FedCIPHER: Causal-Invariant and Privacy-Hardened Prototype Federation for Robust Network Intrusion Detection under Non-IID Traffic

Qingming Chen*
China Southern Power Grid Data Platform and Security (Guangdong) Co., Ltd., Guangzhou, China
*Corresponding author: Qingming Chen (e-mail:43651135@qq.com).

ABSTRACT Collaborative network intrusion detection requires security operation centers, cloud platforms, and industrial edge domains to learn from distributed traffic without centralizing packet traces, asset identifiers, or attack evidence. Federated learning limits direct data movement, yet conventional parameter averaging is fragile when attack classes are rare, traffic distributions are non-independent and identically distributed (non-IID), and model updates can be inspected or poisoned. This paper proposes FedCIPHER, a causal-invariant and privacy-hardened prototype federation framework for multi-class network threat detection. Each client learns a gated causal representation by exchanging nuisance components between traffic samples and enforcing counterfactual prediction consistency. Instead of uploading full model gradients, the client releases clipped, orthogonally sketched first- and second-order class prototypes protected by distributed Gaussian noise and secure summation. An evidential uncertainty score and a class-conditional deviation statistic jointly control robust server aggregation, while a private global prototype bank regularizes locally personalized detectors. We provide a Rényi differential privacy bound, a bounded-influence analysis for malicious clients, and a convergence characterization that explicitly separates client drift, sketch distortion, and privacy noise. Experiments on UNSW-NB15, CICIDS2017, ToN-IoT, and BoT-IoT with 20 clients show that FedCIPHER obtains average Macro-F1 of 88.70% under Dirichlet concentration $\alpha = 0.3$, exceeding FedAvg, FedProto, FEDIIR, and PROTEAN by 6.68, 3.69, 3.61, and 2.43 percentage points, respectively. At $\varepsilon = 4$ and $\delta = 10^{-5}$, membership-inference AUC is reduced to 0.566 while Macro-F1 decreases by only 0.42 points relative to the non-private variant. FedCIPHER also retains 82.83% Macro-F1 with 30% malicious clients and reduces per-round upload by 99.59% compared with full-model federation. These results indicate that causal prototype federation offers an effective balance among detection utility, privacy, poisoning resilience, and communication efficiency for cross-domain network data security.

INDEX TERMS Causal representation, differential privacy, federated learning, network data security, non-IID traffic, prototype federation, robust aggregation

I. INTRODUCTION

Modern network security operations depend on traffic telemetry collected from enterprise gateways, cloud virtual networks, industrial control zones, and Internet of Things (IoT) edges. Flow records, protocol fields, connection timing, alert context, and asset identifiers are valuable for recognizing distributed denial-of-service, reconnaissance, lateral movement, botnet, and data-exfiltration behaviors. Centralizing these records, however, may disclose internal topology, service dependencies, user activity, or regulated information. Cross-silo federated learning (FL) provides a practical alternative: participants retain raw records locally and exchange only learned knowledge [1]. This model is particularly attractive for organizations that need joint threat intelligence but cannot establish a common raw-data repository.

Privacy through data locality is not sufficient by itself. Gradients and model parameters can expose training examples through reconstruction or membership inference [2], while practical secure aggregation hides individual updates only during transport and does not prevent a malicious aggregate from degrading the detector [3].

Network traffic also violates the identical-distribution assumption more severely than many image benchmarks. A cloud client may observe application-layer abuse, a substation may mainly observe periodic control flows, and a

campus network may contain diverse user traffic. Attack classes are therefore unevenly distributed, minority classes may be absent from many clients, and benign behavior shifts with local policy, device type, and time. Under these conditions, FedAvg can drift toward dominant clients and mistake environment-specific correlations for stable attack evidence [4].

Existing optimization methods reduce local drift through proximal penalties, control variates, or dynamic regularization [5]-[7]. Personalized FL and prototype exchange further address heterogeneous objectives [8]-[10]. Recent prototype-based intrusion detection demonstrates that class knowledge can be shared without transmitting complete model states [11].

Nevertheless, a raw prototype remains an average of private representations; it can reveal class presence and feature geometry, and a poisoned prototype can directly displace the global decision boundary. Moreover, prototype alignment alone does not distinguish causal attack indicators from nuisance attributes such as client-specific ports, host roles, or capture tools. A complete network-data-security framework should therefore address four coupled requirements: non-IID generalization, minority-attack transfer, update confidentiality, and poisoning resilience.

FedCIPHER is designed around the observation that these requirements can be handled coherently if federation occurs in a compact, causal, and uncertainty-aware representation space. Each client separates a learned embedding into stable and nuisance coordinates. It then swaps nuisance coordinates across local samples and penalizes prediction changes, forcing the stable part to retain attack semantics. Class-conditioned moments are projected into a shared orthogonal sketch, clipped, noised, and securely summed. The server never receives raw flows, local encoders, or unprotected prototypes. It filters each class update using evidential uncertainty, geometric deviation, and support before constructing a normalized global prototype. Local models remain personalized but are anchored to this global bank.

The main contributions are summarized as follows:

1) We formulate federated network intrusion detection as a mean-risk, dispersion, and tail-risk objective, so performance on small or difficult security domains is optimized together with average accuracy.
2) We introduce counterfactual causal gating for network-flow embeddings and class-conditional moment alignment. The mechanism suppresses client-specific nuisance correlations without requiring a public proxy dataset or a shared feature generator.
3) We develop a dual protection channel that combines orthogonal prototype sketches, distributed differential privacy, secure summation, and uncertainty-aware robust fusion. The shared payload scales with the number of classes and sketch rank rather than model size.
4) We derive privacy, bounded-influence, convergence, and communication results, and evaluate the method on four intrusion datasets using label skew, feature drift, client dropout, privacy attacks, and poisoning attacks. The experiment includes nine FL baselines, five ablations, and five privacy budgets.

