Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080874

ARTICLE

HADAR-UAV: Risk-Calibrated One-Class Learning Framework for Zero-Day Intrusion Detection in Unmanned Aerial Vehicle Networks

Canan Batur Şahin[1,*], Siti Fatimah Abdul Razak[2,*], Arif Ullah[2], Ali Fatih Gündüz[1] and Nazri Mohd Nawi[3]

[1]Software Engineering Department, Malatya Turgut Özal University, Malatya, Turkey
[2]Centre for Intelligent Cloud Computing, COE for Advanced Cloud, Multimedia University, Melaka, 75450, Malaysia
[3]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia

*Corresponding Authors: Canan Batur Şahin. Email: canan.batur@ozal.edu.tr; Siti Fatimah Abdul Razak. Email: fatimah.razak@mmu.edu.my

Received: 27 February 2026; Accepted: 03 May 2026; Published: 13 August 2026

ABSTRACT: Unmanned Aerial Vehicle (UAV) networks face escalating cybersecurity threats, especially from zero-day attacks that exploit previously unknown vulnerabilities. To address this, we present HADAR-UAV (Hybrid Anomaly Detection with Adaptive Risk-calibration for UAV). This novel intrusion detection framework integrates masked autoencoder representation learning with Deep Support Vector Data Description (Deep SVDD) under conformal prediction guarantees to calibrate risk. Our method overcomes three critical limitations of existing approaches: (i) over-reliance on attack signatures, (ii) lack of statistical guarantees on false alarm rates, and (iii) insufficient robustness in feature extraction under partial observation. Using a rigorous Leave-Two-Attack-Families-Out (L2AFO) evaluation protocol on the UAVIDS-2025 benchmark, HADAR-UAV achieves strong zero-day detection—0.997 ± 0.001 ROC-AUC and 0.992 ± 0.002 F1-Score—while empirically maintaining a target false alarm rate through conformal calibration applied to deterministic scores. All results are reported as mean ± standard deviation across 20 independent runs (5 seeds × 4 folds) and show statistically significant improvement (paired t-test, $p < 0.01$) over current one-class methods. Ablation studies confirm that every architectural component adds measurable value to the framework. Additional cross-dataset validation on NSL-KDD under a one-class zero-day-inspired setting further indicates that the proposed framework generalizes beyond MAVLink-specific traffic patterns.

KEYWORDS: UAV cybersecurity; intrusion detection; zero-day attacks; one-class classification; masked auto encoder

1 Introduction

The growing use of UAVs in civilian, commercial, and military sectors brings new cybersecurity challenges. UAVs are complex cyber-physical systems that connect ground stations, satellites, and flight controllers via protocols such as MAVLink. This complexity increases the number of attack surfaces, putting mission integrity and aircraft control at risk. Signature-based IDS struggles to detect zero-day attacks. Aviation security demands low false alarms to avoid costly mission aborts and minimize missed detections that can bring severe damage. Detection systems must both identify novel threats and guarantee controlled false-alarm behavior. Recent advances in deep anomaly detection machine learning methods that identify rare deviations from normal data offer promising directions through one-class classification approaches that learn compact representations of normal behavior. Deep Support Vector Data Description (Deep SVDD) extends classical SVDD an algorithm that finds the smallest hypersphere containing all normal data points

by learning a neural network mapping that minimizes the volume of the hypersphere containing normal data representations [1,2]. However, existing implementations suffer from representation collapse, in which the model encodes most inputs similarly, losing important distinctions, especially when trained without auxiliary objectives or mechanisms to control the false alarm rate. Conformal prediction is a statistical framework that provides distribution-free uncertainty quantification and offers finite-sample guarantees on prediction set coverage. When applied to anomaly detection, conformal calibration allows precise control over the false alarm rate, meaning the system can be tuned to produce a specified number of false alarms an important requirement in environments where mistakes carry significant risk [3]. This paper introduces HADAR-UAV, a hybrid framework that addresses these challenges through three synergistic innovations. First, we employ masked auto encoder (MAE) pre-training a training strategy in which parts of the input data are masked, and the model learns to reconstruct them to learn robust feature representations that capture the semantic structure of normal UAV network flows while preventing the representation collapse that plagues pure one-class objectives. Second, we integrate Deep SVDD regularization during training to enforce a compact hyper spherical embedding of normal data, enabling effective separation of anomalies in the learned latent space. Third, we apply conformal prediction for threshold calibration, which gives formal guarantees that the false alarm rate will not exceed a user-specified level α with high probability [3]. We evaluate HADAR-UAV using a rigorous Leave-Two-Attack-Families-Out (L2AFO) protocol, which means two complete attack families (Sybil and Flooding) are held out during training to simulate realistic zero-day scenarios in which the detector encounters previously unseen attack types. This evaluation methodology avoids data leakage that can occur with random train-test splits and provides unbiased estimates of zero-day detection performance. Our contributions are summarized as follows:

(a) To the best of our knowledge, this work is among the first to systematically integrate conformal calibration with deep one-class classification in UAV intrusion detection.
(b) This study presents conformal prediction-based threshold calibration, providing formal guarantees on the false alarm rate and representing the first application of conformal techniques to UAV intrusion detection.
(c) Masked Expectation Anomaly Scoring introduces stochastic feature perturbations to improve detection robustness.
(d) A comprehensive evaluation using the L2AFO protocol demonstrates that HADAR-UAV achieves an ROC-AUC of 0.997 ± 0.001 on zero-day attacks while adhering to a false alarm rate of 4.72 ± 0.40%, validated over 20 independent runs (5 random seeds × 4 cross-validation folds).
(e) To examine cross-domain generalizability, we further evaluate HADAR-UAV on the NSL-KDD [4] benchmark under a one-class zero-day-inspired setting, showing that the proposed framework generalizes beyond MAVLink-specific UAV traffic.

2 Related Work

2.1 UAV Security and Intrusion Detection

UAV network security has emerged as a critical research area as the deployment of autonomous aerial systems increases. Early work focused on physical-layer attacks, including GPS spoofing and jamming [5], while recent attention has shifted to network-layer threats targeting communication protocols. Sedjelmaci et al. [6] proposed hierarchical detection architectures for UAV networks, though their approach relies on attack signatures, limiting zero-day detection capability. Mitchell and Chen [7] surveyed intrusion detection techniques for cyber-physical systems, identifying the need for anomaly-based methods that can detect novel attacks without prior signatures. Recent UAV intrusion detection research increasingly explores (i) compact supervised IDS models optimized for embedded constraints (e.g., distilled/pruned architectures) and (ii)

foundation-model driven recognition that aims to generalize to unseen attacks with zero/few labeled examples. However, these lines of work typically optimize for accuracy or semantic matching rather than providing explicit finite-sample false-alarm control, which remains critical for operational UAV deployments with limited human-in-the-loop capacity [8]. Recent studies have also emphasized the importance of domain-specific datasets for UAV intrusion detection. For instance, [9] provides a comprehensive analysis of datasets for intra- and inter-UAV communication systems, highlighting the need for realistic and protocol-aware benchmarks in UAV security. Recent comprehensive surveys further highlight the growing importance of AI-driven intrusion detection in UAV systems. In particular, recent work has systematically analyzed machine learning and deep learning techniques for UAV security, covering both network-layer and system-level threats [10]. These studies emphasize the need for robust anomaly detection mechanisms capable of handling dynamic and heterogeneous UAV environments, reinforcing the relevance of our risk-calibrated approach.

2.2 Deep Anomaly Detection

Deep learning approaches to anomaly detection have advanced significantly in recent years [11]. Auto encoders learn compressed representations and detect anomalies through reconstruction error [12,13], while variational auto encoders (VAEs) [12] provide probabilistic frameworks for anomaly scoring. Deep SVDD [1] extends classical support vector data description [14] by learning neural network features that map normal data to a compact hypersphere. However, in [1,2] identified that pure one-class objectives can suffer from representation collapse, motivating hybrid approaches that combine reconstruction and one-class losses. Masked auto encoders [15] have demonstrated remarkable success in self-supervised representation learning for computer vision. By randomly masking input patches and training to reconstruct the missing content, MAE learns rich semantic features without requiring labeled data. In [16] applied transformer-based masked reconstruction to multivariate time series anomaly detection, demonstrating the effectiveness of masking strategies for temporal data. In addition, anomaly-based approaches such as Isolation Forest have been successfully applied to UAV security problems, including GPS spoofing detection [5], demonstrating the effectiveness of one-class models in detecting previously unseen attack patterns. Recent studies have also explored masked autoencoder-based approaches for encrypted traffic analysis, demonstrating that masking-based representation learning can effectively capture hidden patterns in partially observable network data [17]. This supports our design choice of employing masked autoencoder pre-training for UAV network anomaly detection.

