Scientific Research Publishing

Open Journal of Applied Sciences, 2026, 16(8), 2718-2737
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Stress-Testing Explainable Intrusion Detection in Agricultural IoT Networks against Noise and Evasion Attacks

Alexandre Kouamé Kanga[1*], Doffou Jérôme Diako[2], Kouamé Abel Assielou[1], Souleymane Oumtanaga[1], Yao Casimir Brou[1]

[1]Laboratoire de Recherche en Informatique et Télécommunication (LARIT), Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire
[2]École Supérieure Africaine des Technologies de l'Information et de la Communication (ESATIC), Abidjan, Côte d'Ivoire
Email: *kalexane@gmail.com

How to cite this paper: Kanga, A.K., Diako, D.J., Assielou, K.A., Oumtanaga, S. and Brou, Y.C. (2026) Stress-Testing Explainable Intrusion Detection in Agricultural IoT Networks against Noise and Evasion Attacks. *Open Journal of Applied Sciences*, **16**, 2718-2737.
https://doi.org/10.4236/ojapps.2026.168152

Received: July 19, 2026
Accepted: August 21, 2026
Published: August 24, 2026

Abstract

Agricultural Internet of Things networks operate under variable communication conditions and expose intrusion-detection systems to both environmental perturbations and deliberate evasion. This study presents a reproducible multi-seed stress-testing protocol for a CNN-IWHO-Lite-Random Forest intrusion-detection pipeline evaluated on CICIoT2023, Farm-Flow, and UNSW-NB15. Three independent model seeds were used, while Gaussian-noise and adversarial experiments included nested internal repetitions. Preprocessing, representation learning, latent-feature selection, calibration, and threshold optimization were restricted to mutually disjoint non-test partitions, and exact feature-vector overlaps were removed before evaluation. Clean F1-scores were 0.9953 ± 0.0001, 0.4724 ± 0.0131, and 0.7480 ± 0.0183 for CICIoT2023, Farm-Flow, and UNSW-NB15, respectively. At $\sigma = 1.0$, the corresponding F1-scores were 0.9940 ± 0.0002, 0.4118 ± 0.0013, and 0.6768 ± 0.0550. Conditional attack success rates at $\varepsilon = 0.8$ were 16.17% ± 22.66%, 13.58% ± 10.56%, and 98.43% ± 1.17%. TreeSHAP audits and complete false-positive and false-negative summaries were used to interpret selected latent dimensions and decision failures. The results show that robustness is strongly dataset- and seed-dependent: Farm-Flow is limited primarily by baseline false positives and noise sensitivity, whereas UNSW-NB15 is consistently vulnerable to the evaluated score-query evasion attack. These findings support deployment decisions based on explicit stress tests rather than nominal accuracy alone.

Keywords

Agricultural IoT, Intrusion Detection, Evasion Attack, Explainable Artificial

Intelligence, Robustness Audit

1. Introduction

1.1. Background and Motivation

Agricultural production increasingly relies on connected sensors, actuators, edge gateways, and cloud services to support irrigation, greenhouse regulation, equipment monitoring, livestock supervision, and agronomic decision-making. These components form agricultural Internet of Things (AG-IoT) environments in which cyber incidents can affect not only information systems but also physical and biological processes. Denial-of-service attacks, false data injection, unauthorized access, malware, ransomware, and manipulation of connected equipment have therefore become central concerns for Agriculture 4.0 and Agriculture 5.0 [1] [2].

Intrusion detection is an important protective layer in this setting. Edge gateways are natural deployment points because they aggregate device traffic and can identify suspicious activity before it reaches supervisory services or affects control functions. However, agricultural edge environments are constrained by memory, processing capacity, intermittent connectivity, wireless interference, and changing traffic loads [3]. An intrusion-detection system (IDS) must consequently be assessed as an operational component rather than only as a classifier.

Domain relevance is also important. Generic IoT datasets provide broad benchmarks, but they do not necessarily reproduce the periodicity, synchronization patterns, and response-flow variability found in agricultural networks. Farm-Flow was introduced to support network-flow intrusion detection in a smart-agriculture setting and offers a domain-specific complement to large generic benchmarks [4]. A credible AG-IoT evaluation should therefore combine domain-specific and general datasets while avoiding the assumption that strong benchmark performance transfers directly to the field.

1.2. Limits of Accuracy-Only Evaluation

Accuracy, precision, recall, F1-score, and area-under-the-curve measures remain essential, but they do not reveal how an IDS behaves when inputs are perturbed. Reported performance is influenced by dataset composition, preprocessing, feature selection, validation strategy, class imbalance, threshold selection, and leakage-prone variables [5]. A model may obtain a high F1-score while generating a substantial false-positive burden, relying on unstable correlations, or failing under small adversarial changes.

These limitations are particularly consequential in agricultural cyber-physical systems. A false positive may unnecessarily isolate a legitimate sensor or actuator, while a false negative may allow malicious traffic to remain active. Environmental

noise and adversarial manipulation must therefore be examined separately. Gaussian perturbations can approximate sensitivity to continuous feature variability, whereas bounded evasion attacks test whether malicious observations can be moved across the decision boundary [6] [7].

Explainability provides an additional audit layer. SHAP-based explanations can identify which model inputs contribute to individual decisions and have been applied to privacy-aware and federated intrusion detection [8]. In a hybrid pipeline, however, the explained variables may be latent dimensions rather than original network descriptors. Interpretations must therefore remain faithful to the space in which the final classifier operates.

1.3. Research Question and Contributions

This study addresses the following question: How reliably does a compact CNN-IWHO-Lite-Random Forest pipeline behave across clean, noisy, adversarial, and hard-case conditions when the evaluation is repeated with independent model seeds?

The contribution is an audit protocol rather than a claim of universal architectural superiority. Specifically, the study provides:

- a leakage-aware, train-only, overlap-audited evaluation with three independent model seeds and mutually disjoint model-selection, calibration, threshold, and test roles;
- a dataset-specific mutability policy for Gaussian noise and normalized bounded evasion, with protected variables preserved and derived constraints recomputed;
- a conditional Attack Success Rate (ASR) computed only on malicious samples detected under clean conditions, with clean false-negative rates and attack coverage reported separately;
- TreeSHAP global and local audits of the final Random Forest in the selected latent space, together with complete false-positive and false-negative statistics;
- a hierarchical reporting strategy that distinguishes variation between model seeds from repetitions nested within each model.

The evaluated pipeline remains relevant as a compact representation-and-selection design, but it is not presented as uniformly superior to simpler baselines or as a proven adversarial defense.

1.4. Paper Organization

Section 2 reviews related work. Section 3 describes the evaluated architecture, datasets, partitioning, preprocessing, optimization, calibration, and repetition strategy. Section 4 defines the perturbation, evasion, explainability, and hard-case protocols. Section 5 reports clean, noise, evasion, SHAP, and error results. Section 6 discusses the implications, ablations, operational use, and limitations. Section 7 concludes the paper.