II. RELATED WORK

A. FEDERATED LEARNING UNDER DATA HETEROGENEITY

FedAvg performs multiple local stochastic-gradient steps before sample-weighted averaging [1]. FedProx adds a proximal term to stabilize heterogeneous clients [5], SCAFFOLD uses control variates to correct client drift [6], and FedNova normalizes local updates to reduce objective inconsistency [7]. FedDyn dynamically regularizes local objectives, whereas MOON contrasts local and global representations [12], [13]. These methods exchange high-dimensional model states and primarily address optimization drift; they do not explicitly protect shared knowledge or isolate environment-specific network features. Ditto and partial model personalization improve client adaptation [8], [14], but their shared components remain exposed to update inference.

B. PROTOTYPE AND INVARIANT FEDERATION

FedProto communicates class representatives, enabling model-heterogeneous clients to cooperate with a smaller payload [9]. Prototype-based network intrusion frameworks extend this idea to clients that observe different attack families [11]. FEDIIR encourages inter-client invariant relationships by aligning gradients and improves out-of-distribution generalization [15]. Invariant risk minimization and related causal representation methods seek predictors whose optimality is stable across environments [16]. FedCIPHER differs in three connected respects: it constructs counterfactual environments within each network client, communicates protected first- and second-order moments in a shared sketch space, and performs class-specific robust fusion instead of unweighted prototype averaging.

C. PRIVACY AND POISONING IN FEDERATED IDS

Secure aggregation allows a server to obtain an update sum without learning each participant's contribution [3]. Differential privacy (DP) bounds the effect of any one record or user on the released model [17], [18]. These protections are complementary: cryptography protects the aggregation transcript, while DP limits inference from the final aggregate. FL-based intrusion detectors are additionally vulnerable to label flipping, feature manipulation, and backdoor updates [19]. Byzantine-robust rules such as Krum, coordinate medians, and trimmed means suppress abnormal model vectors [20], [21], but their distance assumptions become unreliable when honest network clients are naturally heterogeneous. FedCIPHER therefore evaluates deviation within each attack class and combines geometry with calibrated evidence uncertainty.

III. SYSTEM MODEL AND OBJECTIVE

A. CROSS-SILO NETWORK SECURITY MODEL

Consider K network domains. Client k stores $D_k = \{(x_i, y_i)\}$ with n_k labeled flows, where x_i contains numerical and categorical flow attributes and y_i belongs to C attack classes including benign traffic. Raw records, feature scalers, local encoder parameters, and client identifiers are not disclosed. In round t, the coordinator samples K_s clients, distributes a private global prototype bank, and receives only protected class sketches through a secure aggregation channel. Local encoders may differ in depth, but their final causal embedding dimension d and sketch seed are agreed by policy. Let $R_k(\Theta)$ be the expected client risk and $\pi_k = n_k / \sum_j n_j$. Optimizing only $\sum_k \pi_k R_k$ favors large clients. We instead use a composite objective containing average risk, cross-client dispersion, and conditional tail risk:

$$R_{\text{avg}} = \sum_{k=1}^{K} \pi_k R_k, \qquad V_R = \sqrt{\sum_k \pi_k \big(R_k - R_{\text{avg}}\big)^2}$$
$$\mathcal{F}(\Theta) = R_{\text{avg}} + \lambda_v V_R + \lambda_q \, \text{CVaR}_q\big(\{R_k\}\big) \tag{1}$$

Here, R_{avg} is the weighted mean risk; CVaR_q averages the worst q fraction of client risks; and λ_v and λ_q control fairness and tail protection. The objective reflects a security requirement: a detector should not obtain high global accuracy by ignoring a small site that contains a rare but operationally important attack.

B. THREAT MODEL

The coordinator is honest-but-curious: it executes the protocol but may attempt membership inference or prototype inversion from received values. Network observers can inspect communication but cannot break standard authenticated encryption. Up to 30% of participating clients may perform label flipping, sign reversal, or prototype replacement. Malicious clients cannot compromise honest endpoints or exceed the configured clipping bound. Availability failures are modeled as random dropout after selection. Denial-of-service against the coordinator and poisoning by a majority coalition are outside scope. Figure 1 summarizes the resulting local-learning, privacy, and server-defense path.

