International Journal of Advanced Engineering and Technology Research
ISSN: 3106-8677 (Print), ISSN: 3106-8685 (Online) | Vol. 2, No. 3, 2026

Uncertainty-Aware Evidential Adversarial Defense for Cloud Intrusion Detection under GAN-Based Evasion Attacks

Jiawen Luo *, Samuel Price

Pennsylvania State University, Department of Computer Science and Engineering, USA

* **Corresponding author**: (Email: 238402011@psu.edu)

Abstract: Cloud intrusion detection systems (IDSs) increasingly depend on deep neural classifiers, which are vulnerable to adversarial evasion attacks—including realistic, GAN-crafted traffic—that flip malicious flows to benign with near-imperceptible, functionally consistent perturbations. Adversarial training (AT) hardens the decision boundary but yields overconfident point predictions: a hardened detector still cannot signal when an input is suspicious, and Bayesian or ensemble uncertainty estimators that could do so require many forward passes, which is impractical at cloud line rate. We propose UA-EAD, an uncertainty-aware evidential adversarial defense that (i) equips the detector with an evidential (Dirichlet) head yielding calibrated predictive uncertainty in a single forward pass, (ii) trains it with an uncertainty-weighted adversarial objective plus a consistency regularizer that concentrates robustness on the most uncertain, near-boundary flows, and (iii) uses the resulting uncertainty for selective prediction, abstaining on inputs it cannot confidently classify. On NSL-KDD under FGSM, BIM, PGD, C&W, and a black-box WGAN-GP transfer attack, UA-EAD matches the strongest AT baselines in robust accuracy (97.4%) while attaining the highest clean accuracy (98.1%). Its single-pass evidential uncertainty equals maximum-softmax-probability and surpasses Monte-Carlo Dropout for adversarial detection at 21x lower inference cost, and yields the best adversarial-detection AUROC among robust models. Selective prediction lifts accuracy on 17 novel (out-of-distribution) attack types from 76.9% to 87.0% at 70% coverage. We further report an empirical robustness–detectability trade-off that clarifies why hardening alone is insufficient for trustworthy cloud IDS.

Keywords: Intrusion detection, cloud security, adversarial robustness, adversarial training, uncertainty quantification, evidential deep learning, selective prediction, GAN evasion.

1. Introduction

Machine-learning intrusion detection systems (IDSs) are now a standard component of cloud security stacks, screening network flows at gateways and virtual-network boundaries for malicious activity. Deep neural detectors deliver strong accuracy on canonical benchmarks [1-3], yet a decade of adversarial-examples research has shown that such models are brittle: small, carefully chosen input perturbations can force confident misclassification [4-7]. In the IDS setting this brittleness is a security liability rather than a curiosity—an adversary who nudges a handful of continuous flow statistics can make an attack flow read as benign, evading detection while preserving the attack's function [13, 14].

Generative models sharpen the threat. GAN-based attackers such as IDSGAN [13] learn to synthesize adversarial traffic that is both evasive and realistic, and recent work tailored to cloud environments couples an improved Wasserstein GAN with gradient penalty (WGAN-GP) threat generator to a cost-sensitive minimax defense. In particular, Ding et al. [15] propose GT-CSAT, a game-theoretic cost-sensitive adversarial-training framework that assigns asymmetric misclassification penalties through a two-player zero-sum payoff matrix and adapts the cost weights via a Nash-equilibrium-inspired update, achieving robust cloud intrusion detection against GAN-based evasion while mitigating the classical robustness–accuracy trade-off. Their results demonstrate that carefully designed training objectives can substantially harden cloud IDSs against covert, functionally consistent evasion.

Adversarial training and its cost-sensitive variants, however, share a structural limitation: they harden the decision boundary but still produce overconfident point predictions. A hardened detector classifies every input—including inputs unlike anything seen in training—with high confidence, and offers no principled way to say "I am not sure about this flow." Operationally, a cloud IDS rarely acts in isolation: uncertain or anomalous traffic can be escalated to deep packet inspection, sandboxing, or a human analyst. Realizing this fail-safe behavior requires a detector that not only resists perturbations but also quantifies its own uncertainty and abstains when that uncertainty is high. Classical Bayesian and ensemble estimators (Monte-Carlo Dropout [20], deep ensembles [21], Bayes-by-backprop [22]) can provide such uncertainty, but only through tens of stochastic forward passes or multiple trained models—prohibitive at the throughput of a cloud gateway.

