Poster: Unsupervised Attack Classification in Smart Grid AGC Using Variational Autoencoder Gradient Profiles

Tran L.T. Le, David K.Y. Yau, Song Qun
Singapore University of Technology and Design

***Abstract*—We propose an unsupervised machine learning approach for classifying cyberattacks on smart power grids, with a focus on multi-area AGC systems. Unlike prior work on attack design or detection, our method is the first to classify advanced attacks without labeled data. By analyzing internal gradients from a VAE combined with a TCN, we distinguish among time delay (TDA) and two types of false data injection (FDI) attacks. Simulations using PowerWorld show that K-means clustering on these gradients achieves over 95% accuracy—comparable to supervised methods but without costly labeling. Our approach also detects zero-day attacks as distinct clusters using equilibrium K-Means.**

***Index Terms*—Smart grid cybersecurity; automatic generation control (AGC); Time delay attack; False data injection; Attack classification, gradient clustering.**

I. Introduction

Automatic Generation Control (AGC) maintains grid stability by adjusting generator outputs to balance supply and demand. With increasing digitization, AGC faces cyber threats like Time Delay (TDA) and False Data Injection (FDI), prompting research on attack design and detection [4], [6], [7]. Detection methods include model-based and data-driven approaches: model-based methods struggle with system uncertainty, while data-driven ones require labeled data. Unsupervised methods like VAEs can detect anomalies (e.g., TDA) but cannot classify attack types [3].

Most existing methods treat all attacks as generic anomalies, making it hard to distinguish between types—especially under simultaneous threats. Yet, each attack leaves distinct patterns, and accurate classification is key to targeted responses.

This work proposes an unsupervised classification framework that extracts gradient features from a TCN-VAE trained on normal data and clusters them using K-Means and eK-Means. The main contributions are:

1) Introduces gradient profiles for accurate, label-free classification of AGC attacks.
2) Presents a fully unsupervised framework achieving over 95% accuracy using TCN-VAE gradients and K-Means—without attack labels.
3) Demonstrates adaptability by grouping zero-day attacks into distinct clusters using eK-Means.

II. System and Attack Model

A. AGC Model

We simulate a 37-bus, three-area power system in PowerWorld [1], representative of small to mid-scale national grids (common in over half of 130 grids analyzed). The setup follows POSOCO's Indian Grid Standard [8], with each area running AGC for frequency control via tie-lines. This realistic CPS scenario captures key dynamics, including noise, biases, and setpoint disturbances.

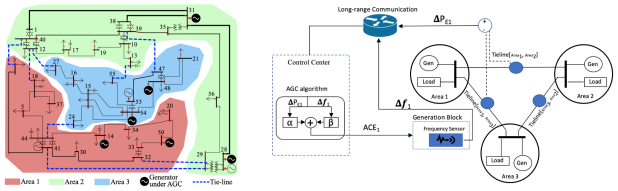

Fig. 1. **System Design and AGC Model: Left: Three-area power grid in PowerWorld simulator, interconnected via tie-lines. Right: AGC model at the control center computes Area Control Error (ACE) using frequency (Δf_i) and power export deviations (ΔP_{Ei}) from each area.**

B. Attack Models and Objectives

AGC's dependence on wide-area communication makes it vulnerable to attacks on sensor data and ACE signals. Using the system in Fig. 1, we simulate TDA, FDI-optimal, FDI-adaptive, and zero-day attacks to assess their distinct impacts [2], [4], [6], [7].

*1) **Time Delay Attack**:* TDA methods—such as channel jamming, router compromise, or Man-in-the-Middle interception—delay control signals between AGC controllers and actuators by buffering packets. This disrupts real-time responses and threatens system stability, especially when delays are introduced after a specific time point.

*2) **False Data Injection (Optimal)**:* As defined in [6], attack effectiveness is measured by Total Time-to-Emergency (TTE). FDI-optimal attacks aim to minimize TTE by carefully structuring attack sequences under constraints. Due to system inertia, a single-cycle attack may not suffice, prompting attackers to adopt persistent, multi-cycle strategies.

2325-1301/25/$31.00 ©2025 IEEE
DOI 10.1109/RTCSA66114.2025.00036

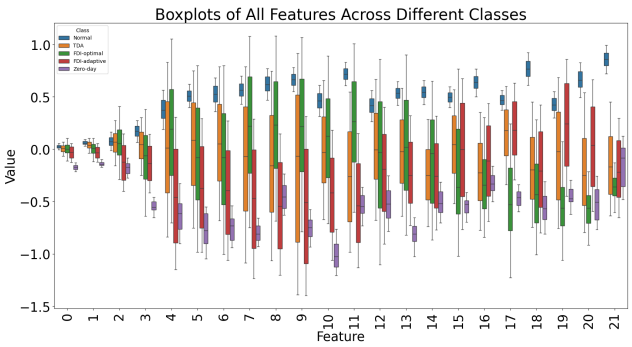

Fig. 2. **Gradient Deviations Across Attack Types:** Normal (0), TDA (1), FDI-optimal (2), FDI-adaptive (3), and Zero-day (4) show distinct gradient profiles with clearly separable mean and variance patterns.

