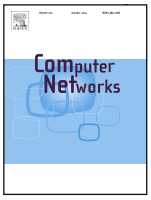
Computer Networks 274 (2026) 111821

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



GIIDS-AR: End-to-end generalized intelligent intrusion detection system with adversarial robustness for heterogeneous UAVs in UAM

Fahmina Kabir [a], Nishat I Mowla [b], Thomas Rosenstatter [c], Inshil Doh [d,*]

[a] Division of Artificial Intelligence and Software, Ewha Womans University, Seoul, Korea
[b] Department of Industrial Systems, RISE Research Institutes of Sweden, Sundsvall, Sweden
[c] Josef Ressel Centre for Intelligent and Secure Industrial Automation, Salzburg University of Applied Sciences, Salzburg, Austria
[d] Department of Cyber Security, Ewha Womans University, Seoul, Korea

ARTICLE INFO

Keywords:
IDS
UAM
Security
AI/ML
Generalizability
Adversarial robustness

ABSTRACT

This study presents GIIDS-AR, an enhanced version of the Generalized Intelligent Intrusion Detection System (GIIDS), developed to enhance robustness and secure diverse Unmanned Aerial Vehicles (UAVs) in Urban Air Mobility (UAM) while preserving generalization. As UAVs grow vital in logistics, emergency response, and disaster relief, their reliance on wireless communication increases exposure to cyber threats. GIIDS leverages machine learning for cross-platform detection but remains vulnerable to adversarial machine learning (AML) attacks. To assess this, GIIDS was tested under black-box, white-box, and transfer attacks. Accuracy dropped to 72 % under black-box and recall to 49.9 % under white-box settings. Adversarial training restored original performance improving accuracy to 99.0 % and F1 to 99.8 %, with AUC reaching 1.00. These evaluations were conducted using cross-dataset splits of live and simulated UAV telemetry, ensuring resilience on previously unseen data. GIIDS-AR retains layered modeling, time-aware feature encoding, and ensemble learning, while incorporating adversarial examples to improve resilience. It demonstrates strong detection performance under diverse attacks and generalizes effectively across heterogeneous UAV platforms. Our findings reveal that generalization techniques inherently contribute to adversarial robustness, positioning GIIDS-AR as one of the first unified UAV IDS frameworks capable of securing UAV networks against evolving cyber threats.

1. Introduction

Unmanned Aerial Vehicles (UAVs) within Urban Air Mobility (UAM) systems have an increasingly significant role across sectors such as agriculture, logistics, disaster management, and public safety, providing critical capabilities including real-time data acquisition, infrastructure inspection, and search and rescue operations [1,2]. These UAVs communicate through UAV-to-UAV, UAV-to-Ground Control Station (GCS), and UAV-to-satellite links, which, due to their inherently open and wireless nature, are vulnerable to cyberattacks such as jamming, spoofing, and denial-of-service (DoS) attacks [3]. Such breaches may lead to unauthorized access, commandeering, crashes, or navigation errors. Given UAVs' growing integration into mission-critical applications, ensuring their security and privacy is paramount. Intrusion Detection Systems (IDSs) are a critical defense mechanism, monitoring system and network activities to identify threats including route manipulation, message forgery, malware injection, UAV hijacking, routing attacks, GPS spoofing and jamming, and DoS attacks [1]. The adoption of machine learning (ML) techniques has enhanced IDS capabilities by automating anomaly detection and response [3–6]. However, ML-based IDSs commonly assume that training and testing datasets are independent and identically distributed (IID) with sufficient variability for generalization. In real-world UAV deployments, this assumption often fails, causing significant performance degradation as shown in Fig. 1.

A key obstacle in developing machine learning-based IDSs, particularly in the UAV domain, lies in achieving robust generalization. While many ML-driven IDS solutions perform well under controlled conditions, they often falter in real-world deployments due to data heterogeneity and distributional shifts between training and testing data [7–9]. Generalization, in this context, refers to an IDS's ability to maintain accurate detection on previously unseen data, ensuring operational reliability across different UAV platforms and dynamic environments [6]. The absence of this capability limits practical deployment. In our earlier work, GIIDS [10], we addressed this by designing techniques that enhance cross-platform generalization within heterogeneous UAV ecosystems.

* Corresponding author.
E-mail addresses: kabirfahmina@ewhain.net (F. Kabir), nishat.mowla@ri.se (N.I Mowla), thomas.rosenstatter@fh-salzburg.ac.at (T. Rosenstatter), isdoh1@ewha.ac.kr (I. Doh).

https://doi.org/10.1016/j.comnet.2025.111821
Received 28 July 2025; Received in revised form 29 September 2025; Accepted 30 October 2025
Available online 8 November 2025

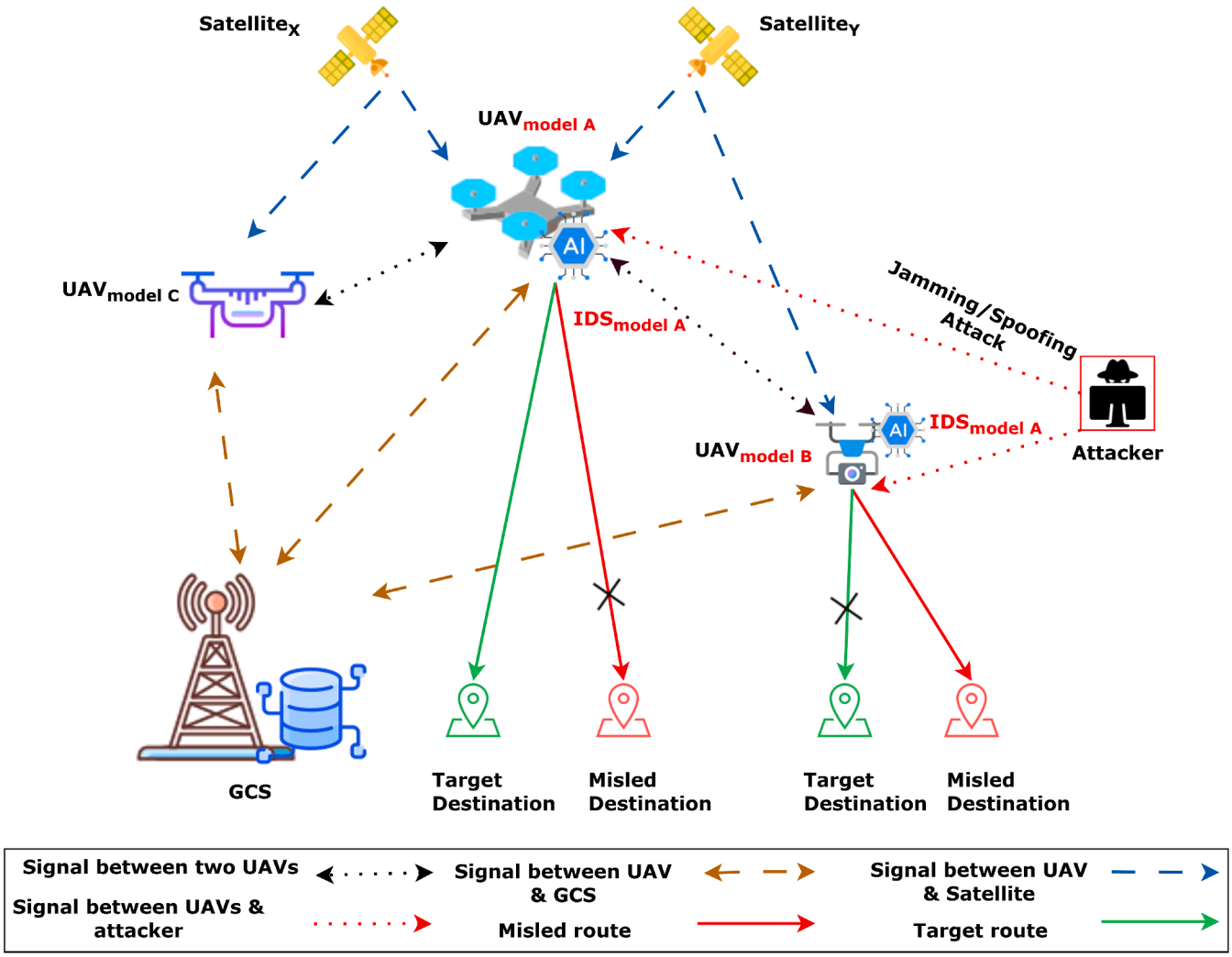


Fig. 1. Challenges in generalization for ML-based IDS in UAV networks. An IDS trained on UAV Model A fails to detect threats on UAV Model B, resulting in vulnerabilities such as misdirection.

However, despite these advancements, ML-based IDSs remain vulnerable to AML attacks. These attacks involve subtly perturbed inputs, known as adversarial examples, which mislead models while maintaining similarity to legitimate data [11]. Although such vulnerabilities have been widely studied in domains like computer vision, their implications for security-critical IDSs, particularly in UAV and UAM contexts, remain underexplored [12–14]. Moreover, Most AML research focuses on models that are trained using gradient-based methods, such as deep neural networks , often assuming static data distributions. In contrast, non-gradient-based models like Random Forests, Decision Trees, and K-Nearest Neighbors receive limited attention, especially under white-box threat models [12,13,15]. This creates a critical gap in understanding the adversarial robustness of many ML-based IDSs deployed in practice.

Building upon the challenges of generalization, our earlier work, the Generalized Intelligent Intrusion Detection System (GIIDS) [10], addressed the first key research question: *how to secure heterogeneous UAV platforms operating within UAM settings characterized by diverse platforms and threat landscapes*. GIIDS demonstrated reliable adaptability and strong detection capabilities across various unseen UAV platforms and data variations by employing novel time-aware feature engineering, ensemble-based learning, and robust multi-stage cross-dataset evaluations, enabling effective generalization in heterogeneous UAV environments.

Recognizing the critical vulnerability of machine learning-based IDSs to adversarial manipulation, this study investigates the second key question: *how do adversarial attacks affect the efficacy of generalized UAV-specific IDSs such as GIIDS, and can enhanced defensive mechanisms recover detection accuracy without compromising generalization?* To address this, we propose GIIDS-AR, an adversarially robust extension of the original GIIDS framework that specifically targets AML threats. GIIDS-AR integrates a comprehensive adversarial evaluation and defense pipeline designed to secure the system against advanced threats while preserving its generalization capabilities across different UAV platforms and operational conditions.

In particular, GIIDS-AR introduces a multi-stage validation and filtering process that ensures only feasible adversarial samples are used during training and evaluation. This feature-wise 3-sigma filtering removes unrealistic perturbations while maintaining the integrity of UAV network traffic sequences. Furthermore, the system undergoes extensive cross-attack evaluations, covering black-box, white-box, and transfer attacks across both gradient and non-gradient-based models. To contextualize GIIDS-AR's performance, we also compare it with three representative non-generalized baselines and both our original GIIDS models, highlighting improvements in detection accuracy, robustness, and operational efficiency. Additionally, a detailed resource and computational analysis demonstrates the system's practical deployability under realistic constraints.

