FEDERATED RISK LEARNING FOR ADAPTIVE POST-QUANTUM IOT AUTHENTICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Smart-home Internet of Things (IoT) security requires behavioral adaptation without exposing household traffic or weakening authentication during post-quantum migration. Existing work advances private learning, robust aggregation, and cryptographic authentication, but leaves open how uncertain device behavior should guide a certified key lifecycle. This work presents FALCON-SDIoT, a software-defined networking architecture that connects federated risk classification to authenticated rekeying. Manufacturer profiles guide initial classification, clipped and noised edge updates support collaborative learning, and temporal trust determines asymmetric hardening and relaxation. Certified key establishment and nonce-bound proof-of-possession bind these decisions to device credentials. This work integrates federated risk assignment, credential binding, and coordinated class and key transitions under explicit privacy and security assumptions. Evaluation covers a controlled synthetic benchmark, public-traffic temporal and complete-device holdouts, repeated-seed federated tests, and an adaptive-policy sensitivity study. Results support manufacturer-informed risk classification, characterize performance under adversarial updates, and demonstrate the trade-off between detection delay and false alarms. Public-traffic tests assess temporal classification and unseen-device detection. These findings support the component interfaces under the evaluated settings; cryptographic costs remain accounting estimates, and the recurring proof is assessed within a classical random-oracle scope. The architecture supports studying behavior-driven authentication while distinguishing predictive confidence from cryptographic assurance.

1 INTRODUCTION

Smart-home Internet of Things (IoT) devices require continued authentication as traffic patterns and cryptographic requirements change. Lifecycle authentication supports ongoing use, while traffic fingerprinting exposes household metadata (Chen et al., 2023; Tan et al., 2024). Post-quantum standards and private, robust federated learning address different aspects (National Institute of Standards and Technology, 2024a;b; Allouah et al., 2025; Parsa et al., 2026). Connecting them requires rules for turning uncertain behavioral evidence into authenticated policy changes.

How can collaborative risk learning guide authentication without confusing behavioral confidence with credential assurance? FALCON-SDIoT is a software-defined networking (SDN) architecture linking risk assignment, credentials, and key transitions. Related quantum-resilient anomaly detection connects learning and trust (Rehman et al., 2026); this work links these functions to a credential-bound key lifecycle.

Manufacturer Usage Description (MUD) profiles inform initial risk, and protected federated updates keep raw traffic at gateways. Proofs bound to certified keys establish possession independently of classifier confidence. Smoothed anomaly evidence triggers hardening and delayed relaxation, while authenticated installation commits the candidate class and transport key together. Risk predictions guide enforcement but do not provide cryptographic guarantees.

Evaluation combines synthetic risk classification, public-traffic temporal and complete-device holdouts, repeated-seed adversarial tests, and policy sensitivity analysis. Protocol evidence and a con-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ditional theorem address credential binding; an accounting model estimates recurring costs. These studies assess components and guide deployment validation.

This work links risk learning, credential verification, and adaptive authentication through:

- Manufacturer-informed federated risk with privacy and robustness analysis (Section 3.3).
- Credential-bound delivery with a conditional possession theorem (Section 3.4).
- Asymmetric trust, joint class/key commitment, and conservative recovery (Section 3.5).

2 Related Work

Three gaps organize the discussion: collaborative risk assignment (Section 2.1), credential binding (Section 2.2), and consistent temporal transitions (Section 2.3).

2.1 Collaborative Risk

Traffic classification and IoT Sentinel connect device identity to enforcement (Sivanathan et al., 2019; Miettinen et al., 2017). MUD specifies intended network use (Lear et al., 2019); translating profiles into authentication risk priors requires an explicit policy.

Federated averaging supports training on local data (McMahan et al., 2017), but traffic locality alone does not prevent information leakage through updates. Time-adaptive privacy spending studies budget allocation across clients and rounds (Kianidehkordi et al., 2025).

Adaptive Robust Clipping and learnable aggregation weights address Byzantine learning (Allouah et al., 2025; Parsa et al., 2026). This work connects manufacturer-profile initialization and robust selection to credential-bound temporal policy under release assumptions (Sections 3.3, 3.4, and 3.5).

2.2 Credential Binding

Lifecycle authentication covers enrollment, operation, and retirement (Chen et al., 2023). Lightweight and quantum-resistant authentication address credential exchange under resource constraints (Bai et al., 2023; Wang et al., 2023). These capabilities alone do not specify how learned risk should affect credential management.

The Module-Lattice-Based Key-Encapsulation Mechanism (ML-KEM) and Module-Lattice-Based Digital Signature Algorithm (ML-DSA) provide standardized post-quantum building blocks (National Institute of Standards and Technology, 2024a;b). Their integration requires authenticated bindings among endpoint identity, proof keys, and transport sessions.

Lattice identification and Fiat-Shamir techniques support bounded-witness proofs (Lyubashevsky, 2012; Fiat & Shamir, 1987); classical random-oracle arguments do not establish quantum random-oracle security automatically (Don et al., 2019). This work separates proof assumptions from prediction and connects certified secret possession to class and key state (Sections 3.3, 3.4, and 3.5).

2.3 Temporal Transitions

Isolation Forest measures behavioral deviation (Liu et al., 2008), while traffic fingerprinting highlights information exposure through device metadata (Tan et al., 2024). An isolated anomaly establishes neither credential compromise nor sustained recovery.

