Impact of Query and Non-Query Black-Box Evasion Attacks on UAV Intrusion Detection Systems

Salma Aboelmagd*, Layhan Mishra†, Muhammad Ismail‡, Abdulrahman Takiddin*
*Department of Electrical and Computer Engineering, Florida State University, Tallahassee, FL, USA
†Center for Advanced Power Systems, Florida State University, Tallahassee, FL, USA
‡Cybersecurity Education, Research, and Outreach and Computer Science Dept., Tennessee Tech Uni, Cookeville, TN, USA
{saboelmagd, lem24a, a.takiddin}@fsu.edu, mismail@tntech.edu

***Abstract*—Unmanned aerial vehicles (UAVs) are increasingly integrated into modern communication and sensing infrastructures, making onboard intrusion detection systems (IDSs) essential for resilience. While recent research has explored adversarial threats to UAV IDSs, most studies focus on white-box settings or assume full access to training data and model parameters. This paper investigates the vulnerability of machine learning (ML)-based UAV IDSs to black-box adversarial evasion attacks, where the adversary lacks access to model parameters. We utilize two fused cyber-physical datasets collected from UAV testbeds: one to train the operator IDS and one as a surrogate dataset to train the attacker model, mimicking black-box conditions. We evaluate ML-based IDS models under two classes of black-box attacks: (i) non-query-based attacks, which rely on transferring perturbations from a surrogate model, and (ii) query-based attacks, which iteratively craft adversarial examples using only output feedback from the operator model, without requiring a surrogate model, making them more practical. Experimental results show that shallow ML models suffer detection rate (DR) degradations of up to 85% when subject to black-box evasion attacks compared to classic attacks (i.e., false data injection attacks), while deep ML models exhibit more moderate DR declines of up to 64%. Query-based evasion attacks result in a greater degradation in DR compared to non-query-based attacks of up to 67%, highlighting the elevated risk posed by query-based threats in black-box settings.**

***Index Terms*—adversarial evasion attacks, black-box attacks, cyber-physical features, machine learning, transferability-based attacks.**

I. Introduction

Unmanned aerial vehicles (UAVs) are increasingly deployed in civilian, commercial, and defense applications, operating as aerial base stations, relays, and edge nodes to extend connectivity and processing in next-generation networks [1]. As cyber-physical systems, UAVs integrate two interconnected layers: a cyber layer and a physical layer. The cyber layer comprises components such as communication protocols, MAC addresses, routing tables, and control commands, while the physical layer includes sensors and actuators responsible for flight stability and navigation, such as GPS, gyroscopes, accelerometers, and barometers. This fusion exposes UAVs to a broader attack surface across both cyber and physical domains [2]. Recent advances in adversarial machine learning have shown that even well-trained intrusion detection systems (IDSs) can be evaded by carefully crafted adversarial examples. However, most UAV IDS research to date has focused on white-box assumptions, individual attack types, or isolated feature sets (either cyber or physical) [3]. In contrast, black-box threat models, where the adversary lacks direct access to the operator model's parameters or gradients, remain underexplored, despite better reflecting realistic attack scenarios. In black-box settings, adversaries may still exploit knowledge of the system's input-output behavior or data format to craft evasion samples using either query-based or non-query-based techniques. Query-based attacks rely on probing the operator model through repeated inputs and observing outputs, while non-query-based attacks generate adversarial examples using surrogate models and transfer them to the black-box target.

Part of this work was supported by NSF under Grant 2220346.

A. Related Works and Limitations

IDSs for UAVs have primarily focused on defending against specific classic attacks that disrupt system behavior, such as false data injection attacks (FDIAs), denial-of-service (DoS), or signal spoofing. In an FDIA, adversaries manipulate telemetry data to mislead the UAV's control system or the IDS itself. More recently, evasion attacks have emerged as a growing concern, where adversarial examples are crafted to bypass trained IDSs at test time without altering the fundamental attack behavior. While IDS research has progressed significantly in detecting classical attacks, most studies assume a white-box threat model, where the adversary has full access to the IDS architecture, parameters, and gradients, or overlook the possibility of black-box evasion attacks, where such access is unavailable.

