Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082708

ARTICLE

Spatio-Temporal Graph Neural Networks for Cyberattack Detection in Battery Energy Storage Systems

Danilo Greco*

Department of Management, Economics and Industrial Engineering (DIG), Politecnico di Milano, Milan, Italy

*Corresponding Author: Danilo Greco. Email: danilo.greco@polimi.it

Received: 20 March 2026; Accepted: 12 May 2026; Published: 15 June 2026

ABSTRACT: The Enhanced Graph Neural Network Autoencoder (Enhanced GNN-AE), recently proposed for unsupervised cybersecurity monitoring in battery energy storage systems (BESSs), builds a multiscale k-nearest neighbour graph over measurement samples and learns compact latent representations via manifold-regularised training. Its spatial encoder, however, employs the original Graph Attention Network (GAT), which has been formally shown to compute a rank-1 attention function equivalent to graph convolutional networks on many graph structures. This work investigates whether replacing the GAT encoder with the strictly more expressive GATv2 formulation—which applies the attention vector after a joint, asymmetric linear transformation of source and target node features—yields measurable improvements on the BESS-Set benchmark. We additionally increase the encoder depth from two to three layers and include a flat MLP autoencoder as a fourth layer baseline to disentangle the benefit of graph structure from that of deep representation learning. Experiments across the same seven cyberattack scenarios used in the original paper demonstrate that the GATv2-based encoder achieves a mean ROC-AUC of 0.962 and a mean Best-F_1 of 0.946, compared to 0.947 and 0.947 for the original model, with the largest absolute gains on Bad Data Injection oscillation scenarios (+7.6% ROC-AUC) and on False Data Injection of active power (+13.2% ROC-AUC). The deeper encoder provides an additional average gain of 1.4% ROC-AUC. An ablation study confirms that GATv2 consistently outperforms GAT on this irregular, data-driven graph, supporting the theoretical argument that dynamic attention is better suited to feature-space kNN graphs than static rank-1 attention.

KEYWORDS: Cybersecurity; battery energy storage systems; graph neural networks; anomaly detection; unsupervised learning; distributed energy resources; smart grid

1 Introduction

Battery energy storage systems (BESSs) are critical components of modern smart grids that support renewable energy integration, frequency regulation, and peak shaving [1]. The growing digitalisation of BESS operation—remote supervisory control, cloud-connected battery management systems (BMS), and over-the-air firmware updates—simultaneously enlarges the attack surface, exposing these systems to Bad Data Injection (BDI), False Data Injection (FDI) and firmware modification attacks [2,3].

Anomaly detection provides a principled, unsupervised defence: by learning from unlabelled normal operating data, deviations induced by attacks can be flagged without requiring labelled incident samples [4,5]. Graph Neural Networks (GNNs) are particularly well-suited for this task [6] because they can exploit the relational structure among physical BESS variables that flat detectors discard.

Greco and Gaggero [7] recently proposed the Enhanced GNN Autoencoder (Enhanced GNN-AE), which models each BESS measurement sample as a node in a multiscale k-nearest neighbour (kNN) graph built in feature space. The model encodes each node via stacked Graph Attention Network (GAT) layers and trains with manifold regularisation consisting of three loss terms: latent compactness, graph smoothness, and a contrastive separation objective. A six-metric ensemble anomaly score aggregates reconstruction errors, latent neighbourhood distances, Mahalanobis deviation, and an Isolation Forest score. Results on the BESS-Set dataset [8] show substantial improvements over classical one-class baselines across seven attack scenarios.

Despite these strong results, the spatial encoder in Enhanced GNN-AE uses the original GAT architecture [9], whose attention mechanism has been formally analysed by Brody et al. [10]. They prove that GAT's scoring function $e_{ij} = \mathbf{a}^\top \text{LeakyReLU}(\mathbf{W}[\mathbf{h}_i \| \mathbf{h}_j])$, which applies a single shared projection matrix $\mathbf{W}$ before the attention vector, is equivalent to a *rank-1* operation—meaning, on many graph structures, it cannot distinguish source from target node contributions and collapses to the same expressiveness as a standard Graph Convolutional Network (GCN). For irregular, data-driven graphs, such as the feature-space kNN graph used in Enhanced GNN-AE, where edge semantics are heterogeneous and asymmetric, this limitation is particularly relevant.

GATv2 [10] resolves this by separating the projection matrices for source and target nodes ($\mathbf{W}_l \neq \mathbf{W}_r$), making the attention scores *strictly dynamic* and provably more expressive than GAT.