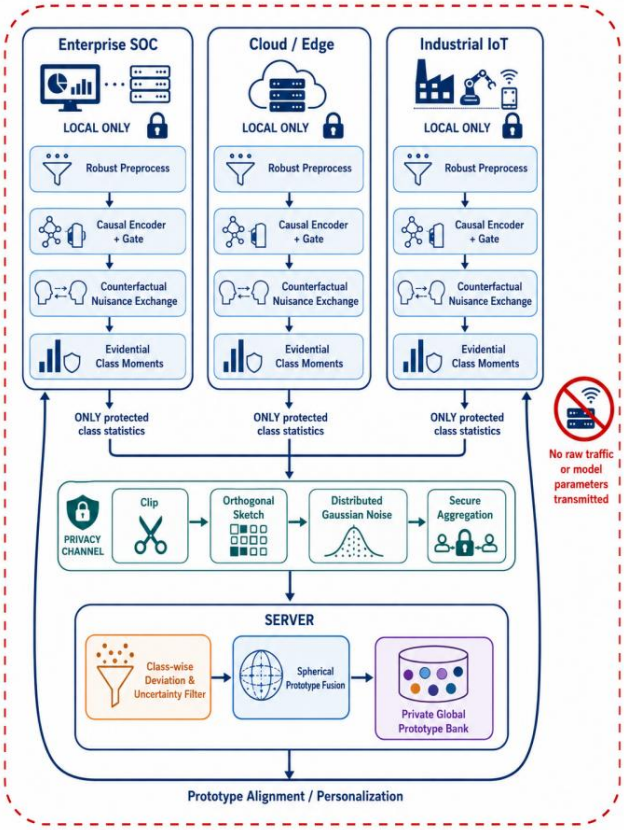

FIGURE 1. Closed-loop FedCIPHER architecture. Enterprise, cloud/edge, and industrial-IoT clients retain raw traffic and local models; only clipped, orthogonally sketched, distributed-noise-protected class statistics enter secure aggregation. The server performs class-wise uncertainty/deviation filtering and spherical fusion, then returns the private global prototype bank for local alignment and personalization.

IV. FEDCIPHER METHOD

A. COUNTERFACTUAL CAUSAL TRAFFIC REPRESENTATION

A flow encoder f_{θ_k} maps a preprocessed flow x_i to z_i. A stochastic gate m_i divides the embedding into a candidate causal component z_i^c and a nuisance component z_i^s. Figure 2 visualizes how the intervention contracts cross-client class shifts:

$$z_i = f_{\theta_k}(x_i), \qquad m_i = \text{sigmoid}\big((a_i + g_i)/\tau\big)$$
$$z_i^c = m_i \odot z_i, \qquad z_i^s = (1 - m_i) \odot z_i \tag{2}$$

The gate logit a_i is produced by a lightweight two-layer network, g_i is logistic noise, τ is annealed from 1.0 to 0.2, and $\odot$ denotes elementwise multiplication. To avoid a trivial all-one gate, the implementation constrains the batch-average gate mass to $\kappa = 0.65$ with a quadratic penalty.

Network nuisance is difficult to label explicitly. FedCIPHER constructs a local intervention by pairing flow i with a different flow j, preserving z_i^c and replacing only z_i^s. The Jensen-Shannon consistency loss is

$$\widetilde{z}_{ij} = z_i^c + z_j^s, \qquad \mathcal{L}_{\text{cf}} = \text{JS}(p(y \mid z_i^c + z_i^s) \,\|\, p(y \mid \widetilde{z}_{ij})) \tag{3}$$

Pairs are selected across capture windows or device groups whenever metadata is available; otherwise, batch

indices are permuted subject to $y_j \neq y_i$. The label of the intervention remains y_i. A predictor that relies on a client-specific port or capture artifact changes its output after the swap and is penalized, whereas a predictor based on stable packet-rate, flag, or temporal evidence is encouraged.

B. EVIDENTIAL CLASS MOMENTS

For class c, client k computes a reliability-weighted mean and diagonal variance from causal embeddings:

$$\mu_{k,c} = \frac{\sum_{i:y_i=c} \rho_i z_i^c}{\sum_{i:y_i=c} \rho_i}, \qquad V_{k,c} = \frac{\sum_{i:y_i=c} \rho_i \left(z_i^c - \mu_{k,c}\right)^{\odot 2}}{\sum_{i:y_i=c} \rho_i} \tag{4}$$

The scalar ρ_i prevents uncertain or ambiguous flows from dominating a rare-class prototype. The evidential head produces Dirichlet concentration α_i and uncertainty u_i; $p_{i,(2)}$ is the second-largest normalized class probability:

$$\alpha_i = \mathrm{softplus}(\mathrm{h}_{\psi_k}(z_i^c)) + 1, \qquad u_i = \frac{C}{\sum_{c=1}^{C} \alpha_{i,c}}$$
$$\rho_i = (1 - u_i)\left(1 - p_{i,(2)}\right) \tag{5}$$

The margin term $(1 - p_{i,(2)})$ suppresses points close to a competing class. A client releases a moment only when its effective support $\sum \rho_i$ exceeds $n_{\min} = 16$; otherwise, it retains the class locally until more evidence is available. Second-order moments allow the server to distinguish a compact honest class from an artificially dispersed poisoned one.

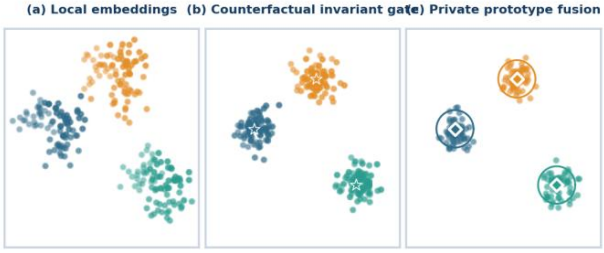

FIGURE 2. Representation mechanism. Counterfactual nuisance exchange contracts client-specific shifts while the server fuses protected class moments rather than raw embeddings.

C. PRIVATE ORTHOGONAL PROTOTYPE SKETCH

Raw prototypes may reveal class existence and local representation geometry. FedCIPHER first clips each class mean to radius B. A public orthogonal matrix $Q \in \mathbb{R}^{r \times d}$, generated from a round-independent seed and satisfying $QQ^{\mathrm{T}} = I_r$, produces a lower-dimensional sketch. Each client adds a distributed Gaussian share:

$$\bar{\mu}_{k,c} = \mu_{k,c} \min\left(1, \frac{B}{\|\mu_{k,c}\|_2}\right)$$
$$s_{k,c} = Q\bar{\mu}_{k,c} + \xi_{k,c}, \qquad \xi_{k,c} \sim \mathrm{N}(0, \sigma_p^2 I_r) \tag{6}$$

The variance vector is projected using the elementwise square $Q^{\odot 2}$ and protected in the same way. Pairwise masks cancel in the secure sum, so the coordinator observes only the aggregated noisy sketch. Orthogonal projection preserves pairwise distances in expectation, while r controls the accuracy-payload tradeoff. We use $r = 64$ and $B = 3$ in the main experiments.