1) IDSs Against Classic Attacks: A support vector machine (SVM) model was proposed to detect GPS spoofing attacks based on signal features [4]. An isolation forest-based anomaly detector was used to detect and mitigate GPS spoofing by learning patterns in navigation data [5]. A supervised machine learning framework was proposed to detect Sybil attacks, using classical classifiers such as J48, Classification via Regression, OneR, and JRip implemented via the Weka platform [6]. An autoencoder combined with extended Kalman filtering was proposed to detect smart attacks [7]. DoS attacks were detected using a hybrid convolutional neural network (CNN)–long short-term memory (LSTM) architecture [8]. Jamming threats were addressed through spectrogram-based learning [9], con-

volutional attention models [10], reinforcement learning [11], and federated learning frameworks [12]. A shallow deep learning model comprising a random forest (RF) classifier and neural networks was applied to detect multi-class threats such as DoS, brute force, infiltration, and botnet attacks [13]. A cyber-physical IDS, based on LSTM recurrent neural networks (RNN), for UAVs was developed in [2], but the system was evaluated under traditional attacks.

2) IDS Against Evasion Attacks: An ensemble learning-based IDS using fused cyber-physical features was proposed in [3] and evaluated under white-box evasion attacks, specifically gradient-based methods. A study on transfer-based black-box evasion was presented in [14], where evasion samples were crafted on one deep learning model and tested on another to evaluate whether the evasion behavior transferred. The study used various vision-based architectures, including ResNet18, DenseNet121, VGG16, and EfficientNet. However, it focused solely on image-based UAV datasets and did not explore telemetry-driven IDSs or query-based evasion strategies. A deep reinforcement learning model for UAV guidance was proposed in [15], but it addresses policy manipulation during training (i.e., poisoning), rather than test-time evasion attacks against IDS classifiers.

3) Limitations: The aforementioned studies primarily focus on isolated attack types such as signal spoofing, Sybil, DoS, and jamming [4], [6], [8], [9], often assuming white-box access to the IDS model [3]. Additionally, only one study focuses on the detection of black-box evasion attacks, but it evaluates vision-based UAV systems that operate within simulation environments [14], which may not reflect the telemetry-driven IDS architectures deployed on real UAVs. Notably, none of the surveyed works assess IDS robustness under black-box evasion scenarios, especially those involving query-based attacks, where adversaries lack model access but can infer behavior through output probes, without the need for training a surrogate model. The aforementioned limitations present a gap in the literature, motivating a comparative evaluation of UAV IDS resilience under both query-based and non-query-based black-box evasion attacks using fused cyber-physical features.

B. Contributions

To address gaps in black-box evaluation of UAV IDSs, we carry out the following contributions:

- We utilize two fused cyber-physical datasets: one collected from a real-world UAV testbed under both benign and FDIA conditions to train the IDS as the operator, and another used as a surrogate by the attacker to simulate black-box transfer for non-query-based evasion attacks.
- We implement a set of black-box evasion attacks, including non-query-based methods such as the fast gradient sign method (FGSM), basic iterative method (BIM), and Carlini & Wagner (C&W), as well as query-based methods including zeroth order optimization (ZOO) and boundary attack, all constrained by fixed perturbation limits. Our investigations reveal that query-based attacks lead to higher DR deterioration compared to non-query attacks by 67%.
- We evaluate five ML-based IDS models—SVM, RF, feedforward neural network (FNN), CNN, and LSTM—against evasion samples, demonstrating detection rate (DR) degradations of up to 85% for shallow ML models and up to 64% for deep learning models under black-box evasion.

The remainder of this paper is organized as follows. Section II describes the UAV cyber-physical testbed and outlines the data collection, preprocessing procedures for generating the fused feature set used in this study, as well as the evasion attack test sets. Section III presents the black-box attack models and threat assumptions, covering both non-query-based (FGSM, BIM, C&W) and query-based (ZOO, boundary attack) settings, along with perturbation. Section IV presents the ML-based IDS models alongside the training and evaluation methodology. Section V reports the experimental results, highlighting the damaging impact of evasion attacks on IDS performance. Finally, Section VI concludes the paper and discusses potential directions for future work.