Our evaluation includes a detailed analysis of system performance before and after applying generalization and robustness measures. Collectively, this approach aims to provide a practical IDS solution capable of detecting both conventional and sophisticated adversarial threats, thereby enabling secure and dependable deployment in complex UAV and IoT environments. The main contributions of GIIDS-AR are as follows:

- We develop an *extensive adversarial testing pipeline* for both gradient and non-gradient-based models, featuring:
  - *Black-box attacks:* Structured, data-dependent perturbations based on autoencoder reconstruction error, a novel approach in IDS research.
  - *White-box attacks:* Explainability-driven attacks that use feature importance to guide targeted perturbations for shallow ML GIIDS.
  - *Transfer attacks:* Evaluation of adversarial examples' transferability across different model types.
- We perform adversarial training on the generalized models. In particular, we incorporate adversarial examples into the training process to improve resilience without sacrificing generalization performance.
- We provide *empirical insights* and evidence that generalized IDS models inherently exhibit improved robustness against adversarial manipulations, highlighting a positive relationship between generalization and adversarial defense.
- We introduce validated filtering of adversarial samples, extensive cross-attack evaluations, comparison with three representative non-generalized baselines and both GIIDS models, and detailed resource and computational analysis, ensuring practical robustness and deployability.

To our knowledge, GIIDS-AR is among the first UAV IDS frameworks to unify generalization and adversarial robustness, offering a deployable solution for securing heterogeneous UAV networks against both conventional and adversarial threats in real-world scenarios.

2. Related works

2.1. UAV security

UAVs process critical information such as imagery, telemetry, and sensor data, making robust security measures essential to ensure data confidentiality, integrity, and availability [16]. To address these needs, UAV security integrates a range of mechanisms including IDSs, end-to-end encryption, secure communication protocols, and data integrity verification. Given their reliance on open wireless communication channels, UAVs are inherently vulnerable to cyber threats such as DoS attacks, signal jamming, and spoofing. Mitigation strategies typically include frequency hopping, signal authentication, and encryption, alongside IDSs for threat detection [16].

To prevent unauthorized access and control, UAV platforms commonly employ multi-factor authentication, role-based access control, and secure digital certificates. Redundant sensor arrays and data validation techniques are also implemented to maintain operational safety and reliability, even in the presence of hardware failures or external attacks [16]. Real-time monitoring and anomaly detection algorithms are integrated into UAV control systems to facilitate prompt identification and response to potential breaches. Furthermore, effective UAV security must comply with relevant legal and aviation standards, which require fault-tolerant architectures, formal operational policies, and comprehensive operator training. *However, traditional UAV security approaches often struggle to adapt to evolving threat patterns and lack the intelligence needed for proactive intrusion detection.* Recent research highlights the importance of designing UAV security frameworks that can not only detect known threats but also maintain resilience against previously unseen attack vectors, including adversarial manipulations of ML-based IDSs.

2.2. AI-based IDS for UAV

Intrusion Detection Systems are vital for maintaining UAV cybersecurity, as they enable the timely identification of anomalous or potentially malicious activity. Conventional IDS methodologies encompass signature-based, specification-based, statistical, anomaly-based, and hybrid detection techniques [16]. With advancements in artificial intelligence, the integration of machine learning (ML) and deep learning (DL) approaches has considerably enhanced IDS performance by enabling the recognition of intricate and previously unseen attack patterns.

Common ML algorithms employed in UAV IDSs include Support Vector Machines (SVM), Decision Trees, and Random Forests among others. These algorithms are capable of learning from labeled datasets to accurately differentiate between benign and malicious behaviors. Model effectiveness can be further improved through advanced feature engineering, which extracts meaningful statistical, temporal, and contextual insights from raw telemetry data [17].

For example, the study in [18] introduces an ensemble-based detection framework for UAV eavesdropping attacks, combining both supervised and unsupervised learning methods. In another work, [19] investigates multiple classifiers including Decision Trees, K-Nearest Neighbors, Logistic Regression, Random Forests, Artificial Neural Networks, and SVMs to detect man-in-the-middle (MITM) attacks in UAV systems.

The study in [20] investigates the application of the Isolation Forest algorithm for detecting GPS spoofing attacks in UAV systems. The authors conduct a comparative evaluation against Random Forest and Naive Bayes classifiers using a dedicated UAV GPS deception dataset. Their results indicate that the Isolation Forest method achieves superior detection accuracy, demonstrating its efficacy as a robust approach for UAV anomaly detection.

To improve accuracy and computational efficiency, optimization strategies such as hyperparameter tuning [6] and genetic algorithms [4] have been employed. By leveraging both real-world and simulated data, AI-driven IDSs can adapt to diverse mission profiles and network conditions. *Despite their strong performance in controlled environments, many AI-based IDSs exhibit limited generalizability and may fail to maintain reliability in complex, dynamic UAM scenarios where unseen threats and platform diversity pose significant challenges.* Moreover, their vulnerability to adversarial examples remains largely untested, with only a few studies incorporating multi-attack evaluations or filtering to ensure realistic adversarial samples, highlighting a gap addressed by GIIDS-AR.

2.3. Generalizability and IDS

There is increasing interest in applying Machine Learning (ML) and Deep Learning (DL) to address challenges in computer systems [21–25]. However, practical adoption remains limited due to concerns regarding generalizability [26–29], interpretability [27,29], and scalability. Generalizability refers to a model's ability to maintain performance on data that differs from its training distribution, while avoiding overfitting to the training data , thereby mitigating violations of the Independent and Identically Distributed (IID) assumption. This capability is particularly vital in dynamic environments such as UAV operations, where data may vary significantly across platforms, missions, or operating conditions [30]. A generalizable model can effectively strike a balance between underfitting and overfitting, ensuring consistent detection performance on previously unseen attack scenarios.

To enhance generalization in IDS, researchers have proposed several strategies. Cross-validation techniques like K-fold validation and cross-dataset evaluation help assess model robustness across distributions [4,30]. Feature engineering methods such as normalization, scaling, and cyclic feature encoding improve temporal pattern recognition and reduce overfitting [31]. Ensemble learning, by combining multiple diverse base models, increases robustness to data-specific noise and improves reliability across different datasets [4,32,33]. Transfer learning and domain adaptation enable models to adapt to new environments by

leveraging knowledge from related domains, even when labeled data is scarce [34]. Additionally, data augmentation methods like noise injection, dropout, and regularization help improve model flexibility. Recent frameworks such as TabPTM [35] further contribute by building standardized meta-representations for tabular data, promoting generalization without retraining. Despite these advancements, many IDS still exhibit overfitting due to intra-dataset evaluations or data leakage. Therefore, evaluating generalizability using disjoint or heterogeneous datasets remains essential, especially for real-world deployment in UAV or IoT systems where evolving threats demand resilient and adaptive detection mechanisms. Importantly, evaluating adversarial robustness alongside generalization, as done in GIIDS-AR, requires filtering unrealistic adversarial samples and performing extensive cross-attack assessments to ensure practical resilience.

2.4. Adversarial machine learning

AML is a growing interdisciplinary field that integrates machine learning with cybersecurity to explore both adversarial attack techniques and the development of robust defenses. It examines how attackers can manipulate models through methods like evasion, poisoning, and inference attacks by introducing subtle input perturbations that lead to misclassification without detection. Although AML originated in the image domain, its significance is now widely recognized in security-critical systems such as intrusion detection, malware detection, and cyber-physical infrastructures. In adversarial contexts where assumptions like data stationarity do not hold, AML becomes essential for modeling attacker behavior, identifying algorithmic weaknesses, and enhancing model resilience under adversarial influence [14,36]. Recent UAV IDS research emphasizes the need for multi-model, multi-attack evaluations and comparisons against representative SOTA baselines to understand both robustness and computational trade-offs, as demonstrated in our GIIDS-AR study.

2.4.1. Types of adversarial attacks based on model access

Adversarial attacks aim to deceive machine learning models, particularly deep networks, by adding carefully crafted perturbations δ to input data x, resulting in an adversarial example x^{adv} defined as:

$$x^{\text{adv}} = x + \delta, \quad \text{where} \quad \delta \in S = \{\delta : \|\delta\|_p \leq \epsilon\} \tag{1}$$

Here, S represents the set of allowable perturbations constrained by a norm-based threshold ϵ, with $\|\cdot\|_p$ denoting the p-norm. This formulation ensures that the perturbations are subtle enough to evade detection while misleading the model.

These attacks are typically categorized by the attacker's knowledge of the model [14]:

- *White-box attacks:* The attacker has full access to the model's architecture, parameters, and gradients, enabling gradient-based adversarial example generation. These are commonly used to assess model robustness under idealized threat conditions [14].
- *Black-box attacks:* The attacker can only query the model and observe outputs. Without internal access, strategies like confidence score probing or label querying are used. No table examples include ZOO, Autoencoder-based attacks, Boundary Attack, and HopSkipJump [12,14].
- *Transfer-based attacks:* These exploit the transferability of adversarial examples by training a surrogate model using queries to the target (oracle). Adversarial inputs crafted using white-box attacks on the surrogate often successfully fool the original model [14].

2.4.2. Defensive approaches against AML attacks

Defensive strategies in AML aim to enhance model robustness against adversarial manipulations. In UAV-based IDS, such defenses are critical to maintaining detection efficacy against both conventional and adversarial threats.

- *Adversarial training:* This method augments model robustness by incorporating both clean and adversarially perturbed samples into the training process. Exposure to these adversarial examples improves generalization and helps the model learn more resilient feature representations [15]. In GIIDS-AR, adversarial training is applied to both shallow and deep models using filtered, realistic adversarial samples across multiple attack types, ensuring practical robustness without sacrificing generalization.
- *Feature squeezing:* A computationally efficient defense that detects adversarial inputs by reducing input complexity. Techniques like bit-depth reduction or smoothing to simplify inputs, and discrepancies in model predictions between original and squeezed inputs are used to flag potential attacks [37]. Originally demonstrated on image datasets (e.g., MNIST [38], CIFAR-10 [39], ImageNet [40]), this method is adaptable to other domains by emphasizing input invariance and robustness.

Despite growing research in AML, most existing defenses are tailored to specific threat models or application domains and fail to generalize effectively across diverse scenarios. Techniques such as adversarial training and feature squeezing have shown promise, but they are typically evaluated under idealized settings and often target image-based domains. Their direct applicability to tabular data, such as telemetry and network traffic used in IDS, specially UAV IDS, remains limited. Moreover, many defenses focus solely on enhancing robustness to adversarial perturbations without accounting for the need to maintain generalization across heterogeneous platforms, unseen attack types, or shifting data distributions. Our work addresses these gaps through extensive cross-attack evaluation, comparisons with non-generalizable SOTA baselines, and detailed resource and computational analysis to provide a realistic picture of IDS robustness.

