IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Byzantine-Resilient Federated Anomaly Detection for Cyber-Physical Critical Infrastructure Protection

[1]Sri Varsha, [1]Yaathra P, [1]Vippin Antony

[1]Department of Computer Science and Engineering,

Sri Venkateswara College of Engineering (Autonomous), Sriperumbudur – 602 117, Tamil Nadu, India

Mentor: Ms. P. Uma, Asst. Professor/CSE, SVCE

Doi: https://doi.org/10.56975/ijcrt.v14i7.309000

Abstract: Industrial control systems governing electrical power distribution, water treatment, and petroleum transport represent a critical attack surface that is both consequential and structurally underprotected. Between 2019 and 2022, recorded intrusions against such systems increased by 140 percent globally, with documented incidents including the 2021 Oldsmar water treatment facility compromise and the Colonial Pipeline ransomware event collectively causing multi-billion-dollar disruptions [1][2]. Two structural inadequacies limit existing defences. Signature-based intrusion detection systems are incapable of detecting zero-day attack patterns absent from their rule repositories. Centralised machine-learning anomaly detectors require raw operational telemetry to be transmitted off-premises, violating data-sovereignty mandates under frameworks including the US NERC CIP standards and India's NCIIPC guidelines. Neither class of system can detect compound cyber-physical intrusions, in which an adversary simultaneously exploits the control network and manipulates physical sensor telemetry to produce contradictory monitoring outputs that cancel detection alerts. This paper presents SENTINEL, a four-layer detection framework that addresses all three limitations within a single architecture. A federated learning layer enables collaborative model training across geographically distributed infrastructure sites without centralising operational data. A Gaussian differential privacy mechanism prevents reconstruction of site-specific parameters from transmitted gradient updates. Byzantine fault-tolerant aggregation via the Krum algorithm provides formal resilience against gradient-poisoning attacks from up to $f < n/2$ compromised federation participants. The central technical contribution is a cyber-physical correlation engine that monitors, in a sliding window, the Pearson correlation between SCADA digital control event logs and physical sensor telemetry streams. Under compound attacks, adversarial sensor spoofing causes this correlation to degrade measurably, producing a detection signal unavailable to any single-modality monitoring system. Evaluated on the BATADAL water distribution attack benchmark across 10 independent experimental trials, SENTINEL achieves mean detection accuracy of 91.4% with a false positive rate of 6.2%, and outperforms centralised baselines by 18.3 percentage points on compound attack scenarios. Experimental configurations will be made available to support reproducibility.

Index Terms: federated learning, Byzantine fault tolerance, anomaly detection, cyber-physical systems, SCADA security, differential privacy, critical infrastructure protection, Krum aggregation, compound attack detection, industrial control systems.

I. INTRODUCTION

Supervisory Control and Data Acquisition (SCADA) systems and Industrial Control Systems (ICS) constitute the operational substrate of modern critical infrastructure, governing physical processes that include electrical power distribution, water treatment, petroleum transport, and rail signalling at national scale. Unlike conventional information technology targets, successful attacks against these systems produce consequences that are simultaneously digital and physical: an adversarially manipulated set-point in a water treatment controller does not result in a data loss event but in a public health emergency. The convergence of operational technology with IP-based networking, combined with the proliferation of commercial off-the-shelf software in ICS environments, has substantially expanded the attack surface of this infrastructure over the past decade [1][4].

The threat landscape has evolved from opportunistic intrusions toward targeted, multi-stage campaigns attributed to nation-state actors. Dragos reported a 140 percent increase in recorded ICS intrusions between 2019 and 2022 [1]. The February 2021 Oldsmar, Florida water treatment facility incident, in which a remote adversary elevated sodium hydroxide concentration to 111 times the permissible limit before an operator intervened manually, demonstrated the absence of automated anomaly detection at the field level. The Colonial Pipeline ransomware event of the same year halted 45 percent of US East Coast fuel supply for six days at an estimated economic cost of USD 4.4 billion [2]. The October 2020 Mumbai grid disruption, subsequently attributed to a foreign threat actor, caused concurrent failure of hospital power systems, rail services, and financial market infrastructure [3]. These incidents share a common characteristic: the attacked systems lacked real-time anomaly detection capable of identifying the intrusion before physical consequences materialised.

Current defensive approaches exhibit three structural limitations. First, rule-based intrusion detection systems such as Snort and Suricata match traffic against curated signature libraries; any intrusion variant not represented in those libraries passes undetected. Dragos confirmed that 65 percent of ICS intrusions recorded in 2022 employed techniques with no corresponding published signature at the time of attack [1]. Second, learning-based anomaly detection overcomes the generalisation problem but introduces a data centralisation requirement that is operationally unacceptable: raw SCADA telemetry reveals the precise topology, set-points, and control logic of national infrastructure, and its off-premises transmission is prohibited under multiple regulatory frameworks. Third, and most critically, no existing system addresses compound cyber-physical intrusions, in which an adversary simultaneously exploits the control network layer and injects false readings into physical sensor streams. Under this attack pattern, the digital anomaly detector observes an anomalous control-layer event while the physical monitoring subsystem observes sensor-reported normality. The two subsystems produce contradictory evidence and their respective alerts cancel, leaving the intrusion undetected regardless of the sophistication of either individual component.

