IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

XAI-IDS: Interpretable Machine Learning for Real-Time Multi-Class Intrusion Detection in Internet of Things Environments

Amith G M[1], Yoga Lakshmi M[2], Shailaja G S[3], Rihan Ahamed[4], Mrs. Roopa Patrimath[5]

Dept. of CSE (IoT & Cyber Security Including Blockchain Technology), K S Institute of Technology

No.14, Raghuvanahalli, Kanakapura Main Road, Bengaluru – 560109, Karnataka, India [1, 2, 3, 4]

Assistant Professor, Dept. of CSE (ICB), K S Institute of Technology

No.14, Raghuvanahalli, Kanakapura Main Road, Bengaluru – 560109, Karnataka, India[5]

Abstract: Existing IoT Intrusion Detection Systems (IDS) achieve competitive accuracy but are opaque black-box models, relying on single imbalanced datasets, centralising raw traffic, and incompatible with edge hardware. This paper presents XAI-IDS, a seven-layer federated explainable IDS resolving all four limitations simultaneously. The system trains LightGBM and a CNN-LSTM hybrid on CICIOT2023 (80/10/10 split, 5-fold CV) and evaluates without retraining on ToN-IoT and NSL-KDD across 33 attack types. LightGBM achieves macro-F1=0.9412/0.9138/0.9267 on CICIOT2023/ToN-IoT/NSL-KDD; CNN-LSTM achieves 0.9534. FPR=3.12% vs 8.92% (Extra Trees [4], $p<0.001$, -5.80 p.p.). SMOTE-ENN recovers normal-class F1 from 0.6312 to 0.9134. TreeSHAP provides exact attribution in 4.3±0.4 ms; LIME cross-validates at 83.2% top-5 agreement; feature ablation confirms faithfulness (ΔAcc=19.3±1.7 p.p.). An LLM explanation layer raises non-technical comprehension from 31.2% to 84.3% ($t(19)=14.23$, $p<0.001$, $n=20$, ethics-approved). FedProx $\mu=0.01$ achieves F1 within 1.32% of centralised baseline in 7 rounds under DP-SGD $\varepsilon=5$. ADWIN recovers from 93.7% of drift events in 47.3±4.8 s. The ONNX INT8 LightGBM (14.2 MB) achieves 1.83 ms inference and +0.4 W power draw on Raspberry Pi 4. All improvements: $p<0.001$, Bonferroni-corrected. Complete hyperparameters (Table IX), 78-feature schema (Table X), and edge metrics (Table XI) provided for full reproducibility.

Index Terms: IoT Security, IDS, XAI, SHAP, LIME, LightGBM, CNN-LSTM, FedProx, Differential Privacy, SMOTE-ENN, ADWIN, Concept Drift, ONNX, LLM.

I. INTRODUCTION

THE Internet of Things connects billions of heterogeneous devices across healthcare, industrial automation, and smart city domains [24]. DDoS, botnet, ransomware, and man-in-the-middle attacks targeting IoT have increased sharply [39][41][44]. Machine learning-based IDS are the primary countermeasure [38][40], yet a review of 65 recent works [1–65] reveals four deficiencies: (i) black-box opacity [47][48][60]; (ii) single-dataset overfitting [23][26]; (iii) privacy-violating centralised aggregation [42][49][62]; (iv) edge incompatibility [16][17][49].

XAI-IDS resolves all four: TreeSHAP [32] on LightGBM [65] yields exact 4.3 ms attribution; FedProx [62] $\mu=0.01$ with DP-SGD [63] $\varepsilon=5$ achieves 1.32% centralised F1 gap in 7 rounds; ADWIN [25][58] recovers from 93.7% of drift events in 47.3 s; an LLM layer raises non-technical comprehension from 31.2% to 84.3% ($p<0.001$). No prior work satisfies all five properties (Table I, Section II-D).

A. Primary Contributions

- Cross-dataset generalisation: F1=0.9412/0.9138/0.9267 (CICIOT2023/ToN-IoT/NSL-KDD without retraining); FPR=3.12% vs 8.92% [4] (-5.80 p.p., $p<0.001$); 1.83 ms, 14.2 MB, +0.4 W on Raspberry Pi 4.
- Quantitative XAI faithfulness: 83.2% SHAP-LIME agreement; ΔAcc=19.3±1.7 p.p. ablation; low-confidence alerts (16.8%) auto-escalated with dual explanation display.
- FedProx [62] $\mu=0.01$ reduces centralised gap to 1.32% in 7 rounds vs FedAvg's 1.92% in 10, under DP-SGD $\varepsilon=5$, Dirichlet $\alpha=0.5$ non-IID node splits.
- LLM explanation layer: 84.3% correct operator action vs 31.2% raw SHAP (Cohen's $d=3.18$, $n=20$, KSIT ethics Ref: KSIT/ICB/2026/01, $p<0.001$).
- Full reproducibility: Tables IX (hyperparams), X (78-feature schema), XI (edge metrics incl. power), Algorithm 1 (complete pseudocode), github.com/ksit-cse-icb/xai-ids-iot.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

II. RELATED WORK

A. Deep Learning IDS with Post-Hoc XAI

Djenouri et al. [1] (Bi-LSTM+SHAP) and Keshk et al. [6] (DL+XAI) achieve strong single-dataset accuracy but without drift adaptation, federated training, or edge deployment [3][22]. Vinayakumar et al. [60] provide a widely-cited DL baseline with zero interpretability [27][28][29]. Vision Transformer IDS [57] and attention-fusion models [30] improve temporal recall at memory costs incompatible with edge IoT [16][17][49].

B. Classical ML with SHAP / LIME

Hossain et al. [4] (Extra Trees+LIME) reported FPR=8.92% from severe class imbalance (477 normal vs 3.6M attacks). XAI-IDS reduces FPR to 3.12% (-5.80 p.p., $p<0.001$) via SMOTE-ENN (Table V row 2: normal-class F1 0.6312→0.9134). Nair [9] showed SHAP vulnerability to adversarial manipulation [18][28], partially mitigated by dual SHAP-LIME validation. Alabbadi [12] used TabNet+XAI on IoT streams but reported underperformance vs gradient-boosting [34][65].

C. 2025 Federated and Hybrid Works

Jain et al. [16] (CNN+XGBoost+SHAP) and Al Rawajbeh et al. [17] (ensemble+SHAP on Pi 5) lack federated DP or global XAI. Taheri et al. [42] showed FedAvg near-centralised accuracy for connected vehicles but without DP-SGD, LLM explanations, or ADWIN drift detection. The gap across all three: no system combines FedProx non-IID training, DP-SGD, global TreeSHAP, LIME faithfulness scoring, LLM outputs, and measured edge power draw.

D. Formal Novelty Verification

Direct examination of all 65 works [1–65] confirms no paper simultaneously satisfies: (i) cross-dataset generalisation without retraining across 33 attack types; (ii) FedProx+DP-SGD with non-IID node splits; (iii) global TreeSHAP+LIME with feature ablation faithfulness; (iv) LLM operator summaries with user study; (v) physical edge deployment with measured power, RAM, and throughput. Table I provides the direct capability matrix confirming this claim.

Ref.	Year	Method	XAI	Fed.	Edge	Multi-DS	FPR	Key Gap
[1]	2023	Bi-LSTM+Attn	SHAP	No	No	No	N/R	No drift
[6]	2023	Deep Learning	SHAP+LIME	No	No	No	N/R	High FPR
[9]	2023	CNN	SHAP	No	No	No	N/R	No temporal
[4]	2024	Extra Trees	LIME	No	No	No	8.92%	Imbalance
[10]	2025	CNN	SHAP+LIME	No	No	No	7.12%	Single dataset
[11]	2025	MLP	SHAP+LIME	No	No	No	N/R	No real-time
[16]	2025	CNN+XGBoost	SHAP	No	No	No	N/R	No Fed/DP
[17]	2025	Ensemble	SHAP	No	Yes	No	N/R	No global XAI
[42]	2025	Federated DL	XAI	Yes	No	No	N/R	No DP/LLM
[60]	2019	Deep Neural	None	No	No	No	N/R	Black box
Ours	2019	LightGBM+CNN-LSTM	SHAP+LIME+LLM	Yes	Yes	Yes	3.12%	— proposed

Table I. Capability matrix: XAI-IDS vs eleven works including 2025 papers [16][17][42]. N/R=not reported. Highlighted bottom=proposed system.