2. Related Work

2.1. Intrusion Detection in IoT and Agricultural Networks

IoT intrusion detection has been studied through signature-based, anomaly-based, machine-learning, deep-learning, and hybrid approaches. Reviews show substantial differences in deployment assumptions, validation strategies, attack coverage, and public datasets [5]. These differences complicate comparisons and make reproducible evaluation as important as model design.

Agricultural environments add domain-specific constraints. Smart farms combine sensing, actuation, wireless links, gateways, and cloud services in a cyber-physical workflow. Recent studies have emphasized both the diversity of agricultural threats [1] [2] and the importance of edge-oriented detection under harsh operating conditions [3]. Hybrid deep-learning approaches have also been proposed for IoT-based smart farming [9]. Nevertheless, most reported evaluations continue to emphasize nominal classification metrics.

2.2. Compact and Hybrid IDS Architectures

Feature selection, dimensionality reduction, compact neural encoders, and ensemble classifiers are commonly used to reduce the computational cost of IoT intrusion detection. A survey by Thakkar and Lohiya identifies feature selection and evaluation methodology as persistent design issues [10]. Dataset heterogeneity further affects feature meaning and model transferability, as illustrated by the ToN-IoT analysis [11].

The present pipeline uses a CNN as a representation encoder, IWHO-Lite to select latent dimensions, and a Random Forest for the final decision. The design separates representation learning from classification and allows the final classifier to operate on a reduced latent subset. This compactness is a design property; it does not by itself establish superiority, edge-device readiness, or adversarial resistance.

2.3. Adversarial Evasion against Network IDS

Adversarial machine learning demonstrates that strong nominal performance does not guarantee robustness. Network IDS attacks differ from image-domain attacks because traffic variables are heterogeneous, constrained, and often interdependent [6]. Qiu *et al*. showed that IoT intrusion detectors can be manipulated through adversarial traffic representations [7]. Valid evaluation must therefore define attacker knowledge, objective, budget, query access, mutable variables, bounds, and validity checks.

For non-differentiable classifiers such as Random Forests, score-query attacks are a practical black-box alternative to gradient-based methods. Their results are meaningful only within the stated oracle and query budget. They do not imply robustness against white-box, transfer-based, poisoning, extraction, or packet-level adaptive attacks.

2.4. Explainability and Leakage-Aware Evaluation

SHAP provides additive feature attributions for global and local model interpretation [12]. Explainable IDS studies have used SHAP to improve transparency and support distributed or privacy-aware detection [8]. In hybrid systems, the explained features must be identified explicitly. When the Random Forest receives selected latent dimensions, SHAP describes contributions in that latent space and cannot be interpreted as a direct causal attribution to raw packet or flow variables.

Leakage-aware evaluation is equally important. Identifiers, duplicate flows, capture artifacts, and test-informed preprocessing can inflate reported performance. Dataset heterogeneity and weak feature standardization increase this risk [11]. A defensible protocol must separate fitting and evaluation roles, audit overlap, and qualify claims when device, session, or temporal identifiers are unavailable.

2.5. Research Gap

Existing work has advanced agricultural IDS design, compact inference, adversarial evaluation, and explainability, but these elements are often assessed separately. Few studies jointly report independent model seeds, disjoint calibration and threshold roles, mutable-feature perturbations, conditional ASR denominators, attack coverage, complete hard-case totals, and latent-space SHAP additivity. This study addresses that gap through a single reproducible stress-testing protocol applied to three complementary datasets.

3. Evaluated Framework and Experimental Protocol

3.1. Agricultural IoT Edge Scenario

The evaluated scenario consists of field sensors and actuators communicating through an edge gateway to supervisory or cloud services. The gateway is the intended observation point for flow-based intrusion detection. The experimental study does not reproduce a specific commercial gateway; it evaluates the behavior of the detection pipeline on flow tables representative of agricultural, IoT, and general network-security contexts.

3.2. CNN-IWHO-Lite-Random Forest Architecture

After preprocessing, each observation is represented as a feature-axis sequence of shape $d_s \times 1$, where (d_s) depends on the dataset and the training-seed-specific feature contract. A CNN-Lite encoder maps the preprocessed vector to a 128-dimensional representation:

$$z_i = g_{\theta_s}(x_i), \quad z_i \in \mathbb{R}^{128}. \tag{1}$$

The encoder comprises Conv1D (32, kernel size 5), batch normalization, ReLU, Conv1D (64, kernel size 3), batch normalization, ReLU, global average pooling, a 128-unit dense latent layer with layer normalization, a 128-unit ReLU layer, dropout of 0.1, and a sigmoid training output. The network contains 31,905 trainable parameters. Adam and binary cross-entropy are used with a batch size of 2048, a

maximum of 15 epochs, learning-rate reduction, and early stopping.

IWHO-Lite is a lightweight binary adaptation of the Improved Wild Horse Optimizer [13]. It searches a binary mask $M_s \in \{0,1\}^{128}$. The selected representation is:

$$\tilde{z}_i = M_s \odot z_i, \quad k_s = \|M_s\|_0. \tag{2}$$

The optimization objective is evaluated on the model-selection partition:

$$J(M_s) = 1 - F1_{MS}(M_s) + 0.05\frac{\|M_s\|_0}{128}. \tag{3}$$

The search uses a population of 16, 25 iterations, three restarts, a sigmoid binary transfer, and successive multi-fidelity Random Forest evaluations with 25, 50, and 100 trees. The final classifier is a 300-tree Random Forest with Gini splitting, bootstrap sampling, square-root feature subsampling, unrestricted depth, and class weights derived from the full training distribution.

A sigmoid calibrator $C_s(\cdot)$ is fitted on the calibration partition. The calibrated malicious score is:

$$p_i = C_s\left(h_{\phi_s}(\tilde{z}_i)\right). \tag{4}$$

The threshold t_s^* is selected on the independent threshold partition by maximizing F1:

$$\hat{y}_i = \mathbb{I}\left[p_i \geq t_s^*\right]. \tag{5}$$

Calibration is treated as an explicit procedural stage rather than as a universally beneficial transformation because its effect on Brier score and expected calibration error varies across datasets and seeds.

3.3. Datasets, Label Mapping, and Partitioning

The evaluation uses CICIoT2023, Farm-Flow, and UNSW-NB15. CICIoT2023 is mapped to binary labels by treating recognized benign traffic as class 0 and all attack categories as class 1. Farm-Flow uses the binary is_attack label, and the traffic field is excluded. UNSW-NB15 uses the binary label field, while id and attack_cat are excluded from the model inputs.

The experiments use model seeds 42, 2027, and 314159. Partitioning precedes all learned transformations. Same-label duplicate vectors are reduced to a single instance. Exact feature-vector overlaps across fitting and test roles are audited and removed. Conflicting-label feature groups are excluded from fitting partitions and retained without label rewriting in the test set for sensitivity analysis. CICIoT2023 contains no ambiguous test rows; Farm-Flow contains 51 rows in three ambiguous groups; and UNSW-NB15 contains eight rows in four groups.