This paper addresses all three limitations through a unified framework, SENTINEL, whose principal technical contributions are as follows. (1) A federated learning architecture for ICS anomaly detection in which each infrastructure site trains a local model on its own operational data and transmits only privatised weight updates, satisfying data-sovereignty constraints without sacrificing the accuracy benefit of federation-wide collective learning. (2) A differential privacy mechanism, calibrated via the moments accountant to achieve (epsilon, delta)-DP per communication round, that ensures no adversary with bounded computational resources can reconstruct any participating site's operational parameters from intercepted gradient transmissions. (3) Byzantine fault-tolerant aggregation via the Krum algorithm, which provides a formal convergence guarantee under the condition $f < n/2$, protecting the global model against gradient-poisoning and label-flipping attacks from compromised participants. (4) A cyber-physical correlation engine, the central novel contribution, that detects compound intrusions by monitoring the Pearson correlation between SCADA digital event logs and physical sensor telemetry in a calibrated sliding window. Evaluated on the BATADAL industrial control system benchmark, SENTINEL achieves 91.4 percent mean detection accuracy and 82.6 percent compound attack detection rate, improving upon the Krum-only federated baseline by 21.7 percentage points on compound scenarios. The remainder of this paper is organised as follows. Section II provides background on federated learning and Byzantine fault tolerance. Section III surveys related work. Section IV formalises the system model and threat model. Section V presents the SENTINEL framework with formal guarantees. Section VI describes the implementation architecture. Section VII presents the experimental evaluation. Section VIII provides security analysis. Section IX discusses limitations and future directions. Section X concludes.

II. BACKGROUND

A. Federated Learning Formulation

Let I = {s1, ..., sn} be a set of n distributed data-holding participants. In standard federated learning [9], each participant si holds a local dataset Di and optimises a local objective Li(theta). The global objective is L(theta) = (1/n) * SUM_i Li(theta). In federated averaging, participants perform E steps of local stochastic gradient descent from a broadcast global model theta_global, then transmit the weight update delta_theta_i = theta_i - theta_global to a central coordinator. The coordinator aggregates via theta_global <- theta_global + (1/n) * SUM_i delta_theta_i. Convergence guarantees for FedAvg hold under assumptions of bounded gradient dissimilarity and sufficiently small learning rates, but these guarantees break down in the presence of adversarial participants [8][10].

B. Byzantine Fault Tolerance in Federated Learning

A Byzantine participant is one that transmits an arbitrary update, possibly optimised to maximise damage to the global model. The Byzantine Generals Problem [Lamport et al., 1982] establishes that agreement among n processes is achievable when at most f < n/3 participants are Byzantine under arbitrary failures, and under more structured adversarial models the bound relaxes to f < n/2. In the federated learning context, Byzantine fault-tolerant aggregation replaces the arithmetic mean with a robust estimator that is statistically insensitive to adversarial outliers. Blanchard et al. [10] proved that Krum achieves (f, epsilon)-Byzantine resilience: with f < n/2 adversarial clients, the update selected by Krum lies within a bounded distance of the optimal honest gradient in expectation, provided honest clients have bounded gradient divergence.

C. Differential Privacy

Differential privacy [17] provides a formal privacy guarantee for mechanisms operating on datasets. A randomised mechanism M satisfies (epsilon, delta)-DP if for all adjacent datasets D, D' differing in one record and all measurable output sets S:

$$Pr[M(D) \text{ in } S] \leq exp(epsilon) * Pr[M(D') \text{ in } S] + delta$$

Smaller epsilon implies stronger privacy. The Gaussian mechanism achieves (epsilon, delta)-DP by adding noise N(0, sigma^2 C^2 I) to a function with L2 sensitivity C. Under composition across T independent applications, the moments accountant [12] provides tighter cumulative privacy bounds than naive composition, making it the standard tool for privacy accounting in differentially private federated learning.

III. RELATED WORK

A. ICS Anomaly Detection

Anomaly detection in ICS environments has been studied extensively. Stojanovic et al. [4] surveyed signature-based and statistical approaches, establishing that signature systems achieve high precision on known attack patterns at the cost of zero generalisation to novel vectors. Anthi et al. [5] deployed supervised classifiers on ICS network flow features and reported 98.2 percent accuracy on known attack categories, but their approach requires centralised data collection and does not evaluate zero-day scenarios. Lin et al. [6] demonstrated LSTM-based temporal anomaly detection for power grid SCADA, showing that sequence-to-sequence models outperform threshold-based methods on temporally correlated attack vectors. Critically, neither centralised approach is deployable in environments subject to data-sovereignty restrictions, a limitation that applies to virtually all national critical infrastructure operators.

B. Federated Learning Security

Federated learning was introduced to the ICS security context by Nguyen et al. [7], who demonstrated that cross-device federated intrusion detection approaches the accuracy of centralised baselines while preserving data locality. Their system, however, assumes all federation participants are honest. Tolpegin et al. [8] subsequently characterised the vulnerability of this assumption: under a label-flipping attack from a single compromised client, standard FedAvg [9] accuracy degrades by up to 40 percentage points in non-IID settings. More recent work has examined backdoor attacks in federated learning; Nguyen et al. proposed FLAME [19], which uses noise injection and model clustering to defend against backdoor submissions, though FLAME assumes a known proportion of malicious clients and does not carry Krum's formal convergence guarantees. Zhang et al. proposed FLDetector [20], a method for identifying malicious updates through consistency checking across rounds. Neither work addresses the cyber-physical compound attack problem.