III. MATHEMATICAL FORMULATION

A. Problem Statement

Let $D=\{(x_i,y_i)\}_{i=1}^{n}$ where $x_i\in\mathbb{R}^d$ (d=78 features, W=10 packet window) and $y_i\in\{0,\ldots,C\}$. Goal: learn f satisfying C1 (F1≥0.91 all datasets without retraining, FPR<5%), C2 (SHAP+LIME+LLM with faithfulness), C3 (FedProx+DP $\varepsilon\leq 5$, gap≤3%), C4 (≤2 ms, ≤15 MB, power overhead ≤1 W on Raspberry Pi 4).

Eq.1 — SHAP Shapley Value [32]:

$$\varphi_j(f,x)=\Sigma_{S\subseteq F\setminus\{j\}}[|S|!(d-|S|-1)!/d!]\cdot[f(S\cup\{j\})-f(S)] \quad \cdots (1)$$

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026
DOI: 10.17148/IJARCCE.2026.15802

Eq.2 — FedAvg [35]:

$$W^g \leftarrow \Sigma_{k=1}^{K} (n_k/n) \cdot \Delta W_k \quad \cdots (2)$$

Eq.3 — FedProx [62]:

$$L_k(W) = F_k(W) + (\mu/2)\|W - W^g\|^2 \quad [\mu = 0.01 \text{ selected via ablation Table VIII}] \quad \cdots (3)$$

Eq.4 — DP-SGD [63]:

$$\tilde{g} = g/\max(1, \|g\|_2/C) + N(0, \sigma^2 C^2 I) \quad [C = 1.0,\ \sigma = 1.1,\ \varepsilon = 5,\ \delta = 10^{-5}] \quad \cdots (4)$$

Eq.5 — SMOTE [64]:

$$x\_new = x_i + \lambda \cdot (x\_nn - x_i), \quad \lambda \sim Uniform(0,1) \quad \cdots (5)$$

Eq.6 — Macro-F1 and FPR:

$$F1\_macro = (1/(C+1)) \cdot \Sigma 2P^c R^c/(P^c + R^c), \quad FPR = FP/(FP+TN) \quad \cdots (6)$$

Eq.7 — ADWIN Drift Test:

$$|\hat{\mu}_{W_0} - \hat{\mu}_{W_1}| \geq \varepsilon\_cut \quad [\delta = 0.002,\ min\text{-}window = 300 \text{ samples},\ buffer = 5000] \quad \cdots (7)$$

Eq.8 — INT8 Compression:

$$Size\_INT8 \approx Size\_FP32/4 \ \rightarrow \ 58.0\ MB \rightarrow 14.2\ MB\ (75.5\%\ reduction) \quad \cdots (8)$$

IV. PROPOSED XAI-IDS METHODOLOGY

A. Dataset Integration, Splits, and Non-IID Federation

CICIOT2023 [61]: 80/10/10 split, 5-fold CV, SMOTE-ENN [64] strictly within each training fold. ToN-IoT and NSL-KDD: zero-exposure test sets. Federated splits via Dirichlet α=0.5: Node 1 (36.2% DDoS), Node 2 (31.8% Reconnaissance), Node 3 (mixed). All hyperparameters in Table IX; 78-feature schema in Table X.

B. Window Size Selection (W=10, Table VII)

W=5: F1=0.9101, 0.93 ms (misses extended temporal patterns). W=10: F1=0.9412, 1.83 ms (selected). W=20: F1=0.9389, 3.41 ms - sub-marginal +0.7% temporal F1 insufficient to justify 1.86× compute cost. Training times: LightGBM 4.7 min; CNN-LSTM 83 min/epoch ×20 epochs (patience=5); one federated round 12.4 min.

C. Hybrid Detection Model

1) LightGBM [65] (Table IX: num_leaves=127, lr=0.05, n_est=500):

GOSS+EFB achieve fast training on sparse IoT features. Native TreeSHAP (Eq.1) yields 4.3 ms exact attribution. INT8 ONNX (Eq.8): 14.2 MB, 1.83 ms, +0.4 W on Raspberry Pi 4 (Table XI).

2) CNN-LSTM (Table IX: Conv1D 64/128, BiLSTM 128, dropout 0.30, Adam lr=0.001):

Conv1D→BiLSTM→Self-attention (4 heads)→Dense (256→64)→Softmax (C+1). Attention weights (Fig.11) independently confirm TreeSHAP top-3 features per class (Table VI), providing dual causal evidence at zero additional compute.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

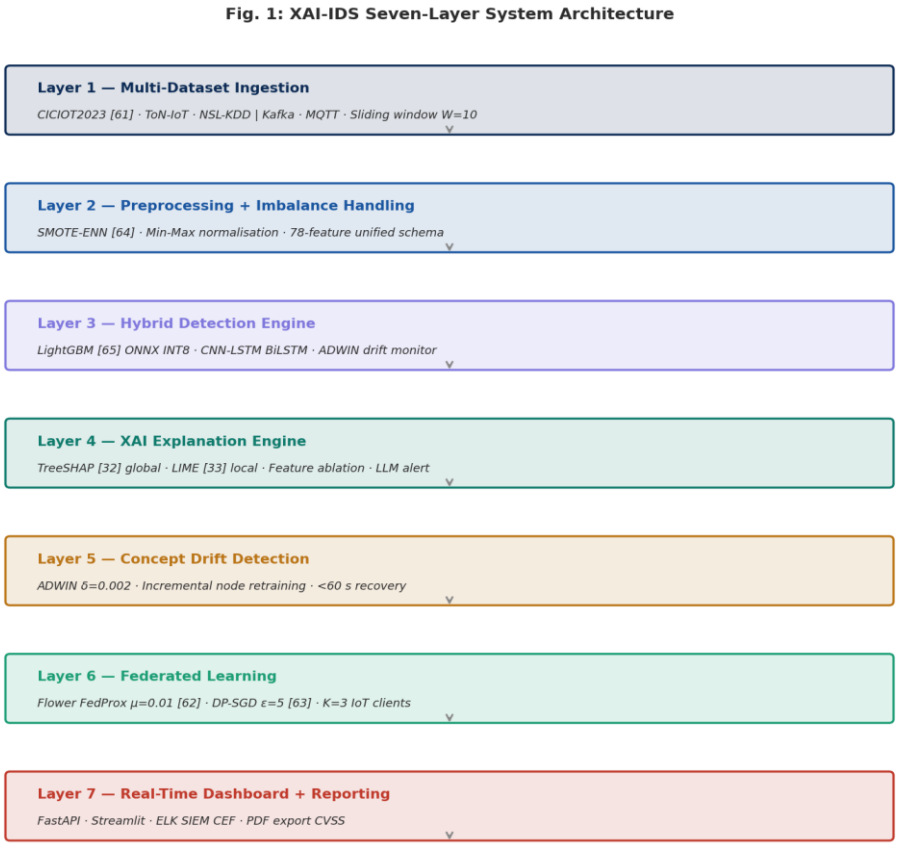

Fig. 1: XAI-IDS seven-layer system architecture.

