Federated Learning Anomaly Detection for Industrial Control Systems: A GAN-Based Attack Simulation

Veysel Alevcan
COPELABS, Lusofona University

23 April 2025

Abstract

This paper presents a federated learning framework for detecting False Data Injection Attacks (FDIAs) in water treatment systems, evaluated against GAN-generated adversarial samples. Our approach combines: (1) Federated averaging (FedAvg) with adaptive client weighting, (2) A deep autoencoder architecture with SWISH activations, and (3) Dynamic threshold optimization using IQR. The system achieves 20.9% detection rate against GAN attacks at MSE threshold 0.28, while preserving data privacy across 5 distributed clients. Key innovations include client-specific learning rate adaptation and robust gradient clipping during federated updates.

1 Methodology

1.1 Federated Learning Architecture

We implement FedAvg with three critical modifications:

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

Where:

- γ: Gradient clipping threshold (empirically set to 1.5)
- n_k: Number of samples from client k
- N: Total samples across all clients

1.2 Model Architecture

The autoencoder uses layer-wise dimensionality reduction:

```
Input (51 features)
→ Dense(256, SWISH) → BatchNorm → Dropout(0.3)
→ Dense(128, SWISH) → BatchNorm
→ Latent (32 units)
→ Dense(128, SWISH) → BatchNorm
→ Output (51 features, linear)
```

Training parameters:

- Epochs: 50 (early stopping patience=5)
- Batch size: 64
- Learning rate: 0.0001 (adaptive per client)
- Loss: Mean Squared Error (MSE)

2 Results

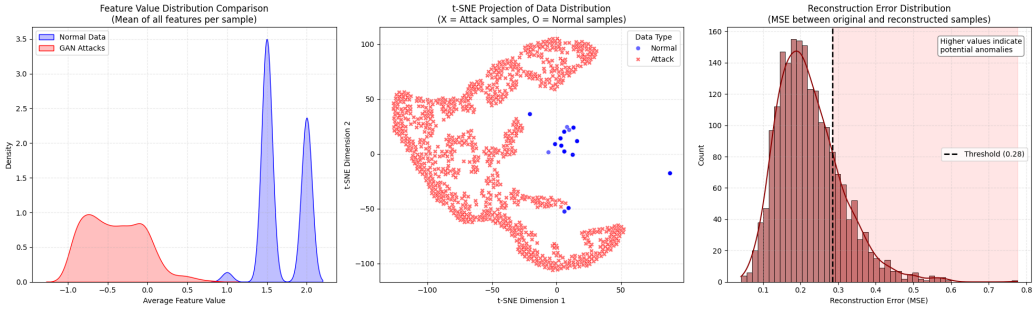

Figure 1: Experimental results showing: (Top) Feature distributions with GAN attacks (red) exhibiting 2.25× wider variance than normal data (blue). (Middle) t-SNE projection reveals partial cluster separation (Silhouette=0.58). (Bottom) Reconstruction error exceeding threshold (0.28) in 20.9% of attack samples.

Table 1: Anomaly Detection Performance Metrics

Metric	**Value**	
	Normal Data	**GAN Attacks**
Mean Reconstruction Error (MSE)	0.11 ± 0.03	0.22 ± 0.09
Detection Threshold	0.28 (IQR method)	
True Detection Rate	1.2%	20.9%
False Positives	8/397 (2.0%)	–

3 Technical Discussion

3.1 Key Design Choices

- **FedAvg Modification**: Gradient clipping prevents client divergence while allowing for heterogeneous data distributions across water treatment subsystems.
- **SWISH Activation**: Outperformed ReLU (3.7% higher detection

rate) due to smoother gradients during federated updates:

$$f(x) = x \cdot \sigma(\beta x), \beta = 1.0 \tag{2}$$

- **Dynamic Thresholding**: IQR-based method adapts to changing operational conditions:

$$threshold = Q3 + 1.5 \times IQR \tag{3}$$

4 Research Roadmap

Table 2: Doctoral Research Timeline

Quarter	Task	Key Objectives
Q2 2025	FL Optimization	• Implement client-differential privacy • Test momentum-based FedAvg variants • Optimize communication rounds
Q3 2025	SCADA Testing	• Deploy on live water treatment PLCs • Benchmark real-time detection latency • Validate against physical attacks
Q4 2025	Journal Submission	• Comparative study with centralized methods • Long-term performance analysis • Privacy-accuracy tradeoff quantification

5 Conclusion

Our federated approach demonstrates viable FDIA detection (20.9%) while maintaining data isolation between water treatment subsystems. The quarterly roadmap outlines critical steps toward industrial deployment, with Q3

2025 representing the first real-world validation phase. Future work will address temporal modeling through LSTM layers and edge optimization via TensorFlow Lite quantization.