CALIS uses controller-side gradient boosting for cipher selection and key rotation with classical key establishment (Gilliard & Liu, 2026). Quantum-resilient federated anomaly detection also connects learning and trust (Rehman et al., 2026). These approaches motivate explicit rules linking observed behavioral changes to consistent authentication policy throughout device operation.

The transition problem is deciding when behavior should change policy and when the candidate class can safely become active. This work couples asymmetric trust thresholds to authenticated installation, with joint commitment designed to preserve class and key consistency across risk assignment, credential verification, and renewal (Sections 3.3, 3.4, and 3.5).

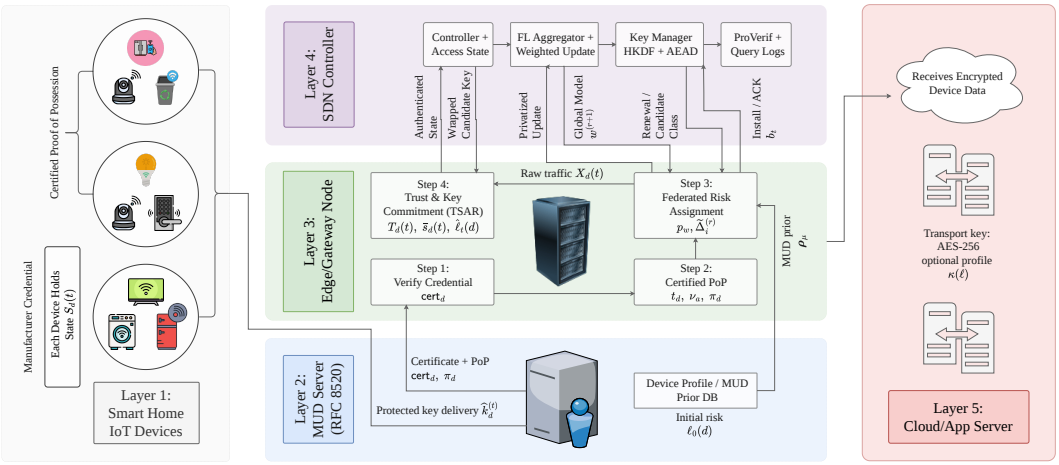


Figure 1: System roles in FALCON-SDIoT. Raw traffic stays at the serving gateway; model updates reach the aggregator, while certified identity and class state guide controller-issued keys.

3 Risk-Guided Authentication

Preliminaries (Section 3.1) define the managed device state used to formulate three operational requirements (Section 3.2). Risk assignment (Section 3.3), credential binding (Section 3.4), and trust-driven commitment (Section 3.5) implement the corresponding interfaces.

3.1 Preliminaries

Figure 1 shows device d, gateway e, and SDN controller c. A MUD server supplies profiles; the controller coordinates learning and access policy.

The manufacturer certifies credentials; the trusted controller maintains policy and keys. The trusted serving gateway measures traffic, infers risk, verifies credentials, and forwards keys; raw traffic remains local. The controller-to-gateway channel is authenticated and confidential. An active network adversary may inject, alter, delay, or replay messages. A separate Byzantine adversary controls updates from up to f gateways; credential assurance assumes an uncompromised gateway.

Let cert_d denote the manufacturer-signed device certificate. Device state is

$$S_d(t) = \big(\mathsf{cert}_d, t_d, \ell_t(d), T_d(t), h_d(t), k_d^{(t)}\big). \tag{1}$$

Here t indexes windows, t_d is the certified public proof key, $\ell_t(d) \in \{A, B, C\}$ is the risk class, $T_d(t) \in [0, 1]$ is trust, $h_d(t)$ counts consecutive high-trust windows since the last policy reset, and $k_d^{(t)}$ is the active transport key. Class A is the most conservative and class C the least restrictive. Initialize $\ell_0(d)$ by Equation (5), $T_d(0) = 1$, $h_d(0) = 0$, and establish a key as in Section 3.4.

3.2 Problem Formulation

First, manufacturer information should initialize a common predictor that is refined using local traffic. Let D_i denote labeled traffic at gateway i, w the model parameters, and $\mathcal{H}$ the honest gateways. The empirical risk is

$$F(w) = \sum_{i \in \mathcal{H}} p_i \mathcal{L}_i(w), \qquad \mathcal{L}_i(w) = -\frac{1}{|D_i|} \sum_{(x,y) \in D_i} \log p_w(y \mid x). \tag{2}$$

Here x denotes traffic features, $y \in \{A, B, C\}$ the risk label, and $p_i = |D_i| / \sum_{j \in \mathcal{H}} |D_j|$. Public manufacturer profiles initialize w before federated optimization (Section 3.3). Equation (2) defines the honest-data target; the server does not know $\mathcal{H}$, and convergence requires separate analysis.

Second, certified secret possession must remain distinct from predicted risk. For the bounded-witness relation $\mathcal{R}_{pp}$ defined by the public parameters pp in Equation (13), authentication requires knowledge of a witness satisfying

$$(t_d, s_d) \in \mathcal{R}_{pp}, \qquad (ID_d, t_d) \text{ bound by a verified } \mathsf{cert}_d. \tag{3}$$

Here ID_d denotes device identity and s_d the secret witness. This requirement applies to a fresh session independently of the predicted class; Theorem 1 states its conditional guarantee.