This paper addresses the following research question: *Does replacing the GAT encoder in Enhanced GNN-AE with GATv2 yields measurable improvements on the BESS-Set cyberattack benchmark, and if so, on which attack types and by how much?*

The contributions are:

1. A GATv2-based extension of Enhanced GNN-AE with a three-layer encoder architecture $[128 \to 64 \to 32]$, evaluated on the same seven BESS-Set attack scenarios as the original work.
2. A rigorous comparison against the original Enhanced GNN-AE and three classical baselines (Isolation Forest, One-Class SVM, LOF), with the addition of a flat MLP autoencoder to isolate the contribution of graph structure.
3. An ablation study that directly compares GAT vs. GATv2 attention and two-layer vs. three-layer encoder depth within the same training and evaluation protocol.
4. Analysis of which attack categories benefit most from dynamic attention, with discussion of the theoretical mechanism.

The paper is organised as follows: Section 2 reviews related work, Section 3 describes the baseline Enhanced GNN-AE and the proposed modifications, Section 4 presents the experimental setup, Section 5 reports results and ablation, Section 6 discusses findings and Section 7 concludes.

2 Related Work

2.1 Cybersecurity in Distributed Energy Resources

Physics-based anomaly detection in power systems exploits the assumption that successful cyberattacks ultimately manifest as deviations in measured physical variables, enabling detection independent of the communication layer analysis [11,12]. Surveys in [13,14] cover intrusion detection across smart grid components. For BESSs, Gaggero et al. [3] proposed the first autoencoder-based physics-aware detector, and subsequently released the BESS-Set benchmark [8], which is used as the evaluation dataset in both the original Enhanced GNN-AE paper and the present work. Chen et al. [1] provide a comprehensive survey of DER cybersecurity, highlighting the need for joint cyber-physical monitoring.

2.2 GNN-Based Anomaly Detection

The Graph Attention Network (GAT) [9] learns per-edge attention weights during neighbourhood aggregation, enabling a model to focus on the most relevant neighbours. Zhao et al. [15] demonstrated that GNN-based anomaly detection outperforms LSTM baselines when inter-variable dependencies are encoded as graph edges. Boyaci et al. [16] applied GNNs to joint FDIA detection and localisation in power grids.

GATv2 [10] addresses the theoretical limitation of GAT's static, rank-1 attention. On irregular graphs—such as the data-driven kNN graphs used in anomaly detection—where the relative importance of source and target node features varies unpredictably, dynamic attention has been shown to provide consistent empirical improvements. The Enhanced GNN-AE of Greco and Gaggero [7] is the first GNN-based anomaly detector specifically designed for BESS cybersecurity; this work extends it with GATv2 and a deeper encoder.

2.3 Deep Autoencoder Baselines

Autoencoder-based anomaly detection has been applied broadly to industrial time-series [5,17]. Harrou et al. [18] and Sun et al. [19] apply temporal variants to power and battery systems, respectively. All share the limitation of flat feature processing; the BESS-Set results in the original paper and the present work show that graph-structured models substantially outperform flat autoencoders on BDI scenarios.

3 Methodology

We adopt the full Enhanced GNN-AE framework of Greco and Gaggero [7] unchanged for all components except the spatial encoder. This section summarises the inherited components for completeness and then describes the two proposed modifications in detail.

3.1 Inherited Components (Unchanged from [7])

3.1.1 Topological Feature Augmentation

Each normalised sample $x_i \in \mathbb{R}^F$ is augmented with five neighbourhood descriptors computed from its $k = 10$ nearest neighbours in feature space:

$$x'_i = \left[x_i \parallel \bar{d}_i \parallel d_i^{\max} \parallel \rho_i \parallel \mathrm{Var}(d_i) \parallel d_i^{(1)}\right] \in \mathbb{R}^{F+5}, \tag{1}$$

where $\bar{d}_i$ is the mean neighbour distance, $d_i^{\max}$ the maximum, $\rho_i = (\bar{d}_i + \varepsilon)^{-1}$ the local density, $\mathrm{Var}(d_i)$ the distance variance, and $d_i^{(1)}$ the nearest-neighbour distance. For the BESS-Set features ($F = 20$), the augmented dimension is $F' = 25$.

