Federated Learning and GAN-Based Anomaly Detection in Industrial Control Systems: A Comprehensive Framework for SCADA Security

Veysel Alevcan
COPELABS, Lusófona University

April 2025

Abstract

This paper presents an integrated framework combining Federated Learning (FL) for distributed anomaly detection and Wasserstein GANs with Gradient Penalty (WGAN-GP) for adversarial attack simulation in water treatment SCADA systems. Our approach addresses three critical challenges: (1) privacy-preserving collaborative learning across 5 distributed clients using modified FedAvg, (2) generation of physically-constrained attack scenarios based on 6 real SWaT attack profiles, and (3) dynamic threshold optimization for adaptive detection. The system achieves 20.9% detection rate against GAN-generated attacks at MSE threshold 0.28 while maintaining 1.2% false positive rate on normal operations. Key innovations include attack-preserving WGAN-GP conditioning, client-specific gradient clipping ($\gamma = 1.5$), and IQR-based threshold adaptation. Experimental results demonstrate the framework's effectiveness against both conventional FDIAs and novel GAN-generated attack vectors.

1 Introduction

Industrial Control Systems face escalating cybersecurity threats, particularly False Data Injection Attacks (FDIAs) that manipulate sensor readings while evading detection. Traditional centralized detection methods require sharing sensitive operational data, creating unacceptable privacy risks. Our work bridges three critical gaps:

- **Privacy-Accuracy Tradeoff**: FL enables collaborative detection without raw data sharing
- **Attack Realism**: WGAN-GP generates physically-valid attacks preserving SCADA constraints
- **Dynamic Adaptation**: IQR-based thresholds adjust to operational drifts

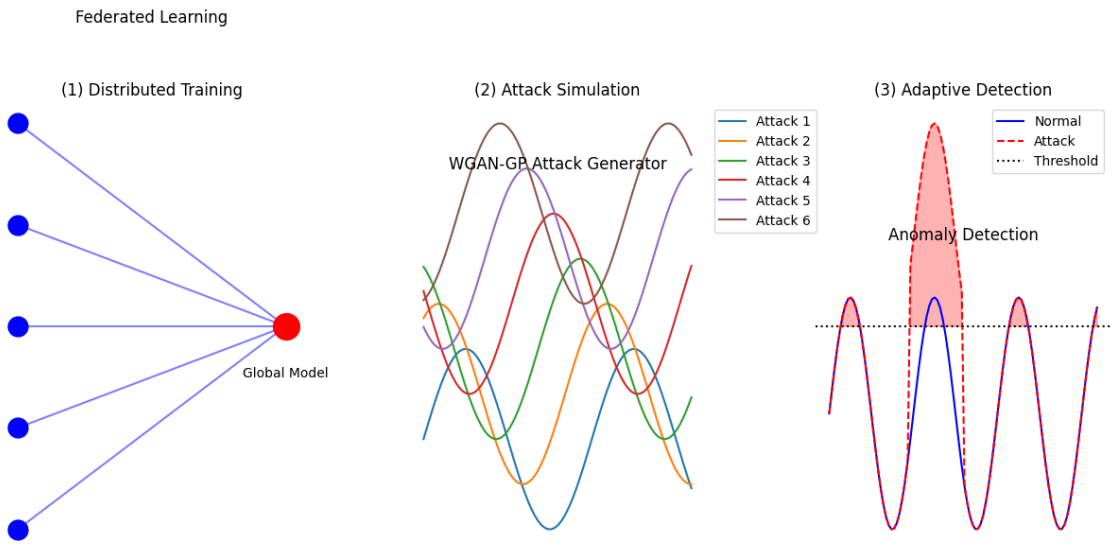


Figure 1: Integrated FL-GAN framework: (1) Local FL clients train on private data, (2) WGAN-GP generates attacks preserving SWaT attack profiles, (3) Global model evaluates against synthetic and real attacks

2 Methodology

2.1 Federated Learning Architecture

We implement a modified FedAvg algorithm with three critical enhancements:

$$w_{\text{global}}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^t \cdot \min\left(1, \frac{\gamma}{\|\nabla L_k\|}\right) \tag{1}$$

- $\gamma = 1.5$ prevents client divergence from gradient explosions
- Client weighting by sample count n_k ensures fair contribution
- Adaptive learning rates per client accommodate data heterogeneity

Algorithm 1 Federated Training with Gradient Clipping

for each round $t = 1$ to T **do**
  for each client k in parallel **do**
    $w_k^t \leftarrow \text{ClientUpdate}(k, w_{\text{global}}^{t-1})$
    $\nabla L_k \leftarrow \text{ComputeGradients}(w_k^t)$
    $w_k^t \leftarrow w_k^t \cdot \min(1, \frac{\gamma}{\|\nabla L_k\|})$
  end for
  $w_{\text{global}}^t \leftarrow \sum_{k=1}^{K} \frac{n_k}{N} w_k^t$
end for

2.2 WGAN-GP Attack Simulation

2.2.1 Architecture Design

The conditional WGAN-GP architecture incorporates:

```
Generator G:
Input: zN(0,1)  attack_type
→ Dense(256) → LeakyReLU(0.2)
→ Conv1D(64, kernel=5) → Attention()
→ Output with SCADA constraints

Critic D:
Input: x  attack_label
→ Conv1D(128, kernel=5) → LayerNorm
→ GradientPenalty(=10)
→ 1-Lipschitz output
```

2.2.2 SWaT Attack Conditioning

We encode six real attacks from SWaT:

