

# Drift-Aware Online Anomaly Detection in Smart Buildings via Temporal Variational Autoencoder Gradient Profile

Tran L. T. Le<sup>1,3</sup>, Heng Chuan Tan<sup>1</sup>, Zhen Wei Ng<sup>1</sup>, David Yau<sup>3</sup>, Zbigniew Kalbarczyk<sup>1,2</sup>, Daisuke Mashima<sup>3</sup>, Xin Lou<sup>4</sup>

<sup>1</sup>Illinois Advanced Research Science Center at Singapore, Singapore

<sup>2</sup>Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA

<sup>3</sup>Singapore University of Technology and Design, Singapore

<sup>4</sup>Singapore Institute of Technology, Singapore

## 1. The Challenge: Security in Evolving Environments

Smart buildings and Cyber-Physical Systems (CPS) are not static. They operate under constantly changing conditions—sensor drift, seasonal updates, and varying occupancy.

- **The Conflict:** Conventional Anomaly Detection (AD) models struggle to distinguish between normal drift and malicious cyber-attacks.
- **The Risk:**
  - **False Alarms:** Models flag normal drift as attacks.
  - **Model Contamination:** Adaptive models blindly update on all new data, accidentally learning from attack data and normalizing the threat.
- **Our Goal:** Develop an online AD framework that explicitly separates *Normal*, *Normal Drift*, and *Attack* behaviors to enable safe model updates.

## 2. Main Methodology: Drift-Aware TCN-VAE

We introduce a framework combining a Temporal Convolutional Network (TCN) with a Variational Autoencoder (VAE), utilizing input-gradients as behavioral fingerprints.

### A. The Backbone: TCN-VAE

We use a TCN encoder/decoder to capture long-range temporal dependencies in sensor data. The VAE regularizes the latent space, making the model robust to minor noise.

### B. The Innovation: Gradient Profiles

Instead of relying solely on reconstruction error, we analyze Input Gradients  $\nabla_x L(x)$

- **Why?** Gradients measure the sensitivity of the reconstruction loss to input features.
- **Insight:** Normal changes induce structured gradient patterns, while attacks produce abrupt, distinctive sensitivity footprints.

### C. Selective & Contamination-Free Updates

**1.Cluster & Classify:** We use K-Means (mapped via the Hungarian algorithm) to cluster gradient profiles into semantic classes: *Normal*, *Drift*, or *Attack*.

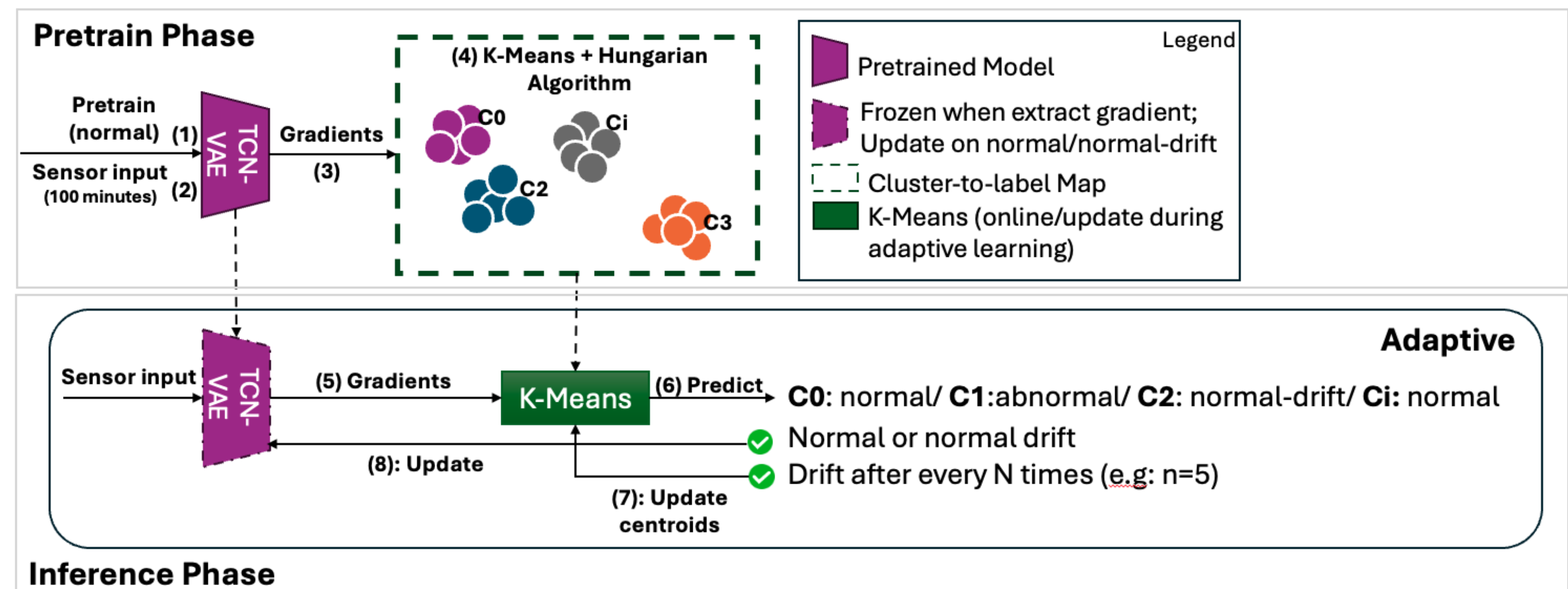
**2.Selective Adaptation:** The model **only updates** on samples classified as *Normal* or *Normal Drift*.