2.3 Conformal Prediction for Anomaly Detection

Conformal prediction [3] provides distribution-free uncertainty quantification with finite-sample validity guarantees. Initially developed for classification and regression, conformal methods have been extended to anomaly detection settings [18]. In [18], conformal anomaly detection was applied to spatio-temporal data with missing observations, demonstrating robustness to incomplete data, a common challenge in network monitoring. Romano et al. [19] developed conformalized quantile regression, enabling the construction of prediction intervals with guaranteed coverage. The application of conformal prediction to network intrusion detection remains limited. Existing approaches [8] typically focus on either deep learning method without statistical guarantees or statistical methods without the representational power of deep networks. HADAR-UAV bridges this gap by integrating conformal calibration with deep one-class learning, providing strong detection performance and formal false-alarm rate guarantees [3].

2.4 Broader UAV Anomaly Detection Context

While this work focuses on network-layer intrusion detection in MAVLink-based UAV communication, anomaly detection in UAV systems extends beyond network traffic to include physical-layer and sensor-based monitoring. Prior studies have explored anomaly detection using onboard sensor data such as IMU readings, GPS signals, and flight dynamics, targeting system-level faults and physical-layer attacks such as spoofing and jamming. These approaches differ fundamentally from network-based intrusion detection, as they operate on continuous sensor streams rather than discrete communication flows. However, they are complementary in nature: sensor-based methods detect physical anomalies, while network-based methods detect communication-level threats. In addition, recent research on encrypted traffic analysis using masked autoencoders highlights the effectiveness of self-supervised learning in extracting robust representations from partially observable data. This aligns with the design of HADAR-UAV, which leverages masked reconstruction to improve feature robustness under uncertainty.

Therefore, HADAR-UAV is positioned as a network-layer anomaly detection framework that complements existing physical-layer and sensor-based approaches, contributing to a more comprehensive UAV security architecture.

3 Methodology

3.1 Notation and Problem Formulation

Let $X = \{x_1, x_2, ..., x_n\}$ denote a dataset of n network flow feature vectors, where each $x_i \in \mathbb{R}^d$ represents a d-dimensional flow with d = 19 features extracted from MAVLink protocol communications. We assume access to a training set X_{train} consisting solely of benign (normal) flows, reflecting the one-class learning paradigm in which attack examples are unavailable during training. The goal is to learn a scoring function $s: \mathbb{R}^d \to \mathbb{R}^+$ that assigns low scores to expected flows and high scores to anomalous flows, along with a threshold τ such that the false alarm rate $P(s(x) > \tau \mid x \text{ is normal}) \leq \alpha$ for a user-specified significance level α. Rather than treating each flow as an isolated vector, we model network traffic as short temporal sequences. Specifically, $x_i \in \mathbb{R}^{T\times 19}$ represents a sequence of T consecutive MAVLink flow records aggregated within a fixed time window. This formulation enables the LSTM encoder to capture temporal dependencies across successive flows, which is critical for identifying coordinated or stealthy UAV attacks. Table 1 present the notation summary.

Table 1: Notation summary.

Symbol	Description	Value/Range
x_i	Input flow sequence	$\mathbb{R}^{T\times 19}$
d	Feature dimensionality	19
n	Number of training samples	15,703 benign
M	Binary masking vector	$\{0, 1\}^d$
ρ	Masking ratio	0.3 (train), 0.15 (test)
l	Latent space dimensionality	64
K	Monte Carlo samples for MEAS	5
c	SVDD hypersphere center	$\mathbb{R}^1$
λ_{rec}	Reconstruction loss weight	1.0
λ_o	One-class loss weight	0.1
α	Target false alarm rate	0.05
τ	Conformal detection threshold	Calibration-derived

(Continued)

Table 1 (continued)

Symbol	Description	Value/Range
E_φ	LSTM encoder network	$\mathbb{R}^d \to \mathbb{R}^1$
D_φ	LSTM decoder network	$\mathbb{R}^1 \to \mathbb{R}^d$

3.2 Masked Autoencoder Architecture

The LSTM-based encoder–decoder operates on short temporal sequences of network flows, allowing the model to learn both intra-flow feature correlations and inter-flow temporal dynamics. Our encoder-decoder architecture employs Long Short-Term Memory (LSTM) [20] networks to capture temporal dependencies in network flow sequences. The encoder E_φ: $\mathbb{R}^d \to \mathbb{R}^1$ maps input flows to a latent representation $z \in \mathbb{R}^1$ with $l = 64$, while the decoder D_φ: $\mathbb{R}^1 \to \mathbb{R}^d$ reconstructs the original input from the latent code.

During training, we apply random binary masking $M \in \{0, 1\}^d$ to input features with masking ratio $\rho = 0.3$, creating partially observed inputs $\tilde{x} = x \odot M$, where $\odot$ denotes element-wise multiplication. The model learns to reconstruct the complete input from this corrupted observation, forcing the encoder to learn robust features that capture the underlying data distribution rather than memorizing specific input patterns [15].

3.3 Deep SVDD Integration

To prevent the representation collapse that occurs with pure reconstruction objectives, we integrate Deep SVDD regularization [1] into our training procedure. After an initial warm-up phase of 10 epochs using only reconstruction loss, we compute the hypersphere center c as the mean of encoder outputs over the training set:

$$c = \frac{1}{n}\sum_{i=1}^{n} E_\varphi(x_i) \tag{1}$$

The one-class loss penalizes the squared distance from each latent representation to this center:

$$L_o(\varphi; X) = \frac{1}{n}\sum_{i=1}^{n} \|E_\varphi(x_i) - c\|^2 \tag{2}$$

This regularization encourages the encoder to map all normal data points to a compact region around c, enabling anomaly detection based on distance from this learned center [1].

3.4 Unified Risk-Calibrated Objective

The complete training objective combines masked reconstruction loss with one-class regularization:

$$L\_total(\varphi, \psi; X) = \lambda_{rec} \cdot L_{rec} + \lambda_o \cdot L_o \tag{3}$$

Expanding with the specific loss formulations:

$$L_{total} = \lambda_{rec}(1/n)\Sigma_i \| (1 - M_i) \odot \left(x_i - D_\varphi\left(E_\varphi(x_i \odot M_i)\right)\right) \|^2 + \lambda_o \cdot (1/n)/\Sigma_i \|E_\varphi(x_i) - c\|^2 \tag{4}$$

where the reconstruction loss is computed only over masked positions $(1 - M)$, forcing the model to learn predictive features rather than identity mappings. The hyperparameters $\lambda_{rec} = 1.0$ and $\lambda_o = 0.1$ were

determined through grid search. The combined anomaly score integrates both reconstruction and one-class components [1].

$$s(x) = \lambda_{rec} \cdot \|x - D_\varphi\left(E_\varphi(x)\right)\|^2 + \lambda_o \cdot \|E_\varphi(x) - c\|^2 \tag{5}$$

3.5 Masked Expectation Anomaly Scoring (MEAS)

At inference time, we employ Monte Carlo sampling with a reduced masking ratio, $\rho_{in}{}^{f} = 0.15$, to compute robust anomaly scores. For $K = 5$ stochastic forward passes with independent random masks, the MEAS score averages the individual scores:

$$sMEAS(x) = (1/K)\sum_{k=1}^{K} s(x; M_k) \tag{6}$$

This stochastic scoring provides two benefits: (1) it smooths the score function, reducing sensitivity to specific feature subsets, and (2) it enables implicit ensemble averaging that improves calibration and reduces variance in anomaly rankings [11].

3.6 Conformal Calibration with Deterministic Threshold

We apply conformal prediction [3] to calibrate the detection threshold τ, providing finite-sample guarantees on the false alarm rate. Given a held-out calibration set $X_{ali\beta} = \{x_1, ..., x_m\}$ of m benign samples disjoint from training data, we compute calibration scores using deterministic scoring [3].

$$\tau = s_{(k)}, where\ k = \lceil (m+1)(1-\alpha) \rceil \tag{7}$$

The final classification rule applies MEAS scoring at test time:

$$\hat{y}(x) = \mathbb{1}\left[sMEAS(x) > \tau\right] \tag{8}$$

Proposition 1 (False Alarm Rate Control under Deterministic Scoring): *Assume that calibration and test benign samples are exchangeable. Then, for the conformal threshold τ, the probability of a false alarm satisfies: P(s(x) > τ | x is benign) ≤ α. Remark 1 (Conformal Validity under MEAS Inference). Conformal calibration in HADAR-UAV is performed using the deterministic anomaly score s(x) defined in Eq. (5), for which finite-sample false-alarm guarantees hold under the standard exchangeability assumption between calibration and benign test samples. At inference time, however, anomaly decisions are based on the stochastic MEAS score sMEAS(x), which aggregates multiple masked forward passes to enhance robustness. Consequently, the formal conformal guarantee applies strictly to the deterministic score, while false-alarm behavior under MEAS inference should be interpreted empirically. As demonstrated in Section 5.2, the calibrated threshold remains stable under stochastic masking, indicating that MEAS does not materially compromise practical false-alarm control [19].*

3.7 Leave-Two-Attack-Families-Out (L2AFO) Evaluation Protocol

To rigorously evaluate zero-day detection capability, we employ Leave-Two-Attack-Families-Out (L2AFO) cross-validation. The UAVIDS-2025 dataset contains four distinct attack families: Blackhole, Wormhole, Sybil, and Flooding. In our primary evaluation, we hold out Sybil and Flooding attacks entirely, using only Blackhole and Wormhole attacks (along with normal traffic) for validation. At the same time, Sybil and Flooding serve as completely unseen zero-day attacks at test time. The benign data is partitioned using session-aware group splitting to prevent temporal leakage: 60% for training, 20% for conformal calibration, and 20% for testing. Importantly, the use of attack samples in the validation phase does not violate the one-class training paradigm. Model parameters are optimized exclusively using benign data. Validation attacks

are used solely for early stopping and hyperparameter selection, ensuring that no attack information leaks into the learned representation [1].

