FALCON: A Cyber-Physical Benchmark Dataset for Intrusion Detection in UAV Swarms - Design, Generation, and Evaluation

Amylia Ait Saadi
ESME Research lab, ESME Sudria
LISV Laboratory, UVSQ, Paris-Saclay
Vélizy-Villacoublay, France
aitsaadi.amylia@outlook.com

Bikram Pratim Bhuyan
LISV Laboratory
UVSQ, University of Paris-Saclay
Vélizy-Villacoublay, France
bikram23bhuyan@gmail.com

Amar Ramdane-Cherif
LISV Laboratory
UVSQ, University of Paris-Saclay
Vélizy-Villacoublay, France
amar.ramdane-cherif@uvsq.fr

***Abstract*—Flying Ad-Hoc Networks (FANETs) are widely deployed in critical missions such as disaster relief, border surveillance, and defense, where a compromised UAV can impact the entire swarm. Existing intrusion detection systems for FANETs rely mostly on network-layer features and ignore the physical state of the UAVs themselves. In this paper, we introduce FALCON (FAnet cyberphysicaL attack COllectioN), a cyber-physical co-simulation dataset that jointly captures flight telemetry and network traffic across six attack scenarios and swarms of 10 to 50 UAVs. We benchmark eleven machine learning classifiers, including linear, tree-based, and neural architectures. Attack detectability depends strongly on attack mechanism: only attacks with a measurable physical effect are detectable from telemetry alone. The remaining attacks are best detected by combining network and physical features. Our best-performing models, LightGBM and XGBoost, achieve an F1-macro score of 0.83 on the combined feature set. We further demonstrate that cyber-physical features capture attack patterns combining both domains. FALCON provides a benchmark for advancing cyber-physical intrusion detection research in UAV swarms.**

***Index Terms*—FANET, UAV swarms, intrusion detection, cyber-physical security, dataset, co-simulation**

I. Introduction

Unmanned Aerial Vehicles (UAVs) increasingly operate as a coordinated and dynamic class of mobile networks: Flying Ad-Hoc Networks (FANETs) in critical missions such as border surveillance, disaster relief, and military operations [1]. Unlike traditional mobile networks, FANETs operate in three-dimensional space with moving nodes, intermittent connectivity, and severe resource constraints such as limited onboard energy [2]. In such environments, a compromised UAV can disrupt an entire mission. Furthermore, intrusion detection systems (IDS) proposed for FANETs rely almost entirely on network-layer features, leaving UAV physical telemetry such as flight dynamics, attitude, and energy consumption unexplored as a source of attack-discriminating information [3].

Existing general-purpose datasets such as KDD Cup 99 [4], NSL-KDD [5], CIC-IDS2017 [6], and CICIoT2023 [7] provide useful network-security benchmarks but do not model UAV mobility or flight state. More recent UAV-specific datasets address attacks such as flooding, blackhole, Sybil, wormhole, GPS jamming, and gray-hole behavior, but they remain predominantly network-centric. Consequently, they provide limited support for studying whether an attack leaves a detectable signature in the physical dynamics of a UAV swarm.

To address these challenges, this paper contributes the following:

- We introduce FALCON, a comprehensive cyber-physical co-simulation dataset for FANET IDS, combining ArduPilot SITL telemetry and ns-3 network traces across 6 attack scenarios and swarms of 10 to 50 UAVs.
- We show that attack detectability depends on the mechanism. Physical telemetry alone detects attacks with a strong physical effect, such as SpeedManipulation. Most attacks are instead revealed through network footprint.
- We show that cyber-physical fusion features add diagnostic value for attack characterization.
- We establish ML baselines using eleven classifiers with grouped five-fold cross-validation and evaluate both feature modality and generalization to an unseen swarm size.

The rest of this paper is organized as follows. Section II describes the FALCON dataset, its co-simulation architecture, attack model, and feature set. Section III presents our experimental evaluation, including the classification benchmark, modality analysis, and generalization to unseen swarm sizes. Section IV concludes the paper and outlines future work.

II. The FALCON Dataset Design

In this section, we present the methodology for creating the FALCON dataset, detailing the co-simulation environment, attack implementation, data collection, synchronization process, and feature extraction.