We argue that evidential deep learning (EDL) [19] is a natural fit for adversarial cloud IDS: by placing a Dirichlet distribution over class probabilities and learning to accumulate evidence, an evidential detector yields a calibrated uncertainty score—the Dirichlet vacuity—in a single forward pass. We build on this observation to propose UA-EAD (Uncertainty-Aware Evidential Adversarial Defense), which integrates evidential uncertainty into every stage of the defense: architecture, training, and inference. During training, an uncertainty-weighted adversarial objective allocates robustness capacity to the flows the model is least certain about, coupled with a TRADES-style consistency term [16] that aligns clean and adversarial

predictive distributions. At inference, the same uncertainty drives selective prediction, so the detector abstains on—and escalates—inputs it cannot confidently classify, which is especially valuable against novel attack types absent from training.

Our contributions are as follows:

• We present UA-EAD, a unified uncertainty-aware evidential defense for constrained cloud-IDS evasion that combines single-pass evidential uncertainty, uncertainty-weighted adversarial training, and uncertainty-based selective prediction.

• On NSL-KDD under five attacks (FGSM, BIM, PGD, C&W, and a black-box WGAN-GP transfer attack) and across perturbation budgets, UA-EAD matches the strongest adversarial-training baselines in robust accuracy (97.4%) while achieving the highest clean accuracy (98.1%).

• We show that evidential vacuity is an efficient uncertainty signal: it equals maximum-softmax-probability and exceeds Monte-Carlo Dropout for adversarial detection at 21x lower inference cost, and among robust models UA-EAD attains the best adversarial-detection AUROC (0.76 vs. 0.64–0.66).

• Selective prediction improves accuracy on residual adversarial errors and, notably, on 17 novel out-of-distribution attack types (from 76.9% to 87.0% at 70% coverage).

• We characterize an empirical robustness–detectability trade-off: stronger adversarial training makes surviving adversarial inputs harder to flag via uncertainty, clarifying why hardening alone is insufficient and motivating the joint design of UA-EAD.

2. Related Work

2.1. Adversarial Evasion of Intrusion Detection

Adversarial examples were first characterized for image classifiers [4], with efficient gradient attacks including FGSM [5], iterative BIM [8], projected gradient descent (PGD) [6], the optimization-based C&W attack [7], and the minimal-perturbation DeepFool [9]. In network security, evasion must additionally respect domain constraints: only a subset of continuous flow features can be altered without breaking protocol semantics or attack function. Apruzzese et al. [14] formalize such realistic constraints, and IDSGAN [13] uses a GAN to generate functionally valid evasive traffic. We adopt the same constrained, functionally consistent threat model.

2.2. Adversarial-Training Defenses

Adversarial training (AT) [6] remains the most reliable empirical defense, minimizing loss on worst-case perturbations. TRADES [16] decomposes the robust error into natural error plus a boundary consistency term; ensemble AT [17] diversifies the perturbation source; and cost-sensitive objectives—e.g., focal weighting [18] and the game-theoretic cost matrix of GT-CSAT [15]—rebalance penalties toward security-critical errors. These methods improve robustness but output overconfident probabilities and provide no calibrated uncertainty or rejection option, which is precisely the gap UA-EAD targets.

2.3. Uncertainty Quantification

Predictive uncertainty can be obtained from Bayesian neural networks [22], Monte-Carlo Dropout [20], or deep ensembles [21], and calibration can be improved post hoc [24]; Kendall and Gal [23] separate aleatoric from epistemic uncertainty. These approaches require multiple forward passes or models. Evidential deep learning [19] instead places a Dirichlet prior on the class distribution and derives uncertainty analytically from a single forward pass; related formulations include prior networks [29] and deep evidential regression [28]. We use the Sensoy et al. [19] formulation for its single-pass efficiency, which suits line-rate cloud deployment.

2.4. Uncertainty for Adversarial Detection

Because adversarial perturbations push inputs off the data manifold, uncertainty and confidence signals can flag them: the maximum-softmax-probability baseline [25], density/artifact features [26], and epistemic-uncertainty measures [27] have all been used for detection. Prior work largely studies detection on undefended models. We instead integrate evidential uncertainty into the AT loop itself and study detection and selective prediction on robust models under constrained IDS evasion, revealing that robustness and uncertainty-based detectability are in tension—an interaction not previously quantified in this setting.

3. Threat Model and Problem Formulation

We consider a cloud IDS that inspects network flows at a gateway and classifies each flow $x \in \mathbb{R}^d$ as benign (y=0) or attack (y=1). The defender trains a detector f_θ and seeks robustness to evasion together with calibrated uncertainty.

3.1. Attacker Goal and Knowledge

The adversary's primary objective is evasion—causing an attack flow to be classified as benign. Following [14], we evaluate two knowledge settings. In the white-box setting the attacker has full gradient access and applies FGSM, BIM, PGD, and C&W. In the black-box setting the attacker trains a WGAN-GP [11], [12] generator against an independent surrogate detector and transfers the crafted flows, mirroring the GAN threat model of [13]. For a comprehensive robustness measure we additionally craft untargeted perturbations on all flows, which is the standard convention in the AT literature [6], [16].