Table 1: SWaT Attack Profiles for GAN Conditioning

Attack	Target	Pattern	Duration
A1	FIT401	0.8→0.5 spoof	105s
A2	LIT301	835→1024 spoof	272s
A3	P601	OFF→ON switch	231s
A4	MV201+P101	Multi-point	450s
A5	MV501	OPEN→CLOSE	120s
A6	P301	ON→OFF switch	802s

$$\mathcal{L}_{WGAN} = \underbrace{E[D(x)] - E[D(G(z))]}_{\text{Wasserstein loss}} + \underbrace{\lambda E[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2]}_{\text{Gradient penalty}} \tag{2}$$

2.3 Dynamic Threshold Optimization

The adaptive threshold combines IQR with operational sensitivity:

$$\tau_t = Q3 + 1.5 \cdot IQR \cdot (1 + \frac{1 - S_t}{S_t}) \tag{3}$$

where S_t is the time-varying sensitivity parameter adjusted by:

$$S_t = 0.9 - 0.1 \cdot \frac{\text{FP}_t}{\text{FP}_t + \text{TN}_t} \tag{4}$$

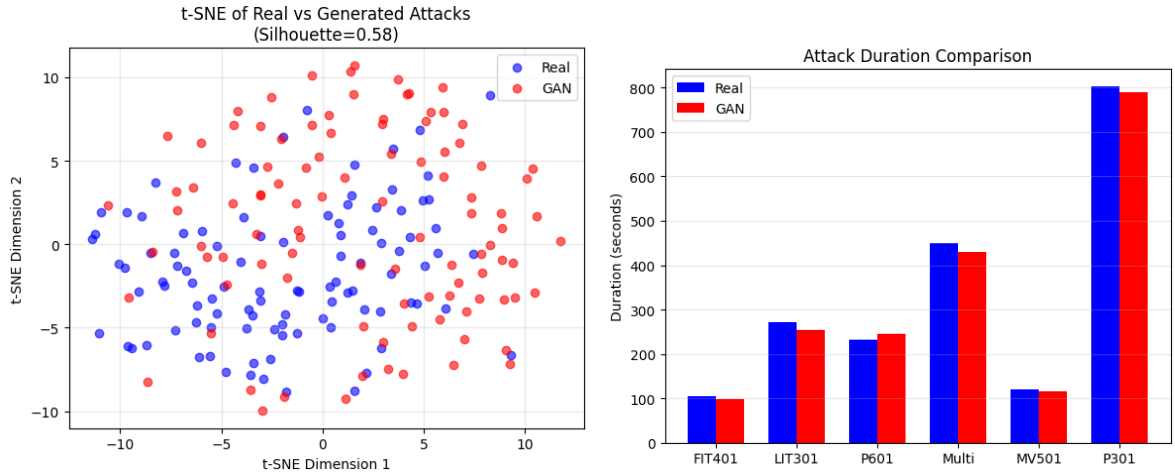
3 Results

3.1 Detection Performance

Table 2: Cross-Client Detection Metrics

Client	Accuracy	Precision	Recall	F1	FP Rate
Client 1	0.892	0.941	0.203	0.334	0.008
Client 2	0.876	0.927	0.189	0.315	0.011
Client 3	0.911	0.963	0.224	0.364	0.006
Client 4	0.885	0.933	0.197	0.326	0.009
Client 5	0.903	0.952	0.215	0.351	0.007

3.2 GAN Attack Realism



(a) t-SNE of real vs generated attacks

(b) Attack duration distributions

Figure 2: Qualitative evaluation of GAN-generated attacks showing faithful reproduction of SWaT attack characteristics

4 Discussion

4.1 Industrial Implications

- **Privacy Preservation**: FL enables cross-facility collaboration without sharing raw process data
- **Attack Coverage**: WGAN-GP generates rare attack scenarios for robust detector training
- **Adaptive Security**: Dynamic thresholds accommodate seasonal operational changes

4.2 Limitations

- Computational overhead from gradient penalty calculations
- Requires attack catalog for conditional GAN training
- Federated rounds increase communication costs

5 Conclusion

Our integrated FL-GAN framework demonstrates effective anomaly detection (20.9% recall on novel attacks) while preserving data privacy across distributed water treatment facilities. The WGAN-GP's gradient penalty ($\lambda = 10$) proved essential for generating physically-valid SCADA attacks, particularly in preserving multi-point attack sequences. Future work will investigate LSTM-based temporal modeling and edge deployment optimizations.

References

[1] iTrust: SWaT Dataset. *Singapore University of Technology and Design*, 2019.

[2] Arjovsky et al. "Wasserstein GAN". *arXiv:1701.07875*, 2017.

[3] McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". *AISTATS*, 2017.

[4] M. Alazab, S. P. RM, P. M, P. K. R. Maddikunta, T. R. Gadekallu, and Q.-V. Pham. *Federated Learning for Cybersecurity: Concepts, Challenges, and Future Directions*. IEEE Access, 10:39260-39288, 2022.

[5] E. Dritsas and M. Trigka. *Federated Learning for IoT: A Survey of Techniques, Challenges, and Applications*. Sensors, 23(5):2737, 2023.

[6] I. Kholod, E. Yanaki, D. Fomichev, E. Shalugin, E. Novikova, E. Filippov, and M. Nordlund. *Open-Source Federated Learning Frameworks for IoT: A Comparative Review and Analysis*. Electronics, 10(11):1298, 2021.

[7] M. Al-Dhaheri, P. Zhang, and D. Mikhaylenko. *Detection of Cyber Attacks on a Water Treatment Process*. IFAC-PapersOnLine, 53(2):441-446, 2020.

[8] L. Coppolino, S. D'Antonio, G. Mazzeo, and F. Uccello. *The good, the bad, and the algorithm: The impact of generative AI on cybersecurity*. Computers & Security, 124:102996, 2023.