C. Byzantine-Resilient Aggregation

Byzantine-resilient aggregation algorithms have been formalised by Blanchard et al. [10], who introduced Krum and Multi-Krum with proofs under the f < n/2 Byzantine fraction. El Mhamdi et al. [11] proposed Bulyan, which achieves stronger guarantees against adaptive adversaries at the cost of a stricter f < n/4 constraint. Cao and Fang [18] proposed FLTrust, which bootstraps trust from a small server-side dataset. These methods have been evaluated primarily in computer vision and NLP settings. Their behaviour under the non-IID sensor distributions characteristic of ICS environments, where different sites monitor different physical processes with different operational profiles, has not been systematically examined.

D. Cyber-Physical Correlation

The cyber-physical correlation concept was studied by Adepu and Mathur [16] for the SWaT water treatment testbed, who demonstrated that deterministic input-output relationships in physical processes can be exploited to detect single-stage multi-point attacks. Their approach is centralised and specific to single-site deployments. No prior work integrates a cyber-physical correlation signal with Byzantine-resilient federated learning into a system capable of detecting compound attacks while preserving data sovereignty. Table 1 summarises the comparison with the most closely related systems.

Table 1. Comparison with Related Systems

Method	Data Private	Byz. Def.	CP Correl.	Eval. Dataset	Acc. (%)
Snort/Suricata IDS [1]	No	No	No	Rule-based	71.2
Lin et al. LSTM [6]	No	No	No	Power SCADA	89.1
Anthi et al. [5]	Yes	No	No	Custom IoT	91.8
Nguyen et al. [7]	Yes	No	No	N-BaIoT	90.3
Tolpegin et al. [8]	Yes	Partial	No	CIFAR (proxy)	N/A
FLAME [19]	Yes	Partial	No	CIFAR-10	N/A
FLDetector [20]	Yes	Partial	No	ImageNet	N/A
SENTINEL (proposed)	Yes	Yes	Yes	BATADAL	91.4

IV. System Model and Threat Model

A. System Model

Let I = {s1, s2, ..., sn} denote a federation of n infrastructure sites. Each site si operates a SCADA system producing a paired time-indexed stream (Ei, Si), where Ei = {(t_k, c_k)} is a log of digital control events (timestamp t_k, command c_k) and Si = {(t_k, v_k)} is the corresponding physical sensor telemetry (reading v_k at time t_k). Each site maintains a local anomaly detection model parameterised by theta_i and participates in a federated training protocol coordinated by an aggregator. Table 2 defines all notation used throughout the paper.

Table 2. Notation Table

Symbol	Definition
I = {s1,...,sn}	Set of n infrastructure federation participants
n	Total number of federation participants
f	Number of Byzantine (adversarially controlled) participants
si	The i-th infrastructure site
(Ei, Si)	Paired data stream at site i: event log and sensor telemetry
Ei	Digital control event log at site i: {(t_k, c_k)}
Si	Physical sensor telemetry at site i: {(t_k, v_k)}
theta_i	Local model parameters at site i
theta_global	Global federated model parameters
delta_theta_i	Weight update from site i: theta_i - theta_global
delta_theta_priv	Differentially privatised weight update from site i
Di	Local training dataset at site i (normal operational windows)
f_theta	Autoencoder reconstruction function parameterised by theta
a(theta, x)	Anomaly score for window x: ‖x - f_theta(x)‖^2
tau	Anomaly detection threshold (99th percentile of clean validation errors)
Li(theta_i)	Local reconstruction loss at site i
C	L2 gradient clipping norm (C = 1.0)
sigma	Gaussian noise multiplier (sigma = 1.1)
epsilon	Differential privacy budget per round (epsilon = 1.0)
delta	DP failure probability (delta = 1e-5)
T	Total number of federation communication rounds (T = 50)
E	Number of local SGD epochs per round (E = 5)
Ni	Set of (n - f - 2) nearest neighbours of update i under Krum
rho(t,W)	Pearson correlation over window W ending at time t
W	Correlation window size (W = 15 minutes)
rho_min	Site-calibrated minimum correlation threshold (rho_min = 0.65)
E(t)	Binary control event indicator at time t
S(t)	Physical sensor reading at time t

We define three system requirements. R1 (Data Sovereignty): The raw stream (Ei, Si) must not leave the network boundary of site si at any stage. Only differentially privatised derivatives of model parameters may be transmitted. R2 (Byzantine Resilience): The global model theta_global must maintain bounded detection accuracy when at most f < n/2 participants transmit adversarially crafted updates. R3 (Compound Attack Detection): The system must detect intrusions in which an adversary simultaneously compromises the digital control layer and spoofs physical sensor readings, causing single-modality detectors to produce contradictory and mutually cancelling alerts. Standard FedAvg satisfies R1 but fails R2 and R3 by construction.

B. Threat Model

We model an adversary A operating across two concurrent attack surfaces. At the federation layer, A controls f < n/2 participant sites and can dictate their transmitted gradient updates with full knowledge of the training history and the federation protocol. A may employ gradient poisoning (arbitrary malicious updates), label flipping (local training on inverted attack labels), or low-norm stealth attacks (crafting updates whose Euclidean norm falls within the honest gradient cluster to evade distance-based filters). At the physical layer, A can execute False Data Injection (FDI) against sensor streams and issue protocol-level intrusion commands via Modbus, DNP3, or IEC 60870-5-104. These capabilities correspond to the threat profile documented in

the Sandworm and Triton/Trisis campaigns, which combined persistent control-network access with manipulation of physical safety systems [1].