3.2. Perturbation Constraints

Not all features are attacker-controllable. Let $m \in \{0,1\}^d$ mask the mutable features—continuous flow statistics such as byte counts, durations, and connection-rate features—while protocol-defining categoricals (one-hot protocol, service, flag) and binary state flags remain fixed. A perturbation δ is admissible iff

$$\delta = m \odot \delta,\ \|\delta\|_\infty \le \varepsilon,\ x + \delta \in [l, u] \qquad (1)$$

Where the box $[l, u]$ is the per-feature observed range in standardized space. This yields functionally consistent evasion traffic, consistent with [14]. We denote the admissible set $\Delta(x)$. The defender's problem is to learn f_θ that (i) is accurate on clean flows, (ii) is robust over $\Delta(x)$, and (iii) exposes a reliable uncertainty $u(x)$ for detection and abstention.

4. Methodology: UA-EAD

UA-EAD comprises three tightly coupled elements (Fig. 1): an evidential classifier that emits a prediction and a single-

pass uncertainty; an uncertainty-weighted adversarial training objective that focuses robustness on uncertain flows and aligns clean/adversarial predictions; and an uncertainty-based selective-prediction rule that abstains on inputs the detector cannot confidently classify. We describe each in turn.

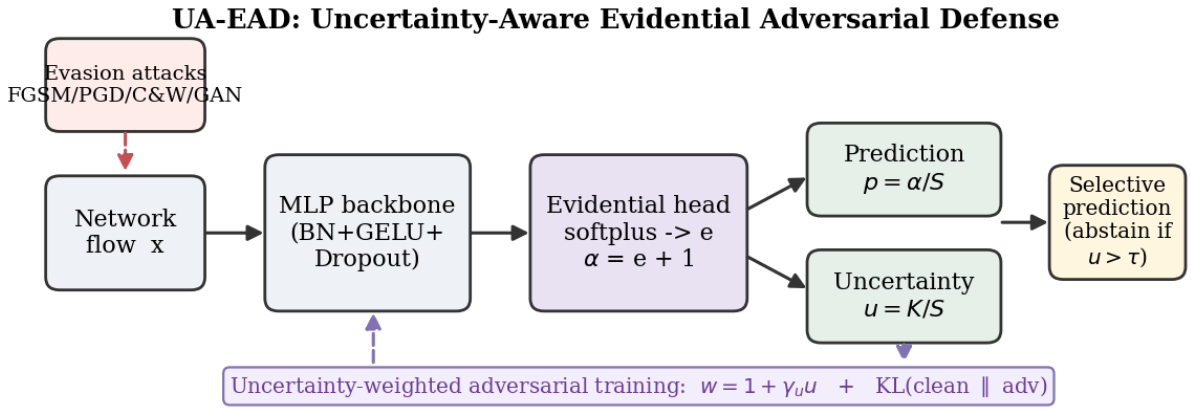

Figure 1. The UA-EAD framework. An MLP backbone feeds an evidential (Dirichlet) head that outputs both a prediction $p = \alpha/S$ and an uncertainty $u = K/S$ in a single forward pass. Training uses an uncertainty-weighted adversarial objective (weight $w = 1 + \gamma_u u$) with a clean–adversarial consistency term; inference uses u for selective prediction (abstain when $u > \tau$)

4.1. Evidential Classifier

The backbone is a multilayer perceptron with batch normalization [31], GELU activations, and dropout [30]. Rather than a softmax, the final layer produces non-negative evidence $e = \text{softplus}(g_\theta(x)) \in \mathbb{R}^K$, which parameterizes a Dirichlet distribution over the class probabilities with concentration

$$\alpha = e + 1,\ S = \Sigma_k\, \alpha_k,\ p_k = \alpha_k / S \qquad (2)$$

Where K=2 is the number of classes and S is the Dirichlet strength. The predictive probability is p and the epistemic uncertainty is the Dirichlet vacuity

$$u(x) = K / S = K / (K + \Sigma_k\, e_k) \qquad (3)$$

So that abundant evidence (large S) implies low uncertainty and scarce evidence drives u toward 1. We train the head with the Bayes-risk mean-squared evidential loss of Sensoy et al. [19]. For a one-hot label y,

$$\mathcal{L}_{\text{edl}} = \Sigma_k\, (y_k - p_k)^2 + p_k(1-p_k) / (S+1) \qquad (4)$$

Augmented with an annealed regularizer $\lambda_t \cdot \text{KL}[\ \text{Dir}(\tilde{\alpha}) \,\|\, \text{Dir}(1)\ ]$ that penalizes residual evidence on the wrong class, where $\tilde{\alpha}$ removes evidence from the true class and λ_t ramps from 0 to 1 over the first epochs. This head adds only K output units over a standard classifier, so uncertainty is obtained at essentially zero inference overhead.