II. Cyber-Physical Datasets

To evaluate the vulnerability of UAV IDSs to black-box adversarial attacks, we use two cyber-physical fused datasets, each containing both benign and labeled FDIA samples. The operator dataset is used to train the IDSs, while the second dataset serves as a surrogate dataset accessible to the attacker. This separation supports a realistic black-box threat model in which the attacker has no access to the defender's model, parameters, or training data. For non-query-based attacks, the adversary leverages the surrogate dataset to train substitute models, then transfers adversarial examples to the target IDS by exploiting the transferability property of machine learning models, which is the tendency for evasion samples to remain effective across different models trained on similar features.

A. Operator Dataset

The dataset is collected using a real-world UAV testbed comprising a DJI Tello EDU drone with mission pad-based autonomous navigation and two host machines [3]. Computer-1 functions as the ground control station, logging physical telemetry and command sequences, while Computer-2 performs wireless traffic monitoring and injects attacks.

Two types of flights are conducted: benign and malicious. In benign flights, the UAV follows predefined autonomous trajectories while physical sensor data and wireless communication logs are recorded. In malicious flights, a stealthy FDIA manipulates the UAV's position estimates mid-flight, diverting it from its intended path. Both physical and cyber data streams are captured in parallel during all flights.

The resulting dataset includes 65 features per sample, composed of 31 physical features (e.g., position, velocity, orientation, and Kalman filter residuals) and 34 cyber features (e.g., MAC addresses, frame lengths, signal quality, and QoS fields). Each sample is labeled based on the flight type,

enabling supervised IDS training. In this study, we use only the fused cyber-physical representation, which has been shown to improve detection performance by capturing joint dynamics across domains. All samples are normalized and structured into fixed-length sequences for use with deep learning models.

B. Surrogate Dataset

To assess cross-dataset generalization and attack transferability, we also employ a publicly available cyber-physical dataset [2]. This dataset is collected using a similar UAV testbed as the one in [3], consisting of a DJI Tello EDU drone, a ground control system, and an adversarial machine equipped for traffic interception and manipulation. The dataset includes four cyber attack classes, which are executed along structured flight paths: deauthentication, replay, evil twin, and FDIA.

The dataset captures telemetry and wireless traffic across 35 benign and 40 malicious flights. Physical data includes features such as roll, pitch, yaw, velocity, battery level, and motor temperature, while cyber features include frame lengths, MAC and IP addresses, protocol-level metadata, and packet timing information. A total of 16 physical and 37 cyber features are extracted and aligned using timestamp interpolation to produce a fused representation.

The final dataset contains over 29,000 labeled samples and is publicly available through IEEE DataPort. For consistency with our primary dataset and to isolate the effect of evasion attacks, we use only the fused feature view and restrict the evaluation to the FDI attack class, which offers a direct comparison to the operator dataset and reflects prior findings on its effectiveness in adversarial settings.

C. Data Preprocessing

Both datasets undergo identical preprocessing, including the removal of incomplete records and feature normalization using standard scaling. Labels are binarized as benign or malicious based on ground truth, and each dataset is partitioned using a 70%, 10%, and 20% stratified split for training, validation, and testing, respectively. All IDS models are trained in a supervised manner using both benign and FDIA-labeled data. To evaluate model robustness, we inject adversarial evasion attacks into the test set using the black-box methods described in Section III. The injection ratio is fixed at 50%, such that the test set consists of 50% benign samples, 25% classic FDIAs, and 25% perturbed evasion samples. This split reflects a realistic operational scenario where IDSs are exposed to known attacks but must remain resilient to unseen evasion strategies. Furthermore, we ensure that attack classes from each attack type, whether query-based, non-query-based, or FDIA, are equally represented in the test set to enable a fair comparison of their individual impact.

III. Attack Generation

In this section, we describe the attack types used to evaluate the resilience of UAV IDSs. We first present the classic FDIA used during IDS training. We then describe two types of black-box evasion attacks: non-query-based and query-based. These evasion attacks are introduced only at test time and are designed to bypass trained IDS models.

A. Classic FDIA

FDIAs aim to compromise the integrity of data during UAV operations by injecting crafted disturbances into selected measurements. Let $\mathbf{X_B} \in \mathbb{R}^n$ denote the original sensor sample (e.g., position, velocity, and orientation), and let $\delta_s \in \mathbb{R}^n$ represent the injected false data vector. Hence, a malicious sample is represented as:

$$X_S = X_B + \delta_s, \tag{1}$$

where δ_s is designed to remain within plausible operational ranges that is tuned to avoid triggering basic threshold-based detection. In our testbed, δ_s primarily targets positional and velocity estimates used by the UAV's navigation system, introducing drift over time while preserving low-level signal characteristics. The goal is to shift the UAV off its intended trajectory without immediate detection. These FDIA samples are labeled as malicious during training.