The released feature tables do not provide a consistent device, farm, session, capture, or timestamp identifier suitable for a harmonized grouped or temporal split across all datasets. Stratified seed-specific partitions and exact-vector overlap auditing were therefore used. The small variation in the CICIoT2023 test size results from overlap removal after seed-specific partitioning. The resulting seed-specific partitions, class distributions, validation roles, overlap audit, and ambiguous test rows/groups are summarized in **Table 1**.

Table 1. Seed-specific partitions, class distributions, validation roles, and overlap audit. B/M denotes benign/malicious; MS/Cal/Thr denotes model-selection/calibration/threshold.

Dataset	Seed	Train N (B/M)	Validation N (B/M)	Test N (B/M)	MS/Cal/Thr N	Exact overlaps before/after	Ambiguous test rows/groups
CICIoT2023	42	498,924 (11,769/487,155)	149,183 (3492/145,691)	1,163,577 (27,395/1,136,182)	74,591/37,296/37,296	7622/0	0/0
CICIoT2023	2027	498,874 (11,768/487,106)	149,200 (3490/145,710)	1,163,579 (27,398/1,136,181)	74,600/37,300/37,300	7581/0	0/0
CICIoT2023	314,159	498,836 (11,773/487,063)	149,272 (3489/145,783)	1,163,586 (27,416/1,136,170)	74,636/37,318/37,318	7513/0	0/0
Farm-Flow	42	3510 (909/2601)	620 (160/460)	2348 (1768/580)	310/155/155	48/0	51/3
Farm-Flow	2027	3510 (909/2601)	620 (160/460)	2348 (1768/580)	310/155/155	48/0	51/3
Farm-Flow	314,159	3510 (909/2601)	620 (160/460)	2348 (1768/580)	310/155/155	48/0	51/3
UNSW-NB15	42	85,689 (43,912/41,777)	15,122 (7749/7373)	52,746 (33,759/18,987)	7561/3780/3781	1204/0	8/4
UNSW-NB15	2027	85,689 (43,912/41,777)	15,122 (7749/7373)	52,746 (33,759/18,987)	7561/3780/3781	1204/0	8/4
UNSW-NB15	314,159	85,689 (43,912/41,777)	15,122 (7749/7373)	52,746 (33,759/18,987)	7561/3780/3781	1204/0	8/4

3.4. Leakage-Aware Preprocessing and Validation Roles

Numeric features are imputed with the training median and standardized with training-only means and standard deviations. For UNSW-NB15, categorical variables (proto, service, and state) are imputed with the training mode and one-hot encoded with unknown categories ignored. Constant columns are identified within the training scope and removed. The fitted transformations are then applied unchanged to the model-selection, calibration, threshold, and test partitions.

The validation partition is divided into three disjoint roles. The model-selection subset is used for CNN monitoring and IWHO-Lite fitness evaluation. The calibration subset is used only to fit the sigmoid score mapping. The threshold subset is used only to select t_s^*. The held-out test set is not used for preprocessing, CNN training, latent-mask selection, calibration, or threshold optimization.

3.5. Model Selection, Calibration, and Computational Profile

Table 2 reports seed-specific input dimensions, selected latent counts, epochs, thresholds, Random Forest sizes, and measured batch-inference latency. The selected latent count ranges from 37 to 39 for CICIoT2023, 59 to 65 for Farm-Flow, and 40 to 44 for UNSW-NB15.

The final Random Forest class weights are approximately 21.19/0.512 for benign/malicious CICIoT2023 observations, 1.931/0.675 for Farm-Flow, and 0.976/1.026 for UNSW-NB15. The resulting model sizes show that the encoder is compact, whereas the Random Forest can occupy several to several tens of mebibytes. The term compact is therefore used for the representation and selected

latent subset, not as evidence of deployment readiness on a specific agricultural gateway.

Table 2. Seed-specific model configuration and measured computational profile. Latency was measured in the experimental environment and is not an edge-device benchmark.

Dataset	Seed	Input dimension	Epochs	Selected latent k	Threshold	RF size (MiB)	Latency (ms/sample)
CICIoT2023	42	43	4	37	0.734	39.3	0.098
CICIoT2023	2027	40	4	37	0.646	40.8	0.097
CICIoT2023	314,159	43	5	39	0.534	38.7	0.096
Farm-Flow	42	28	5	65	0.633	5.5	0.133
Farm-Flow	2027	28	7	65	0.583	5.6	0.135
Farm-Flow	314,159	28	5	59	0.597	5.6	0.135
UNSW-NB15	42	193	4	44	0.231	61.9	0.101
UNSW-NB15	2027	193	15	40	0.366	53.9	0.104
UNSW-NB15	314,159	194	15	41	0.412	55.4	0.103

3.6. Repetition Strategy and Reporting

Clean performance is reported as the mean and sample standard deviation across the three independent model seeds. Gaussian-noise experiments use five noise seeds within each model seed. Evasion experiments use three attack seeds within each model seed. For noise and evasion, internal repetitions are first averaged within each model seed; the reported uncertainty is then the standard deviation between the three model-seed means. This hierarchical procedure prevents nested repetitions from being treated as independent model trainings.

4. Threat, Perturbation, and Explainability Methodology

4.1. Mutable-Feature Gaussian Noise

Perturbations are applied in the cleaned continuous feature space before preprocessing. Let m_d be the dataset-specific binary mutability mask, a_d and b_d the admissible lower and upper bounds, and s_d the training standard-deviation vector. The noisy observation is:

$$\tilde{x}_i(\sigma) = \Pi_{[a_d, b_d]}\left(x_i + m_d \odot \xi_i \odot s_d\right), \quad \xi_i \sim \mathcal{N}\left(0, \sigma^2 I\right). \tag{6}$$

The canonical grid is $\sigma \in \{0, 0.2, 0.5, 1.0\}$. Five noise seeds are used per model seed. CICIoT2023 and UNSW-NB15 are evaluated on fixed test cohorts capped at 50,000 observations for this stage, whereas Farm-Flow uses its complete test set. The $\sigma = 0$ condition uses the same reference cohort as the nonzero conditions.

The mutable continuous features and the validity policy are reported in Table 3(b). Discrete, categorical, label, identifier-like, and protected variables are preserved exactly.

4.2. Bounded Score-Query Evasion

The attacker can query the calibrated malicious score and knows the feature schema, preprocessing procedure, and mutability policy. The attacker does not access CNN parameters, the internal IWHO-Lite mask values, Random Forest trees, gradients, or training data.