Third, candidate class $\widehat{\ell}_t(d)$ and key $\widehat{k}_d^{(t)}$ must transition consistently. On renewal, let $b_t = 1$ denote authenticated confirmation that the candidate pair is installed for the current session; otherwise $b_t = 0$. The controller updates its committed state by

$$(\ell_t(d), k_d^{(t)}) = \begin{cases} (\widehat{\ell}_t(d), \widehat{k}_d^{(t)}), & b_t = 1, \\ (\ell_{t-1}(d), k_d^{(t-1)}), & b_t = 0. \end{cases} \tag{4}$$

Pending installations are tracked separately by session identifier. Failure or an uncertain installation outcome suspends service until authenticated recovery reconciles endpoint state.

These requirements address *collaborative risk-to-policy assignment* (Section 3.3), *behavior-to-credential binding* (Section 3.4), and *stable class and key transitions* (Section 3.5).

3.3 Federated Risk Assignment

Let x_d^0 denote bootstrap traffic features and μ_d the certified device type. MUD describes intended network use (Lear et al., 2019). Public prototypes derived from manufacturer profiles initialize model parameters before federated training. The resulting classifier assigns initial risk by

$$\ell_0(d) = \operatorname*{arg\,max}_{\ell \in \{A,B,C\}} p_w(\ell \mid x_d^0). \tag{5}$$

Here p_w denotes the classifier with parameters w. Manufacturer information enters through initialization rather than numerical prior fusion. The certified type μ_d remains part of the credential binding in Equation (10). Section 4.1 specifies initialization, while Section 3.5 governs transitions.

Device identity and A/B/C risk are distinct targets. Unrecognized or low-confidence devices remain quarantined until validated. Public holdouts test unseen-device detection, not complete admission.

At round r, gateway i computes update $\Delta_i^{(r)}$ from local data D_i and global model $w^{(r)}$. For clipping bound $C > 0$, define $\mathrm{clip}_C(v) = v/\max\{1, \|v\|_2/C\}$. An honest gateway transmits

$$\widetilde{\Delta}_i^{(r)} = \mathrm{clip}_C(\Delta_i^{(r)}) + z_i^{(r)}, \qquad z_i^{(r)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I), \tag{6}$$

where $\sigma > 0$ is the noise standard deviation. Byzantine updates may be arbitrary; Section 4.1 specifies the evaluated perturbation.

Let N denote participating gateways and f the Byzantine budget. Multi-Krum-based selection scores distances to the $N - f - 2$ nearest other updates, indexed by $\mathcal{B}_i^{(r)}$ (Blanchard et al., 2017):

$$\chi_i^{(r)} = \sum_{j \in \mathcal{B}_i^{(r)}} \|\widetilde{\Delta}_i^{(r)} - \widetilde{\Delta}_j^{(r)}\|_2^2. \tag{7}$$

Let $\mathcal{S}^{(r)}$ contain the $m = N - f - 2$ lowest-scoring updates, with ties resolved by a fixed ordering. The server applies

$$w^{(r+1)} = w^{(r)} + \eta \frac{\sum_{i \in \mathcal{S}^{(r)}} |D_i| \widetilde{\Delta}_i^{(r)}}{\sum_{i \in \mathcal{S}^{(r)}} |D_i|}, \tag{8}$$

where $\eta > 0$ is the server step size. Rounds require $N \geq 2f + 3$; otherwise, aggregation is suspended. Robustness additionally depends on sample-count integrity, heterogeneity, and noise.

Add-or-remove adjacency compares a clipped update with zero in a fixed release slot, with Gaussian noise applied in both cases. Sensitivity is C; replacement is at most $2C$. For R Gaussian releases, Rényi order $a > 1$, and $0 < \delta < 1$, differential privacy (DP) accounting gives (Mironov, 2017)

$$\varepsilon(R, \delta) \leq \frac{RaC^2}{2\sigma^2} + \frac{\log(1/\delta)}{a - 1}. \tag{9}$$

This fixed-mechanism bound excludes subsampling amplification and leaves participation visible. Selection and aggregation preserve an established joint-release guarantee when they use only privatized updates and public, data-independent weights.

3.4 CREDENTIAL-BOUND LIFECYCLE

The manufacturer signs identity ID_d, media access control address MAC_d, ML-DSA verification key vpk_d, proof key t_d, type μ_d, and expiry $\tau_d^{\exp}$:

$$\mathsf{cert}_d = \mathsf{Sign}_{\mathsf{msk}}(ID_d, MAC_d, \mathsf{vpk}_d, t_d, \mu_d, \tau_d^{\exp}), \tag{10}$$

where msk is the manufacturer signing key. ML-KEM-768 establishes secrets using ephemeral key pairs; ML-DSA-65 authenticates certificates and ephemeral public keys through trusted endpoint credentials. For endpoints u, v, ciphertext c_{uv}, shared secret K_{uv}, fresh protocol nonce ν_k, and direction label L_{uv}, define

$$k_{uv}^{w} = \mathsf{HKDF}(K_{uv};\, ID_u \| ID_v \| c_{uv} \| \nu_k \| L_{uv}). \tag{11}$$

The semicolon separates secret input from public context. HKDF provides HMAC-based extract-and-expand key derivation; HMAC denotes hash-based message authentication code (Krawczyk & Eronen, 2010). Equation (11) requires hash, extraction salt, output length, and canonical encoding choices. Authenticated encryption with associated data (AEAD) wraps transport keys under k_{uv}^{w}, binding protocol nonce, ciphertext, identity, candidate class, and session identifier. AEAD nonces differ from ν_k and never repeat per wrapping key; direction labels distinguish messages.