A. Co-Simulation Architecture

FALCON builds upon the UAVLnQ co-simulation architecture [8]. UAVLnQ integrates ArduPilot Software-in-the-Loop (SITL) with the ns-3 discrete-event network simulator. A ZeroMQ middleware layer connects the two simulators. While most existing frameworks bypass the simulated network for mobility or position updates, UAVLnQ instead routes every

979-8-3195-0167-7/26/$31.00 ©2026 IEEE

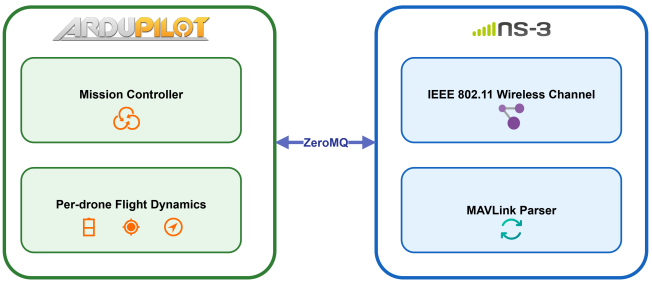


Fig. 1. FALCON co-simulation architecture

MAVLink message through the ns-3 wireless channel. This includes control commands, telemetry broadcasts, and inter-UAV coordination packets. Therefore, network degradation or adversarial manipulation propagates directly into the observable UAV flight behavior.

As shown in Figure 1, the framework operates through three tightly integrated components:

- **ArduPilot SITL:** models realistic flight dynamics for each UAV, including battery consumption, GPS positioning, and attitude control;
- **ns-3 network simulator:** models IEEE 802.11 wireless mesh communications at the packet level across the full swarm topology;
- **ZMQ middleware:** implements bidirectional MAVLink message exchange between the two simulators, guaranteeing synchronization between mobility control and packet-level events.

A mission controller orchestrates swarm behavior and logs timestamped telemetry, while a MAVLink parser relays packets between ns-3 and ArduPilot instances. Table I summarizes the main co-simulation parameters used to generate FALCON.

TABLE I
FALCON SIMULATION PARAMETERS

Parameter	Value
Flight controller simulator	ArduPilot SITL
Network simulator	ns-3 (RealtimeSimulatorImpl)
Synchronization	ZeroMQ (ZMQ), `BestEffort`
Number of UAVs	10, 30, 50
Simulation duration	up to 600 s
MAC standard	IEEE 802.11a (ad-hoc)
Transmit power	40 dBm
Propagation model	LogDistance ($\alpha = 2.0$)
UAV speeds	3, 7, 12 m/s
Attacker ratio	single, 25%, 50%
Attack scenarios	6
Runs per configuration	1
Total runs	144
Window size	30 s
Physical features	28
Network features	39
Cyber-physical features	7
Total features	74

TABLE II
ATTACK SCENARIOS IN FALCON

Scenario	Mechanism
BatterySpoofing	Falsified `BATTERY_STATUS` messages broadcast critically low voltage/charge to all swarm members
GPSSpoofing	Injected `GPS_RAW_INT` packets broadcast fabricated peer positions to the swarm
SpeedManipulation	Spoofed `MAV_CMD_DO_CHANGE_SPEED` degrades mission efficiency
MissionInjection	Forged `MISSION_ITEM` commands attempt to redirect UAV waypoints
HeartbeatFlood	Excessive `HEARTBEAT` packets congest the MAVLink channel
Normal	Benign swarm operation baseline

B. Attack model

FALCON combines two distinct attack models, reflecting realistic scenarios (Table II) in UAV swarm operations:

1) **Distributed attacks:** BatterySpoofing and HeartbeatFlood inject malicious traffic that scales with the entire swarm, simulating an attacker with network-wide reach or capable of compromising the shared communication channel.
2) **Localized attacks:** GPSSpoofing, SpeedManipulation, and MissionInjection target a small, fixed subset of UAVs (2–3 nodes), simulating an attacker that compromises specific high-value targets such as a swarm leader or designated follower nodes, rather than the entire fleet.