The adversary is subject to the following limitations, which are necessary for the system's formal guarantees to apply. (L1) A cannot compromise more than f < n/2 federation participants simultaneously. (L2) A cannot break the cryptographic primitives underlying the (epsilon, delta)-DP mechanism or the encrypted communication channels. (L3) A cannot generate synthetic sensor trajectories that simultaneously preserve Pearson correlation rho(t,W) >= rho_min with the actual control event sequence and mask the physical consequences of an active intrusion across all monitored sensor-actuator pairs. Limitation L3 is the structural constraint exploited by the correlation engine and is justified formally in Proposition 1 (Section VIII).

V. THE SENTINEL FRAMEWORK

A. Layer 1: Federated Anomaly Detection

Each site si trains a local anomaly detection model on windows drawn from its normal operational data (Ei, Si). The local training objective minimises mean squared reconstruction error over the local dataset Di:

$$L_i(\theta_i) = (1 / |D_i|) * \sum_{x \in D_i} ||x - f_{\theta_i}(x)||^2 \quad (1)$$

where ftheta denotes the decoder reconstruction function. The anomaly score for a test window x is a(theta_i, x) = ||x - f_theta_i(x)||^2. A window is flagged as anomalous if its score exceeds a site-calibrated threshold tau, determined from the 99th percentile of reconstruction errors on a held-out clean validation set. After E local training epochs, site si transmits only the weight update delta_theta_i = theta_i - theta_global. No element of the raw stream (Ei, Si), no derived operational statistic, and no intermediate gradient leaves the site network boundary.

B. Layer 2: Differential Privacy

Prior to transmission, the weight update is privatised via the Gaussian mechanism. The update is clipped to bound its L2 norm, then Gaussian noise is added:

$$\Delta\theta_{priv} = Clip(\Delta\theta_i, C) + N(0, \sigma^2 * C^2 * I) \quad (2)$$

where C = 1.0 is the clipping norm and sigma = 1.1 is the noise multiplier, calibrated using the moments accountant [12] to achieve per-round (epsilon, delta)-DP with epsilon = 1.0 and delta = 10^(-5). Theorem 1 formalises the cumulative privacy guarantee.

Theorem 1 (Privacy Composition [12]): *Let each communication round apply the Gaussian mechanism with clipping norm C and noise multiplier sigma to achieve per-round (epsilon_0, delta_0)-DP. Under strong composition across T rounds, the total privacy loss satisfies epsilon_total <= sqrt(2T * ln(1/delta_0)) * epsilon_0 + T * epsilon_0 * (exp(epsilon_0) - 1). Under the moments accountant, a tighter bound is computable numerically. For T=50, sigma=1.1, C=1.0, delta=1e-5, the moments accountant yields epsilon_total < 8.4.*

C. Layer 3: Byzantine Fault-Tolerant Aggregation

Standard FedAvg aggregates the global update as the arithmetic mean of received updates. Under adversarial conditions, a single Byzantine client transmitting an update of sufficiently large norm can dominate the aggregated gradient arbitrarily [10]. SENTINEL replaces FedAvg aggregation with Krum. Given n received updates G = {delta_theta_1, ..., delta_theta_n} and Byzantine tolerance parameter f, Krum scores each update by summing its squared Euclidean distances to its (n - f - 2) nearest neighbours:

$$Krum(G) = \arg\min_{i} \sum_{j \in N_i} ||\Delta\theta_i - \Delta\theta_j||^2 \quad (3)$$

The update with the minimum Krum score is selected as the aggregated gradient. Theorem 2 formalises the Byzantine resilience guarantee.

Theorem 2 (Krum Byzantine Resilience [10]): *Let G = {delta_theta_1, ..., delta_theta_n} be n gradient updates of which at most f satisfy f < (n-2)/2. Let G_h denote the subset of honest updates with bounded pairwise distance. Then the update selected by Krum(G) satisfies E[||Krum(G) - nabla L||^2] <= E[||delta_theta_h - nabla L||^2] + B_f, where B_f is a positive constant depending on the Byzantine fraction f/n and the gradient noise level. For SENTINEL's configuration n=5, f=1, the condition f < (n-2)/2 = 1.5 is satisfied.*

D. Layer 4: Cyber-Physical Correlation Engine

The correlation engine is the central technical contribution of this paper. Its design rests on the observation that under normal ICS operation, physical process responses are causally coupled to digital control events through the physical system dynamics. A pump activation command produces a measurable hydraulic pressure change within a characteristic lag window determined by the process physics. This coupling is stable over normal operating conditions, site-specific in its precise parameters, and quantifiable from historical data.

Let E(t) be the binary control event indicator at time t and S(t) the corresponding physical sensor value. The sliding-window Pearson correlation over window W is:

$$rho(t,W) = Cov(E[t-W:t], S[t-W:t]) / (Std(E[t-W:t]) \times Std(S[t-W:t])) \quad (4)$$

During normal operation, rho(t,W) remains above a site-calibrated threshold rho_min. Under a compound attack, the adversary injects false sensor readings that are statistically decoupled from the actual control event sequence. This decoupling causes rho(t,W) to decrease measurably. The engine issues a compound alert when both the federated model's anomaly score and the correlation metric breach their respective thresholds:

$$Alert\_compound(t) = [rho(t,W) < rho\_min] \ AND \ [a(theta\_global, x\_t) > tau] \quad (5)$$

The two-condition logical conjunction is architecturally important. False positives in either single modality alone, whether from model uncertainty or sensor noise, do not escalate to a compound alert. Only co-occurrence of both conditions triggers the alert, reducing the false positive rate while maintaining sensitivity to the compound attack class that evades all single-modality detectors. Proposition 1 formalises the evasion resistance.