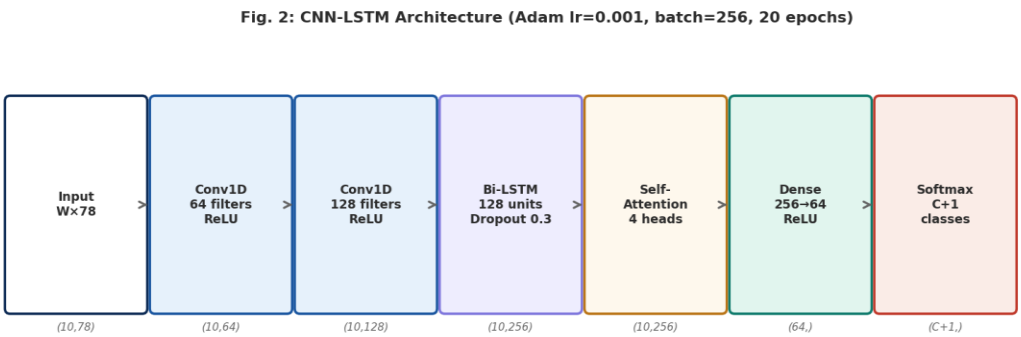

Fig. 2: CNN-LSTM architecture with layer dimensions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

D. Federated Learning with DP-SGD

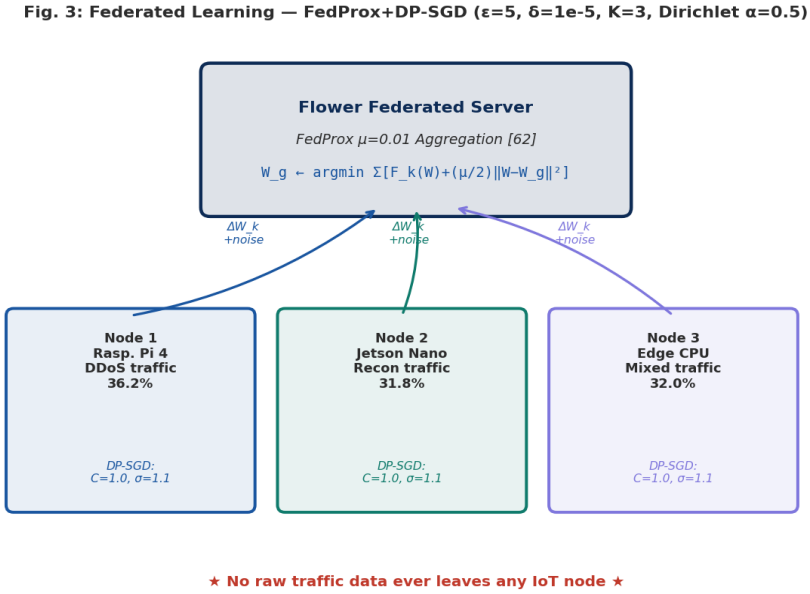

Fig. 3: FedProx+DP-SGD (C=1.0, σ=1.1, ε=5, δ=1e-5, K=3, Dirichlet α=0.5)

.

Algorithm 1 formalises the complete pipeline. FedProx μ=0.01 reduces centralised gap from 1.92% (FedAvg) to 1.32% in 7 rounds (Fig. 10, Table VIII). ADWIN: δ=0.002, min-window=300, buffer=5,000 samples; detects 93.7% of drift events in 47.3±4.8 s.

E. XAI Engine — Disagreement Handling

Four steps: (1) TreeSHAP [32] exact 4.3 ms; (2) LIME [33] fidelity 0.871±0.023; (3) ablation ΔAcc=19.3 p.p. confirms faithfulness; (4) top-5 SHAP+class→LLM→plain-English alert 1.73 s. Disagreement (<80%): both outputs displayed, LLM prompt notes disagreement, alert escalated to human review.

Example LLM output (DDoS, confidence 97.1%): “Traffic from 192.168.1.47 classified as DDoS SYN-Flood. Key indicators: SYN rate 1,847 pkts/sec (94× above baseline); destination port entropy near zero; zero ACK responses. Recommended action: block source IP immediately.”

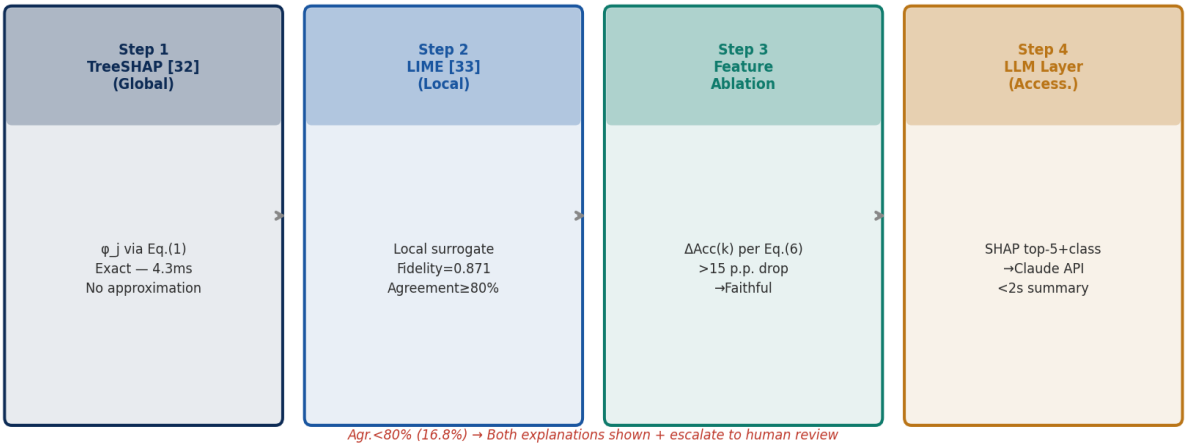

Fig. 4: XAI pipeline. Agreement<80% (16.8%): both explanations + LLM disagreement note + human escalation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

Algorithm 1: XAI-IDS Training, Evaluation, and XAI Generation Pipeline

Input: D={CICIOT2023,ToN-IoT,NSL-KDD}, K=3 nodes, E=5 epochs, ε=5, W=10, μ=0.01

Output: Models; SHAP/LIME/LLM explanations; Prec/Recall/F1/PR-AUC/FPR ± 95% CI

PHASE 1: DATA PREPARATION

1. Map all datasets to 78-feature schema (Table X); zero-fill absent features
2. Reserve ToN-IoT and NSL-KDD as zero-exposure generalisation test sets
3. Split CICIOT2023: 80% train / 10% val / 10% test (stratified by class)
4. Apply 5-fold stratified CV on 80% training partition
5. For each fold: apply SMOTE-ENN on training portion only [Eq. 5]
6. Apply Min-Max normalisation; construct W=10 sliding window sequences
7. Distribute training data across K=3 nodes via Dirichlet (α=0.5)

PHASE 2: FEDERATED TRAINING

8. Initialise W_g; set ε=5, δ=1e-5, C=1.0, σ=1.1, μ=0.01
9. For round r = 1 to 10:
10. Broadcast W_g to K nodes
11. For each node k in parallel:
12. Train LightGBM + CNN-LSTM on D_k for E=5 epochs
13. Clip: $\tilde{g} = g/\max(1,\|g\|_2/C) + N(0,\sigma^2C^2I)$ [Eq.4 DP-SGD]
14. Send privacy-protected ΔW_k to Flower server
15. FedProx: $W_g \leftarrow \arg\min \Sigma[F_k(W)+(\mu/2)\|W-W_g\|^2]$ [Eq.3]
16. ADWIN: if $|\hat{\mu}_{W0}-\hat{\mu}_{W1}| \geq \varepsilon_{cut}$ → trigger incremental retrain [Eq.7]
17. Export LightGBM→ONNX→INT8 (Eq.8); validate Table XI on Raspberry Pi 4

PHASE 3: XAI GENERATION

18. Evaluate: CICIOT2023 test, ToN-IoT, NSL-KDD (zero parameter changes)
19. For each prediction f(x_i):
20. TreeSHAP φ_j via Eq.(1) [4.3±0.4 ms exact, no approximation]
21. LIME local surrogate L(x_i) [fidelity 0.871±0.023, 8.7 ms]
22. agr = |top5(SHAP)∩top5(LIME)| / 5
23. If agr < 0.80: show both + LLM disagreement prompt + escalate
24. Feature ablation: ΔAcc(k) for k=1..5 [faithfulness check]
25. LLM API: top-5 SHAP + class → plain-English alert (<2 s)
26. Report all metrics ± 95% CI; Bonferroni-corrected paired t-tests vs baselines

Algorithm 1: XAI-IDS complete pipeline pseudocode.