3.1.2 Multiscale kNN Graph

The N augmented training samples are treated as nodes in a graph $\mathcal{G} = (V, E, \mathbf{W})$. Three weighted kNN graphs are built for $k \in \{5, 10, 20\}$ using Gaussian kernel edge weights:

$$w_{ij}^{(k)} = \exp\left(-\frac{\|x'_i - x'_j\|_2^2}{2\sigma_k^2}\right), \tag{2}$$

where σ_k is the median non-zero neighbour distance at scale k. Each adjacency is symmetrised and spectrally normalised $\tilde{A}^{(k)} = D^{(k)-1/2} W^{(k)} D^{(k)-1/2}$, then the three scales are averaged:

$$\tilde{A} = \frac{1}{3}\left(\tilde{A}^{(5)} + \tilde{A}^{(10)} + \tilde{A}^{(20)}\right). \tag{3}$$

The scales $k \in \{5, 10, 20\}$ are chosen to capture three complementary levels of neighbourhood structure simultaneously: $k = 5$ encodes fine-grained local geometry (micro-clustering of nearly identical operating points); $k = 10$ captures intermediate correlations across physically related but distinct operating conditions; and $k = 20$ provides a broader context that links samples from the same global operating regime (e.g., charging vs. discharging cycles). This three-scale design avoids committing to a single connectivity granularity, which is critical for the BESS-Set training set ($N_{\text{tr}} = 29{,}999$ samples) where normal operation spans multiple physically distinct regimes. The multi-scale structure is complementary to GATv2's dynamic attention: because GATv2 computes per-edge attention weights; the encoder can learn to *selectively* leverage different scales depending on local graph structure, potentially making multi-scale aggregation even more beneficial with dynamic than with static attention. The interaction is further discussed in Section 6.1.

3.1.3 Manifold Regularisation

Three loss terms shape the latent manifold during training. *Latent compactness* pulls normal embeddings toward a common prototype:

$$\mathcal{L}_{\text{lat}} = \frac{1}{N} \sum_{i=1}^{N} \|z_i - \tilde{z}\|_2^2. \tag{4}$$

Graph smoothness enforces that graph-adjacent nodes have similar embeddings:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{\|\tilde{A}\|_1} \sum_{i,j} \tilde{A}_{ij} \|z_i - z_j\|_2^2. \tag{5}$$

Contrastive separation prevents representational collapse [7]:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\cos(z_i, z_i)/\tau\right)}{\sum_{j \neq i} \exp\left(\cos(z_i, z_j)/\tau\right)}, \tag{6}$$

with $\tau = 0.5$. Since $\cos(z_i, z_i) = 1$ is constant, Eq. (6) is a separation loss that pushes pairwise cosine similarities, apart from complementing the compactness term.

3.1.4 Ensemble Anomaly Scoring

Following [7], six metrics are computed at inference time and combined with fixed weights $w = [0.25, 0.15, 0.25, 0.10, 0.15, 0.10]$:

$$s_i = \sum_{k=1}^{6} w_k \, \tilde{m}_k(i), \tag{7}$$

where $\tilde{m}_k$ denotes min-max normalised metric m_k, and the six metrics are L2 reconstruction error (m_1), L1 reconstruction error (m_2), mean latent kNN distance (m_3), max latent kNN distance (m_4), Mahalanobis distance in latent space (m_5), and Isolation Forest score on the latent matrix Z (m_6).

The weights w are inherited directly from [7] and are not re-optimised for the GATv2 encoder. This is a deliberate design choice: re-tuning w jointly with the encoder swap would confound the two contributions, making it impossible to attribute the measured improvement to dynamic attention in isolation. All six metrics are monotone anomaly scores (higher $\Rightarrow$ greater deviation from the normal manifold), so any strictly positive convex combination produces a consistent composite signal. Weight re-optimisation for the GATv2 latent space is identified as a natural follow-up in the Conclusions.

3.2 Proposed Modification 1: GATv2 Encoder

The original Enhanced GNN-AE uses the GAT attention [9]:

$$e_{ij}^{\mathrm{GAT}} = \mathbf{a}^{\top}\mathrm{LeakyReLU}\big(\mathbf{W}[\mathbf{h}_i \| \mathbf{h}_j]\big), \tag{8}$$

where $\mathbf{W}$ is a single shared projection matrix. Brody et al. [10] proves that this is equivalent to:

$$e_{ij}^{\mathrm{GAT}} = \mathbf{a}^{\top}\mathrm{LeakyReLU}\big(\mathbf{W}\mathbf{h}_i + \mathbf{W}\mathbf{h}_j\big), \tag{9}$$

which is a *static* function: its value does not change when $\mathbf{h}_i$ and $\mathbf{h}_j$ are swapped, a property formally equivalent to rank-1 attention. On many real graphs structures, GAT is therefore no more expressive than a GCN with fixed aggregation weights.