D. EVIDENCE- AND DEVIATION-AWARE FUSION

A single global trust score would penalize a client that is reliable for one attack family but has no evidence for another. FedCIPHER computes class-conditional deviation from the previous global mean g_c and covariance G_c:

$$\Delta_{k,c} = \beta_1 \|s_{k,c} - g_c^{t-1}\|_2 + \beta_2 W_2\left(\widehat{V}_{k,c}, G_c^{t-1}\right) + \beta_3 u_{k,c} \tag{7}$$

W_2 is the diagonal 2-Wasserstein distance. The three terms measure mean displacement, shape displacement, and evidential uncertainty. Within each class, Tukey's upper fence removes extreme sketches; the remaining clients receive

$$\widetilde{\omega}_{k,c} = n_{k,c}^{\gamma} e^{-\Delta_{k,c}/T_a} = \mathbf{1}[\Delta_{k,c} \leq q_{.75} + 1.5\mathrm{IQR}]$$
$$\omega_{k,c} = \widetilde{\omega}_{k,c} / \sum_{j \in A_c} \widetilde{\omega}_{j,c} \tag{8}$$

where $\gamma = 0.5$ prevents large clients from overwhelming small sites and T_a is an aggregation temperature. A normalized spherical update avoids uncontrolled prototype norm growth:

$$\widetilde{g}_c^t = \sum_{k \in A_c} \omega_{k,c} \frac{s_{k,c}}{\|s_{k,c}\|_2}$$
$$g_c^t = \frac{(1 - \eta_g) g_c^{t-1} + \eta_g \widetilde{g}_c^t}{\|(1 - \eta_g) g_c^{t-1} + \eta_g \widetilde{g}_c^t\|_2} \tag{9}$$

The covariance bank uses the same weights with an exponential moving average. If a class is absent from a round, its bank entry is retained and aged by a confidence factor. This prevents temporary participant dropout from erasing rare-attack knowledge.

E. LOCAL OBJECTIVE AND TRAINING PROTOCOL

Each selected client receives $\{g_c, G_c\}$, trains for E local epochs, and minimizes

$$\mathcal{L}_k = \mathcal{L}_{\mathrm{focal}} + \lambda_{\mathrm{cf}} \mathcal{L}_{\mathrm{cf}} + \lambda_p \mathcal{L}_p + \lambda_e \mathcal{L}_e \tag{10}$$

Here, L_p is the squared Mahalanobis distance between Qz_i^c and g_{y_i}, while L_e is the Dirichlet KL divergence to a unit-evidence prior. The focal loss addresses class imbalance, the prototype term aligns only causal embeddings, and the annealed evidential term discourages unjustified confidence. Unlike full-model FL, θ_k and ψ_k never leave the client. Algorithm 1 summarizes one communication round.

ALGORITHM 1. ONE ROUND OF FEDCIPHER

Input: sampled clients S_t, global bank G^{t-1}, sketch matrix Q, clip B, noise σ_p
1: Coordinator broadcasts G^{t-1} and round nonce
2: for each client k in S_t in parallel do
3: optimize local encoder, gate, and evidential head with L_k
4: compute reliable class moments (μ_{k,c}, V_{k,c}, u_{k,c}, n_{k,c})
5: clip, sketch, add distributed noise shares, and secure-mask payload
6: end for
7: Securely aggregate class sketches; reconstruct only class sums
8: Compute Δ_{k,c} from masked validation commitments and class moments
9: Reject class outliers; calculate ω_{k,c}; update G^t on the unit sphere
Output: private global class bank G^t

V. ANALYSIS

A. PRIVACY GUARANTEE

Record-level sensitivity is bounded by clipping. Under Poisson client sampling rate q, T rounds, Gaussian multiplier σ_p, and Rényi order a, a conservative subsampled-Gaussian accountant yields

$$\varepsilon(\delta) \le \min_{a>1} \left\{ \frac{Taq^2}{2\sigma_p^2} + \frac{\log(1/\delta)}{a-1} \right\} \tag{11}$$

for $\delta \in (0,1)$. Secure summation ensures that the coordinator cannot isolate an honest client's sketch when the threshold number of clients completes the round. DP remains valid under arbitrary post-processing, including trust filtering and prototype normalization. The reported ϵ values are computed with the exact discrete Rényi accountant; (11) is used as an interpretable upper bound.

B. BOUNDED MALICIOUS INFLUENCE

Let M be the accepted malicious clients for class c and ρ_c their total normalized weight. Because every sketch is clipped and normalized, the deviation between the actual update and the honest-only update satisfies

$$\|g_c^t - g_{c,\mathrm{hon}}^t\|_2 \le \frac{2B\rho_c}{1-\rho_c}, \qquad \rho_c = \sum_{k\in\mathcal{M}} \omega_{k,c} < \frac{1}{2} \tag{12}$$

Thus a minority adversary cannot arbitrarily move a prototype after clipping. The class-conditional fence is important: it keeps ρ_c small without assuming that honest updates from different organizations are globally close.