3. Proposed generalized intelligent intrusion detection system with adversarial robustness

3.1. Proposed system model

This section presents GIIDS-AR, an enhanced intrusion detection framework tailored for dynamic UAV networks. Previously, GIIDS (Generalized Intelligent Intrusion Detection System) [10] which addressed generalizability limitations in traditional ML-based IDS, was developed to secure heterogeneous UAVs operating within UAM networks. GIIDS incorporates a series of techniques to ensure cross-platform detection capabilities. First, we implemented generalization-driven feature engineering by extracting model-agnostic temporal features from raw timestamp data, including cyclic transformations such as sine and cosine of the minute, time differences, and elapsed time. These features help the model recognize sequential patterns linked to benign and attack behaviors, irrespective of UAV model type. Following feature engineering, we identified key features using random forest and gradient boosting, retaining only commonly important ones and discarding zero-importance attributes. We performed multi-stage cross-validation for model selection, training both shallow and deep individual learners with various scaling techniques to simulate shifts in data distribution. The top-performing models were then combined using weighted voting and stacking ensemble methods. To rigorously assess the generalization capability, we applied two types of cross-data evaluations: (i) *CV1*, where the system was trained on a combination of live and simulated data and tested on live data, and (ii) *CV2*, where training was conducted on three UAV models and testing on five different ones. These evaluations demonstrated the system's robustness to data drift and heterogeneity, with ensemble models, particularly the stacking ensemble, achieving high performance across all key metrics. This robust generalization ensures that GIIDS can operate reliably in real-world UAM environments.

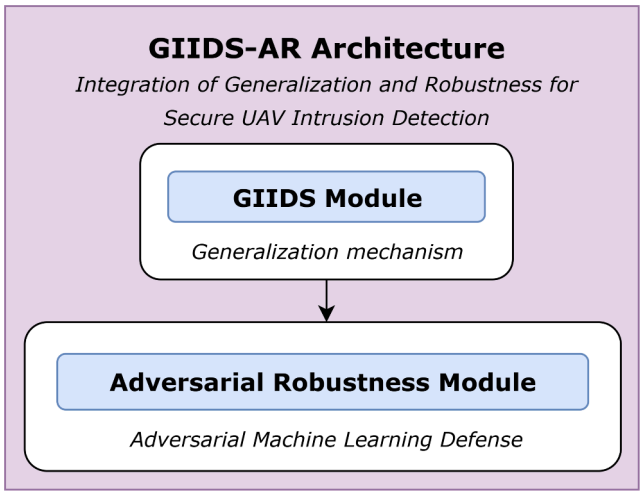


Fig. 2. Overall architecture of GIIDS-AR.

Building on the strengths of the previously developed GIIDS, GIIDS-AR introduces a robust end-to-end two-phase architecture that also mitigates adversarial vulnerabilities. As illustrated in Fig. 2, the system combines a generalization module based on GIIDS with a dedicated adversarial robustness module. This second module is designed to defend against AML threats by incorporating adversarial training, which improves the system's resilience to adversarial attempts without compromising its ability to generalize across heterogeneous UAV platforms. The robustness framework further includes gradient-free black-box attacks that generate structured perturbations using autoencoders, explainability-guided white-box attacks targeting key features, and PGD-based attacks for deep models. Additionally, transfer attacks are executed using FGSM applied to adversarial samples crafted from a surrogate feedforward neural network. Training on a mix of clean and adversarial examples, GIIDS-AR achieves high detection accuracy and robustness, making it a comprehensive and adaptive solution for intrusion detection in adversary-aware UAV environments.

The adversarial robustness module integrated into the AI-based generalized IDS framework (GIIDS) for UAM comprises three primary components: a) threat model, b) system framework, and c) proposed adversarial robustness approach.

3.2. Threat model

This section introduces the susceptibility of GIIDS to adversarial data poisoning attacks within a heterogeneous UAM environment. This environment comprises multiple UAV models (i.e., $UAV_{\text{model A}}$, $UAV_{\text{model B}}$, $UAV_{\text{model C}}$), and satellite communication links among other components. The training dataset for the IDS is collected from diverse UAV platforms operating concurrently in this setting.

A significant threat emerges when an adversary manipulates the training data, either by compromising data streams from a particular UAV model or by injecting malicious samples into the aggregated dataset. As illustrated in Fig. 3, this results in two contrasting outcomes: under ideal conditions, clean training data enables GIIDS to achieve reliable and accurate intrusion detection; conversely, adversarial poisoning leads to a degraded "Poisoned GIIDS" model that exhibits reduced detection performance. This scenario underscores the critical importance of integrating adversarial robustness strategies to protect IDS frameworks deployed in dynamic and security-sensitive UAM environments.

3.3. System framework

The proposed framework, illustrated in Fig. 4, presents a systematic approach for evaluating and enhancing the adversarial robustness of GIIDS in dynamic UAV environments. The process begins with clean input data, which is used to generate adversarial examples through various attack vectors, including white-box, black-box, and transfer-based attacks. These threat models simulate realistic adversarial conditions, establishing a rigorous foundation for assessing GIIDS resilience.

First, the performance of the GIIDS models is measured using clean data to establish benchmark detection metrics. Next, these models are tested against adversarial datasets to assess their vulnerability to different attack strategies. To address identified weaknesses, adversarial training is applied to GIIDS, using a mixture of clean and adversarial samples. This yields a robust model referred to as GIIDS-AR. Adversarial training strengthens the model's generalization ability across heterogeneous UAV platforms, conventional cyber threats, and adversarial perturbations. Finally, a comparative analysis is conducted between the adversarially trained GIIDS-AR and non-generalized state-of-the-art baseline models. This comparison highlights the impact of model generalization on adversarial robustness.

3.4. Proposed adversarial robustness approach

3.4.1. Generating AML attacks

This section outlines the AML attacks developed for evaluating GIIDS. Traditional AML techniques are largely gradient-based, requiring access to model gradients or parameters. These methods typically involve several forward and backward passes, making them computationally expensive.

Such methods are impractical for GIIDS, particularly the shallow variant, which consists of non-gradient models like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and tree-based ensembles. These models do not expose gradients, rendering conventional approaches ineffective. Moreover, these techniques are unsuitable in real-world settings where the model internals are hidden. Addressing these limitations was central to our adversarial strategy.

Assumptions:

In the white-box setting, we assume the adversary has complete access to training data and model internals for the shallow ML GIIDS, deep GIIDS, and baseline model. For shallow ML GIIDS, this includes knowledge of the feature engineering pipeline and the ranked feature importance list.

In contrast, the black-box setting assumes the adversary has limited access to training data and no visibility into model architectures or parameters, reflecting a more realistic and constrained threat model.

Black-box Attack on Shallow ML and Deep GIIDS; Novel Autoencoder Residual Based Perturbation:. To support black-box attacks where gradients and internal model details are inaccessible, we introduce a model-agnostic attack leveraging autoencoders trained solely on clean input data. This is especially applicable to the non-differentiable shallow ML GIIDS.

Attack workflow.

Step 1: autoencoder training. The autoencoder is trained on the raw input data, prior to any feature engineering. Its objective is to learn to reconstruct the inputs by minimizing the reconstruction error. Through this training, it captures the general structure of the input distribution.

Step 2: perturbation generation. Once trained, the autoencoder is used to generate perturbations for each sample. Specifically, the residual between the original input and its reconstruction is computed. This residual represents the discrepancy between the data and the model's learned representation. The adversarial example is then generated by adding this residual back to the original input.

For this process, we chose structured noise over random noise to introduce more meaningful and targeted perturbations. Unlike random

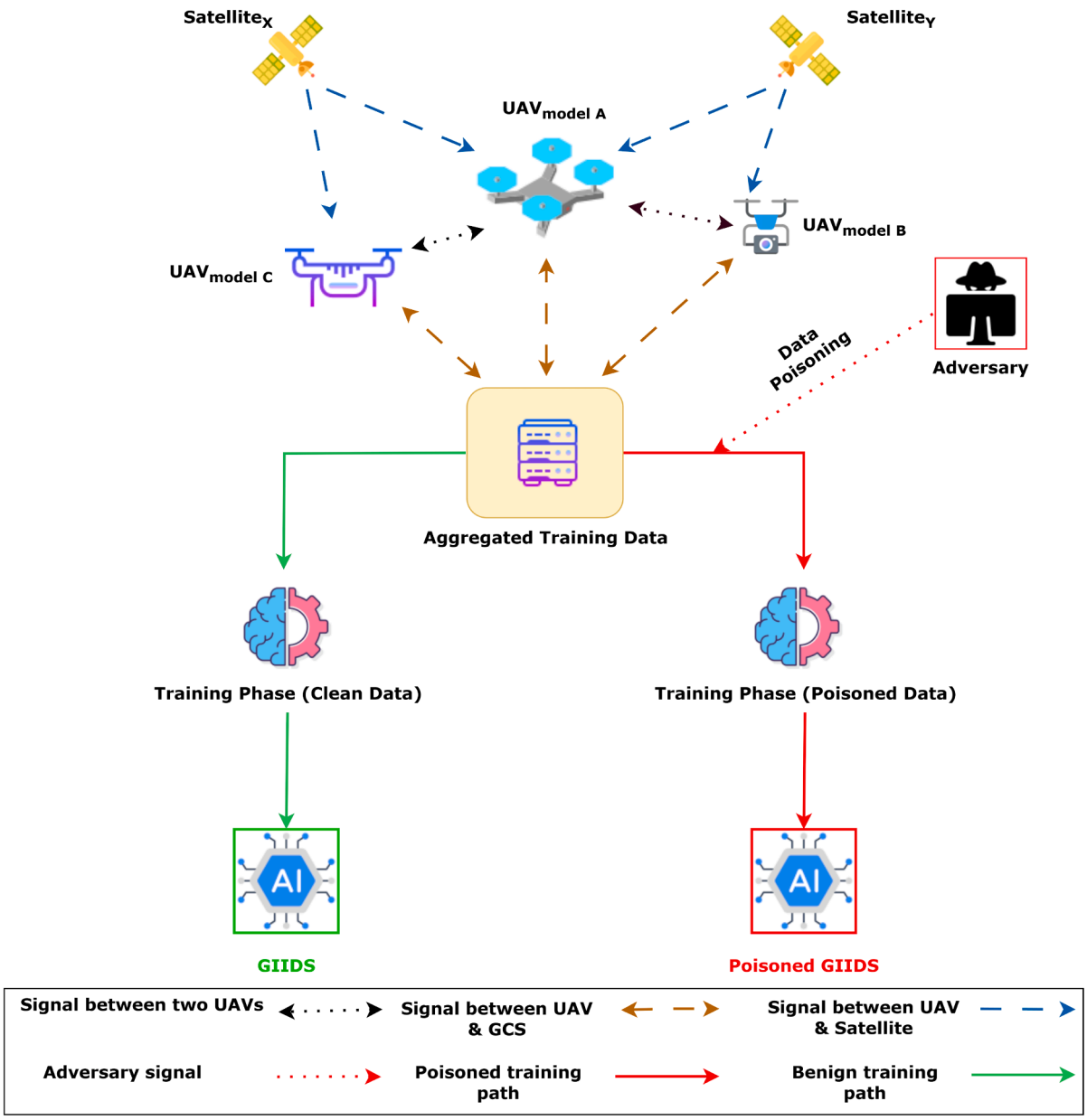


Fig. 3. Adversarial data poisoning in GIIDS.

noise such as Gaussian noise, which does not consider the data distribution and often fails to shift inputs across decision boundaries [41], the structured residual noise is data-driven. It highlights regions with high reconstruction error, often aligning with features critical to classification. This subtle, targeted perturbation leads to more realistic adversarial samples that are more likely to cause misclassification without requiring gradient access. *What makes our approach distinct is that it provides an attack mechanism tailored for non-differentiable models using an unsupervised, gradient-free, black-box strategy, yet it introduces structured, data-informed perturbations that go beyond naive noise.* Traditional AE-based attacks, in contrast, are usually white-box, classifier-aware, and rely on gradient flow or add noise directly to the AE latent space.