Budgets are normalized by the admissible feature ranges. For a mutable feature *j*, the constraint is:

$$\left|\delta_{ij}\right| \le \varepsilon\left(b_{dj} - a_{dj}\right), \quad \delta_{ij} = 0 \text{ for protected features.} \tag{7}$$

The objective is to minimize the calibrated malicious score of the complete pipeline:

$$\min_{\delta_i} C_s\left[h_{\phi_s}\left(M_s \odot g_{\theta_s}\left(x_i + \delta_i\right)\right)\right]. \tag{8}$$

The minimization is subject to Equation (7), admissible feature bounds, protected-field preservation, and the dataset-specific validity rules. The implemented attack is a batched, nested, constrained coordinate score-descent procedure. It uses three restarts, at most three coordinate sweeps, step fractions 0.5, 0.25, and 0.125, and at most 128 score queries per sample and per nonzero budget. The nested budgets are $\varepsilon = \{0.0, 0.2, 0.5, 0.8\}$, and successful adversarial observations are carried forward. Candidate values are projected to admissible bounds, protected variables are restored, and invalid candidates are rejected. For UNSW-NB15, tcprtt is recomputed as synack + ackdat.

Algorithm 1. Nested constrained coordinate score descent

- Select a fixed cohort of malicious test observations correctly detected under clean conditions.
- For each attack seed, initialize the cohort at ε = 0.
- For each increasing nonzero budget, start from the best valid observation obtained at the previous budget.
- Generate coordinate-wise candidates over the approved mutable continuous variables using the scheduled step fractions.
- Project candidates to admissible ranges, restore protected variables, and recompute dataset-specific derived constraints.
- Query the calibrated malicious score and retain a candidate only when it is valid and improves the objective.
- Stop when the score crosses the operating threshold, the query limit is reached, or no further improvement occurs.
- Carry successful observations forward to the next budget and report valid successes, queries, coverage, and rejected candidates.

The experiment characterizes a bounded black-box score-query threat model. It is not a robustness proof against white-box, transfer-based, poisoning, model-extraction, or packet-level adaptive attacks.

4.3. Conditional Attack Success Rate and Clean False-Negative Rate

For dataset *d* and model seed *s*, the eligible population is the set of malicious test

observations detected correctly before perturbation:

$$E_{d,s} = \{i : y_i = 1 \wedge \hat{y}_i(0) = 1\}. \tag{9}$$

When the eligible population exceeds the cohort cap, a fixed sample is drawn from the eligible set. Conditional ASR is then:

$$\mathrm{ASR}_{d,s}(\varepsilon) = \frac{1}{|A_{d,s}|} \sum_{i \in A_{d,s}} \mathbb{I}[\hat{y}_i(\varepsilon) = 0]. \tag{10}$$

Clean false negatives are excluded from the ASR numerator and denominator. The clean false-negative rate is reported separately as FNR = FN/(TP + FN). Under this definition, ASR(0) = 0 for every dataset, model seed, and attack seed.

4.4. TreeSHAP Audit in the Selected Latent Space

TreeExplainer is applied to the final Random Forest with raw model output and tree-path-dependent perturbation. Global explanations use 1000 observations per dataset and seed. Local explanations use representative false-positive and false-negative cases selected from the hard-case sample.

The additive decomposition is:

$$f(\tilde{z}_i) = \phi_0 + \sum_{j \in M_s} \phi_{ij}. \tag{11}$$

Strict additivity checks pass for all nine dataset-seed combinations without fallback. The explanations refer to selected latent dimensions z_j. Because independently trained CNN encoders do not align their latent coordinates, a label such as z_{029} is seed-specific and cannot be mapped directly to a raw flow variable or interpreted causally without a separate linkage analysis.

4.5. Hard-Case Analysis

False positives and false negatives are identified after thresholding. The signed decision margin is:

$$r_i = p_i - t_s^*. \tag{12}$$

A positive margin for a false positive indicates confidence above the alert threshold, whereas a negative margin for a false negative indicates that the malicious score remained below the threshold. Complete error totals are reported separately from the limited samples used for local SHAP explanations.

5. Results

5.1. Clean Classification Performance and Perturbation Policy

Table 3(a) reports clean performance across the three model seeds. CICIoT2023 obtains a very high F1-score, but its specificity is materially lower because benign observations form a small minority. Farm-Flow combines high malicious recall with low precision and specificity, indicating a substantial baseline false-positive burden. UNSW-NB15 occupies an intermediate position, with high recall but lower specificity and seed-dependent F1.

Table 3. (a) Clean test performance, mean ± standard deviation across three independent model seeds; (b) Dataset-specific mutable continuous features and validity controls.

(a)

Dataset	F1	Balanced accuracy	Specificity	MCC	Clean FNR	FPR
CICIoT2023	0.9953 ± 0.0001	0.9052 ± 0.0202	0.8153 ± 0.0411	0.8035 ± 0.0115	0.489% ± 0.079%	18.47% ± 4.11%
Farm-Flow	0.4724 ± 0.0131	0.6345 ± 0.0187	0.3167 ± 0.0560	0.2684 ± 0.0209	4.770% ± 1.891%	68.33% ± 5.60%
UNSW-NB15	0.7480 ± 0.0183	0.8097 ± 0.0183	0.6369 ± 0.0428	0.6041 ± 0.0289	1.754% ± 0.810%	36.31% ± 4.28%

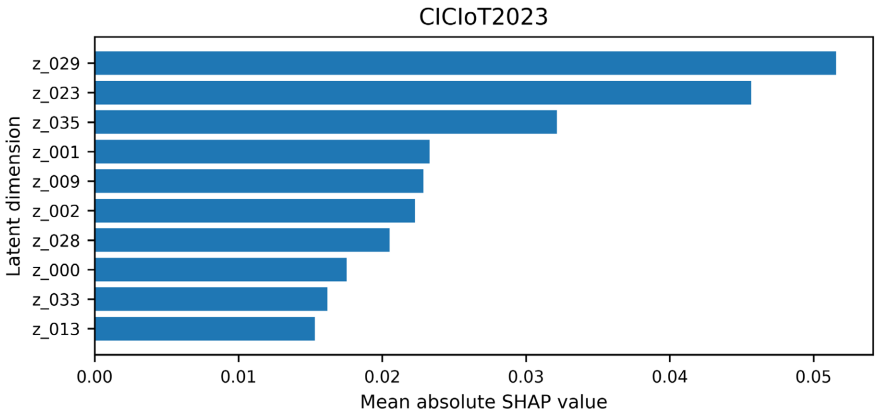
(b)

Dataset	Mutable continuous features	Mutable/ protected N	Derived validity rule	Validity scope
CICIoT2023	flow_duration, Header_Length, Duration, Rate, Srate, Drate, IAT	7/33-36	None	Feature-space bounds and protected-field preservation
Farm-Flow	orig_ip_bytes, resp_ip_bytes, fwd_pkts_per_sec, bwd_pkts_per_sec, flow_pkts_per_sec	5/23-23	None	Feature-space bounds and protected-field preservation
UNSW-NB15	dur, sinpkt, dinpkt, sjit, djit, synack, ackdat	7/34-34	tcprtt = synack + ackdat	Feature-space bounds and protected-field preservation

The selected perturbation variables are conservative subsets of the continuous flow descriptors. All other model inputs remain protected. The resulting validity claim is limited to bounded feature-space plausibility; packet-capture reconstructibility is not asserted.