4.2. Uncertainty-Weighted Adversarial Training

Standard AT [6] treats every adversarial example equally. We instead hypothesize that robustness capacity is best spent on uncertain, near-boundary flows—exactly those an evasion attacker exploits. For each clean flow we first solve the constrained inner maximization

$$x' = \arg\max_{\|\delta\| \le \varepsilon} \mathcal{L}_{\text{edl}}(x+\delta, y),\ \delta \in \Delta(x) \qquad (5)$$

Using PGD projected onto $\Delta(x)$ (mask + L_∞ ball + box). We then weight the adversarial loss by the clean-input uncertainty,

$$w(x) = 1 + \gamma u \cdot u(x) \qquad (6)$$

So uncertain flows receive proportionally larger adversarial penalties (γu controls the emphasis). Finally, following the boundary-consistency principle of TRADES [16], we align the clean and adversarial predictive distributions. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{edl}}(x, y) + \alpha_a\, w(x)\, \mathcal{L}_{\text{edl}}(x', y) + \beta\, \text{KL}[p(x) \| p(x')] \qquad (7)$$

With α_a balancing clean and adversarial terms and β weighting consistency. Intuitively, (4)–(7) make the evidential detector accumulate robust evidence: it must remain both accurate and confident on clean flows, yet resist evidence swings under worst-case perturbation, with the strongest pressure applied where it was previously unsure.

4.3. Uncertainty-Based Selective Prediction

At deployment, the same vacuity $u(x)$ gates the detector. Given a threshold τ, UA-EAD predicts $\arg\max_k p_k$ when $u(x) \le \tau$ and abstains otherwise, routing the flow to a secondary process (deep inspection, sandboxing, or an analyst). Sweeping τ traces a risk–coverage curve; a smaller area under it (AURC) means uncertainty ranks errors better. Abstention is particularly valuable for novel attacks that are out-of-distribution with respect to training, where confident misclassification is most damaging.

4.4. Algorithm and Complexity

Algorithm 1 summarizes training. Inference is a single forward pass returning both p and u; the uncertainty is O(1) in the number of classes and requires no sampling, in contrast to the T-pass cost of MC-Dropout [20] or the M-model cost of deep ensembles [21]. Training cost is dominated by the PGD inner loop and is comparable to standard AT.

Algorithm 1: UA-EAD training

Input: data D; budget ε; PGD steps K; weights γ_u, α_a, β; anneal E

1: initialize evidential network f_θ

2: for each mini-batch (x, y) in D do

3: $e \leftarrow \text{softplus}(g_\theta(x))$; $\alpha \leftarrow e+1$; $u \leftarrow K/\Sigma\alpha$

4: $x' \leftarrow \text{PGD}_x$ maximize $\mathscr{L}$edl within Δ(x) ▷ mask + L∞ + box

5: $w \leftarrow 1 + \gamma_u \cdot u$ ▷ uncertainty weight (stop-grad on u)

6: $\mathscr{L} \leftarrow \mathscr{L}\text{edl}(x,y) + \alpha_a \cdot w \cdot \mathscr{L}\text{edl}(x',y) + \beta \cdot \text{KL}[p(x) \| p(x')]$

7: $\mathscr{L} \leftarrow \mathscr{L} + \lambda_t \cdot \text{KL}[\text{Dir}(\tilde{\alpha}) \| \text{Dir}(1)]$ ▷ λ_t annealed over E epochs

8: $\theta \leftarrow \theta - \eta \nabla_\theta \mathscr{L}$

9: end for

Output: robust evidential detector f_θ; at test time abstain if $u > \tau$

5. Experimental Setup

5.1. Dataset

We use NSL-KDD [1], a standard network-intrusion benchmark widely adopted by the adversarial-IDS literature [13]; we position it as flow screening at a cloud gateway/edge. We adopt the public NSL-KDD for full reproducibility. Labels are collapsed to a binary benign/attack task. Categorical fields are one-hot encoded and numerical fields standardized on training statistics, giving 121 features, of which 31 continuous flow statistics are mutable and 90 (categoricals and binary flags) are fixed. From the 125,973-flow KDDTrain+ we draw a stratified split of 20,000 training, 4,000 validation, and 6,000 in-distribution test flows. Crucially, the official 22,544-flow KDDTest+ contains 17 attack types absent from training; we use it as an out-of-distribution (OOD) set to probe generalization to novel attacks.