Proposition 1 (Compound Attack Detectability): *Under threat model assumptions L1-L3, a compound attack is undetectable by SENTINEL only if the adversary can simultaneously maintain rho(t,W) >= rho_min and produce a physical effect indistinguishable from normal operations across all monitored sensor-actuator pairs. By L3, this is assumed infeasible for any realistic attack causing a measurable physical consequence. The detectability bound is therefore determined by the physical signal-to-noise ratio of the sensor-actuator correlation, which is site-specific and quantifiable from the baseline calibration period.*

VI. Implementation Architecture

SENTINEL is organised into three deployment tiers. At the site tier, each SCADA system provides data to a locally executing anomaly detector and correlation engine. Neither raw event logs nor sensor readings are transmitted beyond the site network boundary. The only outbound communication is the differentially privatised weight update delta_theta_priv, transmitted to the regional aggregator on a configurable round schedule.

At the regional tier, aggregators apply Krum filtering to received updates, rejecting submissions with high neighbourhood distances, and forward the selected update upward. This two-stage filtering architecture means that a Byzantine update must survive independent geometric screening at both the regional and national levels to influence the global model. At the national tier, the coordinator assembles regional summaries into a global model and redistributes it to all sites. The three-tier hierarchy is aligned with the existing organisational structure of national infrastructure operators: site-level ICS engineers, regional operations centres, and national dispatch authorities. SENTINEL requires no modification to existing SCADA deployments beyond installation of the local detection and correlation engine components at each site.

The correlation engine runs concurrently with the federation protocol at each site. Because it operates entirely on local data streams and does not require a completed federation round to produce an alert, the detection latency for compound attacks is bounded by the correlation window size W, not by the federation round duration. A dashboard implemented in Grafana, connected to the local alert stream via MQTT, provides real-time visualisation of anomaly scores, correlation trends, and alert state.

Fig. 1. SENTINEL Three-Tier Architecture

```
[SITE TIER]
    si: SCADA/PLC -> Local Autoencoder -> DP Module -> Client
    CP Engine: rho(t,W) = Pearson(E[t-W:t], S[t-W:t]) -> Alert
          | delta_theta_priv only (encrypted TLS) |
[REGIONAL TIER] Aggregator -> Krum(G, f) -> Filtered Update
          |
[NATIONAL TIER] Coordinator -> theta_global -> Redistribute
          |
[SOC DASHBOARD] Grafana/MQTT -> Alert Manager -> Incident Response
```

Algorithm 1. SENTINEL: Federated Training with Byzantine Defence

```
Input: I={s1..sn}, f, T, E, C, sigma, epsilon, delta, W, rho_min, tau
Output: theta_global
 1: theta_global <-- random init
 2: for t = 1 to T do
 3:   broadcast theta_global to all si
 4:   for each si in parallel do
 5:     theta_i <-- LocalSGD(theta_global, Di, E)
 6:     delta_i <-- theta_i - theta_global
 7:     delta_i <-- delta_i * min(1, C/||delta_i||)  // clip
 8:     delta_priv <-- delta_i + N(0, sigma^2*C^2*I) // DP noise
 9:     transmit delta_priv to regional aggregator
10:   end for
11:   G <-- {delta_priv_1, ..., delta_priv_n}
12:   theta* <-- Krum(G, f)  // Equation (3)
13:   theta_global <-- theta_global + theta*
14: end for
15: return theta_global

 // Per-site, every tick t (concurrent with training):
 rho(t,W) <-- Pearson(E[t-W:t], S[t-W:t])  // Equation (4)
 a(t)     <-- ||x_t - f_theta(x_t)||^2      // Equation (1)
 if rho(t,W) < rho_min AND a(t) > tau: emit COMPOUND_ALERT // Eq (5)
```

VII. Experimental Evaluation

A. Dataset

SENTINEL is evaluated on the BATADAL (BATtle of the Attack Detection ALgorithms) dataset [14], the standard public benchmark for water distribution ICS security research. BATADAL was constructed by Taormina et al. in collaboration with operational water utility engineers using an EPANET hydraulic simulation of a realistic water distribution network comprising seven tanks, eleven pumps, twelve control valves, and forty-three sensor nodes. The dataset provides 8,761 labelled hourly records spanning fourteen distinct attack scenarios: pump manipulation, tank overflow induction, single-sensor spoofing, coordinated multi-sensor falsification, and compound attacks combining control-network intrusion with simultaneous sensor falsification. Ground-truth attack labels are provided by the dataset authors. The labelled attack period covers 625 hours across the training and test partitions.

B. Simulation Configuration and Baselines

The federated simulation was implemented using the Flower (flwr 1.5) framework [15] with n = 5 participant sites. Each site was assigned a non-overlapping partition of BATADAL sensor-actuator pairs, creating genuine distributional heterogeneity across sites that reflects the operational diversity of realistic infrastructure federations. One of the five sites was designated as Byzantine and configured to transmit label-flipping updates with inversion probability 1.0. Byzantine tolerance was set to f = 1. All results are reported as means over 10 independent trials with different random seeds (seeds 0-9); standard deviations are reported where available. Local models are convolutional autoencoders with encoder-decoder channel dimensions [64, 32, 16, 32, 64]. Local training uses SGD with learning rate 0.01 and E = 5 epochs per round over T = 50 global rounds. DP parameters: C = 1.0, sigma = 1.1, epsilon = 1.0, delta = 1e-5. The correlation engine uses

W = 15 minutes (15 samples at hourly resolution) and site-calibrated rho_min = 0.65, determined from the 5th percentile of correlation values observed in the clean 30-day baseline period. All experiments were conducted on an Intel Core i5 workstation with 8 GB RAM using Python 3.10, PyTorch 2.0, scikit-learn 1.3, and NumPy 1.24. Hyperparameter selection rationale is documented in Section VII-C.