**3.Reject Attacks:** Samples classified as *Attack* are discarded to prevent model corruption.

## 3. Generalization Strategy (Cross-Floor)

To deploy the pre-trained model across different building floors without retraining, we address the issue of varying behavioral granularity.

- **Challenge:** A fixed number of clusters (e.g.,  $K=3$ ) works for a known floor but fails when transferring to a new floor with different room usage (e.g., meeting rooms vs. offices).
- **Solution (Adaptive Clustering):** We apply X-means clustering strictly during the cross-floor evaluation phase. This automatically determines the optimal number of clusters to capture the new environment's variability



## 4. Experimental Results

We evaluated the framework on a real-world Building Management System (BMS) dataset, specifically testing against temperature setpoint attacks on HVAC units.

Table 1: Summary of datasets used for model evaluation

Dataset	Collection Period	Total	Normal	Normal Drift	Abnormal
Level 4 Office area (vacant)	Jul 2 – Sep 8, 2025	97,062	27,256 (28.08%)	66,918 (68.94%)	2,888 (2.98%)
Level 14 MR-1 (occupied)	Sep 24 – Oct 8, 2025; Oct 15 – Dec 1, 2025	87,010	26,737 (30.72%)	55,765 (64.09%)	4,508 (5.18%)
Level 14 MR-2 (occupied)	Sep 17 – Oct 8, 2025; Oct 15 – Dec 1, 2025	98,141	23,426 (23.87%)	62,635 (63.82%)	12,080 (12.31%)

Table 2: Performance of the specialized model under different update strategies ( $K=3$ ) (X: no update, ✓: full update, S: selective update)

Datasets	Update	AUC ↑	F2 ↑	Recall ↑	CR ↓
L4	X	0.940	0.862	0.879	–
	✓	0.980	0.949	0.942	0.55%
	S	<b>0.980</b>	<b>0.950</b>	<b>0.943</b>	<b>0.542%</b>
L14-MR1	X	0.810	0.781	0.795	–
	✓	0.830	0.788	0.794	3.1%
	S	<b>0.830</b>	<b>0.788</b>	<b>0.795</b>	<b>2.811%</b>
L14-MR2	X	0.700	0.610	0.690	–
	✓	0.760	<b>0.709</b>	<b>0.757</b>	0.346%
	S	<b>0.760</b>	0.701	0.752	<b>0.293%</b>

Contamination rate (CR) is not applicable for the no-update model.