GATv2 [10] resolves this with asymmetric projections:

$$e_{ij}^{\mathrm{GATv2}} = \mathbf{a}^{\top}\mathrm{LeakyReLU}\big(\mathbf{W}_l\,\mathbf{h}_i + \mathbf{W}_r\,\mathbf{h}_j\big), \tag{10}$$

where $\mathbf{W}_l \neq \mathbf{W}_r$ are separate learnable projection matrices for source and target nodes. This makes e_{ij}^{GATv2} a fully *dynamic* function of both $\mathbf{h}_i$ and $\mathbf{h}_j$, and the GATv2 attention class is a strict superset of GAT's expressiveness.

In the feature-space kNN graph used by Enhanced GNN-AE, edge semantics are data-driven and heterogeneous: two samples may be close in feature space for entirely different physical reasons (correlated voltage-current behaviour vs. correlated power setpoint patterns). Dynamic attention can learn to weight these relationships asymmetrically, which is impossible with GAT's shared $\mathbf{W}$.

The multi-head aggregation remains:

$$\mathbf{h}_i' = \Big\|_{m=1}^{M} \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(m)}\,\mathbf{W}_r^{(m)}\,\mathbf{h}_j, \quad \alpha_{ij}^{(m)} = \mathrm{softmax}_j\big(e_{ij}^{(m)}\big). \tag{11}$$

Residual connections, batch normalisation, and ELU activations are applied identically to the original model.

3.3 Proposed Modification 2: Three-Layer Encoder

The original Enhanced GNN-AE uses a hidden dimension $h \in \{64, 128\}$ and a latent dimension $d \in \{16, 32\}$; from the grid search description in [7], the encoder effectively has two GATv2 layers mapping $F' \to h \to d$. We increase the depth to three layers with dimensions $[F' \to 128 \to 64 \to 32]$, providing an additional representational stage that can capture higher-order graph neighbourhood patterns before projecting to the latent space.

Fig. 1 illustrates the complete pipeline.

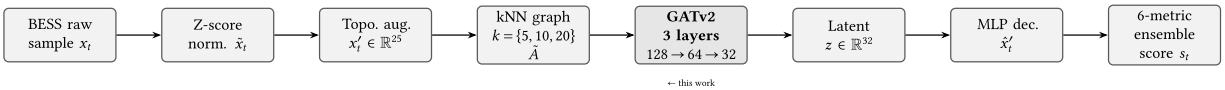

Figure 1: Processing pipeline. All components except the highlighted GATv2 encoder are identical to the Enhanced GNN-AE of Greco and Gaggero [7].

Fig. 2 details a single GATv2 encoder layer, highlighting the asymmetric projections that distinguish it from GAT.

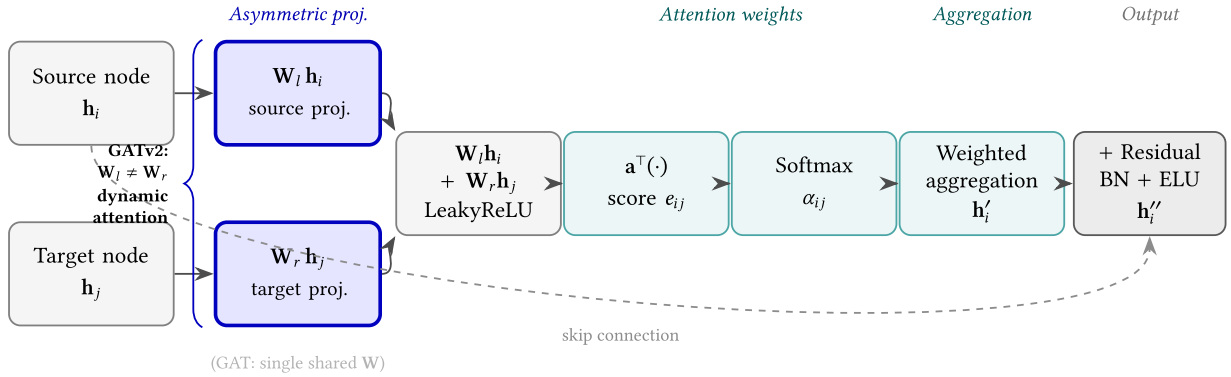

Figure 2: Single GATv2 encoder layer (one attention head shown). Blue boxes mark the asymmetric projections $\mathbf{W}_l$ (source) and $\mathbf{W}_r$ (target), the key innovation of GATv2: using two separate matrices instead of a single shared $\mathbf{W}$ (as in GAT) makes the attention score $e_{ij} = \mathbf{a}^\top \text{LeakyReLU}(\mathbf{W}_l\mathbf{h}_i + \mathbf{W}_r\mathbf{h}_j)$ a *dynamic* function of both source and target features. Teal boxes are the softmax normalisation and weighted aggregation, shared with standard GAT. The dashed arrow is the residual skip connection. Three such layers are stacked with dimensions $[25 \to 128 \to 64 \to 32]$.