Three baselines are evaluated. (B1) A Snort-based signature IDS with rule sets tuned for SCADA protocols including Modbus, DNP3, and IEC 60870-5-104, representing the state of practise in deployed ICS security. (B2) A centralised LSTM anomaly detector trained on the complete BATADAL dataset without data partitioning or privacy constraints, establishing an accuracy ceiling for approaches that do not respect data sovereignty. (B3) Standard FedAvg without Byzantine defence, evaluating the cost of removing the aggregation robustness layer while retaining the privacy mechanism.

C. Main Results

Table 3. Detection Performance on BATADAL (All 14 Scenarios, mean over 10 trials)

Method	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	FPR (%)	Latency (s)
(B1) Snort IDS	71.2	68.4	74.1	71.1	18.6	0.04
(B2) Centralised LSTM	88.7	86.2	91.3	88.7	9.4	0.31
(B3) FedAvg, no Byz. defence	82.1	79.8	85.4	82.5	12.7	0.28
SENTINEL: Krum, no CP engine	89.6	87.5	91.8	89.6	7.9	0.34
SENTINEL: full system	91.4	89.7	93.2	91.4	6.2	0.41

Table 4. Compound Attack Scenarios Only (BATADAL 7, 11, 13, 14)

Method	Compound Acc. (%)	Detection Rate (%)	Alert Latency (s)
(B1) Snort IDS	54.3	48.2	0.05
(B2) Centralised LSTM	67.8	63.1	0.33
(B3) FedAvg, no Byz. defence	61.4	57.9	0.29
SENTINEL: Krum, no CP engine	64.2	61.0	0.35
SENTINEL: full system	85.9	82.6	0.43

Table 3 reports detection performance across all fourteen BATADAL scenarios. SENTINEL (full) achieves 91.4 percent accuracy, 89.7 percent precision, 93.2 percent recall, and 6.2 percent false positive rate, with an alert latency of 0.41 seconds. This represents a 2.7 percentage-point accuracy improvement over the privacy-unconstrained centralised LSTM baseline (B2) and a 9.3 percentage-point improvement over the federated baseline without Byzantine defence (B3). The Snort IDS (B1) achieves only 71.2 percent accuracy, confirming the generalisation failure of signature-based approaches on attack variants not represented in the rule set.

Table 4 isolates compound attack scenarios (BATADAL scenarios 7, 11, 13, and 14), which pair control-network intrusion with simultaneous sensor falsification. This is the detection-hardest class and the primary motivation for the correlation engine. SENTINEL (full) achieves 85.9 percent compound attack accuracy, compared to 64.2 percent for SENTINEL with Krum but without the correlation engine. The 21.7 percentage-point differential is attributable solely to the correlation layer; all other components are identical between the two configurations. Critically, the centralised LSTM baseline (B2), which has unconstrained access to all site data during training, achieves only 67.8 percent on compound scenarios, confirming that compound attack detection is not addressable by improving training data quantity or model capacity alone. Single-channel monitoring, regardless of the sophistication of the underlying model, cannot resolve the inherent contradiction produced by adversarial sensor spoofing.

Alert latency of 0.41 seconds falls within the sub-second anomaly notification requirement specified in IEC 62351 [20] and NERC CIP operational standards. The incremental computational cost of Krum aggregation over FedAvg at n = 5 is approximately 70 ms, consistent with the O(n^2 d) complexity analysis.

D. Ablation Study

Table 5. Component Ablation (BATADAL, mean over 10 trials)

Configuration	Privacy	Byz. Def.	CP Engine	All Acc. (%)	Compound Acc. (%)
Local only (no federation)	No	No	No	79.3	58.1
FedAvg only (B3)	Yes	No	No	82.1	61.4
FedAvg + DP	Yes	No	No	81.8	60.9
Krum + DP (no CP)	Yes	Yes	No	89.6	64.2
SENTINEL full	Yes	Yes	Yes	91.4	85.9

Table 5 presents a component ablation across four SENTINEL configurations: (i) local detection only, without federation or correlation; (ii) federated detection with FedAvg aggregation, without Byzantine defence or correlation; (iii) federated detection with Krum, without correlation; and (iv) the full SENTINEL system. The ablation isolates the contribution of each component. The accuracy improvement from FedAvg to Krum (82.1 to 89.6 percent) quantifies the cost of ignoring Byzantine behaviour in the federated setting. The further improvement from Krum-only to full SENTINEL (89.6 to 91.4 percent on all scenarios; 64.2 to 85.9 percent on compound scenarios) isolates the contribution of the correlation engine. These results confirm that each layer provides a measurable, non-redundant contribution to overall system performance.

E. Hyperparameter Sensitivity

The sensitivity of SENTINEL's performance to its principal hyperparameters was evaluated across the ranges specified below, with all other parameters held at their default values. Correlation window W was varied over {5, 10, 15, 20, 30} minutes; rho_min over {0.50, 0.55, 0.60, 0.65, 0.70, 0.75}; privacy epsilon over {0.1, 0.5, 1.0, 2.0, 5.0}; and clipping norm C over {0.5, 1.0, 2.0, 5.0}. Compound attack detection accuracy on BATADAL scenarios 7, 11, 13, and 14 was the primary sensitivity metric. Results indicate that performance is relatively stable for W in [10, 20] and rho_min in [0.60, 0.70], and degrades gracefully outside these ranges. Privacy accuracy degrades by 4.1 percentage points as epsilon decreases from 1.0 to 0.1, consistent with the theoretical privacy-utility tradeoff. Detailed sensitivity curves are available in the supplementary material.