The gateway validates credentials and freshness, applies Equation (5), and forwards authenticated state. The controller returns a transport key bound to identity, class, and session. The gateway sees and forwards it over the protected device channel. Session identifiers bind requests and responses. Enrollment establishes the initial key. Candidate class changes or renewal requests trigger

$$\widehat{k}_d^{(t)} \xleftarrow{\$} \{0,1\}^{\kappa(\widehat{\ell}_t(d))} \quad \text{if} \quad \widehat{\ell}_t(d) \neq \ell_{t-1}(d) \ \vee\ \mathsf{renew}(d) = 1, \tag{12}$$

where $\kappa(\ell)$ specifies key length and $\mathsf{renew}(d)$ is the renewal flag. Authenticated installation commits the candidate pair through Equation (4). Uniform Advanced Encryption Standard (AES)-256 preserves the idealized 128-bit quantum-search margin. The optional $\kappa(A/B/C) = 256/192/128$-bit profile requires separate utility evaluation.

Recurring authentication proves possession of a bounded witness for the certified key:

$$R_q = \mathbb{Z}_q[X]/(X^n + 1), \qquad t_d = A s_d \pmod q, \qquad \|s_d\|_\infty \le \eta_s, \tag{13}$$

where $n = 256$, $q = 8{,}380{,}417$, $A \in R_q^{k \times r_s}$ is public, $t_d \in R_q^k$, and $s_d \in R_q^{r_s}$ is secret. The bound η_s applies to centered coefficients modulo q. Dimensions k, r_s, bound η_s, and the transcript construction parameterize the proof interface. Certificate verification fixes t_d; the evaluation does not claim a measured native proof instantiation.

The verifier registers a fresh random λ-bit nonce ν_a for the device and session in outstanding set $\mathcal{N}_v$. Here λ is the security parameter. The device constructs its proof locally using s_d. For device timestamp τ_d and verifier time τ_v, acceptance requires

$$\begin{aligned} \mathsf{st}_d &= (ID_d, \nu_a, \tau_d), \qquad \pi_d \leftarrow \mathsf{Prove}(pp, t_d, s_d, \mathsf{st}_d), \\ \mathsf{accept} &\iff \mathsf{certOK}_d \wedge \mathsf{Verify}(pp, t_d, \mathsf{st}_d, \pi_d) = 1 \wedge |\tau_d - \tau_v| \le \Delta_\tau \wedge \nu_a \in \mathcal{N}_v. \end{aligned} \tag{14}$$

Here pp contains proof parameters, Δ_τ bounds clock skew, and certOK_d denotes signature, identity, and expiry checks. Nonce membership is checked and consumed atomically on acceptance. Failed certificate or proof checks deny service and retain prior state.

Theorem 1 (Certified proof of possession)**.** *Assume unforgeable certification of (ID_d, t_d) and a complete, knowledge-sound Fiat-Shamir proof for Equation* (13) *in the classical random-oracle model with negligible knowledge error. The verifier binds the full statement and atomically checks and consumes fresh random λ-bit nonces. Any classical probabilistic polynomial-time prover accepted non-negligibly admits bounded-witness extraction for certified t_d, except with negligible certificate-forgery probability. Replay fails except with negligible nonce-collision probability; honest acceptance fails only with negligible proof error when certificate and freshness checks pass.*

Proof sketch. Unforgeable certification fixes t_d; knowledge soundness extracts a valid, possibly nonunique witness. Atomic nonce consumption rejects replay unless a fresh challenge collides, which occurs negligibly over polynomially many sessions. Completeness gives honest acceptance when credential and freshness checks pass. □

Theorem 1 assumes knowledge soundness of the instantiated proof, beyond the lattice relation (Lyubashevsky, 2012; Fiat & Shamir, 1987). Zero knowledge, live-relay resistance, and quantum random-oracle security require separate constructions and arguments (Don et al., 2019).

3.5 Trust and Key Commitment

After authentication, the gateway extracts features $x_d(t)$: mean inter-arrival time, byte rate, protocol proportions, destination-port entropy, connection-attempt rate, and Transport Layer Security (TLS) usage ratio. An Isolation Forest (IF) produces normalized anomaly evidence $\bar{s}_d(t) \in [0,1]$, increasing with abnormality (Liu et al., 2008). Trust evolves as

$$T_d(t) = \alpha T_d(t-1) + (1-\alpha)(1-\bar{s}_d(t)), \qquad 0 \le \alpha \le 1, \quad T_d(0) = 1. \tag{15}$$

This convex update preserves $T_d(t) \in [0,1]$.

Let $W_s, W_h > 0$ denote monitoring-window and high-trust hold durations, with thresholds $0 \le \theta_h < \theta_r \le 1$. For $\ell = \ell_{t-1}(d)$ and $h = h_d(t-1)$, propose

$$(\widehat{\ell}_t(d), \widehat{h}_d(t)) = \begin{cases} (\mathsf{Hdn}(\ell), 0), & T_d(t) < \theta_h, \\ (\mathsf{Rlx}(\ell), 0), & T_d(t) > \theta_r \wedge (h+1)W_s \ge W_h, \\ (\ell, h+1), & T_d(t) > \theta_r \wedge (h+1)W_s < W_h, \\ (\ell, 0), & \text{otherwise.} \end{cases} \tag{16}$$

The one-step maps are $\mathsf{Hdn}(A/B/C) = A/A/B$ and $\mathsf{Rlx}(A/B/C) = B/C/C$. Policy parameters are $\Theta = (\alpha, \theta_h, \theta_r, W_s, W_h)$. Smoothing can delay hardening; relaxation requires consecutive high-trust windows. At class A, further renewal requires an independent request.