5.2. Global and Local TreeSHAP Audits

Figure 1 shows the top selected latent dimensions for the representative model seed 314159. Importance is concentrated in a subset of the selected representation for every dataset. The identities and magnitudes differ across datasets, and the selected masks contain 39, 59, and 41 dimensions for CICIoT2023, Farm-Flow, and UNSW-NB15, respectively.



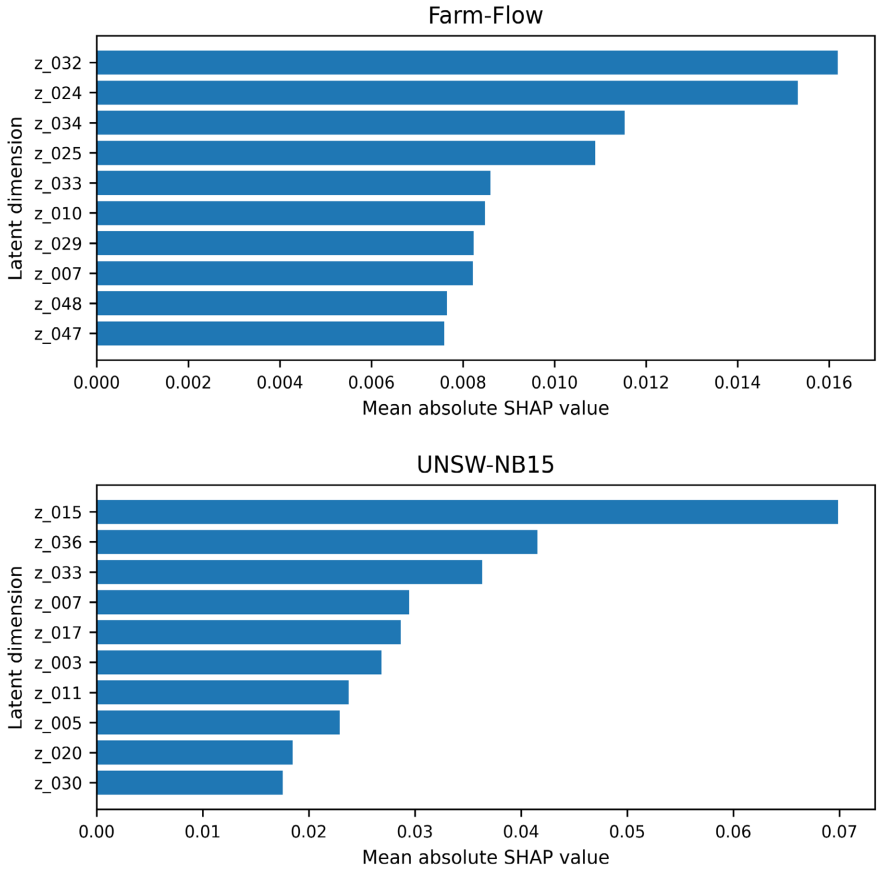
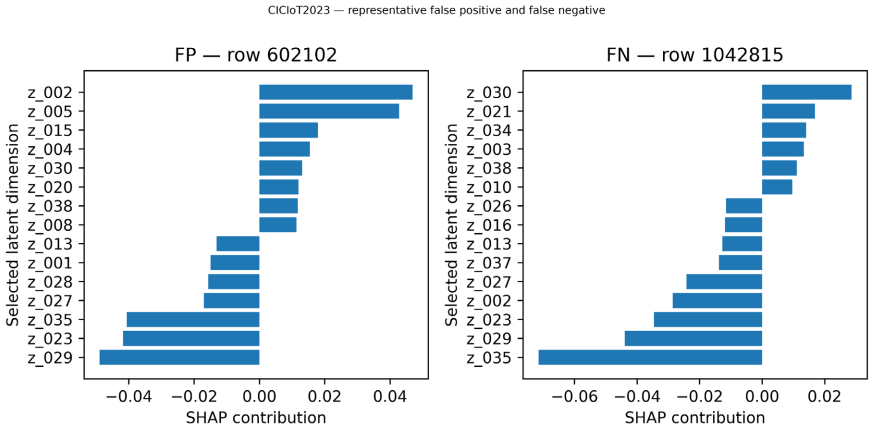


Figure 1. Global TreeSHAP importance of selected latent dimensions for model seed 314159: (a) CICIoT2023, (b) Farm-Flow, and (c) UNSW-NB15. The labels refer to seed-specific latent coordinates and do not denote raw traffic variables.

Figure 2 presents representative false-positive and false-negative explanations for the same model seed. Positive SHAP contributions increase the Random Forest output, while negative contributions decrease it. The plots demonstrate that individual errors can be decomposed additively, but they do not provide a direct semantic mapping from latent coordinates to packet-level mechanisms.



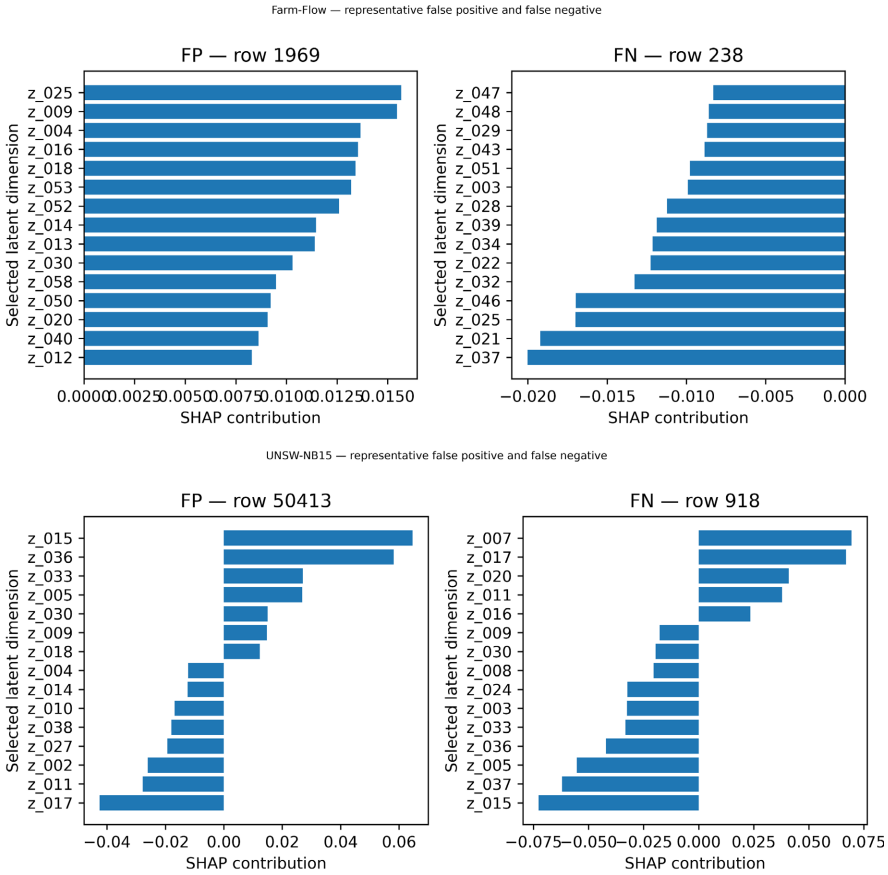


Figure 2. Local TreeSHAP explanations of representative false-positive and false-negative cases for model seed 314159. The explanations are computed in the IWHO-Lite-selected latent space.