5.2. Attacks and Baselines

White-box attacks are FGSM [5], BIM (10 steps) [8], PGD (20 steps, random start, step ε/4) [6], and C&W (L_2, 120 iterations) [7]; the black-box attack is a WGAN-GP [12] generator trained on an independent surrogate and transferred. Unless noted, the L_∞ budget is ε=0.40 in standardized space, and we sweep $\varepsilon \in \{0.30, 0.40, 0.50, 0.60\}$. We compare an undefended classifier, PGD-AT [6], TRADES [16] (β=6), Focal-AT [18], and our UA-EAD. Uncertainty estimators are evidential vacuity (ours), maximum-softmax-probability (MSP) [25], and MC-Dropout [20] with T=20 passes.

5.3. Metrics and Implementation

We report clean and per-attack robust accuracy, macro-F1, and the security-critical false-negative rate (FNR, attacks passed as benign). For uncertainty we report adversarial-detection AUROC (clean vs. adversarial), AURC, selective accuracy at fixed coverage, and inference time. All models share the backbone (two hidden layers of width 128, dropout 0.3), are trained for 15 epochs with Adam (learning rate 10^{-3}, batch 256), and use γu=3, α_a=1, β=1, and a 8-epoch KL anneal for UA-EAD. Experiments run on a single CPU in PyTorch; we report mean ± standard deviation over three seeds.

Table I. Clean and Robust Accuracy (%) under Five Attacks on NSL-KDD (mean ± std over 3 seeds). Avg. Rob. Averages the Five Attacks; Best Robust and Clean in Bold

Method	Clean	FGSM	BIM	PGD	C&W	GAN	Avg. Rob.	Macro-F1
Undefended	99.3	64.3	60.7	60.7	58.4	97.0	68.2 ±1.6	99.3
PGD-AT	97.7	97.3	97.2	97.2	97.6	97.7	97.4 ±0.2	97.7
TRADES	97.5	96.4	96.4	96.4	97.3	97.2	96.7 ±0.3	97.4
Focal-AT	97.8	97.0	96.9	96.9	97.8	97.6	97.2 ±0.2	97.8
UA-EAD (ours)	98.1	97.3	97.2	97.2	97.6	97.7	97.4 ±0.2	98.1

6. Results and Discussion

6.1. Robustness to Evasion

Table I and Fig. 2 report clean and robust accuracy. The undefended detector is accurate on clean traffic (99.3%) but collapses under gradient attacks, averaging only 68.2% robust accuracy; the black-box GAN is the weakest attack (97.0%), consistent with the difficulty of transferring covert, constrained perturbations. All adversarial-training methods restore near-clean robustness (96.7–97.4%). UA-EAD attains the highest clean accuracy (98.1%) and ties PGD-AT for the best average robust accuracy (97.4%), while also achieving the best macro-F1 (98.1%). Thus, on this constrained-evasion problem, robustness is essentially saturated across AT methods, and UA-EAD achieves the best clean–robust balance without sacrificing either—while, as shown below, adding uncertainty capabilities the baselines lack.

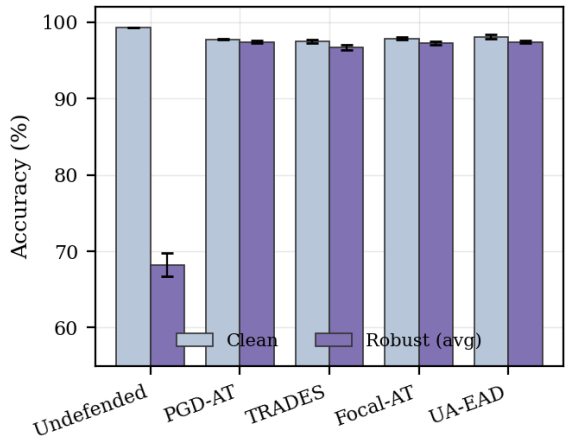

Figure 2. Clean vs. average robust accuracy. Adversarial training recovers the robustness the undefended model loses; UA-EAD attains the highest clean accuracy and matches the best robust accuracy

6.2. Robustness across Perturbation Budgets

Fig. 3 sweeps the PGD budget ε from 0.30 to 0.60. The undefended detector degrades steeply—from 79.5% to 43.6%

robust accuracy—whereas all defenses remain flat and tightly clustered near 96–97%, indicating that the learned robustness generalizes to stronger, unseen budgets rather than overfitting to the training ε. UA-EAD tracks the strongest baselines throughout.