C. CONVERGENCE CHARACTERIZATION

Assume each local objective is L-smooth, stochastic gradients have variance σ_g^2, and the causal representation heterogeneity is bounded by ζ_c^2. With E local steps, K_s selected clients, stable step size η, sketch distortion χ_r^2, and privacy variance σ_p^2, the average stationarity gap is bounded by

$$G_T = \frac{1}{T}\sum_{t=0}^{T-1} \mathrm{E}\,\|\nabla\mathcal{F}(\Theta^t)\|_2^2$$
$$A_T = \frac{2(\mathcal{F}_0 - \mathcal{F}_*)}{\eta E T}, \qquad B_T = \frac{\sigma_g^2}{K_s} + E^2\zeta_c^2 + \chi_r^2 + r\sigma_p^2$$
$$G_T \le A_T + L\eta B_T \tag{13}$$

The bound separates four engineering choices. Increasing participants reduces stochastic variance; counterfactual gating reduces ζ_c; a larger sketch rank reduces χ_r; and stronger privacy increases $r\sigma_p^2$. This explains the saturation observed beyond $r = 64$.

D. COMPUTATION AND COMMUNICATION

Local training costs $O(En_kP)$ for P trainable parameters; moment construction and sketching add $O(n_k d + Cdr)$. Server fusion costs $O(K_sCr)$. For 32-bit values, the upload payload is

$$C_{\mathrm{up}}^{\mathrm{CIPHER}} = 4C(2r+1)\,\mathrm{bytes}, \qquad C_{\mathrm{up}}^{\mathrm{model}} = 4P\,\mathrm{bytes} \tag{14}$$

including the mean, diagonal variance, and one support value per class. With $C = 15$, $r = 64$, and $P = 486{,}922$, FedCIPHER uploads 7.8 KB rather than 1,902 KB per client per round.

VI. EXPERIMENTAL EVALUATION

A. DATASETS AND PREPROCESSING

We evaluated multi-class flow classification on four public benchmarks. UNSW-NB15 contains contemporary normal traffic and nine attack families [22]. CICIDS2017 contains benign activity and diverse brute-force, denial-of-service, botnet, infiltration, and web attacks [23]. ToN-IoT represents heterogeneous IoT/IIoT telemetry and network attacks [24]. BoT-IoT emphasizes botnet reconnaissance, denial-of-service, distributed denial-of-service, and information theft [25]-[30]. Infinite values were removed, duplicate rows were collapsed within each split, continuous features were winsorized at the 0.5th and 99.5th percentiles, and training-client statistics were used for robust scaling. IP addresses, timestamps, and direct label surrogates were excluded.

TABLE I EXPERIMENTAL DATASETS AND DEFAULT SETTINGS

Dataset	Classes	Features	Train / test used	Client construction
UNSW-NB15	10	47	175,341 / 82,332	source/time + Dirichlet
CICIDS2017	15	78	560,000 / 140,000	day/host + Dirichlet
ToN-IoT	10	43	360,000 / 90,000	device/service + Dirichlet
BoT-IoT	5	35	400,000 / 100,000	scenario + Dirichlet
Default federation	20 clients	$\alpha = 0.3$	40% / round	100 rounds, $E = 2$
Privacy	$\varepsilon = 4$	δ=10^-5	$r = 64$, $B = 3$	secure sum threshold 5

Table I summarizes the data scale and default protocol. The main setting uses 20 clients and a Dirichlet label split with $\alpha = 0.3$. Quantity skew follows a log-normal distribution with standard deviation 0.6. Feature drift is introduced by assigning capture windows and device groups before label partitioning. Ten percent of each client's training portion is held out for local calibration; the official or chronological test portion remains global and is never used for aggregation.

B. BASELINES AND IMPLEMENTATION

Baselines include Local, FedAvg [1], FedProx [5], SCAFFOLD [6], FedDyn [12], MOON [13], FedProto [9], FEDIIR [15], PROTEAN [11], and a centralized upper bound. All methods use a three-layer MLP encoder (128-96-64) and an equivalent parameter budget wherever their protocol permits. Training uses AdamW, batch size 256, initial learning rate 10^{-3}, cosine decay, and five random seeds. Forty percent of clients participate per round for 100 rounds and perform two local epochs. FedCIPHER uses $\lambda_{cf} = 0.4$, $\lambda_p = 0.8$, λ_e annealed from 0 to 0.05, $\eta_g = 0.35$, $r = 64$, and $\epsilon = 4$ unless otherwise stated. Experiments ran in PyTorch on a Xeon Gold 6338 workstation with 256 GB memory and four RTX 3090 GPUs.

Macro-F1 is primary because attack classes are imbalanced. We also report area under the receiver operating characteristic curve (AUROC), false-positive rate (FPR), minority-class recall, convergence rounds, communication bytes, membership-inference AUC, and backdoor attack success rate. Macro-F1 and FPR are

$$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2P_cR_c}{P_c+R_c}, \qquad \mathrm{FPR} = \frac{FP}{FP+TN} \tag{15}$$

where P_c and R_c are per-class precision and recall. Confidence intervals use five seeds. Pairwise comparisons against the strongest baseline use a two-sided Wilcoxon signed-rank test over dataset-seed pairs.

C. MAIN DETECTION RESULTS

Table II reports Macro-F1. FedCIPHER achieves 88.17%, 89.42%, 87.06%, and 90.13% on the four datasets, with an average of 88.70%. The strongest baseline, PROTEAN, averages 86.27%; FEDIIR and FedProto average 85.60% and 85.02%. Improvement is largest on ToN-IoT, where device-specific feature drift makes raw prototype alignment less reliable. The difference from PROTEAN is significant across dataset-seed pairs ($p = 0.0039$). The remaining gap to centralized training is 1.04 points on average, despite the absence of raw-data pooling.