B. Non-Query-Based Attacks

Non-query-based attacks rely on evasion samples generated using a surrogate model, where the attacker has access to both the model and the training dataset. These attacks are transferred to a black-box operator model under the assumption that similar decision boundaries exist between the surrogate and the target. This transferability property allows adversaries to evade IDSs without querying the operator model directly.

1) Fast Gradient-Based Attacks: Fast gradient-based attacks, such as FGSM and BIM, craft evasion samples by using the gradient of the loss function with respect to the input to apply minimal but effective perturbations that mislead the model during inference.

a) FGSM Attack: The FGSM attack [16] generates a single-step perturbation using the sign of the loss gradient. We define:

$$X_A = X_B - \epsilon \,\mathrm{sign}(\nabla_{X_B} J(\varphi, X_B, y)), \tag{2}$$

where X_A denotes the evasion sample, X_B the benign input, ϵ the perturbation magnitude, ∇_{X_B} the input gradient of the loss J, φ the model parameters, and y the true label. The perturbation increases the model's loss, pushing the input across the decision boundary. We use ϵ=0.5 to induce misclassification while keeping the perturbation small enough to remain within the normal range of UAV sensor and communication data, making the evasion behavior difficult to detect using thresholding or rule-based defenses.

b) BIM Attack: BIM [16] extends FGSM by applying multiple small steps with projection to maintain a bounded distortion. Let I be the number of iterations and α the step size (with the same ϵ=0.5). The update is expressed as:

$$X_A^{(i)} = \mathrm{Clip}_{X_B,\epsilon}\Big\{ X_A^{(i-1)} - \alpha \,\mathrm{sign}\big(\nabla_{X_A^{(i-1)}} J(\varphi, X_A^{(i-1)}, y)\big) \Big\} \tag{3}$$

where $\mathrm{Clip}_{X_B,\epsilon}\{\cdot\}$ projects the perturbed point back to the ϵ-ball around X_B to keep the attack within bounds.

2) C&W Attack: The C&W attack [17] formulates evasion as a constrained optimization problem that seeks to generate adversarial examples with minimal perturbation under the L_2 norm, while ensuring that the modified input is misclassified. The L_2 norm refers to the Euclidean distance between the original and adversarial inputs, aiming to keep the perturbation as small as possible. The objective can be expressed as

$$X_A = \arg\min_{X'};; |X' - X_B|_2^2 + c \cdot f(X'), \tag{4}$$

where X' is the perturbed candidate (which becomes X_A when successful), $f(X')$ is a loss function that encourages misclassification, and c is a constant balancing distortion and classification objectives.

C. Query-Based Attacks

Query-based attacks assume that the adversary has no access to the target model's architecture, parameters, or gradients, but can issue input queries and observe output labels or confidence scores. These attacks iteratively craft evasion samples by probing the model and adjusting the input based on the observed responses. Unlike non-query-based methods, query-based attacks operate directly on the black-box target, often optimizing perturbations using gradient-free or score-based strategies.

1) ZOO: ZOO estimates gradients via coordinate-wise finite differences using only model outputs [18]. For coordinate k with probe step h, the estimated gradient is:

$$\hat{g}_k = \frac{f(X_B + h\,e_k) - f(X_B - h\,e_k)}{2h}, \tag{5}$$

where $f(\cdot)$ denotes the attack objective derived from the model's scores, and e_k is the k-th basis vector. We then update

$$X_A^{(i)} = \Pi_{\|\cdot - X_B\| \le \epsilon}\Big(X_A^{(i-1)} - \alpha\,\mathrm{sign}(\hat{g})\Big), \tag{6}$$

with step size α, projection Π to enforce the perturbation budget ϵ, and a fixed query budget.