5.3. Robustness under Gaussian Noise

The hierarchical F1 results are shown in **Table 4** and **Figure 3**. CICIoT2023 remains nearly stable across the canonical noise grid. Farm-Flow declines immediately from its already low clean baseline and then approaches a plateau. UNSW-NB15 degrades progressively, with increasing between-seed variability at larger noise levels.

Table 4. F1-score under Gaussian noise, hierarchical mean ± between-model-seed standard deviation; five noise seeds are nested within each model seed.

Dataset	$\sigma = 0.0$	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 1.0$
CICIoT2023	0.9954 ± 0.0002	0.9953 ± 0.0002	0.9949 ± 0.0002	0.9940 ± 0.0002
Farm-Flow	0.4724 ± 0.0131	0.4184 ± 0.0025	0.4130 ± 0.0016	0.4118 ± 0.0013
UNSW-NB15	0.7479 ± 0.0182	0.7439 ± 0.0205	0.7192 ± 0.0358	0.6767 ± 0.0550

At $\sigma = 1.0$, the F1-score is 0.9940 ± 0.0002 for CICIoT2023, 0.4118 ± 0.0013 for Farm-Flow, and 0.6768 ± 0.0550 for UNSW-NB15. These results indicate that en-

vironmental robustness cannot be inferred from clean F1 alone. Farm-Flow is primarily constrained by baseline separability and false positives, while UNSW-NB15 displays a clearer perturbation-dependent decline.

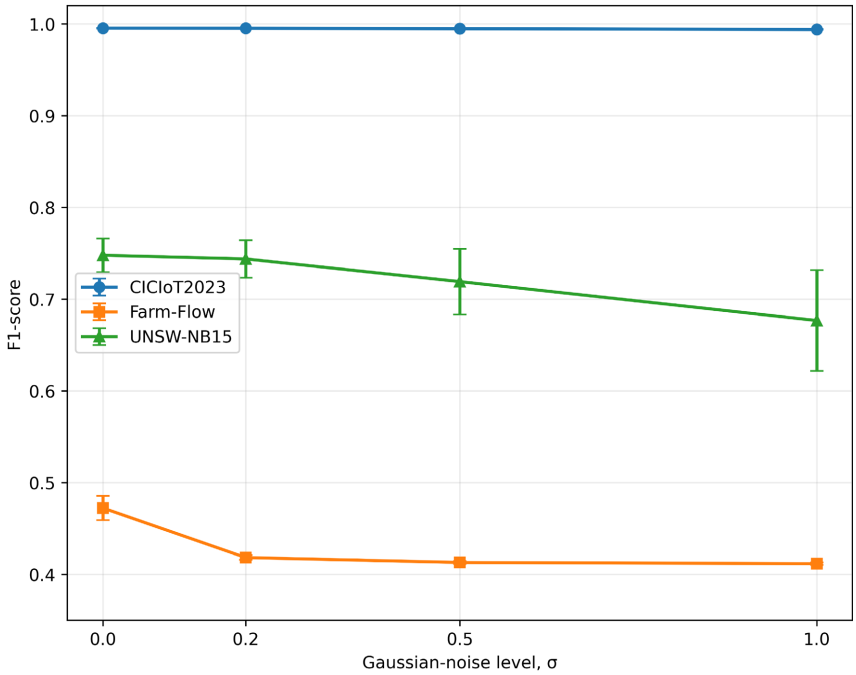


Figure 3. F1-score under Gaussian noise. Points are within-model-seed means and error bars represent the standard deviation between three independent model-seed means.

5.4. Conditional Evasion Results

Table 5(a) distinguishes the full malicious test population, the clean-detected eligible population, the attacked cohort, attack coverage, valid adversarial observations, rejected candidates, and clean FNR. Farm-Flow is evaluated exhaustively on all clean-detected malicious observations. CICIoT2023 and UNSW-NB15 use fixed cohorts capped at 2000 observations per model seed.

Table 5. (a) Conditional-ASR denominators, attack coverage, and clean false-negative rate. Ranges refer to the three model seeds; (b) Conditional ASR under normalized L∞ budgets, hierarchical mean ± between-model-seed standard deviation; three attack seeds are nested within each model seed.

(a)

Dataset	Malicious test N	Clean-detected eligible N	Attacked N	Coverage	Valid adversarial N	Invalid N	Clean FNR
CICIoT2023	1,136,170 - 1,136,182	1,130,001 - 1,131,640	2000	0.177% - 0.177%	2000	0	0.489% ± 0.079%
Farm-Flow	580	540 - 561	540 - 561	100.0%	540 - 561	0	4.770% ± 1.891%
UNSW-NB15	18,987	18,490 - 18,795	2000	10.641% - 10.817%	2 000	0	1.754% ± 0.810%

Continued

(b)				
Dataset	$\varepsilon = 0.0$	$\varepsilon = 0.2$	$\varepsilon = 0.5$	$\varepsilon = 0.8$
CICIoT2023	0.00% ± 0.00%	3.53% ± 2.24%	14.40% ± 19.85%	16.17% ± 22.66%
Farm-Flow	0.00% ± 0.00%	11.64% ± 8.70%	13.05% ± 10.08%	13.58% ± 10.56%
UNSW-NB15	0.00% ± 0.00%	87.93% ± 4.20%	96.70% ± 2.03%	98.43% ± 1.17%

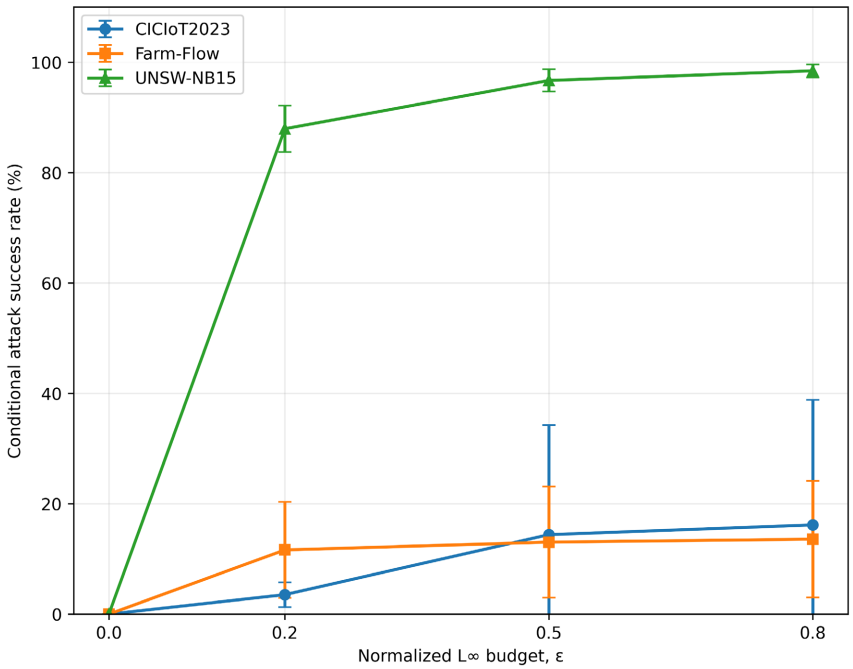


Figure 4. Conditional Attack Success Rate under normalized L∞ budgets. Error bars represent the standard deviation between three independent model-seed means.