3.4.1.1. White-box Attack on Shallow ML GIIDS; Novel Explainability-driven Autoencoder-based Perturbation. In the white-box setting, the adversary exploits access to feature-engineered data and the ranked list of important features. The attack is crafted by perturbing only the top-ranked features using residuals derived from an autoencoder, guided by feature importance. Since the perturbation strategy is informed by model-specific explainability technique, such as feature importance from Random Forest and Gradient Boosting, the method is considered explainability-driven. This method, described in Algorithm 1, is non-gradient and explainability-driven.

3.4.1.2. White-box Attack on Deep GIIDS; PGD Attack. We conduct white-box adversarial attacks on the Deep GIIDS models, specifically the Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN), under the assumption of full knowledge of the model architecture and parameters. Using the Projected Gradient Descent (PGD) method implemented via the Foolbox library, adversarial samples are generated by iteratively perturbing the input data to maximize the model loss while constraining perturbations within an ℓ_∞ norm ball with radius $\epsilon = 0.1$. The attack runs for 20 iterations with a step size of 0.01, and inputs are clipped within the valid data range [0,1]. The adversarial samples generated independently for each model are then combined to create a comprehensive adversarial dataset for evaluation.

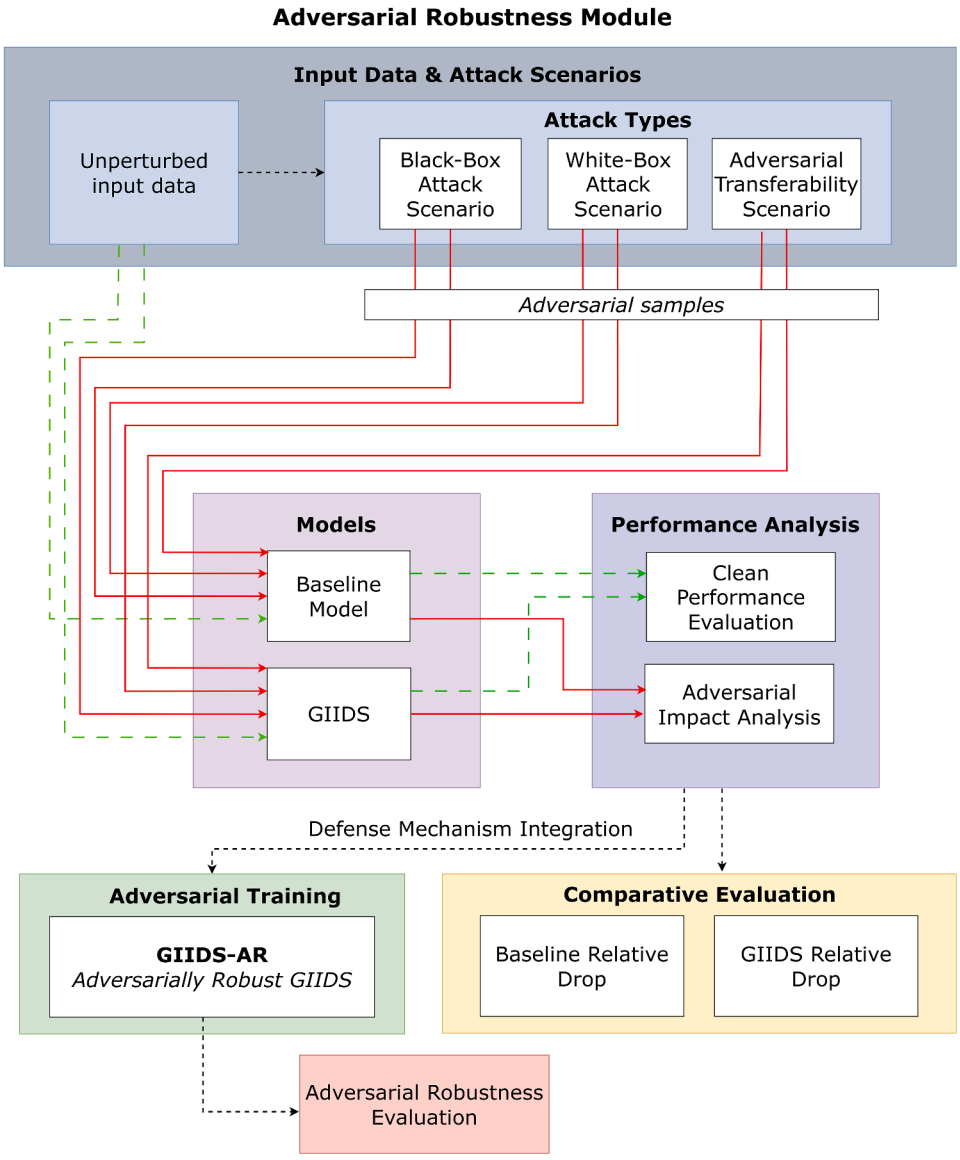


Fig. 4. Proposed adversarial robustness framework.

Algorithm 1 Generate adversarial dataset with noise on important features.

procedure GENERATEADVPARTIAL(df, $important\_features$)
  $original \leftarrow df$ without 'attack' if present
  $adv\_scaled \leftarrow scaled \leftarrow$ MinMaxScaler fit_transform of $original$
  Build and train autoencoder on $scaled$ (5 epochs, batch 32, val_split 0.1)
  $reconstructed \leftarrow$ autoencoder.predict($scaled$)
  $noise \leftarrow reconstructed - scaled$
  for all $f \in important\_features$ **do**
    $idx \leftarrow$ column index of f
    $adv\_scaled[:, idx] \leftarrow adv\_scaled[:, idx] + noise[:, idx]$
  end for
  $adversarial \leftarrow$ scaler.inverse_transform(adv_scaled)
  Convert $adversarial$ to DataFrame with original columns
  if 'attack' in df **then**
    Append 'attack' column to $adversarial$
  end if
  return $adversarial$
end procedure

3.4.1.3. Transfer Attacks on Shallow ML and Deep GIIDS. To evaluate the transferability of adversarial examples, we train a surrogate shallow neural network designed to approximate the decision boundaries of the target IDS models. The surrogate consists of two hidden layers and a softmax output layer and is trained using sparse categorical cross-entropy with the Adam optimizer. Wrapping the model with `TensorFlowV2Classifier` enables compatibility with IBM's Adversarial Robustness Toolbox (ART). Using this surrogate, adversarial examples are crafted via the Fast Gradient Sign Method (FGSM), which perturbs inputs by a single step in the direction of the gradient of the loss with respect to the input features. These adversarial examples are then tested on both shallow ML and Deep GIIDS models to assess the effectiveness of transfer attacks across architectures.

4. Implementation of proposed GIIDS-AR

4.1. Dataset description

The dataset employed in this study [42] comprises both real-time and simulated telemetry data gathered from five heterogeneous UAV platforms: H480 airframe, fixed-wing aircraft, tail-sitter VTOL (Verti-

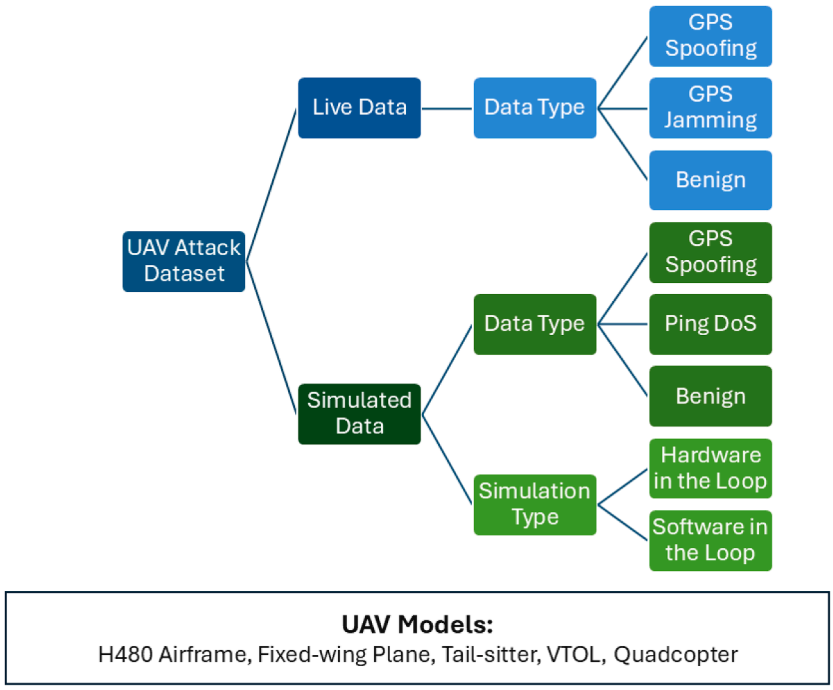


Fig. 5. Overview of the UAV attack dataset.

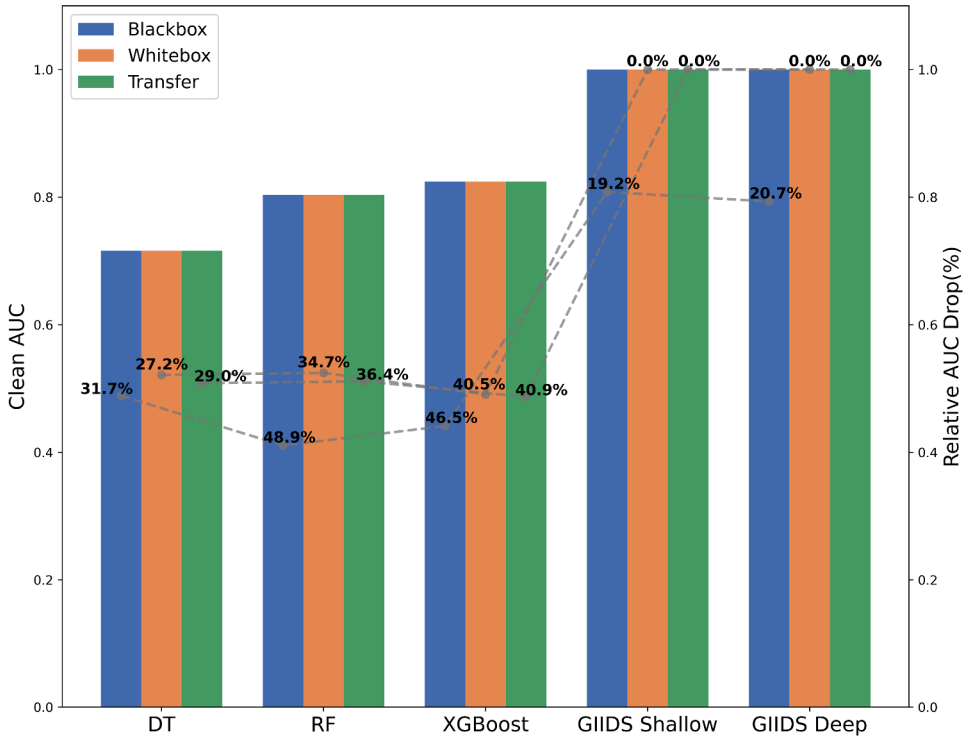


Fig. 6. Comparative analysis of relative AUC drop across GIIDS variants and baseline model.

cal Take-off and Landing), standard VTOL, and quadcopters, as shown in Fig. 5. These platforms were tested under a range of conditions, including normal operations as well as adversarial scenarios such as GPS jamming, GPS spoofing, and Ping-based DoS attacks. The diversity in UAV models and threat scenarios ensures a heterogeneous and realistic dataset, making it highly suitable for evaluating the generalization capability and adversarial robustness of the proposed GIIDS framework.