*3) **False Data Injection (Adaptive)**:* The FDI-adaptive attack [7] uses adaptive control to steer frequency deviation $\omega(t)$ toward a target $d(t)$. It operates in two phases: Phase I mimics normal behavior to remain stealthy; Phase II drives $\omega(t)$ beyond safe limits (e.g., > 0.5 Hz).

*4) **Zero-day Attack** [2]:* We extend our evaluation with a zero-day attack by adding an attack vector $\mathbf{a}$ to power flow measurements $\mathbf{z}$ and injecting a random signal (0.1–1.0) into frequency measurements.

III. Methodology

We train TCN-VAE on normal tie-line data, which responds more quickly to anomalies than frequency signals [3]. TCNs with dilated convolutions efficiently capture long-term temporal patterns [5]. Our model uses a single TCN block (2 stacks, kernel size 3, 32 filters, dilation 2–8), 16 latent dimensions, and 48-point sliding windows. The VAE loss—combining reconstruction error and KL divergence—enables learning informative gradient features. When tested on attack data, these gradients diverge from the normal pattern, revealing attack-type differences (Fig. 2). To reduce noise and retain key features, we apply PCA (95% variance retained), resulting in compact gradient representations.

We then apply clustering algorithms—K-Means, equilibrium K-Means [9], and DBScan—to group these features for unsupervised attack classification.

IV. Evaluation and Results

TABLE I
Unsupervised Clustering Results

Model	Method	Parameters	ARI ↑	NMI ↑
TCN-VAE	**K-Means (ours)**	$K = 4$	**0.9612**	**0.9462**
	eK-Means	$K = 4$, RS = 0.3	0.8793	0.8586
	DBSCAN	$\varepsilon = 0.95$, min_samples = 100	0.2581	0.4710

We evaluate unsupervised clustering on gradient representations from TCN-VAE. Among the methods tested, K-Means with K=4 produces the most distinct and interpretable clusters. TCN-VAE gradients result in coherent cluster assignments, as indicated by significantly higher Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) scores—demonstrating effective separation of anomalies and attack types. Both metrics are particularly useful for evaluating clusters dominated by a single class.

The confusion matrix in Fig. 3 highlights improved clustering performance of eK-Means in a test case involving zero-day attacks with very few samples. While both methods perform similarly in balanced scenarios without zero-day attacks, eK-Means shows clear advantages in handling imbalanced cases, thanks to its adaptability to varying cluster densities and shapes.

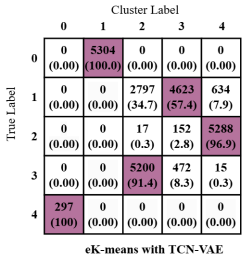

Fig. 3. The eK-Means result shows the zero-day attack forms a distinct cluster labeled 0.

V. Conclusion

We present an unsupervised approach that uses gradients from TCN-VAE to classify cyber-attacks in power systems without relying on attack labels. By applying K-Means to gradient features, the method achieves up to 95% accuracy and successfully detects zero-day attacks using only normal operating data. This demonstrates strong potential for real-world deployment in scenarios where labeled attack data is limited, highlighting the effectiveness of gradient-based features for power system cybersecurity.

References

[1] PowerWorld The visual approach to electric power systems.
[2] Chunyu Chen, Kaifeng Zhang, Kun Yuan, Lingzhi Zhu, and Minhui Qian. Novel detection scheme design considering cyber attacks on load frequency control. *IEEE Transactions on Industrial Informatics*, 14(5):1932–1941, 2018.
[3] G. Shahram et al. Time Delay Attack Detection Using Recurrent Variational Autoencoder and K-means Clustering. In *2021 IEEE PES Innovative Smart Grid Technologies-Asia (ISGT Asia)*. IEEE, 2021.
[4] L. Xin et al. Assessing and mitigating impact of time delay attack: Case studies for power grid controls. *IEEE Journal on Selected Areas in Communications*, 38(1), 2020.
[5] S. Bai et al. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018.
[6] T. Rui et al. Modeling and mitigating impact of false data injection attacks on automatic generation control. *IEEE Transactions on Information Forensics and Security*, 12(7):1609–1624, 2017.
[7] Y. Weili et al. A stealthier false data injection attack against the power grid. In *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2021.
[8] J. Duncan Glover, Mulukutla S. Sarma, and Thomas Overbye. *Power system analysis & design, SI version*. Cengage Learning, 2012.
[9] Yudong He. Imbalanced data clustering using equilibrium k-means. *arXiv preprint arXiv:2402.14490*, 2024.