C. Feature Extraction Methodology

Raw ArduPilot telemetry logs and ns-3 PCAP traces are processed through a four-stage pipeline to produce the final feature set. First, per-drone telemetry CSV files and per-node PCAP captures are parsed independently for each simulation run. Second, both streams are aggregated into fixed 30-second windows, segmenting each 600-second run into 20 temporal samples. Telemetry is averaged across all UAVs within each window (swarm-level aggregation), while PCAP traces are parsed at the IEEE 802.11 frame level to extract source/destination MAC addresses, frame type, and payload size, and to compute window-level traffic statistics. Third, derived cyber-physical fusion features are computed by combining physical and network primary metrics within the same window (e.g., normalizing network throughput by swarm ground speed). Finally, all features are exported to a unified table with no missing values. FALCON contains 28 physical, 39 network, and 7 derived cyber-physical features. Physical features capture position, kinematics, attitude, battery state, and GPS quality; network features describe traffic volume, packet sizes, inter-arrival times, frame types, node diversity, and per-node transmission statistics. The seven cyber-physical

features normalize network activity by physical quantities such as speed or power.

D. Dataset Statistics

Table I summarizes the simulation configuration underlying FALCON. Each run is segmented into fixed 30 s windows, resulting in 20 temporal samples per run. The dataset comprises 2,427 samples across 144 runs and 74 features. The dataset's construction includes a relabeling correction for pre-attack windows and the exclusion of incomplete network captures. As shown in Figure 2, the resulting class distribution is imbalanced. Normal is the largest class with 712 samples (29.3%), since every run contains a pre-attack baseline segment. The attack classes range from 201 (HeartbeatFlood) to 397 (SpeedManipulation) samples.

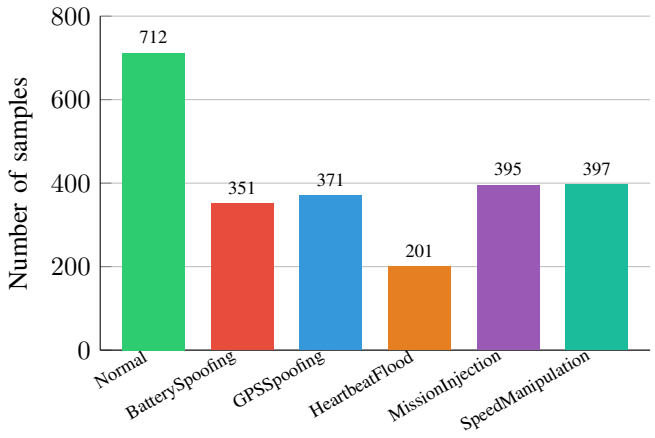


Fig. 2. FALCON class distribution.

1) Feature significance: For each feature, we compute a one-way ANOVA F-test across the six attack classes. Two confounds required correction. Windows from the same run are temporally autocorrelated. Normal windows are also concentrated early in each run, while attack windows always occur after injection. We therefore aggregate to the run level. We also restrict Normal to time-matched windows (t $\geq$ 120s). These windows come only from pure no-attacker baseline runs. Table III reports the top discriminative features per category by F-statistic.

TABLE III
TOP-3 DISCRIMINATIVE FEATURES PER CATEGORY

Category	Feature	F-statistic
Network	pkt_size_mean	42.7
	rx_mean_per_dst	40.4
	tx_mean_per_src	40.3
Cyber-Physical	packets_per_meter	11.5
	byte_per_meter	10.2
	throughput_per_speed	10.2
Physical	battery_level_mean	14.4
	battery_drain_rate	5.9
	battery_voltage_mean	4.5

2) Attack signature characterization: HeartbeatFlood is a distributed denial-of-service attack. It produces a 32$\times$ increase in the maximum per-source transmission rate (tx_max_per_src, 14,354 vs. 462,272 for Normal, +3120.5%). A small number of spoofed source identifiers flood the network. packets_per_meter and byte_per_meter also rise sharply (+408.9% and +357.9%).

BatterySpoofing falsifies battery telemetry. Its strongest signature is tx_max_per_src (+360.8%). Mean inter-arrival time also drops sharply (−88.4%). This is consistent with spoofed traffic being added alongside legitimate telemetry.

GPSSpoofing produces a similar but smaller network signature. tx_max_per_src rises by 181.1%. Mean inter-arrival time drops by 87.7%. This is consistent with high-rate injection of spoofed position packets.

