, visualization approach, and evaluation procedure, but refrains from asserting empirical superiority until experimental results are available. Future studies should clearly differentiate between exploratory observations and confirmatory tests, and should make the aggregation code and anonymized evidence summaries publicly available, subject to licensing and privacy constraints.[20]

Conclusion.

This paper introduces a per-attack-type evidence aggregation methodology for the interpretation of multi-agent Distributed Denial-of-Service (DDoS) detection outcomes. The proposed method systematically retains the outputs of the classification, anomaly detection, normal-behavior, rule-based evidence, and fusion agents for each analysis window, subsequently grouping them by attack scenario and computing a characteristic evidence profile that includes sample counts and measures of dispersion.

The resulting analyst-console visualization serves as a complement to chronological monitoring by mapping detector behavior onto the attack taxonomy. This approach facilitates the identification of agents that exhibit strong responses to specific categories, highlights areas of inter-agent disagreement, assesses the consistency of the fusion output with its constituent inputs, and determines whether benign reference traffic effectively

suppresses component signals. Supplemental dominance and disagreement metrics further inform targeted analytical investigation.

The visualization is deliberately decoupled from assertions regarding detection accuracy. Formal evaluation continues to rely on held-out ground truth, per-class precision and recall, F1 score, confusion matrices, calibration, cross-dataset validation, and leave-one-agent-out ablation experiments. Potential sources of bias, including alert selection, target leakage, unequal sample sizes, correlated analysis windows, and model version changes, must be rigorously controlled.

Accordingly, the principal contribution is an analyst-centered evaluation layer for agent-oriented Distributed Denial-of-Service (DDoS) detection systems. By rendering detector dependencies and inter-agent disagreement observable, this approach facilitates model debugging, fusion validation, bias monitoring, and governance. Future research should focus on implementing the analyst console, evaluating its effectiveness on CIC-DDoS2019 and additional external datasets, and assessing whether the observed evidence profiles consistently predict the empirically measured marginal contributions of individual agents.

REFERENCES

[1] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.

[2] J. Mirkovic and P. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms," ACM SIGCOMM Computer Communication Review, vol. 34, no. 2, pp. 39–53, 2004.

[3] S. T. Zargar, J. Joshi, and D. Tipper, "A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2046–2069, 2013.

[4] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in Proc. International Carnahan Conference on Security Technology, 2019.

[5] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symposium on Security and Privacy, 2010.

[6] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.

[7] Y. Shoham, "Agent-oriented programming," Artificial Intelligence, vol. 60, no. 1, pp. 51–92, 1993.

[8] M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," The Knowledge Engineering Review, vol. 10, no. 2, pp. 115–152, 1995.

[9] J. S. Balasubramaniyan et al., "An architecture for intrusion detection using autonomous agents," in Proc. Annual Computer Security Applications Conference, 1998.

[10] E. H. Spafford and D. Zamboni, "Intrusion detection using autonomous agents," Computer Networks, vol. 34, no. 4, pp. 547–570, 2000.

[11] R. Abu Bakar et al., "An intelligent agent-based detection system for DDoS attacks using automatic feature extraction and selection," Sensors, vol. 23, no. 6, p. 3333, 2023.

[12] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proc. International Joint Conference on Artificial Intelligence, 1995.

[13] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, p. 91, 2006.

[14] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015.

[15] J. Gama et al., "A survey on concept drift adaptation," ACM Computing Surveys, vol. 46, no. 4, article 44, 2014.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017.

[18] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection datasets," Computers & Security, vol. 86, pp. 147–167, 2019.

[19] [19] G. Engelen, V. Rimmer, and W. Joosen, “Troubleshooting an intrusion detection dataset: The CICIDS2017 case study,” in IEEE Security and Privacy Workshops, 2021.
[20] [20] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, “Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset,” Future Generation Computer Systems, vol. 100, pp. 779–796, 2019.