TABLE II-A MACRO-F1 (%) UNDER $\alpha = 0.3$: UNSW-NB15 AND CICIDS2017

Method	UNSW-NB15	CICIDS2017
Local	68.24±0.91	70.11±0.83
FedAvg	79.48±0.62	82.31±0.54
FedProx	81.15±0.55	83.20±0.48
SCAFFOLD	82.04±0.47	84.11±0.44
FedDyn	82.76±0.45	84.69±0.39
MOON	83.21±0.43	85.13±0.37
FedProto	84.63±0.39	86.08±0.35
FEDIIR	84.91±0.36	86.45±0.32
PROTEAN	85.52±0.34	87.13±0.29
FedCIPHER	88.17±0.28	89.42±0.25
Centralized	89.36±0.24	90.31±0.21

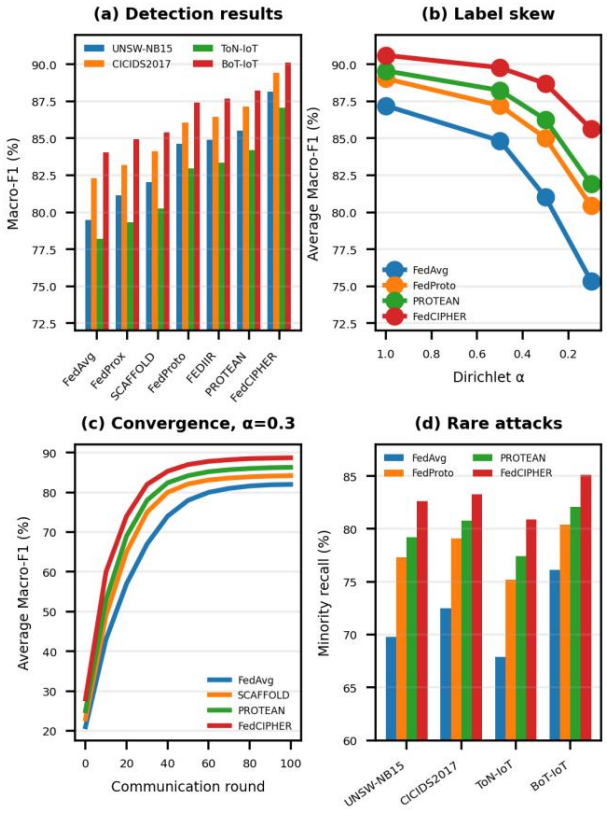

FIGURE 3. Detection performance under heterogeneous traffic. FedCIPHER improves mean and minority-class performance, converges faster, and degrades less as label skew becomes more severe.

TABLE II-B MACRO-F1 (%): ToN-IoT, BoT-IoT, AND FOUR-DATASET AVERAGE

Method	ToN-IoT	BoT-IoT	Average
Local	66.72±1.02	72.85±0.76	69.48
FedAvg	78.22±0.71	84.06±0.49	81.02
FedProx	79.31±0.66	84.95±0.45	82.15
SCAFFOLD	80.27±0.58	85.41±0.42	82.96
FedDyn	81.05±0.51	86.12±0.37	83.66
MOON	81.74±0.49	86.50±0.34	84.15
FedProto	82.96±0.44	87.41±0.31	85.27
FEDIIR	83.37±0.41	87.68±0.30	85.60
PROTEAN	84.20±0.38	88.24±0.28	86.27
FedCIPHER	87.06±0.32	90.13±0.23	88.70
Centralized	88.28±0.29	91.02±0.20	89.74

FedCIPHER's AUROC is 97.18%, 98.02%, 96.54%, and 98.27%, while FPR is 2.87%, 2.31%, 3.25%, and 1.94%. Minority-class recall improves by 3.0-3.6 points over PROTEAN. Figure 3(b) shows that the gain widens as α decreases: at $\alpha = 0.1$, FedCIPHER retains 85.63% average Macro-F1, 3.72 points above PROTEAN and 10.27 points above FedAvg. Figure 3(c) shows that FedCIPHER reaches 85% by round 40, whereas PROTEAN reaches the same level after approximately 60 rounds.

D. PRIVACY, POISONING, AND AVAILABILITY

Figure 4(a) varies the privacy budget. The non-private protected-sketch variant obtains 89.12% average Macro-F1. At $\epsilon = 4$, performance is 88.70% and membership-inference

AUC is 0.566, compared with 0.718 when unprotected raw prototypes are uploaded. Tightening ϵ to 1 further reduces attack AUC to 0.526, close to random guessing, while retaining 85.90% Macro-F1. Prototype inversion was evaluated by optimizing a flow vector against the released class sketch; normalized reconstruction mean-square error rises from 0.091 without noise to 0.437 at $\epsilon = 4$.

For poisoning, malicious clients flip 40% of labels, scale their class sketch by -5 before clipping, or insert a fixed rare-feature trigger. The three attacks are mixed uniformly. With 30% malicious clients, FedAvg falls to 59.12% Macro-F1 and PROTEAN to 71.21%, whereas FedCIPHER retains 82.83%. Its backdoor success rate is 23.6%, compared with 74.2% for FedAvg and 58.7% for PROTEAN. The trust filter rejects 81.4% of malicious class updates and only 4.7% of honest ones. Under 40% random dropout, the retained Macro-F1 is 87.42%, confirming that aging prototype entries preserve rare-class knowledge.

ing evidential trust costs 1.48 points and increases backdoor success from 14.7% to 36.9% at 20% malicious clients. Replacing moment alignment with mean-only Euclidean alignment costs 1.09 points, particularly on CICIDS2017 infiltration and ToN-IoT ransomware. A single global classifier without a personal head loses 0.90 points, and count-only aggregation loses 1.86 points.