MissionInjection shows the weakest signature among all attacks. Its strongest effect is a 155.9% increase in mean inter-arrival time. This is the opposite direction from every other attack. The forged MISSION_ITEM commands never complete their handshake. This likely produces sparser, not denser, traffic.

SpeedManipulation directly reduces ground speed (7.14 vs. 4.34 m/s, −39.2%). It is the only attack with a clear physical signature, not just a network one. Its energy_per_packet also drops by 81.8%, the largest negative shift among all attacks. This is consistent with reduced motor effort at lower speed.

III. EXPERIMENTAL EVALUATION

A. Evaluation protocol

We evaluate FALCON using eleven classifiers: Logistic Regression, SVM, Decision Tree, Random Forest, XGBoost, LightGBM, KNN, MLP, GRU, TabNet, and Kolmogorov-Arnold Network (KAN). We use stratified group k-fold cross-validation with five folds. We group folds by simulation run. Of the 74 features, baseline classifiers use an aggregated subset of 38 (15 physical, 17 network, 6 cyber-physical). This subset includes the high-F-statistic features for each category. All features are standardized per fold. We report accuracy, balanced accuracy, macro-averaged precision/recall/F1, and Matthews Correlation Coefficient (MCC). We compute 95% confidence intervals via the t-distribution across the five folds.

B. Classification benchmark

Table IV reports F1-macro across the three primary feature sets, for all eleven models. LightGBM and XGBoost perform best on Fusion (0.83 each). Random Forest and Decision Tree follow (0.79 and 0.78). KAN leads the neural architectures (0.75), ahead of GRU (0.70) and TabNet (0.67). Every model improves from Physical to Network to Fusion, except KNN. KNN performs worse on Fusion than on Network alone (0.40 versus 0.50): added features dilute its distance metric. Tree ensembles lead in absolute terms, but the margin to the best neural architecture (KAN) is moderate. For LightGBM as shown in Fig 3, cyber-physical features alone achieve an F1-macro of 0.310, whereas Physical+Network reaches 0.829 and full Fusion 0.830, indicating that the engineered cross-domain

TABLE IV
MODEL BENCHMARK ACROSS FEATURE SETS (F1-MACRO, 95% CI)

Model	Physical	Network	Fusion
LightGBM	**0.57 ± 0.03**	**0.73 ± 0.05**	**0.83 ± 0.04**
XGBoost	0.54 ± 0.04	**0.73 ± 0.05**	0.83 ± 0.03
Random Forest	0.56 ± 0.04	**0.73 ± 0.05**	0.79 ± 0.04
Decision Tree	0.49 ± 0.02	0.70 ± 0.04	0.78 ± 0.06
KAN	0.51 ± 0.05	0.63 ± 0.03	0.75 ± 0.05
GRU	0.37 ± 0.04	0.55 ± 0.02	0.70 ± 0.05
TabNet	0.37 ± 0.03	0.52 ± 0.08	0.67 ± 0.06
MLP	0.31 ± 0.06	0.52 ± 0.03	0.62 ± 0.05
Logistic Regression	0.30 ± 0.05	0.48 ± 0.03	0.60 ± 0.03
SVM	0.33 ± 0.06	0.39 ± 0.05	0.51 ± 0.04
KNN	0.37 ± 0.03	0.50 ± 0.00	0.40 ± 0.04

TABLE V
CROSS-SCALE GENERALIZATION TO UNSEEN 50-DRONE SWARMS

Model	Accuracy	Balanced Acc.	F1-Macro	MCC
XGBoost	0.710	0.704	0.693	0.651
Random Forest	**0.724**	**0.742**	0.706	**0.671**
LightGBM	0.718	0.718	**0.719**	0.665
GRU	0.702	0.705	0.705	0.640

ratios provide negligible additional classification gain once both raw modalities are available.