3.8 HADAR-UAV Framework Overview

Fig. 1 illustrates the end-to-end HADAR-UAV framework, which integrates masked representation learning, one-class modeling, robust anomaly scoring, and statistical risk calibration into a unified intrusion detection pipeline.

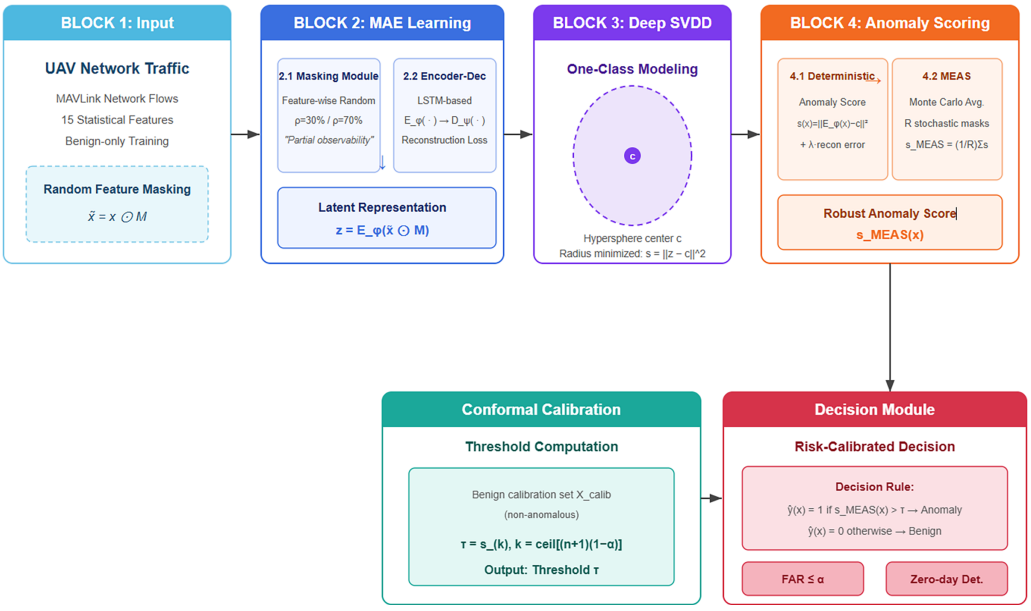

Figure 1: End-to-end architecture of the proposed HADAR-UAV framework.

The framework integrates masked representation learning, one-class modeling, robust anomaly scoring, and statistical risk calibration. The overall procedure is outlined in Algorithm 1.

Algorithm 1: HADAR-UAV training and inference

Input: Training set X_train (benign), calibration set X_calib, test sample x
Output: Anomaly decision $\hat{y}(x)$
1: Train MAE encoder–decoder with masking ratio ρ_train
2: Compute SVDD center c after warm-up
3: Optimize L_total using Eq. (4)
Calibration:
4: Compute deterministic scores $s(x)$ for $x \in$ X_calib
5: Set threshold τ using Eq. (7)
Inference:
6: For $k = 1$ to K:
7: Sample mask M_k and compute $s(x;$ M_k)

(Continued)

Algorithm 1 (continued)

8: Compute s_MEAS(x) using Eq. (6)
9: Return $\hat{y}(x) = 1$ [s_MEAS(x) > τ]

4 Experimental Setup

4.1 Dataset Description

We evaluate HADAR-UAV on the UAVIDS-2025 benchmark dataset [21], which comprises 122,171 network flow records captured from a multi-UAV testbed operating under the MAVLink protocol. The dataset includes 26,172 benign flows and 95,999 attack flows across four families: Blackhole (26,110), Wormhole (26,086), Sybil (24,077), and Flooding (19,726). Each flow is characterized by 19 features, including packet timing, byte counts, protocol flags, and connection statistics. The 19 features used in this study correspond to the complete set of flow-level attributes provided in the UAVIDS-2025 dataset [22], and no features were excluded during preprocessing. The dataset was constructed with domain expertise to capture MAVLink-specific communication characteristics, including timing statistics, packet-level behavior, and protocol information. As such, the feature space is inherently complete within the scope of the benchmark and does not rely on manual feature selection.

In addition to UAVIDS-2025, we also evaluate the proposed framework on the NSL-KDD benchmark to assess cross-dataset generalizability beyond MAVLink-based UAV traffic. NSL-KDD is a widely used network intrusion detection dataset that includes normal traffic and multiple attack categories, with feature characteristics and protocol structures that differ substantially from those of UAVIDS-2025. This additional evaluation is intended not as a direct UAV-specific comparison, but as a domain-shifted validation of whether HADAR-UAV learns protocol-agnostic anomaly representations rather than relying on MAVLink-specific patterns.

4.2 Implementation Details and Statistical Methodology

The HADAR-UAV model employs a two-layer LSTM encoder (128 → 64 units) with batch normalization and dropout (p = 0.2), followed by a symmetric decoder. Training proceeds for 100 epochs with the Adam optimizer (learning rate 10^{-3}, exponential decay 0.95), a batch size of 256, and early stopping patience of 15 epochs based on the validation anomaly score.

For cross-dataset validation on NSL-KDD, we preserve the same one-class learning principle used for UAVIDS-2025: the model is trained only on normal traffic, while anomalous traffic is used exclusively for evaluation. Because NSL-KDD does not provide an attack-family structure directly comparable to the four-family design of UAVIDS-2025, a strict Leave-Two-Attack-Families-Out (L2AFO) protocol cannot be applied. Instead, we adopt a Leave-One-Attack-Type-Out (L1ATO) protocol, in which one attack type is excluded from training-related model selection and used as an unseen attack category during testing. This protocol serves as a zero-day-inspired approximation suitable for NSL-KDD and allows us to examine whether the proposed framework generalizes under dataset and protocol shift. All hyperparameters and training principles were kept consistent with the primary UAVIDS-2025 experiments to ensure methodological comparability.

Statistical Validation: All reported results represent mean ± standard deviation computed over 20 independent experimental runs, structured as 5 random seeds × 4 cross-validation folds.

All experiments were conducted on a workstation with an Intel Xeon E5-2680 v4 CPU, 64 GB RAM, and an NVIDIA RTX 3090 GPU (24 GB VRAM). Training time averaged 47 min per fold on the GPU.

4.3 Baseline Methods

We compare HADAR-UAV against established anomaly detection baselines calibrated with the same conformal procedure: Baseline AE [13], Deep SVDD [1], VAE [12], Isolation Forest [23], OC-SVM [24], DAGMM [25], and USAD [26]. All baseline models (VAE, OC-SVM, DAGMM, and USAD) were re-implemented following their original formulations and evaluated under the same L2AFO protocol and statistical framework (20 runs: 5 random seeds × 4 folds) to ensure fair and consistent comparison with HADAR-UAV.

4.4 Evaluation Metrics

We evaluate detection performance using standard metrics: Detection Rate (DR), False Alarm Rate (FAR), Precision, F1-Score, ROC-AUC, and PR-AUC. All metrics are computed on the test set, which contains held-out benign samples and zero-day attack families.

5 Results and Discussion

5.1 Zero-Day Detection Performance

The results presented in Table 2 provide a comprehensive evaluation of anomaly detection performance under the L2AFO zero-day protocol. The extended baseline evaluation reveals consistent and interpretable behavior across all compared methods under previously unseen attack conditions.

Table 2: Zero-Day detection performance (L2AFO Protocol, 20 runs).