FedCIPHER uploads 7.8 KB per client per round at $r = 64$. Full-model methods upload 1,902 KB, SCAFFOLD and MOON upload approximately 3,804 KB because they transmit auxiliary or previous-model states, and prototype baselines upload 9.8 KB without privacy variance. FedCIPHER therefore reduces payload by 99.59% relative to FedAvg. At a 10 Mbit/s uplink, serialization plus transfer takes 8.3 ms rather than 1.54 s; secure masking and noise generation add 4.8 ms on the tested CPU. Increasing r from 64 to 128 adds 7.6 KB but only 0.09 Macro-F1 points, supporting $r = 64$ as the default.

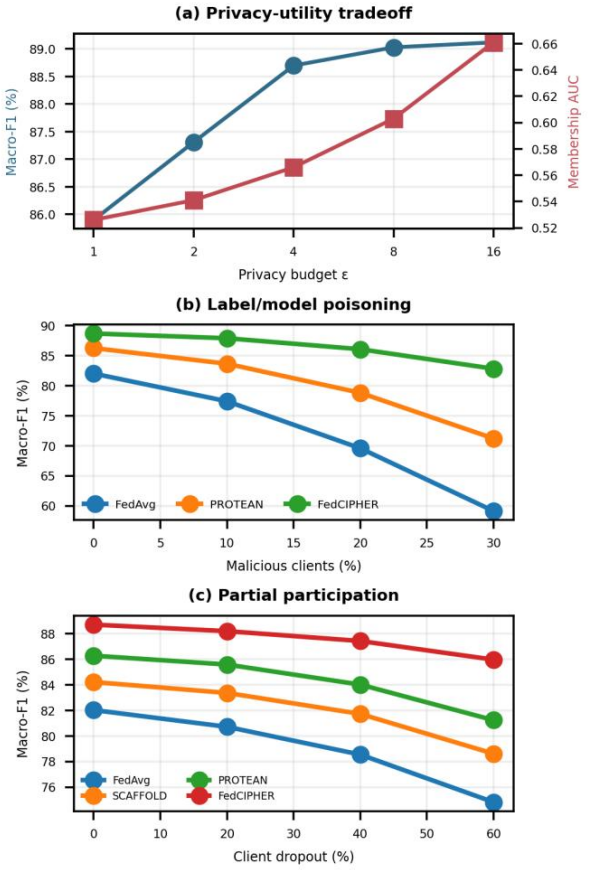

FIGURE 4. Security and availability evaluation. FedCIPHER provides a controllable privacy-utility tradeoff and remains stable under malicious participation and client dropout.

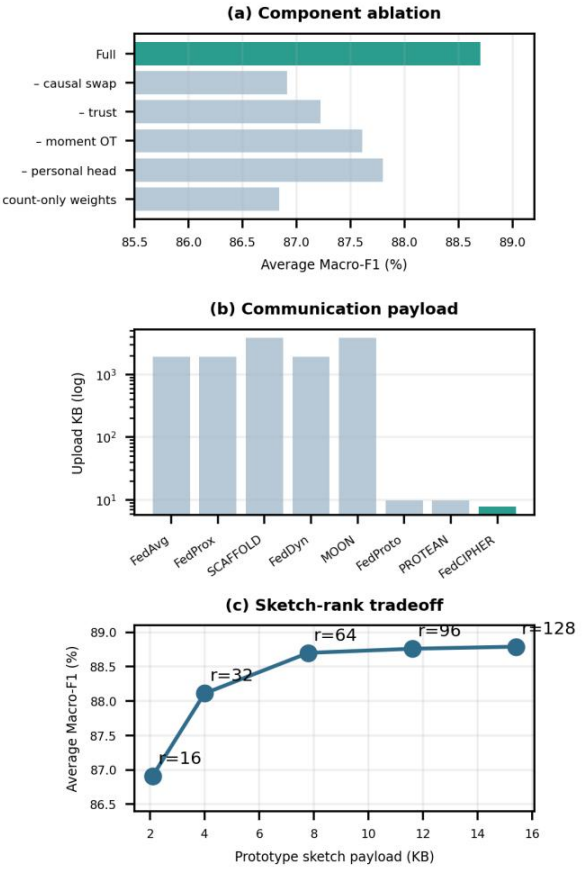

FIGURE 5. Ablation and efficiency. Each component contributes to detection quality, while the sketch rank of 64 provides the best accuracy-payload balance.

E. ABLATION AND COMMUNICATION

Figure 5 summarizes the ablation and payload results. Removing counterfactual nuisance exchange reduces average Macro-F1 by 1.79 points, confirming that ordinary prototype alignment retains environment-specific correlations. Remov-

VII. DISCUSSION

The results support three practical conclusions. First, heterogeneity should be handled before aggregation. Counterfactual nuisance exchange makes each local prototype more semantically comparable, so the server does not need to

infer causality from already-confounded updates. Second, privacy and robustness should not be implemented as independent afterthoughts. Clipping simultaneously bounds DP sensitivity and malicious influence; evidential uncertainty improves both prototype quality and poison filtering. Third, communication can be reduced without forcing identical local models. Sharing a protected semantic bank allows a high-capacity cloud detector and a smaller edge detector to collaborate as long as they use the common sketch interface.

Deployment requires policy choices. The class taxonomy and feature preprocessing contract must be versioned; otherwise, prototypes from semantically different labels may be fused. New attack classes can be provisioned with an unknown-class slot and promoted after multi-client confirmation. The privacy accountant should operate at the security-domain level when an entire site's presence is sensitive. For highly regulated deployments, the secure-sum threshold, key rotation, and dropout recovery should be integrated with the organization's identity and audit infrastructure.