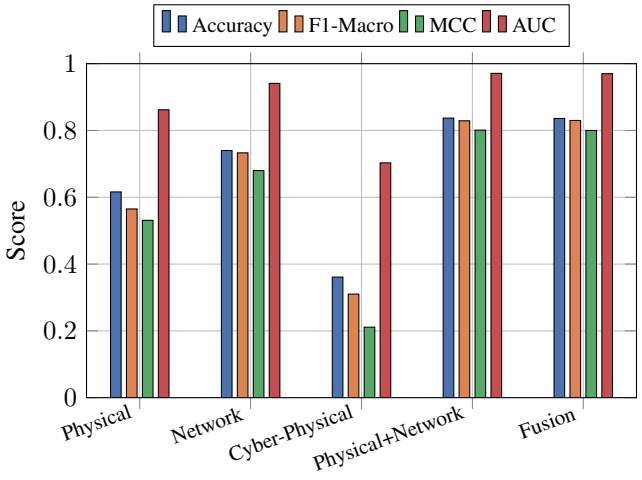


Fig. 3. LightGBM modality ablation on FALCON.

C. Swarm-Size Generalization

Table V reports performance when training on 10- and 30-drone swarms and testing on unseen 50-drone swarms. All four models cluster closely (F1-macro 0.693–0.719). Random Forest and LightGBM perform best. No architecture shows a clear advantage. This contrasts with the larger gap between tree ensembles and neural models in distribution (Table IV). The residual generalization gap for unseen swarm sizes appears to be shared across architectures, not specific to any one model family.

D. Discussion

Network features outperform physical features for most attacks because most attacks in our taxonomy never alter flight behavior. BatterySpoofing, GPSSpoofing, HeartbeatFlood, and MissionInjection all inject falsified or excessive messages. None of these mechanisms changes the UAV's actual flight state. Only SpeedManipulation forces a significant kinematic change, which explains why physical telemetry alone detects it. Physical telemetry fails to detect every other attack class.

For LightGBM, cyber-physical features alone perform poorly (F1-macro = 0.310), indicating that the derived ratios lose information when separated from their raw physical and network components.

Physical+Network matches or exceeds full Fusion for the same reason. Once both raw domains are present, the derived cyber-physical ratios add little new information for classification. They remain useful for interpretation: `tx_max_per_src` and `byte_per_meter` directly expose HeartbeatFlood's traffic volume in a way that raw counts alone do not convey as clearly. Their value lies in characterizing attacks for a human analyst, not in raising aggregate classification accuracy.

IV. CONCLUSION AND FUTURE WORK

We introduced FALCON, a cyber-physical co-simulation benchmark combining ArduPilot SITL telemetry and ns-3 network traces across five attacks, benign operation, and swarms of 10, 30, and 50 UAVs. LightGBM and XGBoost achieve the best in-distribution F1-macro of 0.83, while attack characterization shows that physical telemetry is most informative when the attack directly alters flight behavior. For LightGBM, combining raw physical and network features reaches 0.829 F1-macro compared with 0.830 for full Fusion. Generalization to an unseen 50-UAV swarm remains more difficult, with the best F1-macro of 0.719. Future work will address additional attack types, run-level diversity, varying attack conditions, and validation using real UAV traces.

REFERENCES

[1] S. Abbas, M. Abu Talib, I. Ahmed, and O. Belal, "Integration of uavs and fanets in disaster management: A review on applications, challenges and future directions," *Transactions on Emerging Telecommunications Technologies*, vol. 35, no. 12, p. e70023, 2024.

[2] I. Bekmezci, O. K. Sahingoz, and Ş. Temel, "Flying ad-hoc networks (fanets): A survey," *Ad Hoc Networks*, vol. 11, no. 3, pp. 1254–1270, 2013.

[3] A. A. Saadi, M. D. Hina, B. P. Bhuyan, G. Ivanova, and A. Ramdane-Cherif, "A guided kolmogorov-arnold network for enhanced security in assistive uav technologies," in *2025 International Conference on Computer Systems and Technologies (CompSysTech)*, 2025, pp. 1–8.

[4] UCI KDD Archive, "KDD cup 1999 data," 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[5] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Security Defense Appl.*, 2009, pp. 1–6.

[6] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Security Privacy (ICISSP)*, 2018, pp. 108–116.

[7] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment," *Sensors*, vol. 23, no. 13, p. 5941, 2023.

[8] A. Yousef, C. Tsai, N. N. Mistry, M. Méndez Real, and G. Gogniat, "UAVLnQ: An architecture for security analysis of cyber-physical network behavior in UAV swarms," in *Proc. Workshop Design Archit. Signal Image Process. (DASIP), co-located with HiPEAC*, 2026.