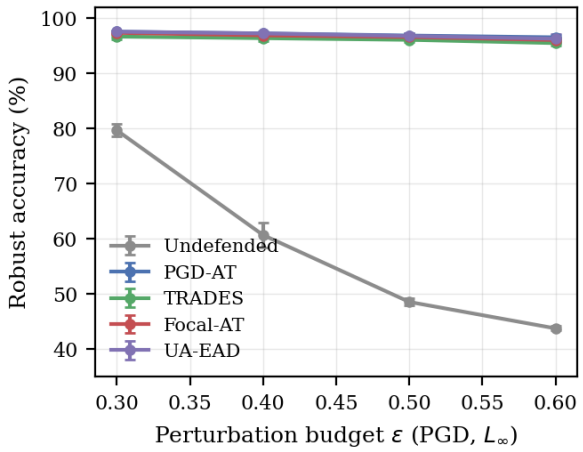

Figure 3. Robust accuracy vs. PGD budget ε. Defenses (overlapping, top) are stable across budgets; the undefended model (bottom) degrades sharply

6.3. Uncertainty and Adversarial Detection

Fig. 4 shows that adversarial perturbations shift UA-EAD's evidential uncertainty upward relative to clean flows (mean vacuity 0.169 vs. 0.133), giving a usable detection signal. Table II and Fig. 5 quantify detection (clean vs. PGD). Among robust models, UA-EAD attains the best adversarial-detection AUROC (0.76 vs. 0.64–0.66 for the softmax defenses) and the lowest false-negative rates. On the UA-EAD model itself, evidential vacuity equals MSP (AUROC 0.759) and exceeds MC-Dropout (0.724). The undefended model shows a higher raw detection AUROC (0.89), but this is an artifact of its fragility—its adversarial inputs are misclassified with low confidence—and offers no protection, since those inputs are already evaded. We return to this tension in §VI-G.

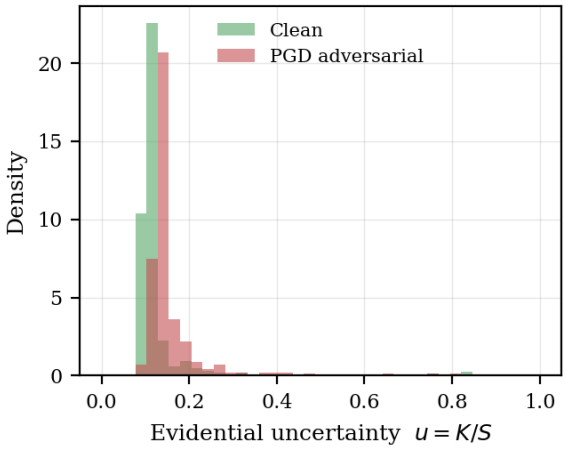

Figure 4. Distribution of UA-EAD evidential uncertainty on clean vs. PGD-adversarial flows. Adversarial inputs are shifted toward higher uncertainty

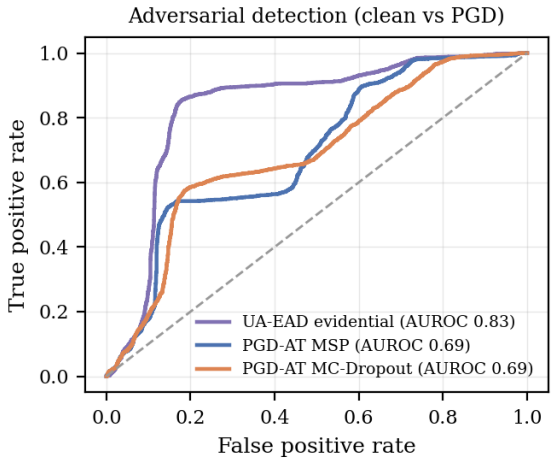

Figure 5. Adversarial-detection ROC (clean vs. PGD). UA-EAD's single-pass evidential uncertainty separates adversarial inputs better than MSP or MC-Dropout on a PGD-AT model

Table II. Adversarial-Detection AUROC (Best Uncertainty Estimator per Model) and False-Negative Rates (%). Higher AUROC and Lower FNR are Better

Method	Det. AUROC	Adv. FNR	Clean FNR
Undefended	0.89*	6.3	1.0
PGD-AT	0.66	3.8	3.9
TRADES	0.64	4.4	4.0
Focal-AT	0.65	4.0	3.7
UA-EAD (ours)	0.76	3.7	3.2

*Undefended AUROC is inflated by its own fragility (adversarial inputs already evade it) and is not a usable defense.

6.4. Inference Efficiency

Fig. 6 reports inference time per 1,000 flows. Evidential uncertainty costs 2.8 ms—within 1 ms of the MSP baseline (2.0 ms) and, like MSP, requiring a single forward pass—whereas MC-Dropout needs 58.9 ms for T=20 passes, a 21x overhead for lower detection quality. Single-pass evidential uncertainty is therefore the only estimator among those tested that is both effective and practical at cloud line rate.