2) The Boundary Attack: Boundary Attack performs decision-based evasion by starting from a sample already classified as the target class and iteratively reducing the distance to X_B while staying adversarial [19]. A typical step combines a small move toward X_B with a random orthogonal exploration:

$$X_A^{(i)} = \Pi_{\|\cdot - X_B\| \le \epsilon}\left(X_A^{(i-1)} + \eta\frac{X_B - X_A^{(i-1)}}{\|X_B - X_A^{(i-1)}\|_2} + \beta\,u_\perp\right), \tag{7}$$

where η and β control the radial and orthogonal components, $u_\perp$ is a random unit vector orthogonal to $X_B - X_A^{(i-1)}$, and projection enforces the perturbation budget under a fixed decision-query limit.

TABLE I
OPTIMAL HYPERPARAMETERS OF IDS MODELS

Model	Parameter	Value
SVM	Kernel	Linear
	Regularization (C)	0.001
RF	Number of Trees	5
	Max Depth	2
	Min Samples per Leaf	10
FNN	Activation Function	ReLu
	Layers	3
	Neurons per Layer	64, 32, 16
	Dropout Rate	0.5
CNN	Filters	64, 128, 256
	Layers	4
	Neurons per Layer	128
	Dropout Rate	0.8
LSTM	Units per Layer	32
	Layers	3
	Dropout Rate	0.85

IV. ML-IDS MODELS

To provide a broad and representative comparison, we evaluate five different ML-IDS models covering both shallow and deep learning approaches with static and dynamic mechanisms. The benchmark models include SVM, RF, FNN, CNN, and LSTM. All models are trained on the operator dataset, with hyperparameters selected via sequential tuning on a validation set to ensure fair comparison. Table I showcases all the hyperparameters for each model.

A. Shallow ML Detectors

Shallow detectors utilize classical learning methods and operate on fixed-length feature vectors. They provide strong discriminative baselines but do not explicitly capture complex patterns or model temporal dependencies.

a) RF: An RF ensemble is trained with bootstrap sampling to reduce variance and improve generalization [20]. We validate the number of trees, maximum depth, and minimum split size, choosing the setting that balances bias and variance on the held-out validation set. Like the SVM, the RF operates on fixed vectors and does not explicitly capture sequence dynamics.

b) SVM: A kernel SVM [4] with a radial basis function is trained on standardized vectors. We tune the regularization strength and kernel bandwidth on the validation set and select the configuration that maximizes validation accuracy. The SVM provides a static decision boundary.

B. Deep Learning Detectors

Deep detectors leverage deep learning-based methods, which makes them capable of learning complex feature interactions and sequential dependencies over time, making them well-suited for fused cyber–physical data.

TABLE II
COMPARISON OF IDS MODELS UNDER DIFFERENT ATTACK SCENARIOS (%)

Model	Metric	Classic Attack	Non-Query-based Attacks		Query-based Attacks	
		FDIA	Fast Gradient-based	C&W	Boundary Attack	ZOO
RF	DR	94.4	8.5	9.2	9.1	9.2
	FAR	0	0	0	0	0
	ACC	97.2	54.3	54.8	54.6	54.6
	F1-score	97.1	15.5	16.8	16.8	16.8
SVM	DR	95.9	68.3	36.2	40.5	36.1
	FAR	10.1	5.5	2.7	2.7	2.7
	ACC	92.9	81.4	66.8	68.9	66.7
	F1-score	93.2	76.8	52.1	56.5	52.1
FNN	DR	99.4	76.7	80.8	89.8	47.5
	FAR	0.29	13.2	26.5	0.3	0.7
	ACC	99.6	81.6	77.2	94.7	73.4
	F1-score	99.5	79.1	78.5	94.4	64.1
CNN	DR	99.7	54.2	73.8	73.3	43.2
	FAR	9.9	26.5	39.5	2.5	0.2
	ACC	94.8	63.8	67.2	85.4	71.5
	F1-score	95.1	61.2	69.2	83.4	60.2
LSTM	DR	99.6	68.6	34.8	39.3	32.0
	FAR	0	4.2	0	0	0
	ACC	99.6	82.2	67.4	69.7	66.0
	F1-score	99.5	75.5	51.7	56.5	48.5

a) FNN: The FNN classifier is a multilayer perceptron with rectified linear units and a softmax output [20]. We tune depth and dropout using validation and train with cross-entropy loss and a first-order optimizer with early stopping. This model remains static (non-recurrent) but captures richer nonlinear interactions than shallow baselines.

b) CNN: The CNN processes fixed temporal windows using one-dimensional convolutions, nonlinearity, and pooling, followed by a dense classification head [20]. Kernel sizes, number of filters, and dropout are selected via validation. The CNN extracts local temporal patterns in the fused or single-modality streams while remaining efficient at inference.

c) LSTM: The LSTM models sequential dependencies through recurrent units with gating. We validate the number of layers, hidden units, dropout, and learning rate, and train with cross-entropy and early stopping [20]. The LSTM captures longer-range temporal correlations across cyber, physical, and fused inputs.