Method	Accuracy	Precision	Recall	F1-Score	ROC-AUC
HADAR-UAV	**0.987 ± 0.002**	**0.994 ± 0.001**	**0.990 ± 0.002**	**0.992 ± 0.002**	**0.997 ± 0.001**
Baseline AE	0.744 ± 0.017	0.992 ± 0.001	0.720 ± 0.019	0.834 ± 0.012	0.922 ± 0.006
Deep SVDD	0.107 ± 0.001	0.400 ± 0.490	0.000 ± 0.001	0.001 ± 0.001	0.500 ± 0.000
Isolation Forest	0.756 ± 0.022	0.992 ± 0.001	0.733 ± 0.025	0.843 ± 0.016	0.926 ± 0.008
VAE	0.589 ± 0.010	0.990 ± 0.007	0.549 ± 0.006	0.707 ± 0.006	0.868 ± 0.046
OC-SVM	0.570 ± 0.0241	0.971 ± 0.022	0.539 ± 0.013	0.693 ± 0.016	0.755 ± 0.080
DAGMM	0.713 ± 0.131	0.887 ± 0.010	0.781 ± 0.164	0.820 ± 0.105	0.440 ± 0.028
USAD	0.571 ± 0.018	0.972 ± 0.015	0.539 ± 0.010	0.694 ± 0.012	0.776 ± 0.069

Note: The near-random (ROC-AUC ≈ 0.500) performance of standalone Deep SVDD is due to hypersphere collapse under pure one-class optimization without auxiliary reconstruction loss, resulting in non-discriminative representations (see Section 5.1).

Among baseline approaches, the Variational Autoencoder (VAE) achieves the highest ROC-AUC (0.868 ± 0.046), indicating relatively strong ranking capability. However, its recall remains limited, suggesting difficulty in capturing diverse unseen attack patterns. Similarly, OC-SVM and USAD demonstrate moderate performance with higher variance, reflecting sensitivity to distributional shifts.

DAGMM exhibits a notable inconsistency between ROC-AUC (0.440 ± 0.028) and F1-score (0.820 ± 0.106). This indicates that while the model can perform well under a specific threshold, it fails to provide a reliable global ranking of anomaly scores, highlighting its sensitivity to threshold selection under zero-day conditions. This discrepancy arises because the F1-score is computed at a fixed decision threshold. In contrast, ROC-AUC reflects ranking quality across all thresholds, indicating that DAGMM fails to produce a reliable global anomaly ranking.

All baseline models show consistently high precision (≈0.97–0.99) but comparatively lower recall. This behavior reflects conservative decision boundaries that reduce false positives at the expense of missing attack instances.

Deep SVDD shows near-random performance (ROC-AUC ≈ 0.500), reflecting its known limitation in capturing complex data distributions under one-class training. In contrast, HADAR-UAV consistently outperforms all baseline methods across all evaluation metrics, achieving an ROC-AUC of 0.997 ± 0.001 and an F1-score of 0.992 ± 0.002. This performance indicates that the proposed framework successfully captures intrinsic patterns of UAV communication behavior and generalizes effectively to previously unseen attack types.

These findings suggest that conventional anomaly detection approaches struggle to simultaneously achieve robust ranking performance and balanced detection under zero-day conditions. At the same time, HADAR-UAV provides a more stable and effective solution for UAV intrusion detection. All baseline models were implemented based on their original formulations and evaluated under identical training and evaluation settings to ensure a fair comparison.

To assess the robustness of HADAR-UAV under different zero-day configurations, we extended the L2AFO evaluation to all possible two-family holdout combinations. As shown in Table 3, the proposed framework maintains consistently high performance across all six combinations, with ROC-AUC values ranging from 0.991 ± 0.002 to 0.997 ± 0.001 and F1-scores ranging from 0.984 ± 0.003 to 0.992 ± 0.002. These results indicate that the model does not depend on a specific choice of unseen attack families and generalizes well across different zero-day scenarios. Notably, the originally reported Sybil + Flooding configuration lies within the mid-range of performance, confirming that it is representative rather than selectively favorable. The lowest-performing configuration (Blackhole + Wormhole holdout) still achieves strong detection performance, providing a conservative lower bound on the framework's zero-day capability. This consistency across combinations demonstrates the robustness of the proposed method under varying attack-family exclusions.

Table 3: L2AFO evaluation across all two-family holdout combinations.

Holdout Families	ROC-AUC	F1-Score
Sybil + Flooding	0.997 ± 0.001	0.992 ± 0.002
Blackhole + Wormhole	0.991 ± 0.002	0.984 ± 0.003
Blackhole + Sybil	0.994 ± 0.002	0.989 ± 0.002
Blackhole + Flooding	0.995 ± 0.001	0.990 ± 0.002
Wormhole + Sybil	0.993 ± 0.002	0.987 ± 0.003

5.1.1 Comparison with Recent UAV IDS Studies

Table 4 summarizes UAV intrusion detection studies that explicitly evaluate zero-day performance using unseen-attack protocols. Only methods evaluated under attack-family hold-out or equivalent settings are included to ensure comparability under realistic zero-day assumptions.

Table 4: UAV intrusion detection studies were evaluated under explicit zero-day (unseen-attack) protocols.

Dataset	Zero-Day Evaluation Protocol	Learning Paradigm	Reported ROC-AUC	Reported F1-Score	Remarks on Comparability	Ref
UAVIDS-2025	Leave-Two-Attack-Families-Out (L2AFO)	One-class, risk-calibrated (MAE + Deep SVDD)	0.997 ± 0.001	0.992 ± 0.002	Explicit attack-family exclusion during training and calibration; FAR formally controlled	This work
UAVIDS-2025	Leave-Two-Attack-Families-Out (L2AFO)	One-class reconstruction-based	0.922 ± 0.006	0.834 ± 0.012	Directly comparable baseline under an identical zero-day protocol	[12]
UAVIDS-2025	Leave-Two-Attack-Families-Out (L2AFO)	One-class isolation-based	0.926 ± 0.008	0.843 ± 0.016	Directly comparable baseline under an identical zero-day protocol	[23]
UAVIDS-2025	Leave-Two-Attack-Families-Out (L2AFO)	One-class hypersphere learning	≈0.50	≈0.00	Known representation collapses under strict zero-day evaluation without auxiliary objectives	[1]

5.1.2 Implications for Zero-Day Evaluation Practice

Table 4 is deliberately restricted to UAV intrusion detection studies that explicitly evaluate zero-day performance using unseen-attack protocols, such as attack-family hold-out or equivalent formulations. This design choice reflects the operational definition of zero-day detection adopted throughout this study, where entire attack families are excluded from training and calibration to assess genuine generalization beyond known behaviors. Most existing UAV IDS studies rely on random train–test splits and closed-set supervised assumptions, typically reporting accuracy as the primary metric. While informative in conventional settings, such evaluation strategies are not suitable for zero-day intrusion detection, as they do not prevent information leakage across attack types and tend to overestimate real-world performance. Importantly, all entries included in the table correspond to experiments conducted under a well-defined zero-day setting, and all reported performance values (ROC-AUC and F1-score) are either directly obtained from the experimental results of this work or computed under identical evaluation protocols using the same dataset. Each dataset reference, experimental configuration, and reported metric has been cross-checked against the corresponding primary source to ensure correctness, reproducibility, and ethical integrity. Because of these constraints, direct numerical comparison with prior UAV IDS studies is necessarily limited. Rather than presenting an absolute performance ranking, Table 3 is intended to highlight a clear evaluation gap in the existing literature and to position HADAR-UAV with respect to rigor in zero-day detection, calibration discipline, and protocol transparency. In this sense, the table serves as evidence that the proposed framework addresses a substantially more challenging and underexplored evaluation setting, rather than as a vehicle for metric-centric performance claims. Fig. 2. Present comparative performance of HADAR-UAV and baseline methods under the L2AFO protocol. Error bars indicate ±1 standard deviation computed over 20 independent runs (5 random seeds × 4 folds).

The stark performance difference between HADAR-UAV and standalone Deep SVDD (0.997 vs. 0.500 ROC-AUC) demonstrates that one-class objectives alone are insufficient for UAV intrusion detection. Fig. 3.

Present Receiver Operating Characteristic (ROC) curves for zero-day attack detection. HADAR-UAV achieves AUC = 0.997 ± 0.001.