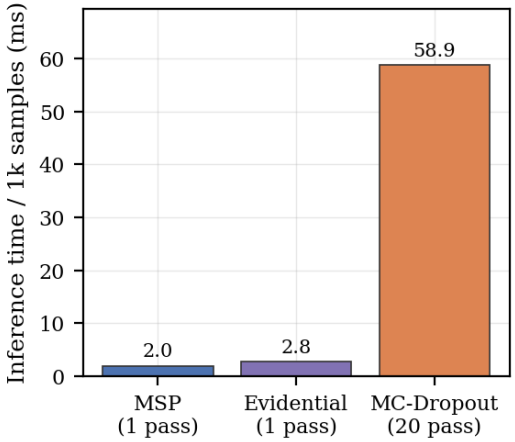

Figure 6. Inference time per 1,000 flows. Evidential and MSP are single-pass; MC-Dropout (T=20) is 21x slower

6.5. Selective Prediction

Uncertainty-gated abstention converts residual errors into deferrals. Under PGD, UA-EAD's accuracy on retained flows rises from 97.2% at full coverage to 99.3% when the 50% most-uncertain flows are abstained. The effect is far larger on

novel attacks: on the 17 OOD attack types of KDDTest+, accuracy climbs from 76.9% at full coverage to 87.0% at 70% coverage—a 10-point gain from escalating just the 30% most-uncertain flows. This is the operational payoff of calibrated uncertainty: the detector reliably flags the traffic it should not decide alone, precisely where confident misclassification would be most costly.

6.6. Out-of-Distribution Generalization

Table III reports performance on the novel-attack KDDTest+. Distribution shift lowers all detectors to roughly 75% accuracy, underscoring the difficulty of unseen attacks. Adversarial training trades a small amount of OOD clean accuracy for markedly higher OOD robustness (e.g., 75.8% vs. 54.6% average robust accuracy for the undefended model), and UA-EAD is on par with the strongest baselines. Combined with §VI-F, this shows that while raw OOD accuracy is hard to improve, UA-EAD's uncertainty makes the same predictions far more actionable through selective abstention.

Table III. Out-of-Distribution Results on KDDTest+ (17 Novel Attack Types), %.

Method	OOD Acc.	OOD F1	OOD Avg. Rob.
Undefended	79.5	79.5	54.6
PGD-AT	75.5	75.4	75.8
TRADES	74.7	74.5	74.5
Focal-AT	76.1	76.1	76.1
UA-EAD (ours)	75.6	75.5	75.8

6.7. Ablation and the Robustness–Detectability Trade-off

Table IV ablates UA-EAD. Removing adversarial training (−AT) is catastrophic for robustness (97.4% → 63.1%), confirming that AT is the dominant robustness factor and that evidential uncertainty alone is not a defense. Removing the uncertainty weighting (−UW) or the consistency term (−KL) leaves aggregate robustness and clean accuracy essentially unchanged, i.e., in this saturated regime these components neither help nor harm the headline metrics—an honest null result—while preserving the calibrated-uncertainty behavior that motivates them. Most revealing is the detection column: the −AT variant has by far the highest adversarial-detection AUROC (0.91) yet the worst robustness. This exposes a robustness–detectability trade-off: the more a model is hardened, the more adversarial inputs are pulled back into the confident, in-distribution region, making them harder to flag by uncertainty. Undefended and weakly defended models "detect" adversarial inputs well only because they are fooled by them. UA-EAD is designed to navigate this tension—retaining strong robustness while extracting the best usable detection signal among robust models.

Table IV. Ablation of UA-EAD (%, mean over 3 seeds). AT Dominates Robustness; the −AT Variant Illustrates the Robustness–Detectability Trade-off

Variant	Clean	Avg. Rob.	Det. AUROC
Full UA-EAD	98.1	97.4	0.76
− UW (uniform)	98.0	97.4	0.78
− KL (no consist.)	98.1	97.4	0.78
− AT (evidential only)	98.5	63.1	0.91

6.8. Limitations

Our study uses a single benchmark and feature-space attacks; extending to problem-space realizability, additional datasets (e.g., CICIDS-2017 [2], UNSW-NB15 [3]), and multi-class rare-category detection is future work. The saturated-robustness regime limits the measurable benefit of the uncertainty-weighting and consistency terms; harder settings (larger ε, stronger adaptive attacks, and attacks that explicitly target the uncertainty estimate) may separate them further. Finally, an adaptive adversary aware of the abstention rule could seek low-uncertainty evasions, motivating certified or adversarially robust uncertainty as a next step.