4.2. Computer configuration

The proposed system was implemented and evaluated on a machine with 32 GB of RAM and an Intel(R) Core(TM) i5-10400 processor, which has 6 cores and 12 threads. The experiments were carried out using the Google Colab CPU runtime environment. Python was used as the primary programming language, with key libraries including TensorFlow, Foolbox, and IBM Adversarial Robustness Toolbox (ART).

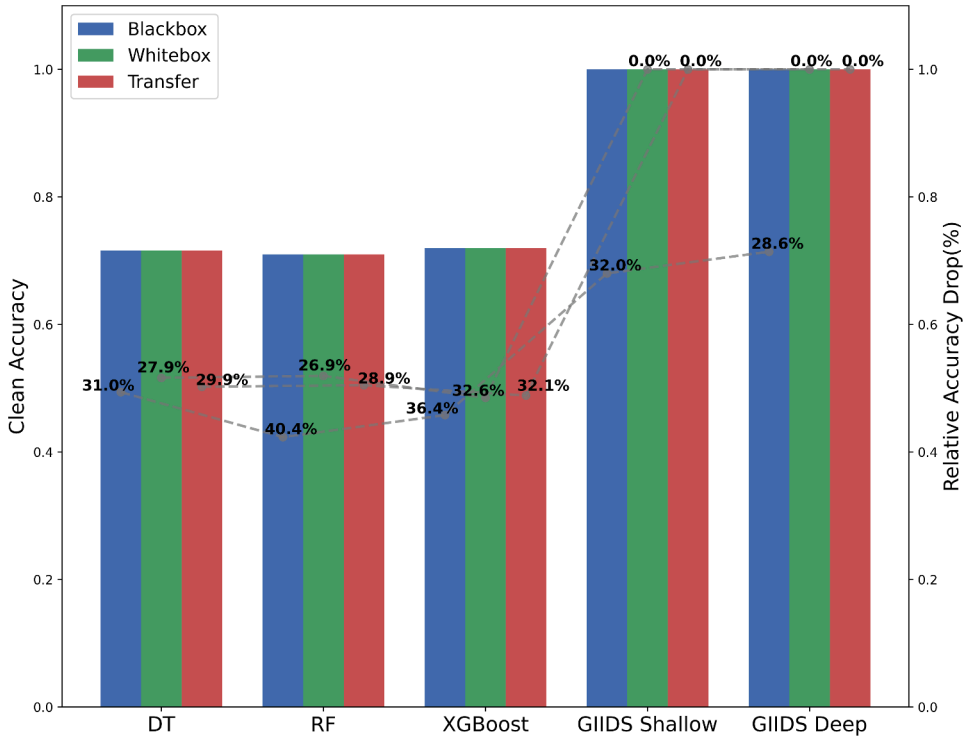


Fig. 7. Comparative analysis of relative accuracy drop across GIIDS variants and baseline model.

4.3. Adversarial robustness module

We evaluate the adversarial robustness of GIIDS under both black-box and white-box settings:

- *Black box attacks:* A structured, non-gradient black-box attack was devised using an autoencoder-based perturbation method. Without access to model internals, the adversary leverages the autoencoder's reconstruction residuals to subtly modify inputs while preserving structure. This attack was applied to both shallow ML and deep GIIDS models to evaluate their resilience under limited-access threat scenarios.
- *White box attacks:* White-box attacks assume complete knowledge of the target model, including its architecture and parameters, enabling the most informed and effective adversarial perturbations.
  - *Shallow ML GIIDS:* An explainability-driven white-box attack exploits the model's internal access, particularly feature-engineered inputs and feature importance rankings. Top-ranked features are selectively perturbed using autoencoder-derived residual noise, creating targeted adversarial samples that maintain semantic validity while misleading the classifier.
  - *Deep GIIDS:* The Projected Gradient Descent (PGD) attack was employed assuming full access to model parameters and gradients. Inputs are iteratively perturbed within an ℓ_∞-bounded region (with $\epsilon = 0.02$ and a step size of 0.005 across 20 iterations) to maximize loss and induce misclassification in both MLP and CNN-based GIIDS variants, while preserving fixed features such as time and sensor-integral values.
- *Transferability analysis:*

  To evaluate the transferability of adversarial examples, a shallow feedforward neural network (FNN) was employed as a surrogate model. This model was trained to approximate the decision boundaries of the target GIIDS systems. Adversarial samples were crafted using the Fast Gradient Sign Method (FGSM) and tested against the original GIIDS models to assess their vulnerability to transferred attacks. Table 1 shows the surrogate model architecture and attack configuration used for this analysis.

Table 1
Adversarial configuration for transferability analysis.

Parameter	Configuration
Model	Shallow Feedforward Neural Network (FNN)
Architecture	2 hidden layers
Optimizer	Adam
Learning Rate	0.001
Attack	Fast Gradient Sign Method (FGSM)
Perturbation Bound	$\epsilon = 0.1$

4.4. Adversarial dataset validation and filtering

To ensure the realism and validity of generated adversarial samples, we employed a feature-wise filtering approach based on the 3-sigma rule. For each feature in the combined clean dataset, the valid range was computed as $\mu \pm 3\sigma$, where μ and σ represent the mean and standard deviation of the feature, respectively. Adversarial samples with feature values outside these ranges were considered unrealistic and removed from further evaluation.

This filtering was applied independently to each adversarial attack dataset: black-box Autoencoder-based perturbations (for both Shallow ML and Deep GIIDS), white-box Explainability-driven Autoencoder perturbations (for Shallow ML GIIDS), white-box PGD perturbations (for Deep GIIDS), and FGSM-based transfer attacks (for both Shallow ML and Deep GIIDS).

The number of adversarial samples before and after applying the 3-sigma filtering procedure is summarized below. This filtering ensures that only valid and realistic UAV network traffic sequences are retained for evaluation:

- Black-box Autoencoder-based perturbations (applicable to both Shallow ML and Deep GIIDS): Training set reduced from 32,000 to 25,613 samples; Testing set reduced from 8000 to 5598 samples.

Table 2
Data splits for different evaluation phases. C = Clean data, A = Adversarial data.

#	Phase	Training Data	Testing Data
1	*GIIDS (Clean Data)*	80 % C	20 % C
2	*Attack on GIIDS*	100 % C / 0 % A	50 % C / 50 % A (unseen)
3	*Retraining (GIIDS-AR)*	70 % C / 30 % A	50 % C / 50 % A (unseen)

- White-box Explainability-driven Autoencoder perturbations (Shallow ML GIIDS only): Training set reduced to 21,862 samples; Testing set reduced to 2908 samples.
- White-box PGD perturbations (Deep GIIDS only): MLP training set reduced from 16,000 to 15,608 samples; CNN training set reduced from 16,000 to 15,917 samples; MLP testing set reduced from 4000 to 3993 samples; CNN testing set reduced from 4000 to 3978 samples.
- FGSM-based transfer attacks (applicable to both Shallow ML and Deep GIIDS): Training set reduced to 21,609 samples; Testing set reduced to 2836 samples.

This filtering ensures that all adversarial examples used in performance evaluation represent **feasible UAV network traffic sequences**. Subsequently, all retraining and testing experiments on GIIDS-AR utilized only these valid adversarial samples, providing a realistic assessment of adversarial robustness.

4.5. Initial experimental setup

The evaluation was conducted using filtered and validated adversarial samples to ensure realism and validity in three distinct phases to assess the performance of GIIDS under clean, adversarial, and retrained conditions. In the first phase, the original GIIDS models were trained and tested exclusively on clean data, using an 80 %-20 % train-test split. In the second phase, to assess robustness, the trained GIIDS models were evaluated on a test set composed of an equal mix of clean and filtered, valid adversarial samples (50 % each), without including any adversarial data during training. Finally, in the third phase, the GIIDS models were retrained using a mixture of clean and filtered, valid adversarial data (70 % clean, 30 % adversarial) to create GIIDS-AR, and tested again on a 50 %-50 % split of clean and filtered, valid adversarial data. Table 2 summarizes the data splits used in each of the three evaluation phases.

The Shallow ML GIIDS ensemble comprised Logistic Regression, KNN, Decision Tree, and Random Forest base models combined via a Logistic Regression stacking meta-model. The Deep GIIDS ensemble included TabNet, MLP, and CNN base models with a Logistic Regression stacking meta-model. The autoencoder used for adversarial generation had a single hidden encoding layer with half the input features. Adversarial attacks were generated using FGSM ($\epsilon = 0.1$) and PGD ($\epsilon = 0.02$, step size = 0.005, 20 iterations) with appropriate scaling and fixed features preserved.

4.6. Proposed generalized models for adversarial robustness

To evaluate the adversarial robustness of our proposed system, we examine the performance of two variants of GIIDS under multiple adversarial attack scenarios: *Shallow ML GIIDS* and *Deep GIIDS*. The Shallow ML GIIDS comprises a stacking ensemble of traditional, non-differentiable machine learning models, while the Deep GIIDS leverages a stacking ensemble of differentiable deep learning architectures.

We evaluate GIIDS-AR against three baselines: a decision tree (DT), Random Forest (RF), and XGBoost (XGB), which represent commonly used intrusion detection architectures with varying levels of complexity and robustness, and in many cases, lack sufficient generalization. All three baselines have been employed in prior works [43–45], ensuring a fair and widely recognized basis for comparison. The decision tree serves as a classical non-generalized IDS, widely adopted in intrusion detection research due to its low computational cost, interpretability, and suitability for heterogeneous data attributes that align well with the constraints of UAV-based environments. Random Forest (RF) is a learning method that constructs multiple decision trees and outputs the majority class, providing stronger robustness than a single DT by mitigating overfitting while requiring minimal additional preprocessing. XGBoost (XGB) is highly effective on tabular security datasets, representing a modern and competitive approach widely used in recent IDS research.

All three baselines are trained and evaluated on the same dataset and preprocessing pipeline as GIIDS-AR, including feature scaling and train/test splits, ensuring a fair comparison across models. The baselines are trained on data obtained exclusively from a single UAV platform (QUAD) and does not incorporate advanced generalization strategies such as feature engineering, advanced ensemble learning, or multi-platform data integration. As such, it serves as a reference point for evaluating the benefits of generalization in the Shallow ML and Deep GIIDS frameworks. By including RF and XGB alongside the classical DT, the evaluation encompasses both traditional and strong contemporary IDS techniques, providing a comprehensive benchmark and allowing for a thorough assessment of GIIDS-AR's performance and adversarial resilience relative to models commonly recognized in the literature.

4.7. Defense mechanism: adversarial training

Adversarial training serves as a fundamental defense strategy to enhance the robustness of the GIIDS framework against adversarial threats. This technique involves augmenting the training process with both clean and adversarially perturbed inputs, enabling the model to learn more robust decision boundaries. By incorporating adversarial examples during training, the system becomes more resilient to input manipulations and better equipped to operate under adversarial conditions.