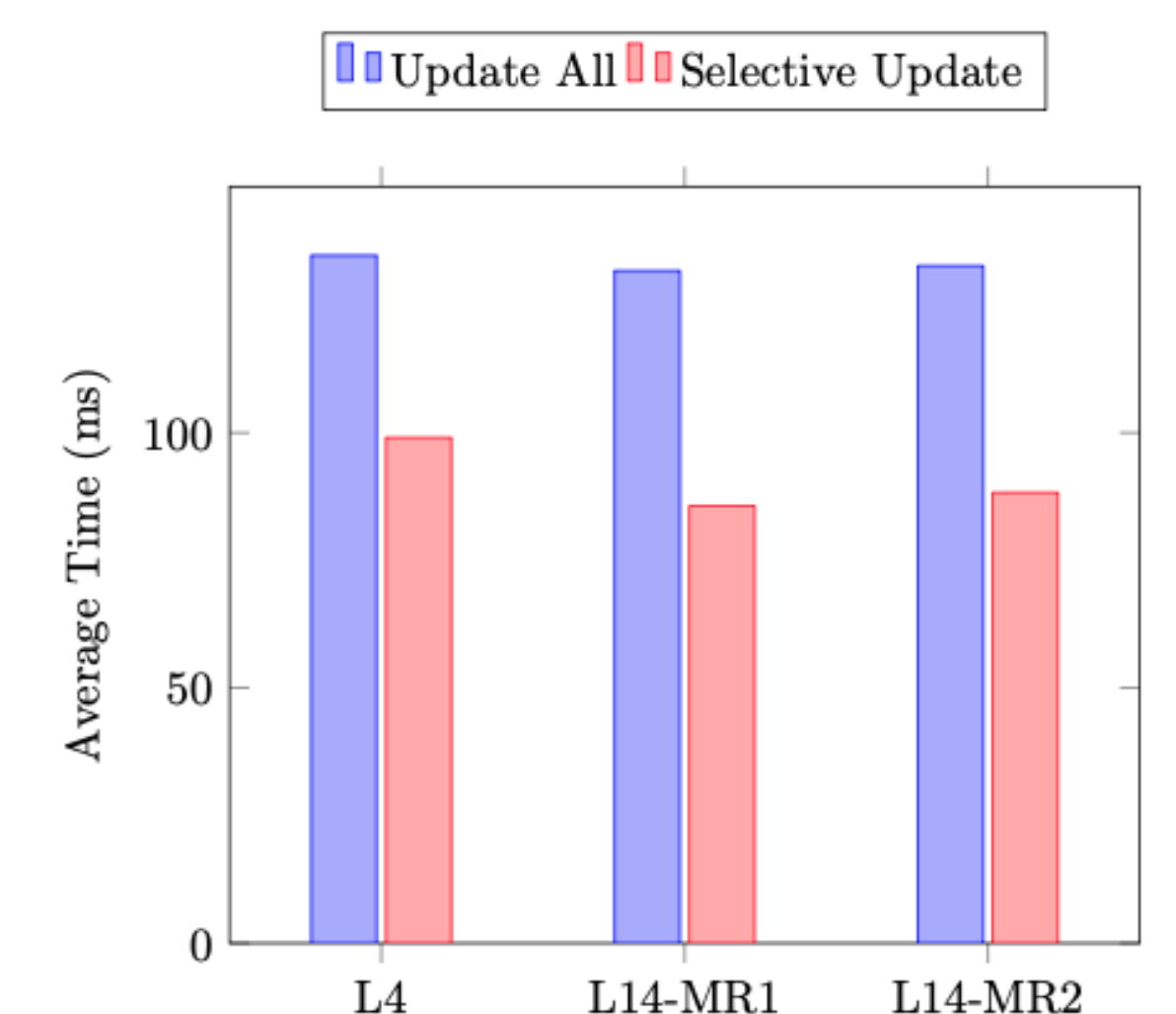


Figure 5: Average model update time under different update strategies.

Table 3: Cross-floor evaluation of pretrained models with adaptive cluster selection (X-means).

Pretrain Floor	Test Floor	K	AUC ↑	F2 ↑	Recall ↑
L4	L14-MR1	3	0.780	0.687	0.731
		15	0.80	0.681	0.726
	L14-MR2	3	0.87	0.761	0.779
		38	0.89	0.835	0.845
L14-MR1	L14-MR2	3	0.82	0.747	0.761
		15	0.87	0.849	0.847
	L4	3	0.98	0.942	0.937
		41	0.98	0.793	0.863
L14-MR2	L14-MR1	3	0.76	0.682	0.708
		17	0.81	0.681	0.711
	L4	3	0.96	0.731	0.785
		25	0.97	0.780	0.827

Highlighted cells are results obtained using adaptive cluster selection via X-means.

Table 4: Comparison of AUC (abnormal vs. rest) Across Datasets

Scheme	L4	L14-MR1	L14-MR2
<b>Our approach</b>	<b>0.980</b>	<b>0.910</b>	<b>0.88</b>
ARCUS (RAPP) [21]	0.479	0.579	0.588
CDAOM [18]	0.463	0.227	0.282
LSTM-ED [13]	0.528	0.583	0.573
REBM [22]	0.979	<u>0.861</u>	0.752
RRCF [7]	0.736	0.510	0.560
STARE [20]	0.672	0.672	0.724
ECOD [12]	<u>0.973</u>	0.831	0.871
COPOD [11]	0.961	0.835	<b>0.902</b>

## Conclusion

- **Novelty:** We propose the first use of gradient profiles to distinguish benign drift from malicious attacks in online CPS streams.
- **Robustness:** The framework achieves **0.98 AUC** with **<1% contamination**, proving effective for mission-critical infrastructure.
- **Scalability:** The X-means extension allows for seamless cross-floor generalization, reducing the need for retraining models for every new zone.