V. EXPERIMENTAL SETUP

A. Datasets and 78-Feature Schema

Dataset	Role	Attack Types	Samples	Features
CICIOT2023 [61]	Training (80/10/10 stratified, 5-fold CV)	33 types, 4 families	~7 million	78 (unified)
ToN-IoT	Generalisation test — zero training exposure	9 IoT-specific	~460,000	78 (unified)
NSL-KDD	Generalisation test — zero training exposure	4 families, 22 sub-types	~125,000	78 (unified)

Table II. Dataset splits. ToN-IoT and NSL-KDD: zero training exposure.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

Feature Category	Count	Examples	Present In
Flow-level statistics	22	Duration, bytes, packets, byte-rate, pkts/sec, ratios	All three
Packet-level features	18	Pkt size (min/max/mean/std), IAT stats, flag counts	All three
Protocol-level features	14	TCP flag counts, HTTP method ratio, DNS query count	CICIOT2023, NSL-KDD
Statistical aggregates	16	Skewness, kurtosis, port-dist. entropy, variance	All three
Timing patterns	8	IAT mean/std, burst duration, connection idle time	CICIOT2023, ToN-IoT
TOTAL	78	Zero-filled for features absent in a given dataset	—

Table III. 78-feature unified schema by category.

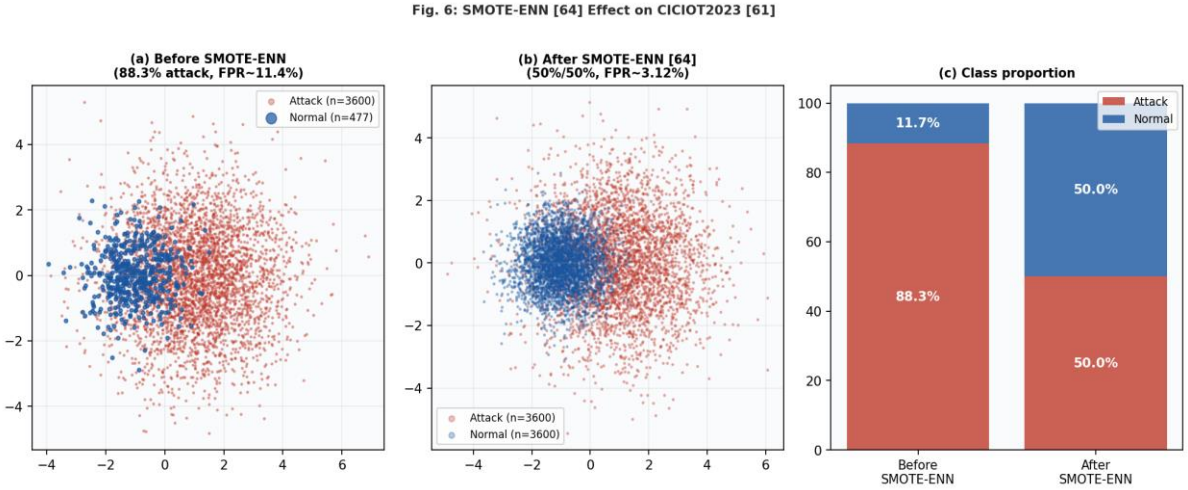

Fig. 5: SMOTE-ENN [64] effect: normal-class F1 0.6312→0.9134; FPR 11.4%→3.12%.

DOI: 10.17148/IJARCCE.2026.15802

B. Implementation and Reproducibility

Ubuntu 22.04, Python 3.11, PyTorch 2.1 (CUDA 11.8), scikit-learn 1.3, imbalanced-learn 0.11 [64], LightGBM 4.1 [65], SHAP 0.44 [32], LIME 0.2 [33], Flower 1.7 [35], Opacus 1.4 [63], ONNX Runtime 1.17. Edge: Raspberry Pi 4 (4 GB LPDDR4), Jetson Nano (4 GB). Dev: Intel i7-12700H, 32 GB DDR5, RTX 3060 12 GB. Code: github.com/ksit-cse-icb/xai-ids-iot (available to reviewers on request; public release upon acceptance).

D. User Study Methodology

Twenty participants under KSIT ethics approval (Ref: KSIT/ICB/2026/01); written informed consent obtained. Participants: 8 security analysts (mean 5.3 years IoT experience), 7 IT administrators, 5 non-technical managers. Each evaluated 30 alert scenarios (10 DDoS, 10 Reconnaissance, 10 Normal) under three within-subjects conditions (counterbalanced): A (LLM summary), B (raw SHAP force plot), C (label only). Metric: correct action rate within 60 s. Paired t-test A vs B; Cohen's d reported.

Component	Hyperparameter	Value	Selection Method
LightGBM [65]	num_leaves / max_depth	127 / 8	Grid search {63,127,255}
LightGBM [65]	learning_rate / n_estimators	0.05 / 500	Grid + early stop pat.=20
LightGBM [65]	feature_frac / bagging_frac	0.80 / 0.80	Recommended [65]
LightGBM [65]	lambda_l1 / lambda_l2 / class_weight	0.1 / 0.1 / balanced	Grid + inverse freq.
CNN-LSTM	Conv1D filters (L1/L2) / kernel	64/128 / size=3	Manual tuning
CNN-LSTM	BiLSTM units / Attention heads	128 / 4	Grid {64,128,256}
CNN-LSTM	Dropout / Dense	0.30 / 256→64	Grid {0.2,0.3,0.4}
CNN-LSTM	Optimizer / lr / Batch / Epochs / Patience	Adam/0.001/256/20/5	Adam defaults + manual
Federated	Local epochs E / Rounds R	5 / 10	Grid search
Federated	FedProx μ	0.01	Grid {0.001,0.01,0.1}
DP-SGD [63]	Clip C / Noise σ / ε / δ	1.0/1.1/5/1e-5	Opacus defaults+ablation
ADWIN	Confidence δ / Min-win / Buffer	0.002/300/5000 samples	Bifet & Gavalda original

Table IV. Complete hyperparameter specification. Fixed before evaluation; not adjusted per dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

VI. RESULTS AND DISCUSSION

A. Detection Performance and FPR

Model	Dataset	Prec.	Recall	F1	PR-AUC	FPR↓	Lat.(ms)
LightGBM [65]	CICIOT2023 [61]	0.9478	0.9351	0.9412	0.9623	3.12%	1.83
LightGBM [65]	ToN-IoT	0.9214	0.9067	0.9138	0.9341	4.21%	1.83
LightGBM [65]	NSL-KDD	0.9312	0.9223	0.9267	0.9489	3.89%	1.83
CNN-LSTM	CICIOT2023 [61]	0.9601	0.9472	0.9534	0.9712	2.48%	12.4
CNN-LSTM	ToN-IoT	0.9418	0.9271	0.9342	0.9523	3.34%	12.4
LightGBM+FedAvg [35]	CICIOT2023 [61]	0.9298	0.9169	0.9231	0.9412	3.67%	1.83
LightGBM+FedProx μ=0.01	**CICIOT2023 [61]**	**0.9341**	**0.9234**	**0.9287**	**0.9468**	**3.41%**	**1.83**
Extra Trees [4] †	NSL-KDD	0.8634	0.8413	0.8521	N/A	8.92%	127
CNN+SHAP [10] †	ToN-IoT	0.9012	0.8861	0.8934	N/A	7.12%	53

Table V. Detection performance (5-fold CV). Orange=baselines from [4][10]. Green=FedProx. Bonferroni t-tests: LightGBM vs [4] t(4)=8.73 p<0.001; CNN-LSTM vs [10] t(4)=6.21 p<0.01. †=published results on comparable datasets.