V. EXPERIMENTAL RESULTS

This section presents the detection results for all evaluated IDS models. We assess performance under classic FDIA, as well as under black-box evasion attacks, which include both non-query-based and query-based methods. The models are evaluated using the following classification metrics: DR $= \frac{\text{TP}}{\text{TP+FN}}$, false alarm rate (FAR) $= \frac{\text{FP}}{\text{FP+TN}}$, accuracy (ACC) $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, and F1-score $= \frac{2\cdot\text{TP}}{2\cdot\text{TP+FP+FN}}$, where TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative predictions, respectively.

A. Evasion Impact on Shallow ML Models

Table II shows that shallow IDS models, RF, and SVM experience the most severe degradation in detection performance under evasion attacks. For RF, DR drops by 85% under both non-query and query-based attacks, the F1-score deteriorates by 81%, and accuracy degrades by over 42%. This decline is due to the reliance of RFs on feature splits and the lack of learned representations. Without gradient awareness or transformation capacity, RF cannot adapt to adversarial inputs, resulting in near-total collapse in detection performance. For SVM, DR drops by 44% under non-query-based attacks and by 57.5% under query-based attacks. The F1-score declines by 28.5% and 38.5% respectively, while accuracy drops by 19% and 25%. The performance of SVMs declines under evasion due to its reliance on fixed linear margins, which can be easily bypassed by adversarial perturbations that subtly shift samples across the decision boundary.

B. Evasion Impact on Deep Learning Models

Deep learning models, FNN, CNN, and LSTM, demonstrate greater robustness than shallow ML models but still suffer measurable degradation. For FNN, DR drops by 20.5% under non-query-based attacks and by 30% under query-based attacks. The F1-score declines by 19.5% and 19%, respectively, while accuracy drops by 18.5% and 14%. The performance of FNNs declines under evasion due to perturbations that shift samples across decision boundaries, though the model's fully connected structure provides moderate resistance by capturing

distributed feature representations. For CNN, DR drops by 35% under non-query-based attacks and by 41% under query-based attacks. The F1-score declines by 30% and 23.5%, respectively, while accuracy drops by 29.5% and 16.5%. The performance of CNNs declines under evasion due to their sensitivity to localized perturbations, which can disrupt the learned spatial features and reduce classification reliability. For LSTM, DR drops by 47% under non-query-based attacks and by 63.5% under query-based attacks. The F1-score declines by 35% and 47%, respectively, while accuracy drops by 24.5% and 31%. The performance of LSTM models declines under evasion due to their vulnerability to input drift over time, where adversarial perturbations accumulate across timesteps and distort temporal dependencies.

C. Impact of Query vs. Non-Query-Based Attacks

Across all models, query-based attacks consistently cause greater performance drops than non-query-based attacks, with DR deterioration reaching 67%. The average F1-score degradation is higher under query attacks, with 81% for RF, 38.5% for SVM, 47% for LSTM, than under non-query-based attacks. This is expected, as query-based attacks like ZOO and Boundary Attack directly optimize against the operator model's outputs, whereas non-query-based attacks rely on surrogate approximations. Moreover, query-based attacks pose a more realistic threat to deployed UAV IDSs. They assume no access to model parameters, gradients, or architecture, and instead rely on output feedback alone, which falls in line with the black-box nature of most real-world scenarios. This practicality, combined with their effectiveness, underscores the importance of designing systems that are robust against these black-box adversarial strategies.

VI. Conclusions and Future Work

This paper investigated the impact of black-box adversarial evasion on machine learning-based UAV IDSs, using a fused cyber-physical feature set collected from a real-world testbed. We evaluated five IDS models, SVM, RF, FNN, CNN, and LSTM, against classic FDIAs, as well as non-query-based and query-based evasion attacks. Our results showed that shallow ML models experienced DR degradations of up to 85%, while deep learning models exhibited more moderate DR deterioration of up to 64%. Compared to non-query-based attacks, query-based evasion attacks lead to a larger drop in DR, reaching as high as 67%. These findings highlight the vulnerability of UAV IDSs to realistic adversarial threats and the need for robust design strategies. Future work includes developing adversarially trained IDS models and exploring certified defense mechanisms to enhance robustness under black-box attack conditions.