Figure 4 shows the conditional Attack Success Rate (ASR) as a function of the normalized L∞ budget for the three datasets. Nested-budget monotonicity holds within every model-seed and attack-seed combination. As shown in Table 5(b), UNSW-NB15 is consistently vulnerable: ASR is 87.93% ± 4.20% at $\varepsilon = 0.2$ and 98.43% ± 1.17% at $\varepsilon = 0.8$. CICIoT2023 and Farm-Flow show much larger between-seed uncertainty relative to their means. CICIoT2023 ranges from very low ASR in two seeds to substantially higher values in seed 42, while Farm-Flow ranges from low to moderate evasion success. These datasets therefore cannot be characterized reliably from a single model initialization.

5.5. Complete False-Positive and False-Negative Statistics

Table 6 reports complete error totals rather than limiting the counts to the observations displayed by SHAP. Farm-Flow produces approximately 1208 false positives per seed against only 27.7 false negatives, which is consistent with its high recall and low specificity. UNSW-NB15 also produces a large false-positive bur-

den. CICIoT2023 has high absolute error counts because of its much larger test set, although its relative F1 remains high.

Table 6. Complete hard-case statistics, mean ± standard deviation across three model seeds. The SHAP sample is a diagnostic subset and is not the total error count.

Dataset	Error	Total cases	Mean score	Mean margin	SHAP sample N
CICIoT2023	FP	5060.3 ± 1129.4	0.856 ± 0.043	+0.218 ± 0.057	100
CICIoT2023	FN	5555.7 ± 895.3	0.280 ± 0.047	-0.358 ± 0.056	100
Farm-Flow	FP	1208.0 ± 99.0	0.746 ± 0.006	+0.142 ± 0.020	100
Farm-Flow	FN	27.7 ± 11.0	0.515 ± 0.025	−0.089 ± 0.015	19 - 40
UNSW-NB15	FP	12259.0 ± 1444.0	0.733 ± 0.031	+0.397 ± 0.063	100
UNSW-NB15	FN	333.0 ± 153.8	0.219 ± 0.079	−0.117 ± 0.016	100

False-positive mean scores remain well above their corresponding thresholds, particularly for CICIoT2023 and UNSW-NB15. False-negative margins are negative for all datasets, showing that some malicious observations remain confidently below the operating threshold. These results support the use of score distributions and local explanations alongside aggregate classification metrics.

5.6. Ablation Perspective

Table 7 compares the evaluated pipeline with a full-latent Random Forest, a Random Forest trained on preprocessed raw features, and an end-to-end CNN. The raw-feature Random Forest obtains the highest mean F1 on CICIoT2023 and UNSW-NB15 and is competitive on Farm-Flow. The evaluated pipeline therefore should not be interpreted as uniformly superior. Its scientific role in this paper is to provide a concrete compact hybrid system for a controlled stress-testing and explainability audit.

Table 7. F1-score ablation audit, mean ± standard deviation across three model seeds.

Dataset	CNN-IWHO-Lite-RF	CNN full-latent RF	Raw-feature RF	CNN end-to-end
CICIoT2023	0.9953 ± 0.0001	0.9949 ± 0.0001	0.9982 ± 0.0001	0.9929 ± 0.0005
Farm-Flow	0.4724 ± 0.0131	0.4550 ± 0.0058	0.4669 ± 0.0276	0.3981 ± 0.0034
UNSW-NB15	0.7480 ± 0.0183	0.7578 ± 0.0045	0.7989 ± 0.0193	0.7071 ± 0.0629

6. Discussion

6.1. Dataset-Specific Failure Modes

The three datasets expose different failure modes. CICIoT2023 combines extremely high malicious recall and F1 with lower benign specificity. Its behavior

under Gaussian noise is stable, but its evasion sensitivity varies markedly across model seeds. This pattern shows that a high aggregate F1 does not imply a stable adversarial boundary or a low operational false-positive rate.

Farm-Flow is the most relevant dataset for the agricultural application, yet its main limitation is visible before adversarial testing. The model detects most malicious observations, but it labels a large proportion of benign flows as malicious. Gaussian perturbation further reduces F1, while conditional ASR remains low to moderate and highly seed-dependent. The central Farm-Flow finding is therefore a baseline separability and false-alarm problem rather than a claim that agricultural traffic is uniquely susceptible to evasion.

UNSW-NB15 shows intermediate clean performance and progressive noise degradation, but its decisive weakness is adversarial. Nearly all attacked clean-detected malicious observations cross the threshold at the largest budget. This result demonstrates that compact representation learning and latent selection do not constitute a proven defense against score-query evasion.

6.2. Interpretation of Latent Selection and Calibration

IWHO-Lite reduces the 128-dimensional representation to smaller seed-dependent subsets. On Farm-Flow, the selected subset is larger than on the other datasets, which is consistent with a more difficult separation problem but does not establish causality. The ablation results show that raw-feature Random Forests can outperform the hybrid pipeline. Latent compression should therefore be understood as a representation choice and an audit target, not as a guaranteed performance or robustness advantage.

Sigmoid calibration and threshold selection are methodologically separated from the test set. Nevertheless, calibration does not improve Brier score or expected calibration error uniformly. The procedure standardizes the score-to-decision workflow and enables a disjoint threshold stage, but its empirical benefit remains dataset- and seed-dependent.

6.3. Operational Implications for Agricultural Security Monitoring

A deployment-oriented IDS should monitor more than a single accuracy value. The following controls are particularly important for agricultural gateways and security operations:

- false-positive rate and alert volume on benign operational traffic;
- clean false-negative rate, reported separately from adversarial success;
- periodic stress tests under the approved mutable-feature policy;
- repeated model training to quantify seed sensitivity;
- threshold review when devices, firmware, seasons, or operational schedules change;
- local latent-space explanations for representative high-confidence errors;
- explicit coverage reporting when adversarial testing uses capped cohorts.

The measured latency values indicate that batch inference is technically feasible

in the experimental environment, but no conclusion is drawn about a specific gateway, memory budget, energy profile, or thermal constraint. Hardware deployment requires direct benchmarking.

6.4. Scientific Positioning

The principal contribution is the evaluation protocol: disjoint fitting roles, overlap auditing, independent model seeds, hierarchical uncertainty, conditional ASR, denominator transparency, feature-space validity controls, and additive explainability checks. The pipeline is the system under examination. The results do not support a general claim that CNN-IWHO-Lite-Random Forest is superior to simpler classifiers, uniformly lightweight, or adversarially robust.