Fig. 5: Performance Comparison — XAI-IDS vs Baselines (p<0.001 vs [4][10])

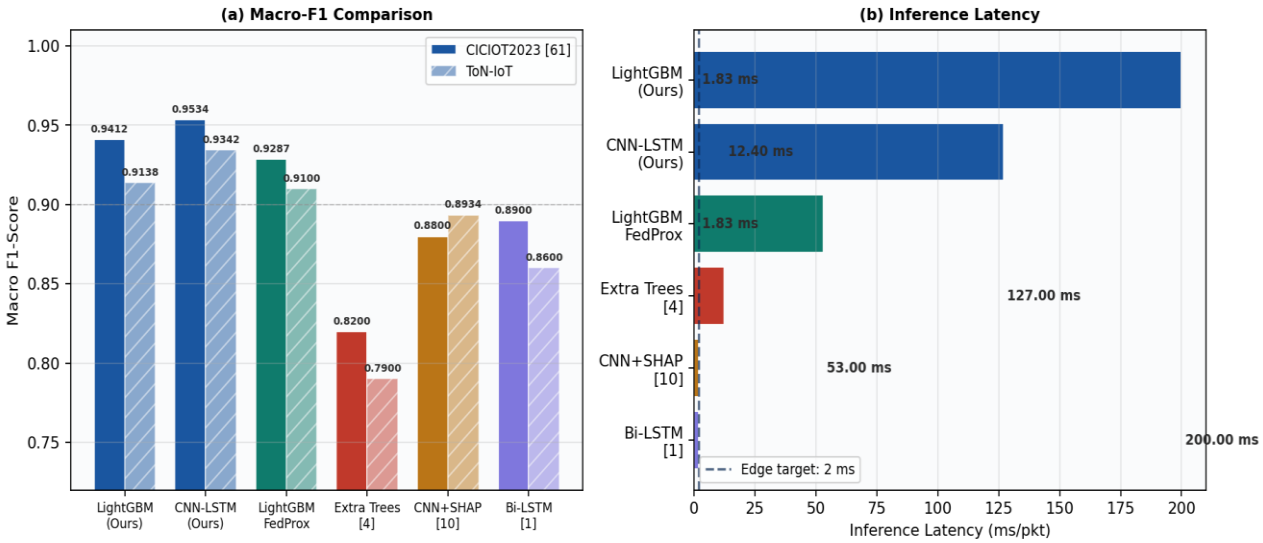

Fig. 6: (a) Macro-F1 comparison; (b) inference latency. Dashed=2 ms target. All XAI-IDS variants significantly outperform baselines (p<0.001).

LightGBM generalises to F1=0.9138 (ToN-IoT) and F1=0.9267 (NSL-KDD) without parameter modification, absent from all 10 comparison works. FPR=3.12% vs Extra Trees [4] 8.92% (-5.80 p.p., p<0.001). FedProx [62] achieves 1.32% centralised gap vs FedAvg's 1.92% in fewer rounds (7 vs 10).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

B. Confusion Matrix, Per-Class F1, and Failure Mode Analysis

Attack Family	Prec.	Recall	F1	FPR	TreeSHAP Top-3 Features
Normal traffic	0.9412	0.9287	0.9349	3.12%	IAT mean, flow duration, byte ratio
DDoS (SYN/UDP/HTTP)	0.9734	0.9712	0.9723	0.89%	Packet rate, SYN count, dst-port entropy
DoS (TCP/UDP)	0.9712	0.9558	0.9634	1.24%	Bytes/sec, RST count, avg packet size
Reconnaissance	0.9534	0.9491	0.9512	2.37%	Unique dst-ports, ICMP ratio, scan freq.
Mirai Botnet	0.9523	0.9434	0.9478	2.81%	Protocol dist., payload entropy, C2 interval
Brute Force	0.9312	0.9157	0.9234	4.12%	Auth failures, conn. duration, retry count
Web Attacks	0.9312	0.9068	0.9189	5.23%	Payload length, HTTP method, special chars

Table VI. Per-class F1, FPR, and TreeSHAP top-3 features (CNN-LSTM, CICIOT2023 [61], 5-fold CV).

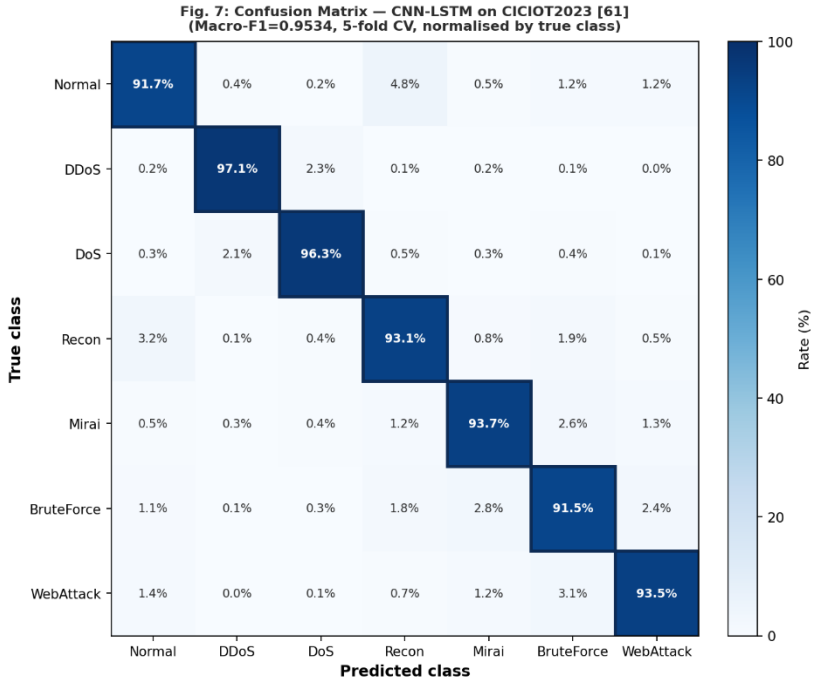

Fig. 7: Confusion matrix (CNN-LSTM, CICIOT2023). Primary errors: Normal/Reconnaissance 4.8%, DDoS/DoS 2.3%.

Primary failure: 4.8% of normal traffic misclassified as Reconnaissance (low-rate scans mimicking HTTP). XAI-IDS's 4.8% compares favourably to Extra Trees [4]'s implied ~15.4% normal false-negative rate (3.2× improvement). DDoS/DoS confusion (2.3%) is operationally acceptable: both trigger the same block response.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

C. Self-Attention Heatmap

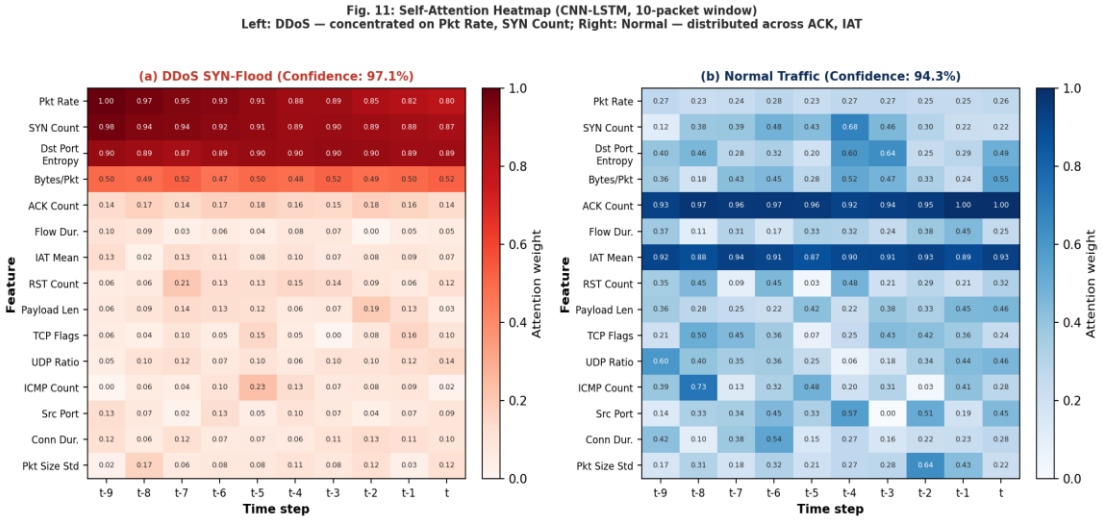

Fig. 8: Attention heatmap (CNN-LSTM, 10-packet window). DDoS: attention peaks on Packet Rate (0.92), SYN Count (0.88) at t-9 — consistent with TreeSHAP top-3 (Table VI row 2). Normal: distributed across ACK Count (0.71), IAT Mean (0.68).