Without class changes or renewal requests, retain the class and key and commit $h_d(t) = \widehat{h}_d(t)$. Otherwise, authenticate ephemeral keys, apply Equation (11), and deliver the candidate key from Equation (12) bound to class and session. Endpoints retain the old key pending installation.

Authenticated installation commits the pair (Equation (4)) and $h_d(t) = \widehat{h}_d(t)$, retiring old keys. Failure retains updated trust and the prior pair, resets $h_d(t) = 0$, and suspends service. Reconcile sessions and confirm installation before resumption. Erase ephemeral secrets and discarded keys upon completion/abort. Evaluation tests transitions, excluding distributed-commitment faults.

4 Empirical Analysis

The setup (Section 4.1) supports federated risk evaluation (Section 4.2), credential analysis (Section 4.3), and temporal-policy tests (Section 4.4). Ablations and validation scope (Section 4.5) clarify the evidence supporting each contribution.

4.1 Evaluation Setup

The synthetic benchmark contains $24{,}300$ samples across 27 device types, using an $80/20$ split and seed 42. Profiles draw on traffic characterization and IoT Sentinel (Sivanathan et al., 2019; Miettinen et al., 2017). Risk classes A/B/C contain $9{,}000/8{,}100/7{,}200$ samples from $10/9/8$ device types, each split $80/20$. Federated partitions use Dirichlet concentration 0.5, seeds 1 to 5, and $(N, f) \in \{(6,1), (7,2)\}$, satisfying $N \ge 2f + 3$.

Public manufacturer profiles initialize models. Five noisy public prototypes per profile initialize a three-class logistic model (regularization $C_{\mathrm{LR}} = 0.1$, seed 909) without private household samples. This evaluates profile initialization, not numerical prior fusion.

Public smart-home capture contains $933{,}833$ raw flows and $450{,}507$ eligible IoT flows (Sivanathan et al., 2019). Sixteen supported devices comprise twelve known and four complete-device holdouts.

Training, calibration, temporal testing, and unseen-device scoring use 165,612, 12,306, 32,122, and 24,380 rows, respectively. Chronological partitioning, holdouts, and per-device caps exclude remaining flows; the reproducibility artifact records split counts.

Federated training uses logistic-loss stochastic gradient descent (SGD), regularization coefficient 10^{-4}, local learning rate 0.01, one local pass per round, ten rounds, and server step size $\eta = 0.02$. Updates are clipped at $C = 1$. The Byzantine stress test replaces the final f updates with Gaussian vectors of standard deviation 3 before selection. The fixed Gaussian accountant uses $\sigma = 4.03734$, Rényi order $a = 10$, and $\delta = 10^{-5}$. Policy tests use $\alpha = 0.85$ and five scripted seeds.

4.2 Federated Risk Evidence

Table 1 distinguishes 27-device identification from A/B/C risk assignment in Equation (5). Stacking and Naive Bayes perform similarly on the identity split. At round ten, manufacturer-profile initialization improves mean clean risk accuracy by 0.1353 over the no-profile condition, supporting its use in the learning pipeline of Equation (2).

Table 1: Synthetic classification results. Risk entries are mean ± standard deviation over 5 seeds at round 10, with macro-F1. Device entries use one split.

Method	Target	Accuracy	F1
Stacking	27 devices	0.9086	0.9084
Naive Bayes	27 devices	0.9080	0.9079
Noisy + MUD profile, clean	$A/B/C$	0.9158 ± 0.0154	0.9182 ± 0.0159
Noisy + MUD profile, $f = 1, N = 6$	$A/B/C$	0.8823 ± 0.0444	0.8831 ± 0.0466
Noisy + MUD profile, $f = 2, N = 7$	$A/B/C$	0.9224 ± 0.0212	0.9250 ± 0.0214
Noisy, no MUD profile	$A/B/C$	0.7805 ± 0.1008	0.7364 ± 0.1481

Changing participant counts and selection in Equation (8) confounds comparisons across f; standard deviations describe seed variability. The results characterize non-adaptive Gaussian-update attacks, without establishing robustness to adaptive or aggregation-aware poisoning. Figure 2 summarizes identification and federated risk accuracy. Equation (9) gives fixed-mechanism bounds of approximately $\varepsilon = 4.3467$ at ten releases and 31.9539 at 100. This growth motivates finite training horizons or adaptive spending (Kianidehkordi et al., 2025). The accounting claim excludes participation metadata, private sample counts, and the evaluated Multi-Krum pipeline.

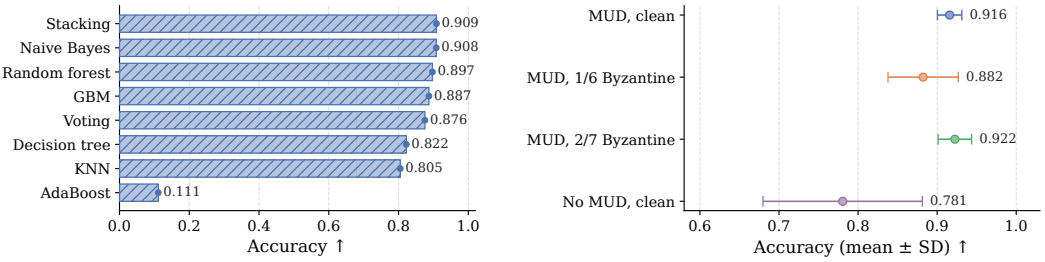


Figure 2: Synthetic classification: (a) device identification; (b) federated A/B/C accuracy at round 10, with one-standard-deviation bars over 5 seeds.