This positioning is consistent with the broader IDS literature, which emphasizes that validation strategy and dataset structure can affect conclusions as strongly as the classifier itself [5] [11]. A stress-testing study is valuable when it identifies both strengths and failure boundaries without converting dataset-specific observations into universal claims.

6.5. Limitations

Several limitations remain. First, reliable device, session, farm, capture, and timestamp identifiers are not consistently available across the released feature tables. Group-based and temporal splits could therefore not be imposed across all datasets. The protocol removes exact feature-vector overlaps and separates all learned stages from the test set, but it does not claim that every possible source of deployment-level leakage has been eliminated.

Second, CICIoT2023 and UNSW-NB15 use fixed adversarial cohorts capped at 2000 clean-detected malicious observations per model seed. Their conditional ASR estimates characterize the sampled cohorts, and coverage is reported explicitly. Farm-Flow is evaluated exhaustively because its eligible population is smaller.

Third, validity is assessed in the cleaned feature space. Continuous mutable values are bounded, protected variables are restored, and the UNSW-NB15 tcprtt relationship is recomputed. No packet-capture reconstruction or protocol execution is performed. In particular, the released Farm-Flow representation includes cleaned model-input values, so perturbations cannot be claimed to correspond to original physical network units.

Fourth, the evaluated attack is a bounded black-box score-query coordinate search. White-box, transfer-based, poisoning, model-extraction, and packet-level adaptive attacks remain outside its scope. The results therefore describe a defined threat model rather than a complete security proof.

Fifth, SHAP explanations operate on selected latent dimensions. The latent coordinates are seed-specific and are not directly mapped to raw network variables. The explanations establish additive attribution for the final Random Forest but not causal or packet-level semantics.

Finally, the computational measurements were obtained in the experimental environment rather than on an agricultural gateway. Model size, latency, memory, energy consumption, thermal behavior, and packet-processing overhead require hardware-specific validation.

7. Conclusions

This study evaluated a CNN-IWHO-Lite-Random Forest intrusion-detection pipeline through a reproducible multi-seed protocol that combines clean testing, Gaussian perturbation, conditional score-query evasion, TreeSHAP explanation, and complete hard-case analysis. The protocol separates model selection, calibration, threshold optimization, and final testing; removes exact feature-vector overlaps; constrains perturbations to approved continuous variables; and reports attack denominators and coverage explicitly.

The results reveal distinct failure modes. CICIoT2023 maintains a very high F1-score and strong noise stability, but benign specificity and adversarial sensitivity vary across seeds. Farm-Flow is limited mainly by baseline false positives and degrades under Gaussian noise. UNSW-NB15 is consistently vulnerable to the evaluated evasion procedure despite reasonable clean performance. These findings demonstrate that robustness cannot be inferred from nominal F1, latent compression, or a single model initialization.

The main practical implication is that AG-IoT IDS evaluation should include repeated training, denominator-correct adversarial testing, feature-valid perturbations, and decision-level error analysis before deployment. Future work should extend the protocol to grouped and temporal agricultural captures, packet-realizable adversarial traffic, multiclass detection, online drift, and direct benchmarking on resource-constrained gateways.

Author Contributions

Conceptualization, A.K.K.; methodology, A.K.K., D.J.D. and K.A.A.; software, A.K.K.; formal analysis, A.K.K.; investigation, A.K.K.; data curation, A.K.K.; writing—original draft preparation, A.K.K.; writing—review and editing, A.K.K., D.J.D. and K.A.A.; visualization, A.K.K.; supervision, S.O. and Y.C.B. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Maraveas, C., Rajarajan, M., Arvanitis, K.G. and Vatsanidou, A. (2024) Cybersecurity Threats and Mitigation Measures in Agriculture 4.0 and 5.0. *Smart Agricultural Technology*, **9**, Article 100616. https://doi.org/10.1016/j.atech.2024.100616

[2] Campoverde-Molina, M. and Luján-Mora, S. (2025) Cybersecurity in Smart Agriculture: A Systematic Literature Review. *Computers & Security*, **150**, Article 104284.

https://doi.org/10.1016/j.cose.2024.104284

[3] Javeed, D., Gao, T., Saeed, M.S. and Kumar, P. (2024) An Intrusion Detection System for Edge-Envisioned Smart Agriculture in Extreme Environment. *IEEE Internet of Things Journal*, **11**, 26866-26876. https://doi.org/10.1109/jiot.2023.3288544

[4] Ferreira, R., Bispo, I.A., Rabadão, C., Santos, L. and Costa, R.L.D.C. (2025) Farm-Flow Dataset: Intrusion Detection in Smart Agriculture Based on Network Flows. *Computers and Electrical Engineering*, **121**, Article 109892. https://doi.org/10.1016/j.compeleceng.2024.109892

[5] Khraisat, A. and Alazab, A. (2021) A Critical Review of Intrusion Detection Systems in the Internet of Things: Techniques, Deployment Strategy, Validation Strategy, Attacks, Public Datasets and Challenges. *Cybersecurity*, **4**, Article 18. https://doi.org/10.1186/s42400-021-00077-7

[6] He, K., Kim, D.D. and Asghar, M.R. (2023) Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials*, **25**, 538-566. https://doi.org/10.1109/comst.2022.3233793

[7] Qiu, H., Dong, T., Zhang, T., Lu, J., Memmi, G. and Qiu, M. (2021) Adversarial Attacks against Network Intrusion Detection in IoT Systems. *IEEE Internet of Things Journal*, **8**, 10327-10335. https://doi.org/10.1109/jiot.2020.3048038

[8] Fatema, K., Dey, S.K., Anannya, M., Khan, R.T., Rashid, M.M., Su, C., *et al*. (2025) Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP. *Future Internet*, **17**, Article 234. https://doi.org/10.3390/fi17060234

[9] Kethineni, K. and Pradeepini, G. (2024) Intrusion Detection in Internet of Things-Based Smart Farming Using Hybrid Deep Learning Framework. *Cluster Computing*, **27**, 1719-1732. https://doi.org/10.1007/s10586-023-04052-4

[10] Thakkar, A. and Lohiya, R. (2022) A Survey on Intrusion Detection System: Feature Selection, Model, Performance Measures, Application Perspective, Challenges, and Future Research Directions. *Artificial Intelligence Review*, **55**, 453-563. https://doi.org/10.1007/s10462-021-10037-9

[11] Booij, T.M., Chiscop, I., Meeuwissen, E., Moustafa, N. and Hartog, F.T.H.D. (2022) Ton_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets. *IEEE Internet of Things Journal*, **9**, 485-496. https://doi.org/10.1109/jiot.2021.3085194

[12] Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, New York, 4-9 December 2017, 4768-4777.

[13] Zheng, R., Hussien, A.G., Jia, H., Abualigah, L., Wang, S. and Wu, D. (2022) An Improved Wild Horse Optimizer for Solving Optimization Problems. *Mathematics*, **10**, Article 1311. https://doi.org/10.3390/math10081311