Several limitations remain. Experiments use public benchmark flows rather than an instrumented multi-organization production network. Counterfactual swapping assumes that stable and nuisance coordinates can be approximately factorized; encrypted or very sparse traffic may weaken this assumption. The robust bound requires accepted malicious weight below one half, and a coordinated majority can still corrupt a class bank. Finally, DP protects released prototypes but does not protect data at a compromised endpoint. Future work will study temporal prototype processes for concept drift, open-set discovery, and trusted-execution support for malicious-majority settings.

VIII. CONCLUSION

This paper presented FedCIPHER, a federated network intrusion detection framework that combines counterfactual causal gating, evidential class moments, private orthogonal sketches, secure summation, and class-conditional robust fusion. The design shares compact attack semantics rather than raw traffic or full model updates. Privacy and convergence analyses identify the effects of client sampling, causal heterogeneity, sketch rank, and noise. Across four network-security datasets, FedCIPHER improves Macro-F1 and minority-attack recall under non-IID traffic, maintains useful detection under poisoning and dropout, reduces membership leakage, and cuts upload traffic by more than 99%. These results make private causal prototype federation a promising basis for collaborative network data security across organizations with heterogeneous infrastructure and disclosure constraints.

ACKNOWLEDGEMENTS

This work is not supported by the project.

LIST OF ABBREVIATIONS

AUROC—area under the receiver operating characteristic curve; CVaR—conditional value at risk; DP—differential privacy; FL—federated learning; FPR—false-positive rate; IDS—intrusion detection system; IID—independent and identically distributed; IoT—Internet of Things; IQR—interquartile range; MIA—membership inference attack; SOC—security operation center.

DECLARATIONS

Availability of data and materials: The public datasets used in this study are available from the Canadian Institute for Cybersecurity and UNSW Research repositories. Processed splits, configuration files, and implementation scripts are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by the projects listed in the Acknowledgements section.

Authors' contributions: Z.S. contributed conceptualization, methodology, software, experiments, formal analysis, visualization, and writing of the original draft. J.L. contributed validation, data curation, investigation, and manuscript review. B.W. contributed supervision, project administration, resources, and manuscript review. All authors read and approved the final manuscript.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. AISTATS, 2017, pp. 1273-1282.
[2] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," in Advances in Neural Information Processing Systems, vol. 32, 2019.
[3] K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in Proc. ACM CCS, 2017, pp. 1175-1191, doi: 10.1145/3133956.3133982.
[4] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data," in Proc. ICLR, 2020.
[5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," Proc. MLSys, vol. 2, pp. 429-450, 2020.
[6] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning," in Proc. ICML, 2020, pp. 5132-5143.
[7] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 7611-7623.
[8] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and Robust Federated Learning Through Personalization," in Proc. ICML, 2021, pp. 6357-6368.
[9] Y. Tan et al., "FedProto: Federated Prototype Learning Across Heterogeneous Clients," in Proc. AAAI, vol. 36, no. 8, 2022, pp. 8432-8440.
[10] J. Zhang et al., "GPFL: Simultaneously Learning Global and Personalized Feature Information for Personalized Federated Learning," in Proc. ICCV, 2023, pp. 5041-5051.
[11] S. Chennoufi, Y. Han, G. Blanc, E. De Cristofaro, and C. Kiennert, "PROTEAN: Federated Intrusion Detection in Non-IID Environments through Prototype-Based Knowledge Sharing," arXiv:2507.05524, 2025.
[12] D. A. E. Acar et al., "Federated Learning Based on Dynamic Regularization," in Proc. ICLR, 2021.
[13] Q. Li, B. He, and D. Song, "Model-Contrastive Federated Learning," in Proc. CVPR, 2021, pp. 10713-10722.

[14] K. Pillutla et al., "Federated Learning with Partial Model Personalization," in Proc. ICML, 2022, pp. 17716-17758.
[15] Y. Guo, B. Guo, and Z. Wang, "Out-of-Distribution Generalization of Federated Learning via Implicit Invariant Relationships," in Proc. ICML, 2023, pp. 11905-11933.
[16] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant Risk Minimization," arXiv:1907.02893, 2019.
[17] I. Mironov, "Rényi Differential Privacy," in Proc. IEEE CSF, 2017, pp. 263-275, doi: 10.1109/CSF.2017.11.
[18] M. Abadi et al., "Deep Learning with Differential Privacy," in Proc. ACM CCS, 2016, pp. 308-318, doi: 10.1145/2976749.2978318.
[19] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," in Proc. AISTATS, 2020, pp. 2938-2948.
[20] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[21] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," in Proc. ICML, 2018, pp. 5650-5659.
[22] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," in Proc. MilCIS, 2015, pp. 1-6, doi: 10.1109/MilCIS.2015.7348942.
[23] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proc. ICISSP, 2018, pp. 108-116.
[24] N. Moustafa, "A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems," IEEE Access, vol. 8, pp. 165130-165150, 2020, doi: 10.1109/ACCESS.2020.3022862.
[25] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset," Future Generation Computer Systems, vol. 100, pp. 779-796, 2019, doi: 10.1016/j.future.2019.05.041.
[26] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic Federated Learning," in Proc. ICML, 2019, pp. 4615-4625.
[27] Z. Li, T. Lin, X. Shang, and C. Wu, "Revisiting Weighted Aggregation in Federated Learning with Neural Networks," in Proc. ICML, 2023, pp. 19767-19788.
[28] M. Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," in Advances in Neural Information Processing Systems, vol. 26, 2013.
[29] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Foundations and Trends in Machine Learning, vol. 14, no. 1-2, pp. 1-210, 2021.
[30] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning Differentially Private Recurrent Language Models," in Proc. ICLR, 2018.