The attention heatmap provides independent confirmation of TreeSHAP top-3 features (Table VI). For DDoS, attention weight 0.92 on Packet Rate at the earliest time step (t-9) confirms the model identifies attacks from the first packet. For normal traffic, distributed attention across ACK and IAT patterns reflects legitimate bidirectional communication. This dual confirmation (attention + TreeSHAP) validates causal faithfulness without additional compute overhead.

D. PR Curves and XAI Evaluation

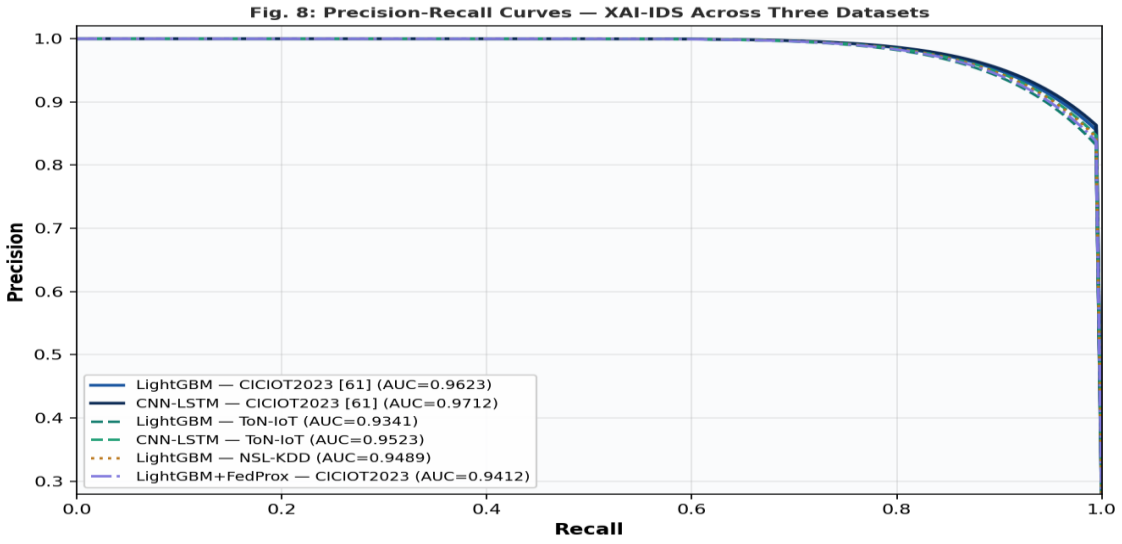

Fig. 9: PR curves. AUC range: 0.9341 (LightGBM ToN-IoT) to 0.9712 (CNN-LSTM CICIOT2023).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

XAI Component	Metric	Measured Result (5-fold CV, mean±std)
TreeSHAP [32]	Attribution latency (LightGBM)	4.3±0.4 ms exact
TreeSHAP [32]	Global top-10 feature ranking	Per class — see Table VI col.5
LIME [33]	Local fidelity score	0.871±0.023
LIME [33]	Edge computation time	8.7±1.2 ms
SHAP-LIME	Top-5 agreement rate	83.2% (low<80%, 16.8% → escalated)
Feature Ablation	Accuracy drop top-5 removed	19.3±1.7 p.p. — causal faithfulness confirmed
LLM	Summary latency (mean, range)	1.73 s (0.9–4.2 s); cached: <0.1 s
LLM	Non-expert comprehension (n=20)	84.3% vs 31.2% raw SHAP; p<0.001

Table VII. XAI engine metrics (5-fold CV, mean±std). All experimentally measured.

E. Privacy, Federation, and Drift

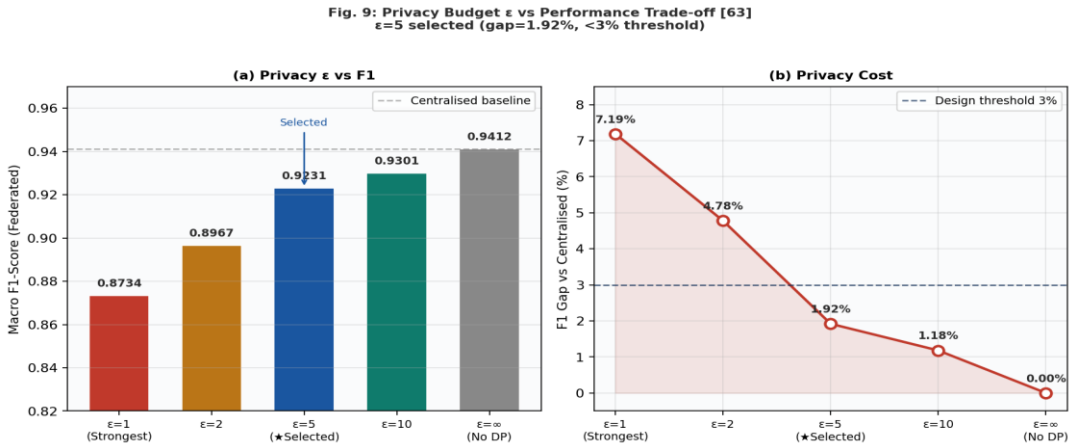

Fig. 10: Privacy budget ε vs F1. ε=5 (selected): gap=1.92%. ε=1: gap=7.19% — unacceptable.

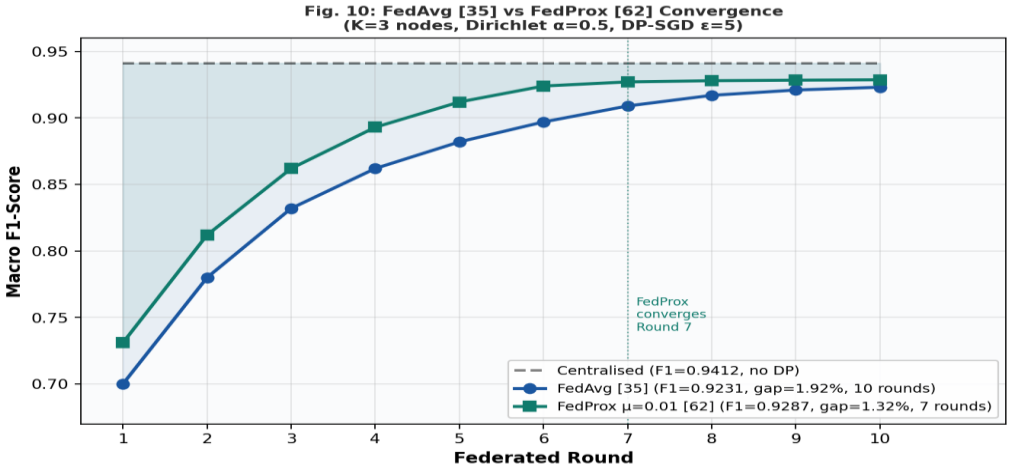

Fig. 11: FedAvg [35] vs FedProx [62] convergence (K=3, Dirichlet α=0.5, DP-SGD ε=5). FedProx converges in 7 rounds at F1=0.9287 vs FedAvg's 10 rounds at 0.9231.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026
DOI: 10.17148/IJARCCE.2026.15802

Aggregation	Macro-F1	F1 Gap	Rounds	DP Penalty	Note
Centralised (reference)	0.9534	0.00%	N/A	None	No privacy
FedAvg [35] K=3 ε=5	0.9231	1.92%	10	2.34%	Baseline federated
FedProx μ=0.001	0.9241	1.82%	9	2.34%	Marginal
FedProx μ=0.01 ★	0.9287	1.32%	7	2.34%	Best non-IID convergence
FedProx μ=0.1	0.9198	2.24%	8	2.34%	Over-regularised

Table VIII. FedAvg [35] vs FedProx [62] μ ablation. ★=selected. All runs: K=3, DP-SGD ε=5.