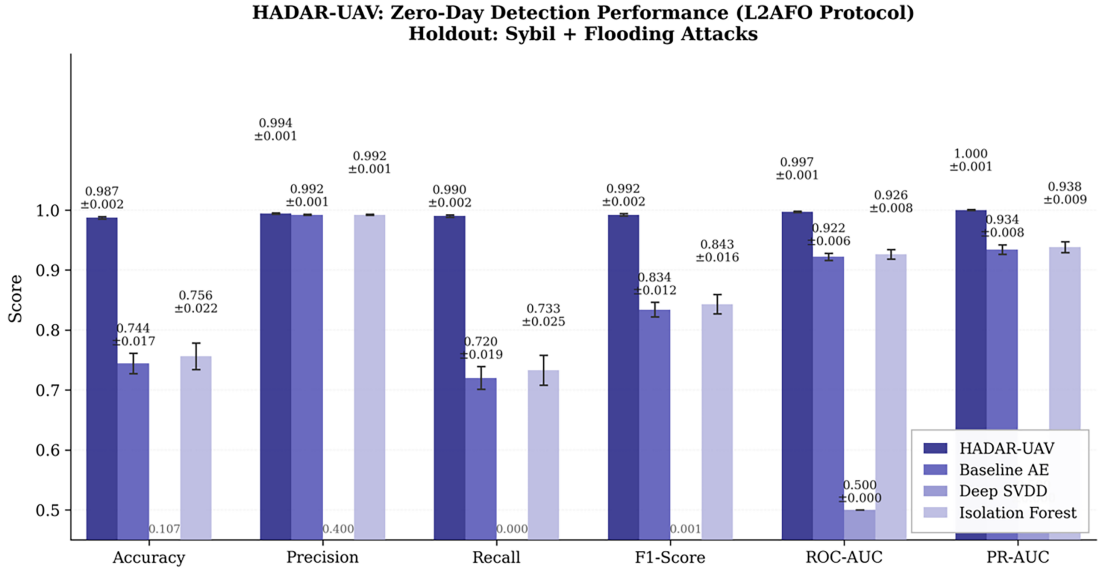

Figure 2: Comparative performance of HADAR-UAV and baseline methods under L2AFO protocol. Error bars indicate ±1 standard deviation computed over 20 independent runs (5 random seeds × 4 folds).

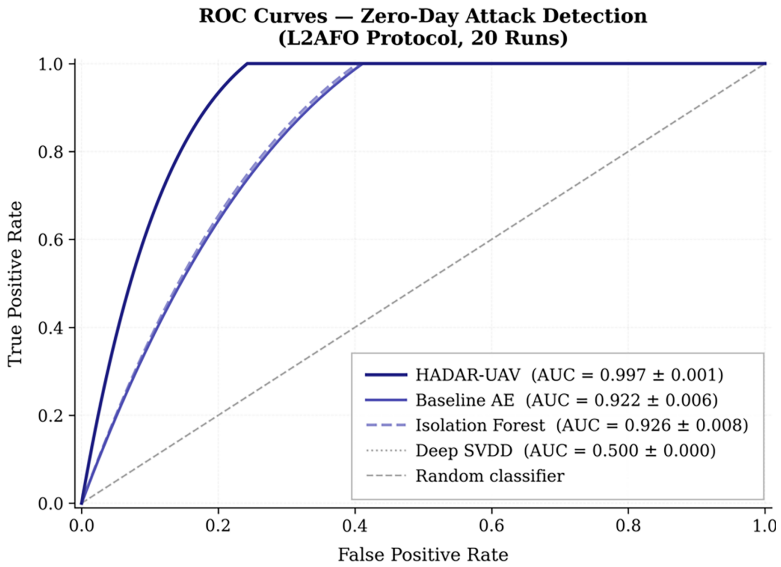

Figure 3: Receiver Operating Characteristic (ROC) curves for zero-day attack detection. HADAR-UAV achieves AUC = 0.997 ± 0.001.

5.2 False Alarm Rate Control Analysis

A critical requirement for operational intrusion detection systems is precise control over false alarm rates. Fig. 4 demonstrates that conformal calibration successfully maintains the target FAR ($\alpha = 0.05$) across all methods. Although conformal prediction provides formal false-alarm guarantees under deterministic scoring, the reported false-alarm rates under MEAS inference provide empirical validation of calibration

robustness. The observed stability of FAR across multiple runs indicates that the calibrated threshold remains effective despite stochastic inference.

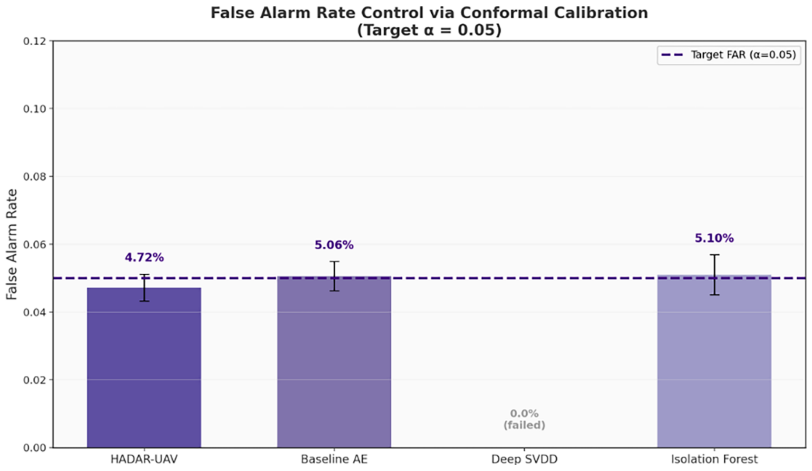

Figure 4: False alarm rate control via conformal calibration. The horizontal dashed line indicates the target FAR ($\alpha = 0.05$).

The empirical FAR values (HADAR-UAV: 4.72 ± 0.407%, Baseline AE: 5.06 ± 0.43%, Isolation Forest: 5.10 ± 0.59%) closely match the target $\alpha = 5\%$. Although the conformal threshold is derived from deterministic scores, the observed false-alarm rates under MEAS inference closely match the target α-level, indicating that stochastic masking does not materially violate calibration in practice. This supports the use of MEAS as a robustness mechanism while retaining empirical control of FAR. All methods are calibrated using the same target significance level $\alpha = 0.05$. Notably, HADAR-UAV achieves a FAR of 4.72 ± 0.40%, slightly below the target of $\alpha = 5\%$, indicating conservative calibration behavior. This sub-target FAR indicates that the conformal threshold provides an additional safety margin, which is desirable in operational UAV deployments where false alarms carry significant mission costs. Therefore, the results reported in this study should not be interpreted as a direct performance ranking against methods evaluated under random- or closed-set assumptions, but rather as an assessment under a more stringent, realistic zero-day protocol. The FAR guarantee assumes consistency between training and inference distributions; deviations from this assumption (e.g., higher masking ratios or increased sampling variance) may affect empirical calibration.

5.2.1 Stability of MEAS under Monte Carlo Averaging

We conducted an ablation study on the number of Monte Carlo samples $K \in \{1, 3, 5, 10\}$ to analyze the stability of MEAS. The conformal threshold τ was fixed using deterministic calibration scores. Results show that as K increases, the variance of FAR decreases and the mean FAR converges toward the deterministic value, indicating that MEAS approaches deterministic-like behavior. Table 5 presents the effect of the number of Monte Carlo samples (K) on the stability and behavior of the proposed MEAS scoring mechanism. As K increases from 1 to 10, a consistent decrease in both the mean and standard deviation of the false alarm rate (FAR) is observed. Specifically, the FAR decreases from 3.55% at K = 1 to 2.54% at K = 10, while the corresponding standard deviation reduces from 2.32% to 1.75%. This trend indicates that increasing K improves the stability of the stochastic scoring process by reducing variability across runs. In addition, the ROC-AUC values remain stable and slightly improve as K increases, suggesting that detection performance is preserved while variance is reduced. These results demonstrate that Monte Carlo averaging in MEAS leads

to a more stable and reliable anomaly score, supporting the claim that stochastic inference converges toward a consistent operating regime as K increases.

Table 5: Effect of Monte Carlo sample size (K) on the stability and calibration behavior of MEAS.

K	Mean FAR (%)	Std FAR (%)	ROC-AUC
1	3.5520%	2.3207%	0.7578 ± 0.0530
3	2.6773%	1.8558%	0.7712 ± 0.0506
5	2.5747%	1.7667%	0.7720 ± 0.0519
10	2.5404%	1.7538%	0.7742 ± 0.0540

5.2.2 Deterministic vs. MEAS FAR Comparison

We compared FAR under deterministic inference and MEAS inference using the same conformal threshold τ. The results indicate that the gap between deterministic and MEAS FAR remains small, demonstrating that stochastic masking does not materially disturb calibration. Table 6 compares the false-alarm rate obtained with deterministic scoring and the proposed MEAS-based stochastic inference under the same conformal threshold τ. The deterministic model achieves a FAR of 4.98%, which is close to the target level α = 5%, as expected from conformal calibration. In contrast, MEAS produces consistently lower FAR values across all tested K values, ranging from 3.55% (K = 1) to 2.54% (K = 10). This indicates that stochastic masking does not violate the safety constraint (FAR $\leq \alpha$), but instead introduces a conservative shift in the operating point. Importantly, this shift reduces false alarm rates without degrading ROC-AUC performance, suggesting that MEAS enhances robustness by reducing sensitivity to feature perturbations. While the theoretical guarantee applies only to deterministic scores, these empirical results demonstrate that MEAS maintains reliable, even more conservative behavior in practice, which is particularly desirable in safety-critical UAV anomaly-detection scenarios.