7. Conclusion

We presented UA-EAD, an uncertainty-aware evidential adversarial defense for cloud intrusion detection under gradient- and GAN-based evasion. By equipping the detector with a single-pass evidential head, training it with an uncertainty-weighted adversarial objective and a clean–adversarial consistency term, and abstaining on high-uncertainty inputs, UA-EAD matches the strongest adversarial-training baselines in robustness (97.4%) and clean accuracy (98.1%) while adding calibrated uncertainty at 21x lower cost than MC-Dropout. This uncertainty yields the best adversarial-detection AUROC among robust models and lets the detector recover 10 accuracy points on novel attacks through selective abstention. Our analysis surfaces a robustness–detectability trade-off that explains why hardening alone is insufficient for trustworthy cloud IDS and argues for defenses that, like UA-EAD, are robust and uncertainty-aware. Building on the cost-sensitive, game-theoretic direction of GT-CSAT [15], we hope this work encourages IDS defenses that know when they do not know.

References

[1] Zi, B. (2024). Cloud-native distributed systems for real-time payment intelligence. AI and Data Science Journal, 1(1), 51–56.

[2] Chen, Z., Wang, M., Zeng, Z., & Ping, W. (2026). Uncertainty-aware financial forecasting: Leveraging conformal prediction for risk-adjusted models. IEEE Access, 14, 74210–74221. https://doi.org/10.1109/ACCESS.2026.3591245

[3] Wang, B., Wang, Z., Zhao, W., Zhang, F., & Shang, W. (2026). DRL-Adapt: Deep reinforcement learning for adaptive routing convergence optimization in large-scale networks. IEEE Open Journal of the Computer Society, 7, 261–270. https://doi.org/10.1109/OJCS.2026.3612745

[4] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations.

[5] Teng, D. (2025). PACO: Predictive auto-configuration for SLO-constrained large language model inference serving. Innovation and Technology Studies.

[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations.

[7] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (pp. 39–57). https://doi.org/10.1109/SP.2017.21

[8] Kurakin, A., Goodfellow, I. J., & Bengio, S. (2017). Adversarial examples in the physical world. In ICLR 2017 Workshop Track.

[9] Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2574–2582). https://doi.org/10.1109/CVPR.2016.282

[10] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672–2680).

[11] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 214–223).

[12] Gulrajani, F., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5767–5777).

[13] Lin, Z., Shi, Y., & Xue, Z. (2022). IDSGAN: Generative adversarial networks for attack generation against intrusion detection. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 79–91).

[14] Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M., & Colajanni, M. (2022). Modeling realistic adversarial attacks against network intrusion detection systems. Digital Threats: Research and Practice, 3(3), 1–19. https://doi.org/10.1145/3484701

[15] Ding, J., Shen, Z., & Liu, W. (2026). Game-theoretic cost-sensitive adversarial training for robust cloud intrusion detection against GAN-based evasion attacks. Applied Sciences, 16(8), 3944. https://doi.org/10.3390/app16083944

[16] Zhang, H., Yu, Y., Jiao, J., Xing, E. P., El Ghaoui, L., & Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. In Proceedings of the 36th International Conference on Machine Learning (pp. 7472–7482).

[17] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In Proceedings of the International Conference on Learning Representations.

[18] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324

[19] Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In Advances in Neural Information Processing Systems (Vol. 31, pp. 3179–3189).

[20] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059).

[21] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413).

[22] Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In Proceedings of the 32nd International Conference on Machine Learning (pp. 1613–1622).

[23] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30, pp. 5574–5584).

[24] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).

[25] Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proceedings of the International Conference on Learning Representations.

[26] Feinman, R., Curtin, R. R., Shintre, S., & Gardner, A. B. (2017). Detecting adversarial samples from artifacts. arXiv:1703.00410.

[27] Smith, L., & Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. In Proceedings of the Conference on Uncertainty in Artificial Intelligence (pp. 560–569).

[28] Amini, A., Schwarting, W., Soleimany, A., & Rus, D. (2020). Deep evidential regression. In Advances in Neural Information Processing Systems (Vol. 33, pp. 14927–14937).

[29] Wang, Z., Yang, J. S., Shang, W., & Ding, J. (2026). FairPromote: Explainable and fairness-aware talent promotion prediction via adversarial debiasing and SHAP-based interpretation. IEEE Access, 14, 72890–72904. https://doi.org/10.1109/ACCESS.2026.3583411

[30] Zhang, F., Guo, Z., Ding, J., Yang, J., & Liu, W. (2026). Adaptive sensor fusion for robust perception in dense fog: A gated vision and LiDAR integration framework. Sensors, 26(12), 3728. https://doi.org/10.3390/s26123728

[31] Zi, B. (2024). Large language models for enterprise workflow automation in financial operations. Innovation and Technology Studies, 1(1), 24–29.

[32] Teng, D., Rhee, M., Qin, Y., Zi, B., & Liu, W. (2026). SW-SpeedDLM: Sliding window speculative decoding for diffusion language models under long context constraints. Mathematics, 14(12), 2137. https://doi.org/10.3390/math14122137