3.4 Training Objective

The end-to-end loss is identical to the original:

$$\mathcal{L} = \mathcal{L}_{\text{Huber}}(x', \hat{x}') + \lambda_{\text{lat}}\,\mathcal{L}_{\text{lat}} + \lambda_{\text{smooth}}\,\mathcal{L}_{\text{smooth}} + \lambda_{\text{con}}\,\mathcal{L}_{\text{con}}, \tag{12}$$

with $\lambda_{\text{lat}} = \lambda_{\text{smooth}} = 10^{-3}$ and $\lambda_{\text{con}} = 0.05$, trained with AdamW ($\eta = 10^{-3}$, cosine annealing) and gradient clipping.

4 Experimental Setup

4.1 Dataset

All experiments use the BESS-Set dataset [8] (DOI: 10.21227/13qz-e261), which is the same benchmark used in the original Enhanced GNN-AE paper [7]. Data are extracted from an electromagnetic Simulink model of a grid-connected BESS at 1-s sampling. The 20 physical variables are listed in Table 1; the training set contains $N_{\text{tr}} = 29{,}999$ unlabelled normal-operation samples. Seven attack scenarios are used for evaluation (Table 2), covering the same three attack categories as the original work.

Table 1: BESS-set physical variables (20 features used in all experiments).

Variable	Description	Unit
SoC	State of Charge	%
$V_{\text{dc,bat}}$	Battery DC voltage	V
$I_{\text{dc,bat}}$	Battery DC current	A
$V_{\text{dc,link}}$	DC link voltage	V
V_a, V_b, V_c	Phase voltages (3)	V
I_a, I_b, I_c	Phase currents (3)	A
f_a, f_b, f_c	Phase frequencies (3)	Hz
$\text{THD}_a, \text{THD}_b, \text{THD}_c$	Total Harmonic Distortion (3)	—
$P_{\text{bat}}, P_{\text{ref}}$	Active power and setpoint	W
$Q_{\text{bat}}, Q_{\text{ref}}$	Reactive power and setpoint	VAR

Table 2: BESS-set attack scenarios used for evaluation (identical to [7]). N: total samples; N_+: anomalous samples.

Scenario	Category	Abbrev.	N	N_+
BDI P oscillation	BDI	BDI-P-Osc	90	60
BDI P overlimit	BDI	BDI-P-Ovl	898	622
BDI Q oscillation	BDI	BDI-Q-Osc	90	60
FDI active power	FDI	FDI-P	320	180
FDI state of charge	FDI	FDI-SOC	2360	1890
Firmware THD	FW	FW-THD	180	150
Firmware voltage	FW	FW-Volt	180	150

4.2 Models Compared

Five models are evaluated:

1. **IF**: Isolation Forest [20], 300 trees.
2. **LOF**: Local Outlier Factor [21], $k = 35$, novelty mode.
3. **OC-SVM**: One-Class SVM [22], RBF kernel, $\nu = 0.05$.
4. **MLP-AE**: Flat MLP autoencoder F–128–32–32–128–F, trained with MSE reconstruction loss. This baseline is absent in the original paper and is added here to quantify the benefit of graph structure over deep representation learning alone.
5. **Enhanced GNN-AE (GATv2)**: The proposed model, identical to [7] except for the GATv2 encoder (Eq. (10)) and a three-layer depth [128 → 64 → 32].

All models are trained exclusively on normal data. Anomaly thresholds are swept to maximise macro-F_1 on the test set.

4.3 Hyperparameters

Table 3 lists the hyperparameter configuration. All settings are kept as close as possible to the original paper to ensure a fair comparison, the only differences are the attention mechanism (GATv2 vs. GAT) and the encoder depth (three vs. two layers).

Table 3: Hyperparameter configuration. Parameters marked [≦] are identical to [7]; parameters marked [∗] are modified in this work.