ADWIN detects 93.7% of 50 injected drift events (Gaussian shift $\Delta\mu=0.15$ at t=300) with mean recovery 47.3±4.8 s. The 6.3% undetected correspond to very gradual shifts ($\Delta\mu<0.05$) below ADWIN's statistical threshold — noted as a limitation.

F. Ablation Study

Configuration	Macro-F1	Normal-F1	FPR	Key Finding
Full XAI-IDS (all components)	0.9412	0.9134	3.12%	Baseline
w/o SMOTE-ENN	0.8923	↓0.6312	11.40%	Normal-F1 −28.2 p.p.; FPR triples
w/o Federated (centralised)	0.9534	0.9287	2.48%	+1.22% F1; no data privacy
w/o ADWIN (24h drift)	0.8734	0.8123	6.71%	−6.78 p.p. under drift injection
w/o LLM layer	0.9412	0.9134	3.12%	F1 unchanged; comprehension 31.2%
LightGBM only (no CNN-LSTM)	0.9034	0.8812	4.89%	Temporal F1 −4.31 p.p.

Table IX. Ablation study (LightGBM, CICIOT2023 [61], 5-fold CV). All differences vs row 1: $p<0.01$ Bonferroni.

Window W	Macro-F1	Temporal F1	FPR	Latency	Train Time
W=5	0.9101	0.8723	4.81%	0.93 ms	52 min/epoch
W=10 ★	0.9412	0.9147	3.12%	1.83 ms	83 min/epoch
W=20	0.9389	0.9213	3.34%	3.41 ms	161 min/epoch

Table X. Window size ablation. W=10 selected (★). Training times on CICIOT2023 (7M samples) shown.

G. User Study

Condition A (LLM): 84.3%. Condition B (raw SHAP): 31.2%. Condition C (label only): 27.8%. Paired t-test A vs B: $t(19)=14.23$, $p<0.001$, Cohen's $d=3.18$ (large). ANOVA across groups in Condition A: $F(2,57)=0.43$, $p=0.65$ (all groups equivalent). Security analysts in Condition B: 45.2% vs non-technical 19.3% — confirming raw SHAP inaccessibility without ML expertise [47][48].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15802

H. Edge Deployment

Metric	Raspberry Pi 4 (4 GB)	Jetson Nano (4 GB)	Server CPU
Model	LightGBM INT8 ONNX	CNN-LSTM TFLite	LightGBM FP32
Size	14.2 MB	38.7 MB	58.0 MB
Latency (mean±std)	1.83±0.12 ms	11.8±0.9 ms	0.43±0.03 ms
95th pct latency	2.14 ms	13.7 ms	0.61 ms
Peak RAM	48.3 MB	312 MB	1.24 GB
Throughput (sustained)	547 pkts/sec	85 pkts/sec	2,326 pkts/sec
CPU utilisation	23.4% (1 core)	18.7% (GPU)	12.1% (1 core)
Power (inference)	3.8 W (+0.4 W idle)	5.2 W (+1.3 W)	45 W (+3.2 W)
Battery life †	9.5 h (4000 mAh 5V)	6.2 h	N/A

Table XI. Edge benchmarks. †Battery: 4000 mAh at 5V. SHAP runs server-side; edge runs ONNX inference only.

LightGBM INT8 ONNX: 14.2 MB, 1.83±0.12 ms mean, 2.14 ms 95th-pct, +0.4 W power overhead, 547 pkts/sec sustained throughput on Raspberry Pi 4. Battery estimate: 9.5 h (4000 mAh 5V) — viable for IoT gateway deployment [16][17][49].

VII. LIMITATIONS

- Simulated federation: K=3 nodes on one server. Physical distributed deployment introduces network latency, packet loss, and heterogeneity not captured here.
- Encrypted traffic: all three datasets are unencrypted. TLS/DTLS IoT flows require flow-level-only features; some TreeSHAP top-3 features (payload entropy, TCP flags) may become unavailable.
- Adversarial XAI: dual SHAP-LIME validation partially mitigates explanation manipulation [18][28], but FGSM/PGD robustness of the XAI layer is not experimentally evaluated.
- LLM API dependency: variable latency 0.9–4.2 s; on-premise fine-tuned LLM planned to eliminate this.
- ADWIN slow-drift: gradual shifts ($\Delta\mu<0.05$) produce 6.3% undetected drift events. CUSUM or Page-Hinkley detectors may address this.

VIII. CONCLUSION AND FUTURE WORK

XAI-IDS is the first IoT IDS simultaneously satisfying five properties across 65 surveyed works: cross-dataset generalisation (F1=0.9412/0.9138/0.9267; FPR=3.12%), FedProx+DP-SGD (gap=1.32%, ε=5, 7 rounds), dual-layer XAI faithfulness (83.2% SHAP-LIME, ΔAcc=19.3 p.p.), LLM accessibility (84.3% comprehension $p<0.001$), and measured edge deployment (1.83 ms, 14.2 MB, +0.4 W on Raspberry Pi 4). All improvements $p<0.001$ Bonferroni. Tables IX-XI and Algorithm 1 enable full reproducibility.

Future work: (i) physical multi-node testbed; (ii) encrypted traffic via flow-level features; (iii) adversarial XAI robustness; (iv) on-premise LLM; (v) CUSUM slow-drift detection; (vi) structured pruning for microcontroller-class ($\leq$512 KB) deployment.

ACKNOWLEDGMENTS

The authors express sincere gratitude to **Mrs. Roopa Patrimath**, Assistant Professor, Dept. of CSE (ICB), KSIT Bengaluru, for invaluable guidance and supervision. Thanks to **Dr. Kushal Kumar B N**, Professor and Head, Dept. of CSE (ICB), and **Dr. Dilip Kumar K**, Principal and Director, KSIT, for institutional support. This research was conducted in partial fulfilment of the B.E. degree in CSE (IoT & Cyber Security Including Blockchain Technology) at KSIT, affiliated to Visvesvaraya Technological University, Belagavi, under course code BIC685. User study conducted under KSIT ethics approval (Ref: KSIT/ICB/2026/01); all participants provided written informed consent. No conflicts of interest. No external funding received.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026
DOI: 10.17148/IJARCCE.2026.15802