In our implementation, adversarial examples are generated using several attack methods and are integrated into the training set in a fixed ratio: *70 % clean data and 30 % adversarial samples*. These perturbed instances are designed to reflect realistic adversarial scenarios pertinent to UAV networks. The impact of each attack type on model performance is evaluated independently. For each attack variant, GIIDS is retrained on a hybrid dataset combining clean and attack-specific adversarial samples to restore its original detection capability.

This retraining is performed separately for each adversarial method. To assess post-training performance, both the Shallow ML and Deep GIIDS models are evaluated on a dataset consisting of *50 % clean and 50 % adversarial inputs*, simulating a realistic deployment environment where adversarial interference is likely. This evaluation measures the effectiveness of adversarial training in recovering and maintaining detection performance in compromised settings.

The adversarial training process is seamlessly integrated into the existing training pipeline, requiring no architectural changes to the underlying models. However, it introduces an additional objective: minimizing loss on adversarial inputs alongside clean data. This dual-objective optimization enhances the system's resistance to adversarial attacks while preserving accuracy on benign inputs.

Crucially, our approach aims not only to improve adversarial robustness but also to maintain the generalization capability of GIIDS across heterogeneous UAV platforms. To assess this, we apply the Cross dataset eValuation technique CV1 [6,8,10]. This evaluation ensures that the adversarially trained GIIDS (GIIDS-AR) retains strong generalization across diverse UAV environments, thereby jointly addressing two key challenges: robustness and generalizability, that are essential for the development of trustworthy and field-deployable IDS solutions.

4.8. Performance metrics

To comprehensively evaluate the effectiveness and robustness of the proposed GIIDS, we employ the following performance metrics:

- *Accuracy:* Measures the proportion of correctly classified instances among all samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- *Precision:* Indicates the proportion of true positive predictions among all predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

- *Recall (Sensitivity):* Represents the proportion of actual positives correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

- *F1-score:* The harmonic mean of Precision and Recall, providing a balanced evaluation.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

- *AUC (Area Under the Curve):* Evaluates the ability of the model to distinguish between classes by calculating the area under the ROC curve. AUC does not have a single closed-form equation like the others, but it is generally computed using:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR} \tag{6}$$

where TPR is the true positive rate and FPR is the false positive rate.

- *Relative drop (%):* Measures the performance degradation under adversarial conditions [46].

$$\text{Relative Drop (\%)} = \left(\frac{\text{Performance}_{\text{clean}} - \text{Performance}_{\text{attack}}}{\text{Performance}_{\text{clean}}} \right) \times 100 \tag{7}$$

5. Performance evaluation of proposed GIIDS-AR

The adversarial robustness of GIIDS models was evaluated through experiments using various adversarial example (AE) generation methods. The evaluation involved four stages: 1) measuring baseline performance on clean data; 2) assessing impact of adversarial attacks on models trained with only clean samples; 3) retraining models with combined clean and adversarial data to enhance robustness; and 4) cross-attack evaluation, testing models trained on one attack type against other unseen attacks.

Subsequent subsections detail performance under black-box, white-box, and transfer-based attacks, highlighting the differences between shallow and deep variants. Cross-attack evaluation results are also included to demonstrate generalization beyond matched threat settings.

5.1. Black-box adversarial attack on GIIDS

5.1.1. Autoencoder residual-based adversarial attack on shallow ML GIIDS

Table 3 shows that the Shallow ML GIIDS model achieves perfect scores on clean data across all evaluation metrics, including Accuracy, Precision, Recall, F1 Score, and AUC. However, its performance significantly declines under black-box adversarial attacks, with accuracy dropping to 0.680 and similar decreases observed in other metrics, demonstrating the susceptibility of the clean-trained model.

Adversarial training using a dataset composed of 70 % clean and 30 % adversarial samples substantially restores the model's robustness. When evaluated on a balanced test set of 50 % clean and 50 % adversarial samples, the retrained model (GIIDS-AR) recovers near-perfect performance with accuracy of 0.999 and corresponding improvements across all metrics.

To further assess the robustness of GIIDS-AR, we performed a cross-attack evaluation where the model trained with black-box adversarial samples was tested against previously unseen attack types, including white-box autoencoder, PGD, and transferability-based attacks. As shown in Table 4, the model maintains exceptional resilience across most attack types, achieving perfect or near-perfect scores for autoencoder-driven and transfer-based attacks. While PGD attacks result in a slightly lower Accuracy of 0.814, the model still demonstrates strong robustness considering that it was never explicitly trained on this perturbation. These results highlight that GIIDS-AR generalizes well to diverse adversarial scenarios, confirming the effectiveness of our adversarial training strategy.

Table 3
Performance of shallow ML GIIDS under residual-based black-box adversarial attack and adversarial training.

Category	Accuracy	Precision	Recall	F1 Score	AUC
GIIDS (Clean)	1.000	1.000	1.000	1.000	1.000
Attack on GIIDS (No Defense)	0.680	0.680	0.680	0.679	0.808
GIIDS-AR (Adversarial Training)	0.999	0.999	0.999	0.999	1.000

Table 4
Cross-attack evaluation of shallow ML GIIDS trained with black-box adversarial samples.

Attack Tested	Accuracy	Precision	Recall	F1 Score	AUC
White-box Autoencoder	1.000	1.000	1.000	1.000	1.000
White-box PGD	0.814	0.834	0.814	0.810	0.950
Transferability Attack	1.000	1.000	1.000	1.000	1.000

Table 5
Performance of deep GIIDS under residual-based black-box adversarial attack and adversarial training.

Category	Accuracy	Precision	Recall	F1 Score	AUC
GIIDS (Clean)	1.000	1.000	1.000	1.000	1.000
Attack on GIIDS (No Defense)	0.714	0.745	0.714	0.707	0.793
GIIDS-AR (Adversarial Training)	0.999	0.999	0.999	0.999	1.000

5.1.2. Autoencoder residual-based adversarial attack on deep GIIDS

Table 5 shows that the Deep GIIDS model achieves perfect scores on clean data across all evaluation metrics, including Accuracy, Precision, Recall, F1 Score, and AUC. However, its performance significantly declines under black-box adversarial attacks, with recall remaining relatively high at 0.714, while precision falls to 0.745. This results in a reduced F1 score of 0.707 and an AUC of 0.793, indicating vulnerability to adversarial perturbations even in deep architectures.

Adversarial training using a dataset composed of 70 % clean and 30 % adversarial samples substantially restores the model's robustness. When evaluated on a balanced test set of 50 % clean and 50 % adversarial samples, the retrained model (GIIDS-AR) recovers near-perfect performance with accuracy of 0.999 and corresponding improvements across all metrics.

To further assess the robustness of Deep GIIDS-AR, we performed a cross-attack evaluation where the model trained with black-box adversarial samples was tested against previously unseen attack types, including white-box autoencoder, PGD, and transferability-based attacks. As shown in Table 6, the model maintains strong resilience across most attack types, achieving perfect scores for autoencoder-driven and transfer-based attacks. While PGD attacks result in a slightly lower Accuracy of 0.808, the model still demonstrates robust defense despite not being explicitly trained on this perturbation. These results indicate that Deep GIIDS-AR generalizes well to diverse adversarial scenarios, confirming the effectiveness of our adversarial training strategy.

Table 6
Cross-attack evaluation of deep GIIDS trained with black-box adversarial samples.

Attack Tested	Accuracy	Precision	Recall	F1 Score	AUC
White-box Autoencoder	1.000	1.000	1.000	1.000	1.000
White-box PGD	0.808	0.833	0.808	0.805	0.746
Transferability Attack	1.000	1.000	1.000	1.000	1.000

Table 7
Performance of shallow ML GIIDS under white-box adversarial evaluation.

Category	Accuracy	Precision	Recall	F1 Score	AUC
GIIDS (Clean)	1.0000	1.0000	1.0000	1.0000	1.0000
Attack on GIIDS (No Defense)	1.0000	1.0000	1.0000	1.0000	1.0000
GIIDS-AR (Adversarial Training)	1.0000	1.0000	1.0000	1.0000	1.0000

Table 8
Cross-attack evaluation of shallow ML GIIDS trained with white-box autoencoder adversarial samples.

Attack Tested	Accuracy	Precision	Recall	F1 Score	AUC
Black-box	0.667	0.667	0.667	0.666	0.810
White-box PGD	0.841	0.871	0.841	0.838	0.985
Transferability Attack	1.000	1.000	1.000	1.000	1.000

Table 9
Performance of deep GIIDS under white-box PGD adversarial evaluation.

Category	Accuracy	Precision	Recall	F1 Score	AUC
GIIDS (Clean)	1.000	1.000	1.000	1.000	1.000
Attack on GIIDS (No Defense)	0.548	0.639	0.548	0.466	0.512
GIIDS-AR (Adversarial Training)	0.991	0.991	0.991	0.991	0.998

Table 10
Cross-attack evaluation of deep GIIDS trained with PGD adversarial samples.

Attack Tested	Accuracy	Precision	Recall	F1 Score	AUC
Black-box Attack	0.734	0.737	0.734	0.733	0.835
White-box Autoencoder	1.000	1.000	1.000	1.000	1.000
Transferability Attack	1.000	1.000	1.000	1.000	1.000

Table 11
Transferability-based adversarial evaluation of shallow ML GIIDS.

Category	Accuracy	Precision	Recall	F1 Score	AUC
GIIDS (Clean)	1.0000	1.0000	1.0000	1.0000	1.0000
Attack on GIIDS (No Defense)	1.0000	1.0000	1.0000	1.0000	1.0000
GIIDS-AR (Adversarial Training)	1.0000	1.0000	1.0000	1.0000	1.0000

Table 12
Cross-attack evaluation of shallow ML GIIDS trained with transfer-based adversarial samples.

Attack Tested	Accuracy	Precision	Recall	F1 Score	AUC
Black-box	0.999	0.999	0.999	0.999	1.000
White-box Autoencoder	1.000	1.000	1.000	1.000	1.000
White-box PGD	0.814	0.834	0.814	0.810	0.950

Table 13
Transferability-based adversarial evaluation of deep GIIDS.

Category	Accuracy	Precision	Recall	F1 Score	AUC
GIIDS (Clean)	1.0000	1.0000	1.0000	1.0000	1.0000
Attack on GIIDS (No Defense)	1.0000	1.0000	1.0000	1.0000	1.0000
GIIDS-AR (Adversarial Training)	1.0000	1.0000	1.0000	1.0000	1.0000

Table 14
Cross-attack evaluation of Deep GIIDS trained with transfer adversarial samples.

Attack Tested	Accuracy	Precision	Recall	F1 Score	AUC
Black-box (Residual)	0.734	0.737	0.734	0.733	0.835
White-box Autoencoder	1.000	1.000	1.000	1.000	1.000
White-box PGD	0.810	0.863	0.810	0.804	0.880

5.2. White box adversarial attack on GIIDS

5.2.1. Explainability-driven autoencoder-based adversarial attack on shallow ML GIIDS

This experiment evaluates the robustness of the Shallow ML GIIDS model under a white-box adversarial attack. The model was tested on a dataset containing an equal mix of clean and adversarial samples (50 % each). As shown in Table 7, the model maintained perfect scores across all metrics, demonstrating strong resistance to the attack.