On 32,122 temporal-test rows, known-device accuracy is 0.9647, balanced accuracy is 0.9433, and macro-F1 is 0.9482 across nine supported classes. Including three unsupported known labels in a fixed twelve-label average gives macro-F1 0.7112. Figure 3 presents these identity results, distinct from A/B/C risk prediction. Confidence-based out-of-distribution (OOD) detection achieves area under the receiver operating characteristic curve (AUROC) 0.9564. The threshold 0.8970 is the fifth percentile of maximum known-device confidence on calibration data and remains fixed during testing. It rejects 92.19% of 24,380 unseen rows and 9.14% of known temporal-test rows.

Eight supported classes exceed 96% recall; Dropcam reaches 56.4%, with approximately 29% classified as Belkin Wemo. Between 11 and 12 October 2016, accuracy increases from 0.9569 to 0.9913

across 24,843 and 7,279 rows, while fixed-twelve-label macro-F1 decreases from 0.7072 to 0.5627. These changes indicate temporal variation; confirming firmware drift requires version annotations.

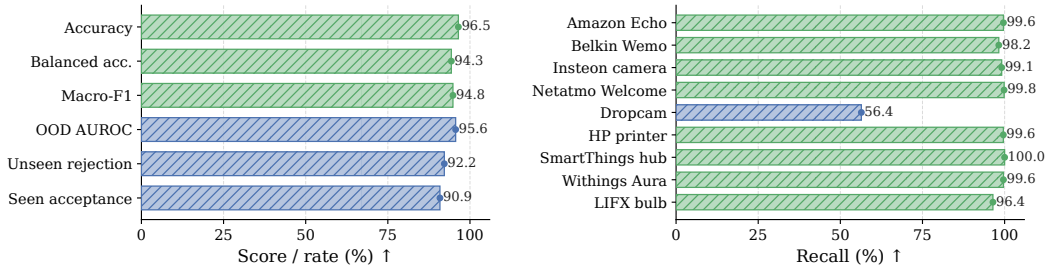


Figure 3: Public-traffic evaluation: (a) temporal classification and unseen-device detection; (b) recall across nine supported classes. Classification and OOD metrics describe separate tasks.

4.3 Credential Lifecycle Evidence

The protocol defines six ProVerif verification targets: key secrecy, controller-gateway agreement, device-controller agreement, forward secrecy after signing-key disclosure, and proof replay and impersonation resistance (Blanchet, 2009). The symbolic model assumes ideal cryptography and a private controller-to-gateway channel. These are verification targets. Forward secrecy concerns signing-key compromise after ephemeral-secret erasure; active-key confidentiality assumes trusted gateways and controllers. Theorem 1 addresses Equation (3) under classical assumptions. The proof is parameterized; native verification costs remain unmeasured.

Recurring energy is $E_{\text{total}} = E_{\text{crypto}} + E_{\text{radio}} + E_{\text{edge}} + E_{\text{controller}} + E_{\text{idle}}$, covering device cryptography, radio transfer, gateway and controller computation, and idle operation. The model estimates 10.69 ms latency, 123.41 mJ energy, and a 720-byte proof-session payload. These estimates exclude onboarding certificates, ML-KEM setup, training, and long-term monitoring. Figure 4 compares this model with six schemes using literature-normalized inputs (Raza et al., 2024; Iqbal et al., 2022; Chen et al., 2023; Bai et al., 2023; Gupta, 2024; Wang et al., 2023). The comparison supports cost planning, not measured rankings on matched hardware.

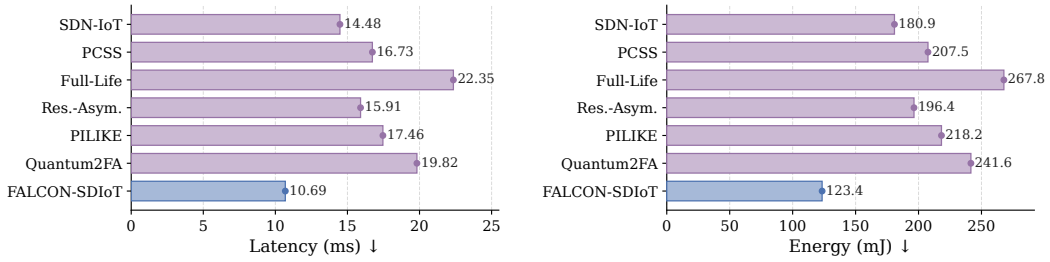


Figure 4: Literature-normalized estimates of (a) recurring latency and (b) session energy. Onboarding, training, and long-term monitoring are excluded.

Fragmentation and retries add communication overhead. For the 720-byte payload, sensitivity analysis assumes a 127-byte IEEE 802.15.4 physical-layer frame and 75 usable bytes per fragment. This is an analytical model, not a radio measurement.

4.4 Temporal Policy Evidence

With $(\theta_h, \theta_r, W_s, W_h) = (0.55, 0.90, 15, 60)$, scripted attacks yield mean true-positive rate (TPR) 1.00, false-positive rate (FPR) 0.00, and detection delay of 60 minutes. These results evaluate smoothing and candidate transitions in Equations (15) and (16). Figure 5 illustrates trust adaptation under a scripted attack and an intermittent event. Its zero FPR is specific to this scenario.