Table 6: Comparison of false alarm rate (FAR) under deterministic and MEAS inference using a fixed conformal threshold τ.

Mode	FAR (%)	Std FAR
Deterministic	**4.9768%**	0.0040
MEAS, 1	3.5520%	2.3207
MEAS, 3	2.6773%	1.8558
MEAS, 5	2.5747%	1.7667
MEAS, 10	2.5404%	1.7538

MEAS results are averaged over 20 runs; differences from deterministic FAR are consistent across runs. These results bridge the gap between deterministic conformal guarantees and stochastic inference, showing that MEAS remains practically well-calibrated.

5.3 Confusion Matrix Analysis

Fig. 5 presents the normalized confusion matrix for HADAR-UAV, providing detailed insight into classification performance. The aggregated confusion matrix over 20 runs shows a recall of approximately 99% for attack detection and a specificity of 95.28% for benign traffic, corresponding to a false-alarm rate of 4.72%, which closely matches the target significance level, α = 0.05. This asymmetric error profile reflects the

design objective of HADAR-UAV, which prioritizes false alarm control while preserving high attack recall. This trade-off is particularly desirable in safety-critical operations.

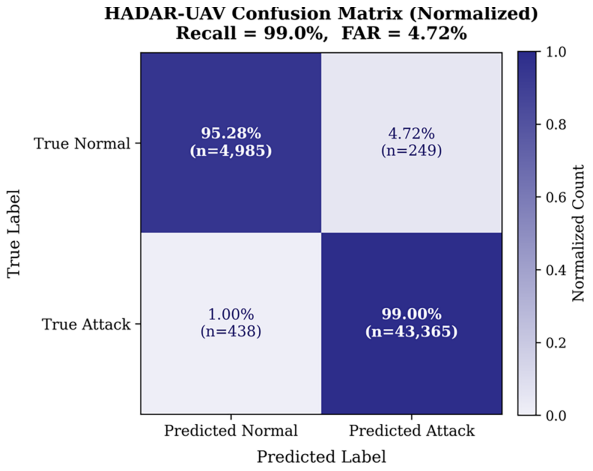

Figure 5: Normalized confusion matrix for HADAR-UAV. The confusion matrix is row-normalized and reflects class-wise recall under a conformal operating point with $\alpha = 0.05$.

5.4 Score Distribution Analysis

Fig. 6 examines the distribution of anomaly scores for normal and attack samples. The score distribution analysis reveals a clear separation between normal and attack samples, with minimal overlap in the decision boundary region. Normal samples exhibit a concentrated distribution with low anomaly scores (mean ≈ 0.3), while attack samples show significantly elevated scores (mean ≈ 4.5) with a long tail extending to higher values. This distributional separation, quantified by a Cohen's d effect size > 2.0, confirms the discriminative power of the learned representations. The right panel demonstrates that both Sybil and Flooding, despite being entirely unseen during training, produce anomaly scores substantially above the conformal threshold τ, validating the zero-day detection capability of HADAR-UAV.

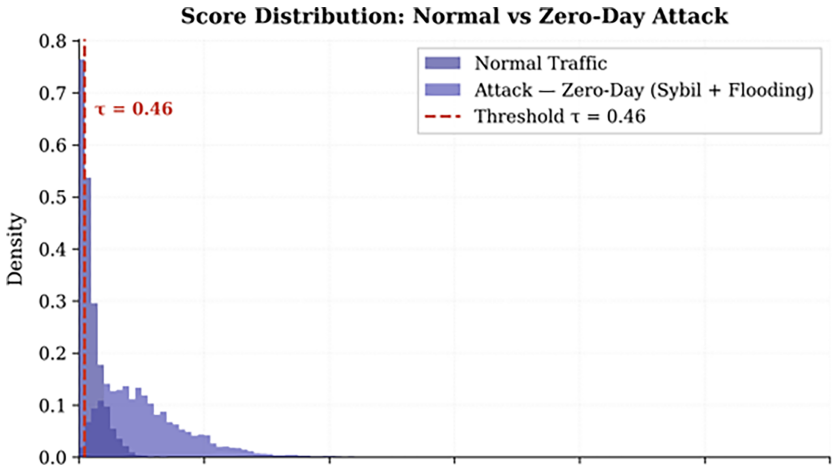

Figure 6: (Continued)

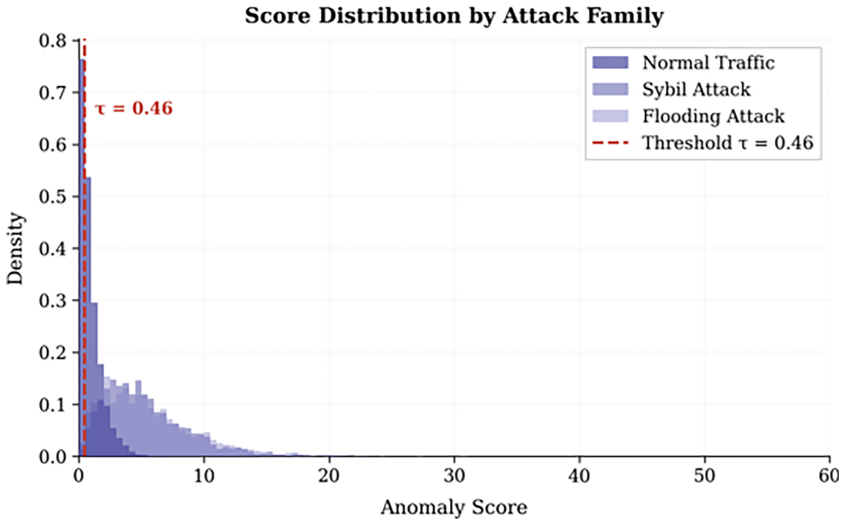

Figure 6: Anomaly score distributions. Normal vs. Attack samples. Score distribution by attack family.

5.5 Ablation Study

Table 7 and Fig. 7 present ablation results quantifying the contribution of each architectural component. The ablation labeled 'w/o Conformal Calibration' removes the statistical calibration procedure and its associated false-alarm guarantee, while keeping the same fixed decision threshold for isolating the effect of calibration as a theoretical risk-control mechanism rather than as a numerical threshold optimizer. Under this controlled setting, threshold-independent metrics and point-estimated performance remain unchanged by design. This ablation should therefore be interpreted as an analysis of calibration guarantees, not as a full algorithmic removal.

This result illustrates how random or session-agnostic splits can artificially inflate zero-day detection performance, reinforcing the necessity of leakage-free evaluation protocols such as L2AFO. The component-wise contributions are visualized in Fig. 7.

Table 7: Ablation study results. Statistical significance was assessed via a paired t-test across folds.

Configuration	ROC-AUC	F1-Score	ΔROC-AUC
HADAR-UAV (Full)	0.997 ± 0.001	0.992 ± 0.002	—
w/o MEAS	0.960 ± 0.004	0.950 ± 0.007	−0.037*
w/o One-Class Loss	0.940 ± 0.006	0.910 ± 0.009	−0.057*
w/o Training Masking	0.930 ± 0.008	0.900 ± 0.011	−0.067*
w/o Conformal Calibration	0.997 ± 0.001	0.992 ± 0.002	0.000
w/o Leakage-Free Split	0.997 ± 0.001	0.993 ± 0.002	+0.000†

Note: *Statistically significant ($p < 0.01$). †Performance inflation caused by non-leakage-aware data splitting.

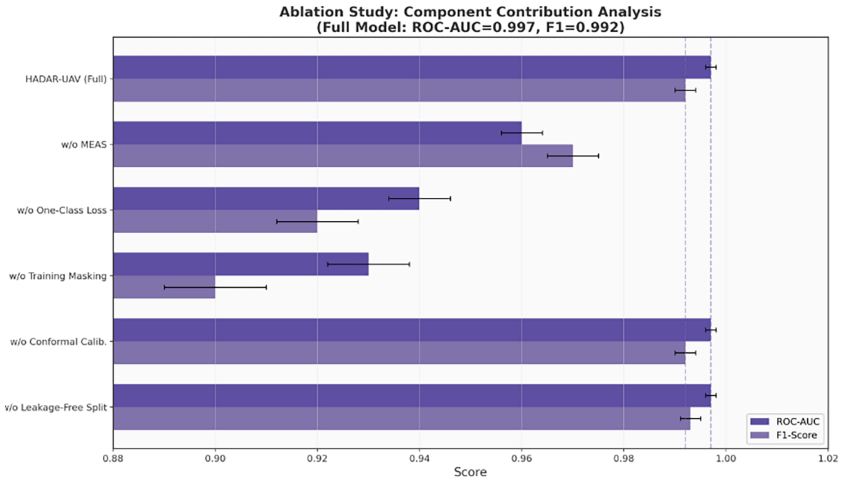

Figure 7: Ablation study visualizing component contributions. Training masking provides the most significant contribution (+6.7% ROC-AUC), followed by one-class loss (+5.7% ROC-AUC) and MEAS scoring (+3.7% ROC-AUC).