Notably, the generalized Shallow GIIDS model maintained perfect performance during both the attack and adversarial retraining phases, with accuracy, precision, recall, F1-score, and AUC all remaining at 1.0000. This outcome demonstrates the strength and inherent robustness of the generalized system, which was resilient enough to withstand white-box adversarial attacks without any degradation. In contrast, the same attack significantly deteriorated the performance of the non-generalized baseline model, as discussed in Section 5.6. These results suggest that the applied generalization techniques in the system, leading to strong generalization capabilities, combined with the non-differentiable nature of its traditional ML components, contributed to its resistance against this form of adversarial manipulation.

To further assess the robustness of Shallow ML GIIDS-AR, we performed a cross-attack evaluation where the model trained with white-box autoencoder adversarial samples was tested against previously unseen attack types, including black-box, PGD, and transferability-based attacks. As shown in Table 8, the model maintains strong resilience across most attack types, achieving perfect scores for transfer-based attacks. While black-box and PGD attacks result in slightly lower Accuracy of 0.667 and 0.841, respectively, the model still demonstrates robust defense despite not being explicitly trained on these perturbations. These results indicate that Shallow ML GIIDS-AR generalizes well to diverse adversarial scenarios, confirming the effectiveness of our adversarial training strategy.

5.2.2. PGD-based adversarial attack on deep GIIDS

This section presents the white-box adversarial evaluation of the Deep GIIDS model. Table 9 shows that the initial clean performance of Deep GIIDS is perfect across all metrics. However, when subjected to a white-box PGD attack, the performance of the clean-trained model drops substantially, with accuracy falling to 0.548 and the F1 score decreasing to 0.466, indicating that the model is highly vulnerable to PGD perturbations.

After applying adversarial retraining using PGD-generated samples, the model demonstrates remarkable resilience. The retrained Deep GIIDS (GIIDS-AR) achieves near-perfect performance, with accuracy and all other metrics reaching 0.991 or higher. This confirms that adversarial training effectively restores robustness against PGD attacks while preserving the model's detection capabilities.

To further assess the robustness of Deep GIIDS-AR, we performed a cross-attack evaluation where the model trained with PGD adversarial samples was tested against previously unseen attack types, including white-box autoencoder, transferability-based, and black-box attacks. As shown in Table 10, the model maintains perfect performance on autoencoder-driven and transfer-based attacks. While black-box attacks result in a slightly lower Accuracy of 0.734, the model still demonstrates strong resilience despite not being explicitly trained on this perturbation. These results indicate that PGD-trained Deep GIIDS-AR generalizes effectively to diverse adversarial scenarios.

5.3. Transferability analysis of GIIDS

In this experiment, we examine the transferability of adversarial examples, originally crafted to attack deep models, against the Shallow ML GIIDS model. This black-box evaluation assesses whether attacks generated from a separate model can still affect a model with a different architecture. Such a scenario mimics a realistic adversary who does

Table 15
Execution time (in seconds) of GIIDS-AR models across adversarial attack scenarios.

Attack Type	Model Type	Feature Engineering & Analysis Time (s)	Training Time (s)	Inference Time (s)
Black-Box	Shallow ML GIIDS-AR	44.00	1.63	0.03
	Deep GIIDS-AR	21.00	651.20	0.95
White-Box	Shallow ML GIIDS-AR	31.00	62.87	6.47
	Deep GIIDS-AR	26.00	1084.43	2.04
Transfer Attacks	Shallow ML GIIDS-AR	30.00	177.01	3.25
	Deep GIIDS-AR	27.00	1026.73	1.56

not have access to the target model but leverages a surrogate model to generate attacks.

5.3.1. Transfer attacks on shallow ML GIIDS

As shown in Table 11, the Shallow ML GIIDS maintained perfect performance under transfer attacks, with Accuracy, Precision, Recall, F1 Score, and AUC all equal to 1.000 during both the attack phase and after adversarial retraining. This demonstrates the strong inherent robustness of the generalized Shallow ML GIIDS model against transferability-based adversarial attacks.

To further assess the robustness of GIIDS-AR trained with transfer-based adversarial samples, we performed a cross-attack evaluation where the model was tested against previously unseen attack types, including black-box, white-box autoencoder, and PGD attacks. As shown in Table 12, the model maintains strong resilience across most attack types, achieving perfect scores for black-box and white-box autoencoder attacks. While PGD attacks result in a slightly lower Accuracy of 0.814, the model still demonstrates robust defense despite not being explicitly trained on this perturbation. These results highlight that GIIDS-AR generalizes well to diverse adversarial scenarios, confirming the effectiveness of transfer-based adversarial training.

5.3.2. Transfer attack on deep GIIDS

This section investigates the transferability of adversarial examples from other models to the Deep GIIDS. Transferability is a key threat model in adversarial machine learning, where an attacker generates adversarial samples using a surrogate model and tests them on a different target model without direct access. This white-box to black-box transfer scenario is realistic and practical in deployment environments.

As shown in Table 13, Deep GIIDS demonstrates perfect performance under transfer attacks, with all metrics remaining at 1.0000 during the adversarial phase. Following adversarial retraining, the model maintains perfect scores across all metrics, indicating strong inherent robustness in the deep model against transfer attacks. These results reaffirm the effectiveness of incorporating adversarial examples during training and highlight that Deep GIIDS remains highly effective in both white-box and black-box transfer scenarios.

To further assess the robustness of Deep GIIDS-AR, we performed a cross-attack evaluation where the model trained with transfer adversarial samples was tested against previously unseen attack types, including black-box residual, white-box autoencoder, and PGD attacks. As shown in Table 14, the model maintains strong resilience across most attack types, achieving perfect scores for white-box autoencoder attacks. While black-box and PGD attacks result in slightly lower Accuracy of 0.734 and 0.810, respectively, the model still demonstrates robust defense despite not being explicitly trained on these perturbations. These results indicate that Deep GIIDS-AR generalizes well to diverse adversarial scenarios, confirming the effectiveness of our adversarial training strategy.

5.4. Execution time analysis of GIIDS-AR models

This section presents a comparative analysis of the execution time of GIIDS-AR models under different adversarial attack scenarios. The evaluation considers three key components of execution: feature engineering and analysis time, training time, and inference time. Both shallow ML and deep variants of GIIDS-AR are assessed against black-box, white-box, and transfer attack settings, as shown in Table 15.

5.4.1. Feature engineering and analysis time

The feature engineering and analysis process was conducted prior to model training for both shallow ML and deep variants, and is therefore independent of the model architecture. Specifically, we combined time-based and raw UAV features, then retained only those deemed important by both Random Forest and Gradient Boosting models. However, since the retained features varied across attack types due to differing importance rankings, the resulting feature sets fed to the models had different dimensionalities. This variation explains the slight differences observed in feature engineering and analysis time across attack scenarios, rather than any model-specific preprocessing overhead.

5.4.2. Training time

Training time shows a sharp contrast between the shallow ML and deep GIIDS-AR models. The deep ensemble consistently requires significantly longer training durations. For example, under white-box attacks, the deep model takes 1084.43 seconds to train, compared to just 62.87 seconds for the shallow model. Even in the transfer attack scenario, the deep model's training time remains high at 1026.73 seconds. This difference is expected, as deep learning models typically involve a larger number of parameters, require iterative backpropagation, and often depend on GPU acceleration for efficient optimization. Additionally, the stacking ensemble structure amplifies this cost, since multiple deep base models and a meta-learner must be trained. In contrast, the shallow ML ensemble relies on traditional machine learning algorithms with simpler optimization processes and shorter convergence times, resulting in substantially faster training.

5.4.3. Inference time

In the black-box setting, shallow ML GIIDS-AR is the fastest, with an inference time of just 0.03 seconds, compared to 0.95 seconds for the deep model. However, in the white-box and transfer scenarios, the shallow models are noticeably slower (6.47 and 3.25 seconds respectively) than their deep counterparts (2.04 and 1.56 seconds). This trend can be attributed to the nature of the stacking ensemble used in both models. In the shallow ML ensemble, each base learner must complete its prediction before the meta-learner can aggregate the outputs. Under adversarial conditions with increased input complexity, base models such as K-Nearest Neighbors and decision trees become more computationally intensive, leading to higher inference latency. In contrast, the deep ensemble benefits from faster base learner predictions due to efficient vectorized computations. Despite the deep model's longer training time, its inference remains faster and more stable across complex scenarios because the operations involved in forward passes are more optimized for high-dimensional inputs.

5.5. Computational resource usage of GIIDS-AR models

The computational resource usage of GIIDS-AR was measured during adversarial training across black-box, white-box, and transfer attack sce-

narios. CPU utilization was consistently modest, ranging between 22 % and 24 % for both the Shallow ML and Deep GIIDS-AR models. Memory consumption varied depending on model complexity: the Shallow ML model required approximately 565–650 MB of RAM, while the Deep GIIDS-AR model required about 1.6–1.7 GB. These measurements reflect resource usage over the entire adversarial training datasets, which contain tens of thousands of samples. In practical UAV deployment, however, inference is performed on small streaming batches rather than full datasets, reducing per-sample requirements to well below 100 ms and minimizing CPU and memory demands. Training remains an offline process carried out on dedicated workstations, while the operational resource footprint on UAVs stays lightweight. This confirms the feasibility of deploying GIIDS-AR in resource-constrained UAV environments without compromising real-time performance.

5.6. Generalization techniques enhance adversarial robustness: empirical evidence

A key insight emerging from our experiments is that generalization techniques not only improve performance on unseen data but also enhance resilience to adversarial attacks. Compared to conventional baselines such as Decision Tree (DT), Random Forest (RF), and XGBoost (XGB), well-generalized models demonstrate stronger adversarial robustness even before the application of dedicated defense mechanisms. This suggests that careful design choices aimed at improving generalization, such as diverse training data, feature engineering, and ensemble learning, can implicitly harden the system against adversarial threats.

To provide evidence, we assess the effect of AML attacks on GIIDS (generalized) and three baselines: Decision Tree (DT), Random Forest (RF), and XGBoost (XGB), which might incorporate some generalization but are not as robustly generalized as our proposed models, by comparing their relative performance drops. In the white-box setting, we employed the explainability-driven autoencoder-based attack uniformly across all models to ensure a fair comparison, rather than using a PGD-based white-box attack solely for the Deep GIIDS.

Figs. 6 and 7 present the clean Accuracy and AUC of DT, RF, XGBoost, GIIDS Shallow, and GIIDS Deep along with their relative drops under Black-box, White-box, and Transfer adversarial attacks. In the figures, bars represent clean metrics, dotted lines indicate post-attack performance for each attack type, and text labels show relative drop percentages. Both GIIDS variants start with perfect clean Accuracy and AUC (1.0), substantially higher than the baselines, which range from approximately 0.49-0.72 for Accuracy and 0.72-0.82 for AUC. Under Black-box attacks, GIIDS experience moderate relative drops (Accuracy: 29-32 %, AUC: 19-21 %), whereas the baselines suffer larger degradations (Accuracy: 28-41 %, AUC: 32-49 %). For White-box and Transfer attacks, GIIDS maintain near-perfect performance with negligible drop, while the baselines show minor decreases under Transfer (Accuracy: 3-4 %, AUC: 27-41 %). Consequently, post-attack Accuracy and AUC remain high for GIIDS (Accuracy: 0.68-1.0, AUC: 0.79-1.0), consistently outperforming all baseline models and demonstrating both superior detection capability and strong resilience under adversarial conditions.