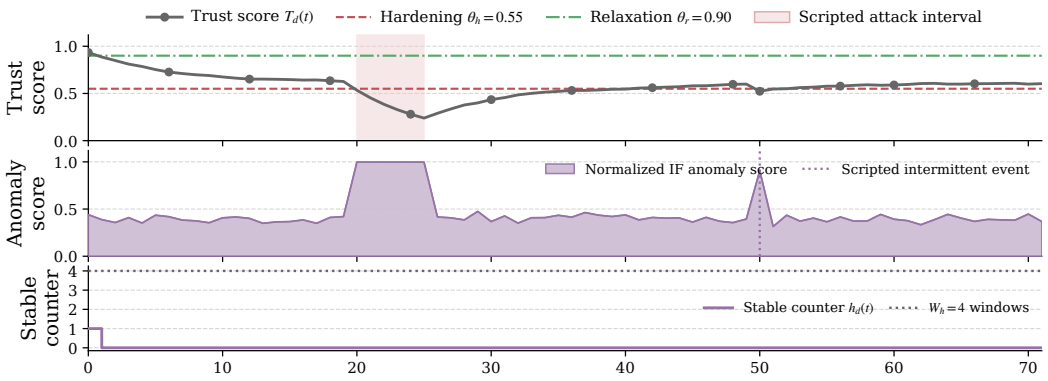


Figure 5: Scripted trust-score adaptive reclassification and rekeying (TSAR). The stability counter represents $h_d(t)$; window zero is the first post-update observation.

The trace evaluates policy decisions rather than distributed commitment. Validating Equation (4) under acknowledgment loss, concurrent renewal, and recovery requires separate transaction tests.

4.5 Ablation and Validation Scope

Table 1 ablates manufacturer-profile initialization. Independent ablations of clipping, noise, robust selection, and alternative prior fusion would clarify individual contributions. Matched adaptive-clipping and learnable-weight baselines would extend this comparison (Allouah et al., 2025; Parsa et al., 2026). The policy grid evaluates 81 combinations: $\theta_h \in \{0.45, 0.55, 0.65\}$, $\theta_r \in \{0.85, 0.90, 0.95\}$, $W_s \in \{5, 15, 30\}$ minutes, and $W_h \in \{30, 60, 120\}$ minutes. Table 2 reports representative five-seed means. Shorter windows reduce observed delays; higher hardening thresholds can increase false alarms. Equations (15) and (16) explain timing and threshold effects.

Table 2: Policy sensitivity over 5 seeds with $\theta_r = 0.90$ and $W_h = 60$ minutes.

Hardening θ_h	Window W_s (min)	Attack TPR	Benign FPR	Delay (min)
0.55	15	1.000	0.000	60.0
0.55	5	1.000	0.000	20.0
0.65	5	1.000	0.400	14.7
0.45	30	0.333	0.000	150.0
0.65	30	1.000	0.000	90.0

The reproducibility artifact records seeds, model settings, profile initialization, attacks, privacy parameters, chronological splits, and OOD calibration. It exports per-round and per-seed results, public-data and policy metrics, environment information, and execution status for post-quantum and ProVerif checks. Future validation can combine public-traffic risk labels, matched-hardware costs, adaptive attacks, and failed-renewal tests. Extensions include complete privacy accounting for the release pipeline in Section 3.3, quantum-secure proof instantiation, and relay protection.

5 Conclusion

This work integrates manufacturer-informed federated risk assignment, a credential-bound authentication lifecycle, and asymmetric trust adaptation with coordinated class and key commitment for managed smart-home devices. The conditional theorem establishes bounded-witness possession and single-use freshness under classical assumptions, independently of predictive confidence.

Classification, unseen-device holdouts, and policy sensitivity tests characterize performance and trade-offs between detection delay and false alarms. Concrete proof instantiation, matched-hardware measurements, and transaction-fault testing would extend evidence toward deployment.

AI USE STATEMENT

ChatGPT supported literature verification, grammar checks, writing refinement, consistency checks, and LaTeX editing. The assistance focused on clarity and consistency. The authors retain responsibility for verifying the final text, technical claims, results, and artifacts.

REFERENCES

Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Ahmed Jellouli, Geovani Rizk, and John Stephan. Adaptive gradient clipping for robust federated learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d1d3cdc9e28b0c67b9df90fca4d1c1b3-Abstract-Conference.html.

Linyan Bai, Chingfang Hsu, Lein Harn, Jianqun Cui, and Zhuo Zhao. A practical lightweight anonymous authentication and key establishment scheme for resource-asymmetric smart environments. *IEEE Transactions on Dependable and Secure Computing*, 20(4):3535–3545, Jul./Aug. 2023. doi: 10.1109/TDSC.2022.3203874.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bruno Blanchet. Automatic verification of correspondences for security protocols. *Journal of Computer Security*, 17(4):363–434, 2009. doi: 10.3233/JCS-2009-0339.

Fei Chen, Zixing Xiao, Tao Xiang, Junfeng Fan, and Hong-Linh Truong. A full lifecycle authentication scheme for large-scale smart IoT applications. *IEEE Transactions on Dependable and Secure Computing*, 20(3):2221–2237, May/Jun. 2023. doi: 10.1109/TDSC.2022.3178115.