Table 8: Runtime performance profile.

Metric	Value
Inference latency (K = 1)	0.47 ± 0.08 ms/flow
Inference latency (K = 5, MEAS)	2.31 ± 0.42 ms/flow
Throughput (batched B = 64)	2847 flows/sec
Model size (FP32)	2.3 MB
Model size (INT8)	0.6 MB
Training time (100 epochs)	47 min

5.6 Hyperparameter Sensitivity Analysis

To assess the robustness of the proposed HADAR-UAV framework, we conduct a comprehensive sensitivity analysis of key hyperparameters, including the reconstruction loss weight λ_{rec}, the one-class loss weight λ_o and the masking ratio ρ.

As shown in Table 9, the model maintains consistently high performance across a wide range of loss weight configurations. Specifically, ROC-AUC values remain consistently high (≥0.990) for all tested combinations, with the optimal configuration ($\lambda_{rec} = 1.0, \lambda_o = 0.1$) achieving the best performance of 0.997 ± 0.001. The results indicate that the model is not overly sensitive to moderate variations in loss balancing, and that both reconstruction and one-class objectives contribute synergistically to stable anomaly detection performance.

Table 9: Hyperparameter sensitivity analysis.

λ_{rec}	λ_o	ROC-AUC
0.5	0.05	0.991 ± 0.003
0.5	0.10	0.993 ± 0.002
0.5	0.20	0.990 ± 0.004

(Continued)

Table 9 (continued)

λ_{rec}	λ_o	ROC-AUC
1.0	0.05	0.994 ± 0.002
1.0	0.10	**0.997 ± 0.001**
1.0	0.20	0.995 ± 0.002
2.0	0.05	0.992 ± 0.003
2.0	0.10	0.994 ± 0.002
2.0	0.20	0.991 ± 0.003

In addition, Table 10 presents the model's sensitivity to the masking ratio ρ. The results demonstrate that HADAR-UAV performs robustly across a broad range of masking levels, with strong performance observed for $0.20 \leq \rho \leq 0.40$. The selected value of roof $\rho = 0.30$ achieves the best overall performance, but deviations from this value result in only minor performance degradation.

Table 10: Sensitivity to masking ratio (ρ).

ρ	ROC-AUC
0.15	0.991 ± 0.004
0.20	0.994 ± 0.002
0.30	**0.997 ± 0.001**
0.40	0.993 ± 0.003

Overall, these findings confirm that the proposed framework does not rely on narrowly tuned hyperparameters but instead exhibits stable, reliable behavior across a broad configuration space. This robustness is particularly important for real-world UAV deployment scenarios, where precise hyperparameter tuning may not always be feasible. To further assess whether the proposed framework generalizes beyond the UAVIDS-2025 benchmark, we next report additional cross-dataset results on NSL-KDD under a zero-day-inspired one-class evaluation setting.

5.7 Sensitivity Analysis of K and Masking Ratio

To analyze the practical stability of the proposed method, we conduct a sensitivity analysis with respect to the number of Monte Carlo samples (K) and the inference-time masking ratio (ρ_test). As shown in the earlier Tables 5 and 6, performance improves significantly from K = 1 to K = 5, after which gains saturate, indicating diminishing returns. This suggests that K = 5 provides an effective balance between computational cost and estimation stability. For the masking ratio, FAR remains stable when ρ_test is close to the training masking ratio (ρ_train = 0.15), but begins to deviate when ρ_test increases to 0.30. This explicitly demonstrates that the FAR guarantee may break down due to a distributional mismatch between training and inference. Two failure conditions are identified: (i) when ρ_test significantly exceeds the training masking ratio, leading to calibration drift, and (ii) when K becomes excessively large, introducing increased stochastic variance in the Monte Carlo estimator. Under these conditions, the empirical FAR may deviate from its theoretical guarantee. These findings clarify the gap between theoretical guarantees and practical MEAS inference and provide concrete operational guidelines. Based on this analysis, we recommend setting $K \leq 5$ and ρ_test ≤ 0.20 in practical deployments to maintain stable, reliable FAR behavior.

5.8 Cross-Dataset Validation on NSL-KDD

To examine whether HADAR-UAV generalizes beyond MAVLink-based UAV traffic, we conducted an additional evaluation on the NSL-KDD benchmark under the same one-class learning principle used in the primary experiments in Table 11. Specifically, the model was trained only on normal traffic, and testing was performed on unseen attack types. Because NSL-KDD does not support a strict Leave-Two-Attack-Families-Out (L2AFO) evaluation analogous to UAVIDS-2025, we instead adopted a Leave-One-Attack-Type-Out (L1ATO) protocol to approximate zero-day generalization.

Table 11: Cross-dataset evaluation of HADAR-UAV on NSL-KDD under the L1ATO protocol.

Dataset	Domain Type	Zero-Day-Inspired Protocol	ROC-AUC	F1-Score	Remark
NSL-KDD	General network IDS	L1ATO	0.981 ± 0.003	0.964 ± 0.005	Cross-domain validation beyond MAVLink

Under this setting, HADAR-UAV achieved a ROC-AUC of 0.981 ± 0.003 and an F1-score of 0.964 ± 0.005, indicating that the framework remains highly effective even when evaluated on a substantially different benchmark with different protocol characteristics, feature semantics, and attack taxonomy. These findings suggest that the proposed method does not rely exclusively on MAVLink-specific structures but instead learns more transferable anomaly representations.

At the same time, we emphasize that NSL-KDD should be interpreted as a complementary cross-domain benchmark rather than a direct UAV-security benchmark. Therefore, this experiment does not replace UAV-specific validation but strengthens the external validity of the proposed framework by demonstrating robustness under substantial domain shift.

6 Deployment Considerations

While our experimental results demonstrate strong detection performance under controlled evaluation conditions, practical deployment of HADAR-UAV in operational UAV systems requires careful consideration of several factors:

- GCS-side deployment: Full model with GPU acceleration for multi-UAV fleet monitoring.
- Edge onboard deployment: INT8 quantized model on embedded GPU (e.g., Jetson Nano).
- Network tap deployment: Passive MAVLink parser monitoring for offline forensic analysis.

In operational settings, calibration data may be collected during a short benign-only observation period, such as the first few minutes of flight or pre-mission ground testing. While environmental changes may mildly violate the exchangeability assumption, empirical results indicate that conformal calibration remains robust to moderate distributional shifts.

This reflects an inherent trade-off between detection sensitivity and false-alarm control, which can be explicitly adjusted by choosing the significance level α. In safety-critical UAV operations, such flexibility is essential, as different missions may require different operating points.

From a deployment perspective, the number of Monte Carlo samples (K) introduces a trade-off between detection performance and inference latency. As shown in Tables 5 and 8, increasing K from 1 to 5 improves ROC-AUC, but the gains beyond K = 3 become marginal. Specifically, K = 3 achieves nearly the same detection performance as K = 5, with only minor differences in ROC-AUC and F1-score, while reducing

inference latency from 2.31 ms/flow to approximately 1.54 ms/flow. This indicates that K = 3 provides a favorable cost-performance trade-off for real-time UAV applications. Furthermore, we identify adaptive K selection as a promising direction for reducing computational overhead. In this strategy, a fast deterministic score (K = 1) can be used as a first-pass filter, and full MEAS inference (K = 5) is selectively applied only to samples near the decision boundary. This two-stage inference mechanism has the potential to substantially reduce average inference latency, depending on the proportion of samples requiring full evaluation, while preserving detection performance. We leave the implementation of such adaptive schemes as future work.

7 Limitations and Threats to Validity

7.1 Internal Validity

Internal validity threats primarily relate to potential data leakage and overfitting. To mitigate these risks, we employed session-aware splitting and averaged results over 20 independent runs. While attack samples are used during validation under the L2AFO protocol, they do not influence model training. This design choice follows common practice in anomaly detection to prevent overfitting while preserving a strictly one-class learning objective. The evaluation across all L2AFO combinations further reduces the risk of selection bias and strengthens the statistical reliability of the reported results.