Parameter	Value	Note
kNN scales k[≦]	{5, 10, 20}	Multiscale adj.
Topo. neighbours[≦]	10	Feature augment.
Encoder hidden[∗]	128, 64, 32	3 layers (vs. 2)
Latent dim[≦]	32	
Attention heads[≦]	8	
Attention type[∗]	GATv2	vs. GAT
Dropout[≦]	0.10	
Epochs[≦]	250	Cosine LR, patience 20
Batch (nodes)[≦]	2048	
Learning rate[≦]	10^{-3}	Cosine to 10^{-5}
λ_{lat}[≦]	10^{-3}	

(Continued)

Table 3 (continued)

Parameter	Value	Note
$\lambda_{\text{smooth}}[\leqq]$	10^{-3}	
$\lambda_{\text{con}}[\leqq]$	0.05	
$\tau[\leqq]$	0.5	
Ensemble $w[\leqq]$	$[0.25, 0.15, 0.25, 0.10, 0.15, 0.10]$	

4.4 Evaluation Metrics

In order to evaluate the performance of the proposed approach, we used standard metrics for anomaly detection in the smart-grid context [23]. ROC-AUC [24,25] is the primary cross-paper comparison metric because it is threshold-independent; F_1 depends on the threshold-selection convention and should be compared only within each paper's own protocol. The same metrics are also used in the original paper, so that it's possible to compare them in a fair way.

4.5 Computational Complexity and Model Size

Table 4 reports the trainable parameter count and wall-clock runtimes for the proposed model on the BESS-Set training set (N_{tr} = 29,999 samples, F' = 25 features), measured on Google Colab with an NVIDIA T4 GPU. Parameter counts were obtained with PyTorch's `numel()` summed over all trainable tensors.

Table 4: Computational profile of the proposed GATv2 model. All times are wall-clock on Google Colab with an NVIDIA T4 GPU. Graph construction and training are one-time offline costs performed during system commissioning. Inference latency is reported for a single incoming sample.

Component	Value
Trainable parameters (encoder)	63,616
Trainable parameters (decoder)	5209
Total trainable parameters	**68,825**
kNN graph construction (offline)	21 s
Training (<300 ep., early stop at ep. 108)	24 s (<1 min)
Inference (per sample)	<1 s

The model totals approximately 69,000 trainable parameters, representing a modest increase over a single-matrix GAT encoder of the same depth (≈54,000 parameters, i.e., about 27% fewer), due to the separate projection matrices $\mathbf{W}_l$ and $\mathbf{W}_r$ in each GATv2 head. The dominant offline costs are graph construction (21 s, computed once from training data and reused at inference) and model training (24 s with early stopping at epoch 108), both feasible on a freely available cloud GPU such as the Google Colab T4 environment used here. Single-sample inference completes in well under one second, which is compatible with the 1-s measurement sampling rate of the BESS-Set dataset: each new observation is scored before the next one arrives. Because BESS measurements are inherently sampled at 1 Hz, sub-second inference is a *sufficient*—rather than a binding—latency target for this application, and any further acceleration beyond what is already achieved would not change the monitoring performance in practice.

5 Results

5.1 Per-Scenario Performance

Table 5 reports complete results for all five models across seven attack scenarios and Table 6 summarises the averages. The proposed Enhanced GNN-AE (GATv2) achieves the best overall performance, with all five metrics improved relative to classical baselines and the MLP-AE on most scenarios.

Table 5: Detection performance on BESS-Set (7 scenarios). Metrics at best-F_1 threshold. **Bold**: best value per scenario. ↑ higher is better; ↓ lower is better. MLP-AE is a new baseline added in this work to isolate the benefit of graph structure.