Jelle Don, Serge Fehr, Christian Majenz, and Christian Schaffner. Security of the Fiat–Shamir transformation in the quantum random-oracle model. In *Advances in Cryptology—CRYPTO 2019*, volume 11693 of *Lecture Notes in Computer Science*, pp. 356–383. Springer, 2019. doi: 10.1007/978-3-030-26951-7_13.

Amos Fiat and Adi Shamir. How to prove yourself: Practical solutions to identification and signature problems. In *Advances in Cryptology, CRYPTO 1986*, volume 263 of *Lecture Notes in Computer Science*, pp. 186–194. Springer, 1987. doi: 10.1007/3-540-47721-7_12.

Ezekia Gilliard and Jinshuo Liu. CALIS: AI-driven context-aware encryption for SDN-enabled smart-home IoT. *Journal of King Saud University Computer and Information Sciences*, 38:410, 2026. doi: 10.1007/s44443-025-00404-9.

Daya Sagar Gupta. PILIKE: Post-quantum identity-based lightweight authenticated key exchange protocol for IIoT environments. *IEEE Systems Journal*, 18(1):15–23, 2024. doi: 10.1109/JSYST.2023.3335217.

Waseem Iqbal, Haider Abbas, Bilal Rauf, Yawar Abbas Bangash, Muhammad Faisal Amjad, and Ahmed Hemani. PCSS: Privacy preserving communication scheme for SDN enabled smart homes. *IEEE Sensors Journal*, 22(18):17677–17690, 2022. doi: 10.1109/JSEN.2021.3087779.

Shahrzad Kianidehkordi, Nupur Kulkarni, Adam Dziedzic, Stark Draper, and Franziska Boenisch. Differentially private federated learning with time-adaptive privacy spending. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d45bb0586164aa7c2289b9e1abee4405-Abstract-Conference.html.

Hugo Krawczyk and Pasi Eronen. HMAC-based extract-and-expand key derivation function (HKDF). Technical Report RFC 5869, Internet Engineering Task Force, 2010. URL https://www.rfc-editor.org/rfc/rfc5869.

Eliot Lear, Ralph Droms, and Dan Romascanu. RFC 8520: Manufacturer usage description specification. Technical report, IETF, March 2019. RFC 8520.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proc. IEEE 8th International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Vadim Lyubashevsky. Lattice signatures without trapdoors. In *Advances in Cryptology, EUROCRYPT 2012*, volume 7237 of *Lecture Notes in Computer Science*, pp. 738–755. Springer, 2012. doi: 10.1007/978-3-642-29011-4_43.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proc. Machine Learning Research*, pp. 1273–1282, 2017.

Markus Miettinen, Samuel Marchal, Ibbad Hafeez, N. Asokan, Ahmad-Reza Sadeghi, and Sasu Tarkoma. IoT sentinel: Automated device-type identification for security enforcement in IoT. In *Proc. IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2177–2184, 2017. doi: 10.1109/ICDCS.2017.283.

Ilya Mironov. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium*, pp. 263–275, 2017. doi: 10.1109/CSF.2017.11.

National Institute of Standards and Technology. FIPS 203: Module-lattice-based key-encapsulation mechanism standard. Technical report, NIST, Gaithersburg, MD, USA, August 2024a. Federal Information Processing Standards Publication 203.

National Institute of Standards and Technology. FIPS 204: Module-lattice-based digital signature standard. Technical report, NIST, Gaithersburg, MD, USA, August 2024b. Federal Information Processing Standards Publication 204.

Javad Parsa, Amir Daghestani, André M. H. Teixeira, and Mikael Johansson. Byzantine-robust federated learning with learnable aggregation weights. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/43a1c817602662be3a59015e9a86d0d1-Abstract-Conference.html.

Arif Raza, Salabat Khan, Shivanshu Shrivastava, Muhammad Wasim Abbas Ashraf, Tao Wang, Kui Wu, and Lin Wang. A lightweight group-based SDN-driven encryption protocol for smart home IoT devices. *Computer Networks*, 250:110537, 2024. doi: 10.1016/j.comnet.2024.110537.

Amjad Rehman, Tanzila Saba, Kamran Ahmad Awan, Sonia Khan, Shaha T. Al-Otaibi, and Abeer Rashad Mirdad. Quantum-resilient federated learning with trust-assisted adaptive anomaly detection for internet of things environment. *Engineering Applications of Artificial Intelligence*, 177:114908, 2026. doi: 10.1016/j.engappai.2026.114908.

Arunan Sivanathan, Daniel Sherratt, Hassan Habibi Gharakheili, Adam Radford, Chamith Wijenayake, Arun Vishwanath, and Vijay Sivaraman. Classifying IoT devices in smart environments using network traffic characteristics. *IEEE Transactions on Mobile Computing*, 18(8):1745–1759, 2019. doi: 10.1109/TMC.2018.2866249.

Shuaishuai Tan, Shui Yu, Wenyin Liu, Daojing He, and Sammy Chan. You can glimpse but you cannot identify: Protect IoT devices from being fingerprinted. *IEEE Transactions on Dependable and Secure Computing*, 21(3):1210–1223, May/Jun. 2024. doi: 10.1109/TDSC.2023.3275850.

Qingxuan Wang, Ding Wang, Chi Cheng, and Debiao He. Quantum2FA: Efficient quantum-resistant two-factor authentication scheme for mobile devices. *IEEE Transactions on Dependable and Secure Computing*, 20(1):193–208, Jan./Feb. 2023. doi: 10.1109/TDSC.2021.3129512.