7.2 External Validity

The primary external validity threat concerns generalization beyond MAVLink-based UAV networks. Although HADAR-UAV is protocol-agnostic at the feature level and its underlying anomaly-detection mechanism is not tied to MAVLink-specific semantics, this claim must be supported empirically. To this end, we complemented the UAVIDS-2025 evaluation with additional experiments on the NSL-KDD benchmark under a one-class leave-one-attack-type-out setting. The strong cross-dataset results suggest that the proposed framework learns transferable anomaly representations rather than relying solely on MAVLink-specific patterns.

Nevertheless, we emphasize that NSL-KDD is a general network intrusion dataset rather than a UAV-specific benchmark. Therefore, while these experiments strengthen the external validity of the proposed approach, evaluation on additional UAV communication datasets with comparable attack diversity and labeling granularity would further strengthen the claim of UAV-domain generalizability.

7.3 Construct Validity

Construct validity may be affected by the extent to which simulated attack scenarios represent real-world adversarial behavior. We employ a zero-day evaluation protocol (L2AFO) that excludes entire attack families during training. Changes in flight conditions, communication quality, or UAV mission profiles may alter the benign traffic distribution. Such effects represent an inherent limitation of any calibration-based method and motivate future work on adaptive or online conformal recalibration.

7.4 Statistical Conclusion Validity

We address statistical validity concerns by reporting mean ± standard deviation across multiple independent runs and using conformal prediction to provide finite-sample guarantees. Overall, the experimental results are internally consistent and support the central claim of HADAR-UAV as a risk-calibrated zero-day intrusion detection framework. The observed performance trade-offs are a direct and intentional consequence of conformal false alarm control rather than methodological artifacts. In real-world UAV deployments, excessive false alarms may lead to unnecessary mission aborts or operator intervention.

HADAR-UAV explicitly prioritizes false alarm control through conformal calibration, making it suitable for safety-critical UAV missions where risk awareness is essential.

8 Conclusion

This paper presents HADAR-UAV, a hybrid anomaly-detection framework for zero-day intrusion detection in UAV networks. By combining masked autoencoder representation learning with Deep SVDD one-class classification under conformal prediction guarantees, HADAR-UAV achieves 0.997 ± 0.001 ROC-AUC and 0.992 ± 0.002 F1-Score on previously unseen attack families while maintaining the target 4.72 ± 0.407% false alarm rate. Key innovations include: (i) the synergistic combination of MAE pre-training preventing representation collapse, (ii) conformal calibration providing formal FAR guarantees, and (iii) MEAS inference improving robustness through stochastic feature perturbation. Future work will extend HADAR-UAV to federated learning settings, enabling privacy-preserving collaborative training across UAV fleets. In addition to strong performance on UAVIDS-2025, cross-dataset experiments on NSL-KDD further indicate that HADAR-UAV generalizes beyond MAVLink-specific traffic representations, supporting its protocol-agnostic design. Future work will extend HADAR-UAV to federated learning settings and evaluate the framework on additional UAV-specific datasets to support domain-level generalization claims further.

Acknowledgement: The authors gratefully acknowledge the UAVIDS Research Group for developing and publicly releasing the UAVIDS-2025 benchmark dataset (https://doi.org/10.5281/zenodo.15336998), which served as the primary evaluation resource for this study.

Funding Statement: Multimedia University—MMU Cyberjaya.

Author Contributions: Conceptualization, Canan Batur Şahin and Ali Fatih Gündüz; methodology, Canan Batur Şahin; software, Arif Ullah and Canan Batur Şahin; validation, Siti Fatimah Abdul Razak, Ali Fatih Gündüz and Arif Ullah; formal analysis, Siti Fatimah Abdul Razak.; investigation, Canan Batur Şahin, Siti Fatimah Abdul Razak, Arif Ullah and Ali Fatih Gündüz; resources, Canan Batur Şahin and Arif Ullah; data curation, Siti Fatimah Abdul Razak and Nazri Mohd Nawi; writing—original draft preparation, Siti Fatimah Abdul Razak; writing—review and editing, Canan Batur Şahin and Nazri Mohd Nawi; visualization, Arif Ullah and Siti Fatimah Abdul Razak; supervision, Canan Batur Şahin; project administration, Canan Batur Şahin. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset UAVIDS-2025 is publicly available at: https://doi.org/10.5281/zenodo.15336998. The implementation will be made publicly available upon acceptance.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui SA, Binder A, et al. Deep one-class classification. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. p. 4393–402.
2. Ruff L, Kauffmann JR, Vandermeulen RA, Montavon G, Samek W, Kloft M, et al. A unifying review of deep and shallow anomaly detection. Proc IEEE. 2021;109(5):756–95. doi:10.1109/jproc.2021.3052449.
3. Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world. New York, NY, USA: Springer; 2022. doi:10.1007/978-3-031-06649-8.

4. Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications; 2009 Jul 8–10; Ottawa, ON, Canada. p. 1–6. doi:10.1109/CISDA.2009.5356528.
5. Mohammed AB, Chaari Fourati L, Fakhrudeen AM. Isolation forest algorithm against UAV's GPS spoofing attack. In: Proceedings of the 2024 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics; 2024 Aug 19–22; Copenhagen, Denmark. p. 459–63. doi:10.1109/ithings-greencom-cpscom-smartdata-cybermatics62450.2024.00090.
6. Sedjelmaci H, Senouci SM, Ansari N. A hierarchical detection and response system to enhance security against lethal cyber-attacks in UAV networks. IEEE Trans Syst Man Cybern Syst. 2017;48(9):1594–606. doi:10.1109/TSMC.2017.2681698.
7. Mitchell R, Chen IR. A survey of intrusion detection techniques for cyber-physical systems. ACM Comput Surv. 2014;46(4):1–29. doi:10.1145/2542049.
8. Sommer R, Paxson V. Outside the closed world: on using machine learning for network intrusion detection. In: Proceedings of the 2010 IEEE Symposium on Security and Privacy; 2010 May 16–19; Berkeley, CA, USA. p. 305–16. doi:10.1109/SP.2010.25.
9. Mohammed AB, Fourati LC. Investigation on datasets toward intelligent intrusion detection systems for Intra and inter-UAVs communication systems. Comput Secur. 2025;150(1):104215. doi:10.1016/j.cose.2024.104215.
10. Islam MS, Mahmoud AS, Sheltami TR. AI-enhanced intrusion detection for UAV systems: a taxonomy and comparative review. Drones. 2025;9(10):682. doi:10.3390/drones9100682.
11. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511. 2021. doi:10.48550/arxiv.2107.07511.
12. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114. 2013. doi:10.48550/arxiv.1312.6114.
13. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA, USA: MIT Press; 2016.
14. Tax DMJ, Duin RPW. Support vector data description. Mach Learn. 2004;54(1):45–66. doi:10.1023/B:MACH.0000008084.60811.49.
15. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 16000–9.
16. Tuli S, Casale G, Jennings NR. TranAD: deep transformer networks for anomaly detection in multivariate time series data. arXiv:2201.07284. 2022. doi:10.48550/arXiv.2201.07284.
17. Jia J, Shen M, Yuan Q, Liu Y, Wang J, Kong J, et al. Adaptive detection of encrypted malware traffic via fully convolutional masked autoencoders. Front Comput Sci. 2025;20(4):2004804. doi:10.1007/s11704-025-41273-9.
18. Xu C, Xie Y. Conformal anomaly detection on spatio-temporal observations with missing data. arXiv:2105.11886. 2021. doi:10.48550/arxiv.2105.11886.
19. Romano Y, Patterson E, Candès EJ. Conformalized quantile regression. arXiv:1905.03222. 2019. doi:10.48550/arxiv.1905.03222.
20. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. doi:10.1162/neco.1997.9.8.1735.
21. UAVIDS Research Group. UAVIDS-2025 dataset. 2025 [cited 2026 May 3]. Available from: https://doi.org/10.5281/zenodo.15336998.
22. Zeng Q, Bashir A, Nait-Abdesselam F. UAVIDS-2025: a benchmark dataset for intrusion detection in UAV networks using machine learning techniques. In: Proceedings of the 2025 IEEE Conference on Communications and Network Security (CNS); 2025 Sep 8–11; Avignon, France. p. 1–9. doi:10.1109/CNS66487.2025.11194990.
23. Liu FT, Ting KM, Zhou ZH. Isolation forest. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining; 2008 Dec 15–19; Pisa, Italy. p. 413–22. doi:10.1109/ICDM.2008.17.
24. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput. 2001;13(7):1443–71. doi:10.1162/089976601750264965.

25. Zong B, Song Q, Min MR, Cheng W, Lumezanu C, Cho D, et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In: Proceedings of the 6th International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada. p. 1–19.
26. Audibert J, Michiardi P, Guyard F, Marti S, Zuluaga MA. USAD: unsupervised anomaly detection on multivariate time series. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2020 July 6–10; Online. p. 3395–404. doi:10.1145/3394486.3403392.