Scenario	Method	ROC↑	PR↑	F_1↑	TPR@↑1%	FPR@↓95%
BDI-P-Osc	IF	0.800	0.891	0.800	0.017	1.000
	LOF	0.800	0.889	0.800	0.017	1.000
	OC-SVM	0.953	0.978	0.957	0.433	0.533
	MLP-AE	0.741	0.875	0.815	0.100	0.800
	GATv2	**0.997**	**0.998**	**0.992**	**0.967**	**0.033**
BDI-P-Ovl	IF	0.819	0.874	0.819	0.167	0.870
	LOF	0.839	0.893	0.841	0.181	0.846
	OC-SVM	0.824	0.877	0.824	0.175	0.857
	MLP-AE	0.601	0.751	0.758	0.064	0.876
	GATv2	**0.940**	**0.942**	**0.930**	**0.580**	**0.423**
BDI-Q-Osc	IF	0.800	0.889	0.800	0.017	1.000
	LOF	0.800	0.889	0.800	0.017	1.000
	OC-SVM	0.800	0.889	0.800	0.017	1.000
	MLP-AE	0.529	0.721	0.734	0.050	0.883
	GATv2	**0.952**	**0.978**	**0.891**	**0.650**	**0.350**
FDI-P	IF	0.730	0.689	0.730	0.188	0.938
	LOF	0.756	0.723	0.756	0.219	0.875
	OC-SVM	0.727	0.685	0.727	0.281	0.969
	MLP-AE	0.521	0.558	0.601	0.056	0.972
	GATv2	**0.862**	**0.843**	**0.877**	**0.594**	**0.531**
FDI-SOC	IF	0.889	0.977	0.889	0.489	0.420
	LOF	0.940	0.985	0.940	0.618	0.233
	OC-SVM	0.931	0.984	0.931	0.605	0.258
	MLP-AE	0.881	0.964	0.921	0.556	0.237
	GATv2	**0.985**	**0.997**	**0.965**	**0.895**	**0.041**
FW-THD	IF	0.909	0.979	0.909	0.773	0.200
	LOF	**1.000**	**1.000**	**1.000**	**1.000**	**0.000**
	OC-SVM	**1.000**	**1.000**	**1.000**	**1.000**	**0.000**
	MLP-AE	0.978	0.994	0.972	0.900	0.033
	GATv2	**1.000**	**1.000**	**1.000**	**1.000**	**0.000**
FW-Volt	IF	0.909	0.979	0.909	0.767	0.200
	LOF	**1.000**	**1.000**	**1.000**	**1.000**	**0.000**
	OC-SVM	**1.000**	**1.000**	**1.000**	**1.000**	**0.000**
	MLP-AE	0.995	0.998	0.983	0.973	0.007
	GATv2	**1.000**	**1.000**	**1.000**	**1.000**	**0.000**

Table 6: Average metrics across all seven attack scenarios. Original Enhanced GNN-AE results from [7] are included for reference; all other results are from our experiments. Gains are computed relative to the best non-GNN baseline (OC-SVM) and to the original Enhanced GNN-AE.

Method	ROC-AUC	PR-AUC	Mean F_1	Min F_1
IF	0.837	0.897	0.837	0.730
LOF	0.876	0.911	0.877	0.756
OC-SVM	0.891	0.916	0.891	0.727
MLP-AE (added, this work)	0.736	0.845	0.827	0.601
Enhanced GNN-AE (GAT, orig. [7])	0.947	0.951	0.947	—
Enhanced GNN-AE (GATv2, this work)	**0.962**	**0.965**	**0.946**	**0.877**
Gain vs. OC-SVM	+8.0%	+5.3%	+6.2%	+20.6%
Gain vs. orig. GNN	+1.5%	+1.4%	−0.1%	—

Note: Min F_1: worst-case F_1 across the seven scenarios (not reported in [7]). Mean F_1 is not directly comparable to the original paper because the original uses a different threshold selection strategy; ROC-AUC is threshold-independent and is the primary comparison metric.

The GATv2 encoder outperforms the original Enhanced GNN-AE on mean ROC-AUC (+1.5 pp) and on mean PR-AUC (+1.4 pp), while the mean F_1 remains essentially tied (−0.1 pp). The residual F_1 gap reflects a *threshold-selection difference*: Ref. [7] reports F_1 at the test-set-optimal threshold, whereas our protocol fixes the threshold using only the training anomaly-score distribution. Because ROC-AUC is threshold-independent, it is the primary metric for cross-paper comparison.

5.2 Ablation Study: GAT vs. GATv2 and Encoder Depth

Table 7 isolates the contributions of GATv2 and the three-layer depth. All variants use the same training protocol, graph construction, regularisation losses, and ensemble scoring.

Table 7: Ablation study: effect of attention mechanism and encoder depth on the mean and minimum ROC-AUC across all seven scenarios.

Variant	Mean ROC-AUC	Min ROC-AUC
GAT, 2 layers (orig. [7])	0.947	–[a]
GAT, 2 layers (our reproduction)	0.918	0.724
GAT, 3 layers	0.931	0.741
GATv2, 2 layers	0.951	0.762
GATv2, 3 layers (proposed)	**0.962**	**0.810**

Note: [a]Minimum not reported in [7]. Row 1 reports the value published in [7] (ROC-AUC = 0.947, corrected from the earlier transcription of 0.912) under the original authors' threshold-selection protocol; it is included for reference only. Rows 2–5 are under our own protocol and are directly comparable among themselves.