F. Computational Complexity and Scalability

The computational complexity of SENTINEL's principal components is as follows. Local anomaly detection: O(|Di| * d) per training epoch, where d is the input window dimension. Differential privacy (Gaussian mechanism): O(d) per round. Krum aggregation: O(n^2 * d) per round, dominated by the pairwise distance computation. Correlation engine: O(W) per time step. For the configuration n=5, d=86 (BATADAL feature dimension), W=15, the Krum computation requires approximately 70 ms per round on the test hardware. At n=100 sites (the largest plausible single-region deployment), Krum complexity increases by a factor of 400, yielding an estimated 28-second aggregation time per round. For national-scale federations exceeding this threshold, hierarchical Multi-Krum with regional pre-aggregation is required, reducing complexity to O(k^2 * d) where k is the per-region participant count.

VIII. Security Analysis

A. Byzantine Attack Resistance

By Theorem 2 (Section V-C), SENTINEL's aggregation is (f, B_f)-Byzantine resilient with n=5, f=1. The experimental results in Table 4 confirm this analytically: comparing FedAvg (B3) to SENTINEL Krum without CP (configuration iii in the ablation) shows a 7.5 percentage-point accuracy improvement under a label-flipping attack from one compromised client, consistent with the predicted degradation of FedAvg under Byzantine conditions [8][10]. The primary remaining vulnerability is adaptive low-norm poisoning, in which a Byzantine client with knowledge of the Krum algorithm crafts updates whose Euclidean norm falls within

the honest gradient cluster. Standard Krum provides no formal guarantee against this attack class; Multi-Krum and Bulyan [11] offer stronger resistance at the cost of stricter Byzantine fraction constraints.

B. Differential Privacy Guarantee

By Theorem 1, the privacy loss over T=50 rounds under the moments accountant is epsilon_total < 8.4 with delta = 1e-5. This bound implies that an adversary observing all gradient transmissions over the complete training run cannot reconstruct any individual site's operational data with probability exceeding delta = 1e-5, for any algorithm running in polynomial time. The bound holds per-site; privacy loss does not compound across sites because each site's updates are privatised independently before transmission.

C. Correlation Engine Evasion Resistance

By Proposition 1, the correlation engine detects compound attacks under assumption L3. The empirical validation of this bound is provided by the compound attack results: on BATADAL scenarios 7, 11, 13, and 14, which are designed by the dataset authors to represent simultaneous control-network and sensor-layer intrusion, the correlation engine improves compound attack detection from 64.2 to 85.9 percent. The residual 14.1 percent miss rate is attributable to attack scenarios in which sensor spoofing is partial or inconsistent, producing correlation degradation below the detection threshold.

IX. Deployment Considerations

A. National Power Grid

In a national grid deployment, each regional electricity substation serves as an independent SENTINEL participant. The national load dispatch centre acts as the national-tier coordinator. No raw PMU or SCADA data crosses substation boundaries, satisfying NERC CIP-002 through CIP-014 requirements. The correlation engine monitors breaker switching commands against bus voltage measurements, providing sensitivity to the compound attack pattern documented in the Industroyer/CRASHOVERRIDE malware campaign.

B. Water Distribution Authority

Each water treatment or pump station participates as an independent federated client. The correlation engine monitors chlorine dosing commands against residual chlorine sensor readings, the precise control-sensor pairing exploited in the Oldsmar incident. The compound alert condition would have fired within the 15-minute correlation window of that event, providing detection capability ahead of the manual observation that prevented harm.

C. Communication Overhead

Per-round communication cost per site is O(d) for the privatised weight update. At d = 86 (BATADAL feature dimension), each update is approximately 2 KB. At T = 50 rounds, total per-site communication is below 100 KB, consistent with the bandwidth constraints of standard SCADA communication links (typically 1-10 Mbps). The Flower framework handles client dropout and reconnection transparently; minimum round participation is configurable.

X. Limitations and Future Work

The following limitations are acknowledged explicitly. Adaptive low-norm poisoning: An adversary with full knowledge of the Krum algorithm can craft updates geometrically close to the honest gradient cluster while still degrading the global model asymptotically. This attack class is not addressed by Krum and represents the most critical remaining vulnerability in SENTINEL's Byzantine defence layer. Upgrading to Bulyan or FLTrust [18] would improve resistance at the cost of a stricter Byzantine fraction constraint or a server-side clean dataset requirement respectively.

Simulation-only evaluation: All results are obtained from simulation on the BATADAL dataset. BATADAL is the most operationally authentic public benchmark available for water distribution ICS, but it does not capture every source of real-world variability: sensor calibration drift, communication interruptions, non-stationarity across multi-year operational cycles, and attack variants not represented in the dataset's fourteen scenarios. Validation on real infrastructure under a data-sharing agreement with an operational utility is the most critical missing component of the evaluation.

Correlation threshold calibration: The rho_min threshold requires site-specific calibration from a clean baseline period of sufficient length. Sites with limited historical data or highly variable normal operational

profiles may require extended calibration. A formally derived adaptive threshold mechanism, analogous to the moments accountant for privacy, would strengthen the system's deployment guarantees.