These results strongly support the hypothesis that incorporating generalization techniques inherently enhances adversarial robustness. The superior resilience of GIIDS Shallow and Deep models, particularly against White-box and Transfer attacks, indicates that learning generalizable and robust features reduces vulnerability even when adversaries have full or partial knowledge of the model. The moderate performance degradation under Black-box attacks further demonstrates GIIDS's capacity to withstand adversarial perturbations crafted without direct access to the target model. Overall, the consistent outperformance of GIIDS compared to all three baselines, Decision Tree, Random Forest, and XGBoost confirms that efforts to improve model generalization naturally translate into greater robustness, reinforcing the strong link between these two desirable properties in UAV intrusion detection systems.

6. Conclusion

This paper presented GIIDS-AR, an adversarially robust extension of the GIIDS framework designed for heterogeneous UAVs operating in Urban Air Mobility (UAM) environments. By integrating generalization strategies such as diverse training data, feature engineering, and ensemble learning, GIIDS-AR demonstrated strong detection performance and adaptability across multiple UAV platforms. Furthermore, the framework incorporates adversarial robustness techniques, including adversarial training with FGSM and PGD attacks, as well as a feature-wise filtered adversarial dataset to ensure realistic and valid perturbations. Our empirical analysis revealed that these generalization and robustness strategies collectively enable GIIDS-AR models to consistently outperform both shallow and deep baseline models under white-box, black-box, and transfer attack scenarios. Comparative evaluations against non-adversarially-robust IDS baselines, such as Decision Trees, Random Forests, and XGBoost, further highlight the unique advantages of GIIDS-AR in terms of resilience and general performance. These results establish a strong link between generalization and adversarial robustness, reinforcing the importance of designing IDSs that are both adaptive and secure.

Future work: Moving forward, we plan to enhance the explainability of GIIDS-AR, particularly its deep learning components, by incorporating model-agnostic and causal explainability techniques. This will allow stakeholders to understand the decision-making process of the IDS, interpret anomaly detections, and perform counterfactual reasoning for potential attacks. In addition, we aim to optimize GIIDS-AR for deployment in resource-constrained UAV environments by exploring model compression, pruning, and lightweight architectures. These efforts will further improve the system's efficiency while maintaining high detection accuracy and robustness against adversarial threats.

CRediT authorship contribution statement

Fahmina Kabir: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization; **Nishat I Mowla:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Thomas Rosenstatter:** Writing – review & editing, Formal analysis; **Inshil Doh:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (RS-2023-NR076673). Inshil Doh is the corresponding author.

References

[1] Z. Yu, Z. Wang, J. Yu, D. Liu, H.H. Song, Z. Li, Cybersecurity of unmanned aerial vehicles: a survey, Proc 39 (2023) 182–215.
[2] B.S. Mohammed, K.K. Basha, Unmanned aerial vehicles for search and rescue in natural disasters: a review, Proc. IEEE Access 8 (2020) 19073–19089.
[3] J. Whelan, A. Almehmadi, K. El-Khatib, Artificial intelligence for intrusion detection systems in unmanned aerial vehicles, Comput. Electr. Eng. 99 (2022) 107784.

[4] K. Cengiz, S. Lipsa, R.K. Dash, N. Ivković, M. Konecki, A novel intrusion detection system based on artificial neural network and genetic algorithm with a new dimensionality reduction technique for UAV communication, IEEE Access 12 (2024) 4925–4937.
[5] M.S. Munir, S.H. Dipro, K. Hasan, T. Islam, S. Shetty, Artificial intelligence-enabled exploratory cyber-physical safety analyzer framework for civilian urban air mobility, Appl. Sci. 13 (2) (2023) 755.
[6] S.C. Hassler, U.A. Mughal, M. Ismail, Cyber-physical intrusion detection system for unmanned aerial vehicles, IEEE Trans. Intell. Transp. Syst. 24 (7) (2023) 4123–4135.
[7] F. Maleki, K. Ovens, R. Gupta, C. Reinhold, A. Spatz, R. Forghani, Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls, 2022. arXiv preprint arXiv:2202.01337
[8] M. Benchekroun, P.E. Velmovitsky, D. Istrate, V. Zalc, P.P. Morita, D. Lenne, Cross dataset analysis for generalizability of HRV-based stress detection models, Sensors 23 (4) (2023) 1807.
[9] V. Gohil, S. Dev, G. Upasani, D. Lo, P. Ranganathan, C. Delimitrou, The importance of generalizability in machine learning for systems, IEEE Comput. Archit. Lett. 24 (1) (2024) 1–4.
[10] F. Kabir, N.I. Mowla, I. Doh, GIIDS: generalized intelligent intrusion detection system for heterogeneous UAVs in UAM, in: Proc. 2025 27th Int. Conf. on Advanced Communications Technology (ICACT), 2025, pp. 377–382.
[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, 2013, pp. 1–10. arXiv preprint arXiv:1312.6199
[12] S. Alkadi, S. Al-Ahmadi, M.M.B. Ismail, Robens: robust ensemble adversarial machine learning framework for securing IoT traffic, Sensors 24 (8) (2024) 2626.
[13] M.S. Haroon, H.M. Ali, Adversarial training against adversarial attacks for machine learning-based intrusion detection systems, Computers 73 (2022).
[14] K. He, D.D. Kim, M.R. Asghar, Adversarial machine learning for network intrusion detection systems: a comprehensive survey, IEEE Commun. Surv. Tutor. 25 (1) (2023) 538–566.
[15] W. Zhao, S. Alwidian, Q.H. Mahmoud, Adversarial training methods for deep learning: a systematic review, Algorithms 15 (8) (2022) 283.
[16] J. Whelan, T. Sangarapillai, O. Minawi, A. Almehmadi, K. El-Khatib, Novelty-based intrusion detection of sensor attacks on unmanned aerial vehicles, in: Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks, 2020, pp. 23–28.
[17] A. Burhan, Mohammed, L.C. Fourati, A.M. Fakhrudeen, Comprehensive systematic review of intelligent approaches in UAV-based intrusion detection, blockchain, and network security, Comput. Netw. 239 (2024) 110140.
[18] K. Das, C. Ghosh, R. Karmakar, Eavesdropping attack detection in UAVs using ensemble learning, in: 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2023, pp. 1–07.
[19] K. Das, R. Basu, R. Karmakar, Man-in-the-middle attack detection using ensemble learning, in: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2022, pp. 1–6.
[20] A. Burhan, Mohammed, L.C. Fourati, A.M. Fakhrudeen, Isolation forest algorithm against UAV's GPS spoofing attack, in: 2024 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, 2024, pp. 459–463.
[21] R. Yu, A. Li, C.-F. Chen, J.-H. Lai, V.I. Morariu, X. Han, M. Gao, C.-Y. Lin, L.S. Davis, Nisp: pruning networks using neuron importance score propagation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9194–9203.
[22] K.A. Dhanya, S. Vajipayajula, K. Srinivasan, A. Tibrewal, T.S. Kumar, T.G. Kumar, Detection of network attacks using machine learning and deep learning models, Procedia Comput. Sci. 218 (2023) 57–66.
[23] H. Slimane, S. Ould, T. Benouadah, N.T. Khoei, Kaabouch, A light boosting-based ml model for detecting deceptive jamming attacks on uavs, in: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2022, pp. 328–0333.
[24] V. Ihekoronye, S. Ukamaka, D.-S.O. Ajakwe, J.M. Kim, Lee, Cyber edge intelligent intrusion detection framework for uav network based on random forest algorithm, in: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2022, pp. 1242–1247.
[25] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy (Sp), IEEE, 2017, pp. 39–57.
[26] K. Zhou, Y. Yang, Y. Qiao, T. Xiang, Domain adaptive ensemble learning, IEEE Trans. Image Process. 30 (2021) 8008–8018.
[27] S. Hassler, U. Chase, M.A. Mughal, Ismail, Cyber-physical intrusion detection system for unmanned aerial vehicles, IEEE Trans. Intell. Transp. Syst. 25 (6) (2023) 6106–6117.
[28] O. Sagi, L. Rokach, Ensemble learning: a survey, Wiley Interdiscip. Rev. 8 (4) (2018) 1249.
[29] R. Beltiukov, W. Guo, A. Gupta, W. Willinger, In search of netunicorn: a data-Collection platform to develop generalizable ML models for network security problems, in: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, 2023, pp. 2217–2231.
[30] G.L. Peterson, B.T. Mcbride, The importance of generalizability for anomaly detection, Knowl. Inf. Syst. 14 (3) (2008) 377–392.
[31] R.H. Shumway, D.S. Stoffer, Time Series Analysis and Its Applications: With R Examples, New York, Springer, 2017. 4th ed.
[32] M. Nasir, A.R. Javed, M.A. Tariq, M. Asim, T. Baker, Feature engineering and deep learning-based intrusion detection framework for securing edge IoT, J. Supercomput. 2022, pp. 1–15.
[33] P. Machado, B. Fernandes, P. Novais, Benchmarking data augmentation techniques for tabular data, in: International Conference on Intelligent Data Engineering and Automated Learning, Cham, Springer International Publishing, 2022, pp. 104–112.
[34] F. Tlili, S. Ayed, L.C. Fourati, Advancing UAV security with artificial intelligence: a comprehensive survey of techniques and future directions, Internet Things 19 (2024) 101281.
[35] H.-J. Ye, Q. Zhou, D.-C. Zhan, Training-free generalization on heterogeneous tabular data via meta-representation, Technical Report, preprint, 2023.
[36] L. Huang, A.D. Joseph, B. Nelson, B.I. Rubinstein, J.D. Tygar, Adversarial machine learning, in: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, 2011, pp. 43–58.
[37] W. Xu, D. Evans, Y. Qi, Feature squeezing: Detecting adversarial examples in deep neural networks,2017. arXiv preprint arXiv:1704.01155
[38] Y. Lecun, C. Cortes, C.J.C. Burges, The MNIST Database of Handwritten Digits, 1998. http://yann.lecun.com/exdb/mnist/.
[39] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images, Technical Report, University of Toronto, 2009. https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 2009, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848
[41] B. Huang, Y. Wang, W. Wang, Model-agnostic adversarial detection by random perturbations, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), the 28th International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 4689–4696.
[42] P. Uav, Dataset, 2024. https://ieee-dataport.org/open-access/uav-attack-dataset.
[43] Z. Chen, L. Zhou, W. Yu, ADASYN- random forest based intrusion detection model, in: Proceedings of the 2021 4th International Conference on Signal Processing and Machine Learning, 2021, pp. 152–159.
[44] A. Almeida, M. Asif, M.T. Rahman, M.A. Rahman, Side-Channel-Driven intrusion detection system for mission critical unmanned aerial vehicles, in: 2024 25th International Symposium on Quality Electronic Design (ISQED), IEEE, 2024, pp. 1–9.
[45] S.S. Dhaliwal, A.-A. Nahid, R. Abbas, Effective Intrusion Detection System Using XGBoost, 9, MDPI, 2018.
[46] K. Sharma, Imperceptible adversarial attacks on discrete-time dynamic graph models, in: NeurIPS 2022 Temporal Graph Learning Workshop, 2022.