References

[1] H. Seong *et al.*, "Hierarchical multi-agent reinforcement learning-based uav control for wireless covert communications," in *IEEE Consumer Communications and Networking Conference (CCNC)*. Las Vegas, NV, USA, 10 –13 Jan. 2025, pp. 1–6.

[2] S. C. Hassler *et al.*, "Cyber-physical intrusion detection system for unmanned aerial vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 6106–6117, June 2024.

[3] J. Richeson *et al.*, "Ensemble learning-based intrusion detection system for aerial base stations against adversarial evasion attacks," in *IEEE International Conference on Communications (ICC)*. Montreal, QC, Canada, 8–12 June 2025, pp. 1–6.

[4] A. Shafique *et al.*, "Detecting signal spoofing attack in uavs using machine learning models," *IEEE Access*, vol. 9, pp. 93 803–93 815, June 2021.

[5] R. S. Tucker *et al.*, "Real-time detection and mitigation of gps spoofing in uav systems," in *IEEE International Conference on Information Technology (ICIT)*. Amman, Jordan, 27 – 30 May 2025, pp. 154–160.

[6] D. Chulerttiyawong *et al.*, "Sybil attack detection in internet of flying things-ioft: A machine learning approach," *IEEE Internet of Things Journal*, vol. 10, no. 14, pp. 12 854–12 866, July 2023.

[7] A. Aladi *et al.*, "Uav attack detection and mitigation using a localization verification-based autoencoder," *IEEE Access*, vol. 11, pp. 117 752–117 764, Oct. 2023.

[8] R. Fu *et al.*, "Machine-learning-based uav-assisted agricultural information security architecture and intrusion detection," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 18 589–18 598, Nov. 2023.

[9] Y. Li *et al.*, "Jamming detection and classification in ofdm-based uavs via feature- and spectrogram-tailored machine learning," *IEEE Access*, vol. 10, pp. 16 859–16 870, Feb. 2022.

[10] J. Viana *et al.*, "Deep attention recognition for attack identification in 5g uav scenarios," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 1, pp. 131–146, Jan. 2024.

[11] J. Ghelani *et al.*, "Gradient monitored reinforcement learning for jamming attack detection in fanets," *IEEE Access*, vol. 12, pp. 23 081–23 095, Feb. 2024.

[12] Z. A. El Houda *et al.*, "A privacy-preserving collaborative jamming attacks detection framework using federated learning," *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 12 153–12 164, April 2024.

[13] Y. Wu *et al.*, "Intrusion detection for unmanned aerial vehicles security: A tiny machine learning model," *IEEE Internet of Things Journal*, vol. 11, no. 12, pp. 20 970–20 982, June 2024.

[14] S. Zhou *et al.*, "Transferability of adversarial attacks on tiny deep learning models for iot unmanned aerial vehicles," *IEEE Internet of Things Journal*, vol. 11, no. 12, pp. 21 037–21 045, Dec. 2024.

[15] T. Hickling *et al.*, "Robust adversarial attacks detection based on explainable deep reinforcement learning for uav guidance and planning," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 10, pp. 4381–4394, Oct. 2023.

[16] A. Takiddin *et al.*, "Robust data-driven detection of electricity theft adversarial evasion attacks in smart grids," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 663–676, July 2023.

[17] A. Takiddin *et al.*, "Spatio-temporal graph-based generation and detection of adversarial false data injection evasion attacks in smart grids," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 6601–6616, Dec. 2024.

[18] P.-Y. Chen *et al.*, "Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *ACM Workshop on Artificial Intelligence and Security*. Dallas, TX, USA, 3 Nov. 2017, pp. 15–26.

[19] W. Brendel *et al.*, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *International Conference on Learning Representations*. Vancouver, Canada, 30 April – 3 May 2018, pp. 1–15.

[20] A. Takiddin *et al.*, "Robust graph autoencoder-based detection of false data injection attacks against data poisoning in smart grids," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 3, pp. 1287–1301, March 2024.