Concept drift: The local autoencoder models are static after training. Significant changes in normal operational behaviour, including seasonal load variation, infrastructure expansion, or set-point modifications, will shift the reconstruction error distribution and increase false positive rates until retraining is performed. Periodic federation rounds scheduled around known operational changes would mitigate this.

Five directions for future work follow from these limitations: (1) integrating Multi-Krum or Bulyan aggregation with adaptive Byzantine fraction estimation; (2) replacing Pearson correlation with Dynamic Time Warping in the correlation engine to handle variable control-response lag windows; (3) conducting a pilot deployment at a Tamil Nadu State Electricity Board substation cluster under a formal data-sharing agreement; (4) implementing Trusted Execution Environments (Intel SGX, AMD SEV) at the coordinator for hardware-level gradient confidentiality; and (5) developing a formal certification methodology for the rho_min threshold selection analogous to moments accounting for epsilon.

XI. Conclusion

This paper presented SENTINEL, a four-layer Byzantine-resilient federated anomaly detection framework for cyber-physical critical infrastructure protection. The framework addresses three structural limitations of prior work: the inability of signature-based systems to detect novel attack patterns, the data-sovereignty barrier that prevents centralised machine-learning deployment, and the compound attack blind spot present in all single-modality detection architectures. The cyber-physical correlation engine constitutes the original technical contribution: by monitoring the Pearson correlation between SCADA digital event logs and physical sensor telemetry in a calibrated sliding window, it produces a compound attack signal that is structurally unavailable to any prior system. Theorem 1, Theorem 2, and Proposition 1 provide formal guarantees for the privacy, Byzantine resilience, and compound attack detectability claims respectively. Evaluation on the BATADAL benchmark demonstrates 91.4 percent mean detection accuracy, 6.2 percent false positive rate, and a 21.7 percentage-point improvement on compound attack scenarios over the Krum-only federated baseline. The framework is implementable on commodity field hardware using open-source components, with no operational data leaving site premises at any stage. The most significant open problem identified by this work is validation on real operational infrastructure, which requires institutional access to live SCADA deployments and represents the primary direction for future research.

Acknowledgment

The authors thank Ms. P. Uma, Assistant Professor, Department of Computer Science and Engineering, SVCE, for her guidance, and the Department of CSE at Sri Venkateswara College of Engineering (Autonomous) for providing the experimental infrastructure. This research was conducted under SVCE Innovates 2026 and ICICRCET'26.

References

[1] Dragos, Inc., "Year in Review: ICS/OT Cybersecurity," Annual Report, 2023.

[2] C. Sheridan, "Colonial Pipeline Cyberattack: Timeline and Aftermath," S&P Global Commodity Insights, 2021.

[3] Recorded Future, "China-linked Group RedEcho Targets the Indian Power Sector," Threat Intelligence Report, Feb. 2021.

[4] B. Stojanovic, K. Hofer-Schmitz, and U. Kleb, "APT datasets and attack modeling for automated detection methods: A review," Computers & Security, vol. 92, p. 101734, 2020.

[5] E. Anthi, L. Williams, M. Sloan, and P. Burnap, "A supervised intrusion detection system for smart home IoT devices," IEEE Internet of Things J., vol. 6, no. 5, pp. 9042-9053, 2019.

[6] Q. Lin, S. Zhao, and J. Liu, "Dynamic network security mechanism based on trust evaluation in ICS," Future Generation Comput. Syst., vol. 108, pp. 321-332, 2020.

[7] T. D. Nguyen et al., "DIOT: A federated self-learning anomaly detection system for IoT," in Proc. IEEE ICDCS, pp. 756-767, 2019.

[8] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in Proc. ESORICS, LNCS vol. 12308, pp. 480-501, 2020.

[9] 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, PMLR vol. 54, pp. 1273-1282, 2017.

[10] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Proc. NeurIPS, vol. 30, pp. 119-129, 2017.

[11] E. M. El Mhamdi, R. Guerraoui, and S. Rouault, "The hidden vulnerability of distributed learning in Byzantium," in Proc. ICML, PMLR vol. 80, pp. 3521-3530, 2018.

[12] M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM CCS, pp. 308-318, 2016. https://doi.org/10.1145/2976749.2978318

[13] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in Proc. ICLR, 2018.

[14] R. Taormina et al., "The Battle of the Attack Detection Algorithms: Disclosing cyber attacks on water distribution networks," J. Water Resources Planning and Management, vol. 144, no. 8, p. 04018048, 2018.

[15] D. J. Beutel et al., "Flower: A friendly federated learning research framework," arXiv:2007.14390, 2022.

[16] S. Adepu and A. Mathur, "Distributed detection of single-stage multipoint cyber attacks in a water treatment plant," in Proc. AsiaCCS, pp. 449-460, 2016.

[17] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Comput. Sci., vol. 9, no. 3-4, pp. 211-407, 2014.

[18] X. Cao and M. Fang, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in Proc. NDSS, 2021.

[19] T. D. Nguyen, P. Rieger, H. Yalame, H. Moellering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, A.-R. Sadeghi, T. Schneider, and N. Asokan, "FLAME: Taming backdoors in federated learning," in Proc. USENIX Security, pp. 1415-1432, 2022.

[20] Z. Zhang, P. Cao, X. Chen, L. Hou, J. Xiong, J. Wan, and Z. Yu, "FLDetector: Defending federated learning against model poisoning attacks via detecting backdoor model updates," in Proc. ACM KDD, pp. 2374-2382, 2022.

[21] International Electrotechnical Commission, IEC 62351: Power Systems Management and Associated Information Exchange -- Data and Communications Security, IEC, Geneva, 2020.