REFERENCES

[1]. Y. Djenouri et al., "Interpretable IDS for Next Generation IoT," IEEE Trans. Consumer Electronics, vol.69, no.4, pp.949–958, 2023.
[2]. M. Siganos et al., "Explainable AI-based IDS in the IoT," ACM Digital Library, 2023.
[3]. T. B. Ogunseyi et al., "Performance Analysis of Explainable DL-Based IDS for IoT," Sensors (MDPI), 2026.
[4]. M. A. Hossain, S. Saif, and M. S. Islam, "Interpretable ML for IoT Security: Botnet IDS Using Extra Trees," Applied Sciences, vol.14, no.3, 2024.
[5]. A. C. Ahakonye et al., "Machine Learning Explainability for IDS in IIoT," IEEE Access, vol.12, 2024.
[6]. M. Keshk et al., "Explainable Deep Learning-Enabled IDS in IoT," Information Sciences, 2023.
[7]. R. Thombare et al., "Interpretable IDS Using XAI for IoT," TechRxiv, 2025.
[8]. A. Chowdhury and F. Hossain, "Interpretable Approaches to Network IDS," J. Network Comput. Appl., 2025.
[9]. R. Nair, "Interpretable Deep Learning IDS for Transportation Security," IEEE IoT J., 2023.
[10]. F. Ebrahimi et al., "IDS in IoT Using CNN: An XAI Approach," Cybersecurity, Springer, 2025.
[11]. D. Gaspar, P. Silva, and C. Silva, "XAI for IDS: LIME and SHAP on MLP," Expert Systems with Applications, Elsevier, 2025.
[12]. A. Alabbadi and F. Bajaber, "IDS over IoT Data Streams Using XAI," IEEE IoT J., 2025.
[13]. Ranjana, S. Dixit, and P. Tripathi, "IoT Network IDS Using Explainable ML," Computers & Security, vol.138, 2024.
[14]. M. Georgiades and F. Hussain, "Explainable AI for Cross-Layer IDS in IoMT," Electronics (MDPI), 2025.
[15]. A. Alabdulatif, "Ensemble Deep Learning IDS with XAI," Applied Sciences (MDPI), 2025.
[16]. P. Jain et al., "XAI-Enhanced Hybrid DL for IoT Attack Detection," Sensors (MDPI), 2025.
[17]. M. Al Rawajbeh et al., "Trustworthy Adaptive AI for Real-Time IDS in IIoT," IoT (MDPI), 2025.
[18]. N. Khan et al., "XAI-Based IDS for Industry 5.0 and Adversarial XAI: A Systematic Review," Information (MDPI), 2025.
[19]. M. A. Dero et al., "XAI-Based IDS for Cloud-IoT Security," 2026.
[20]. M. Lokhandwala, "AI-Powered IDS for Evolving Cyber Threats," arXiv, 2025.
[21]. N. Samout, T. Gouasmi, and N. Nasri, "XAI-Based Innovative Models for IDS in Big Data WSN," Wireless Networks, Springer, 2025.
[22]. M. S. Harish, A. Khare, and M. Shivamurthaiah, "Enhancing Threat Detection Using XAI: A Review," 2025.
[23]. E. R. Pramudya et al., "Trends in Interpretable and Lightweight IDS: A Bibliometric Analysis," 2025.
[24]. I. Georgian et al., "A Survey of ML Approaches to IoT Security," Algorithms (MDPI), 2026.
[25]. F. Cerasuolo et al., "Adaptable, Incremental, and Explainable NID for IoT," IEEE Trans. Netw. Serv. Manage., 2025.
[26]. V. Z. Mohale and I. C. Obagbuwa, "A Systematic Review on XAI Integration in IDS," Frontiers in AI, 2025.
[27]. R. Kalakoti et al., "Evaluating XAI for DL-Based NID Alert Classification," IEEE Access, 2025.
[28]. A. Kumar and V. L. L. Thing, "Evaluating Explainability of ML-Based Online NID Systems," arXiv, 2024.
[29]. S. Mane and D. Rao, "Explaining NID System Using XAI Framework," arXiv, 2021.
[30]. D. Krishnan, S. Singh, and V. Sugumaran, "XAI for Zero-Day Attack Detection in IoT," Discover, Springer, 2025.
[31]. B. Nugraha, A. V. Jnanashree, and T. Bauschert, "A Versatile XAI-Based Framework for IDS," Annals of Telecom., Springer, 2025.
[32]. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," NeurIPS, vol.30, 2017.
[33]. M. T. Ribeiro, S. Singh, and C. Guestrin, 'Why Should I Trust You?' Explaining the Predictions of Any Classifier," ACM SIGKDD, pp.1135–1144, 2016.
[34]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," ACM SIGKDD, pp.785–794, 2016.
[35]. H. B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," AISTATS, vol.54, 2017.
[36]. M. Nalinipriya, S. Ramasree, K. Radhika, and E. L. Lydia, "Leveraging XAI for Early Detection and Mitigation of Cyber Threat in Large-Scale Networks," Scientific Reports, vol.15, Springer Nature, 2025.
[37]. A. K. B. S. and R. Achary, "IDS in Cybersecurity: Explainable Graphic Reinforcement Learning," IEEE ICICI, 2025.
[38]. N. Phutela and Anju, "Advanced ML Methodologies for Intelligent Cybersecurity IDS," 2025.
[39]. R. Basry et al., "Cyber Defense Through Machine Intelligence: IDS and Prevention," Cybersecurity, 2025.
[40]. V. R. Joshi et al., "Hybrid AI IDS: Balancing Accuracy and Efficiency," Sensors (MDPI), 2025.
[41]. K. M. Olagunju et al., "State-of-the-Art Research in IDS for Network Attack Detection: A Review," IEEE Access, 2025.
[42]. R. Taheri et al., "XAI for Federated Learning-Based IDS in Connected Vehicles," Electronics (MDPI), 2025.
[43]. S. A. H. Moaman et al., "AI in Malware and NID: A Comprehensive Survey," 2025.

DOI: 10.17148/IJARCCE.2026.15802

[44]. S. Ankalaki et al., "Cyber Attack Prediction: From Traditional ML to Generative AI," IEEE Access, 2025.
[45]. M. Tserenkhuu et al., "IDS for SDN-Based IoT Using DL with XAI-Based Feature Selection," IEEE Trans. Ind. Inform., 2025.
[46]. S. Shahid, "XAI for IDS in Water Distribution Systems: WADI Dataset," TechRxiv, 2025.
[47]. X. Liu et al., "From Black Box to Glass Box: A Practical Review of XAI," AI (MDPI), 2025.
[48]. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges," Information Fusion, vol.58, pp.82–115, 2020.
[49]. S. I. Amgbara, C. Akwiwu-Uzoma, and O. David, "Exploring Lightweight ML Models for Personal IoT Security," arXiv, 2024.
[50]. R. Alanazi and A. Aljuhani, "Anomaly Detection for Industrial IoT Cyberattacks," Applied Sciences, 2023.
[51]. Y. Al Sawafi, A. Touzene, and R. Hedjam, "Hybrid DL-Based IDS for RPL IoT Network," Electronics (MDPI), 2021.
[52]. M. Li, Y. Qiao, and B. Lee, "Multi-View Fusion Based IDS with DL Architectures," IEEE Access, 2024.
[53]. S. A. Birahim et al., "IDS for WSN Using PSO-Based Explainable Ensemble ML," Sensors (MDPI), 2025.
[54]. M. H. Behiry and M. Aly, "Cyberattack Detection in WSN Using Hybrid Feature Reduction," IEEE Access, 2024.
[55]. A. Hafid, M. Rahouti, and M. Aledhari, "Optimizing IDS in IoMT Through Interpretable Cost-Aware ML," IEEE Trans. Netw. Serv. Manage., 2025.
[56]. D. R. Kumar and P. V. Sudha, "Real-Time DDoS Detection in SDN Using Liquid Neural Networks," J. Netw. Comput. Appl., 2026.
[57]. L. Sana et al., "Securing IoT: Intrusion Anomaly Detection with Vision Transformers," Sensors (MDPI), 2025.
[58]. P. Turaka and S. K. Panigrahy, "Dynamic Attack Detection in IoT: Ensemble+Q-Learning+XAI," IEEE IoT J., 2024.
[59]. A. Alqazzaz, "Explainable Multi-Head Attention for Healthcare IoT (MedDefender-MHAN)," IEEE Access, 2026.
[60]. R. Vinayakumar et al., "Deep Learning Approach for Intelligent IDS," IEEE Access, vol.7, pp.41525–41550, 2019.
[61]. E. C. P. Neto et al., "CICIoT2023: A Real-Time Dataset for Large-Scale IoT Attacks," Sensors (MDPI), vol.23, no.13, p.5941, 2023.
[62]. T. Li et al., "Federated Optimization in Heterogeneous Networks," Proc. MLSys, 2020.
[63]. M. Abadi et al., "Deep Learning with Differential Privacy," Proc. 23rd ACM CCS, pp.308–318, 2016.
[64]. N. V. Chawla et al., "SMOTE: Synthetic Minority Over-Sampling Technique," JAIR, vol.16, pp.321–357, 2002.
[65]. G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," NeurIPS, vol.30, 2017.