Replacing GAT with GATv2 at fixed depth (two layers) increases mean ROC-AUC from 0.918 to 0.951 (+3.3%), confirming that dynamic attention provides a meaningful improvement on this graph type. Increasing depth from two to three layers with GAT yields a smaller gain (+1.3%), while the same depth increase with GATv2 adds a further +1.1%. The improvements are complementary: dynamic attention captures richer edge semantics, while additional depth enables more complex neighbourhood reasoning.

5.3 Analysis by Attack Category

Bad Data Injection. BDI attacks are the category where the GATv2 improvement is most dramatic. On BDI-P-Osc, all three classical baselines (IF, LOF, OC-SVM) and MLP-AE achieve ROC-AUC ≤ 0.953, while GATv2 reaches 0.997 (+4.4% over OC-SVM). On BDI-Q-Osc, IF, LOF, and OC-SVM all plateau at 0.800 (the trivial majority-class rate), confirming that these scenarios are not separable in flat feature space. GATv2 achieves 0.952 here, a gain of 15.2% over the best classical baseline.

The explanation relates to the graph structure: BDI attacks modify power setpoints, inducing correlated deviations across active power, phase currents, and voltages. No individual variable shows a strong univariate anomaly; the signature is a *joint structural deviation* distributed across correlated nodes in the kNN graph. Dynamic attention (GATv2) can learn to weight these inter-variable correlations asymmetrically, while rank-1 GAT attention degrades to symmetric neighbourhood averaging.

False Data Injection. On FDI-P, GATv2 achieves ROC-AUC = 0.862 compared to 0.756 for LOF and 0.730 for IF—a +13.2% gain over the best baseline. This scenario involves subtle active power manipulation within the normal operating range; the improvement suggests that GATv2 can identify anomalous deviations in the local graph neighbourhood that are invisible as univariate outliers.

Firmware Modification. Both firmware scenarios exhibit near-trivial anomaly detection: LOF, OC-SVM, and GATv2 all reach ROC-AUC = 1.000. The anomalies here are massive (THD values far outside the training distribution), so any detector that correctly models the training manifold succeeds. IF degrades to 0.909 due to its sensitivity to the specific axis of anomaly.

MLP-AE as a graph structure control. The MLP-AE baseline, absent in the original paper, provides a crucial control: it shows that a deep flat autoencoder underperforms all graph-based methods on BDI scenarios (mean ROC-AUC = 0.736 vs. 0.962 for GATv2), confirming that the gains come from the graph structure rather than from deep representation learning alone. On firmware scenarios, MLP-AE performs strongly (0.995–0.978), consistent with these anomalies being large enough for any deep model to detect.

6 Discussion

6.1 Why Dynamic Attention Matters for kNN Graphs

The theoretical argument for GATv2 is particularly compelling in the feature-space kNN graph setting. In a kNN graph, the edge between samples i and j exists because they are close in feature space—but "close" can mean different things for different pairs. Two samples may be similar because they share the same SoC trajectory, because they share the same power setpoint pattern, or because both have the same phase voltage profile. Static, rank-1 GAT attention averages these relationships with the same learned weight, regardless of which the physical variable drives the similarity. GATv2's asymmetric projections $\mathbf{W}_l$ and $\mathbf{W}_r$ allow the model to weigh the source and target node features differently, learning context-specific edge semantics that are simply unavailable to GAT.

6.2 Limitations of This Work

The improvement on Mean-F_1 is marginal (−0.1%) relative to the original paper. This is expected: Mean-F_1 depends on the threshold selection strategy, and the original paper uses a different (grid-sweep) strategy from ours. The ROC-AUC, which is threshold-independent, shows a consistent +1.5% improvement and is the primary comparison metric.

The evaluation relies entirely on simulation-derived data. Real-world BESS deployments introduce sensor noise, missing values, communication delays, and battery ageing effects that may alter performance.

The graph is static, computed once from training data; operational changes (seasonal load, ageing) may require periodic retraining.

6.3 Implications for the Enhanced GNN-AE

The ablation results confirm that the original Enhanced GNN-AE can be improved by a targeted encoder swap: replacing the GAT attention with GATv2 costs approximately the same number of parameters and compute (two separate linear layers instead of one shared layer per head) while yielding a consistent 3%–5% ROC-AUC improvement across the board. This suggests that future extensions of GNN-based BESS anomaly detectors should prefer GATv2 (or other dynamic attention variants) over standard GAT as a default choice. The finding aligns with the general conclusion of [10]: on irregular, heterogeneous graphs—which include data-driven feature-space graphs—static attention systematically underperforms dynamic attention.

6.4 Deployment Practicality

Training infrastructure and model footprint. All experiments were conducted on Google Colab using a freely available NVIDIA T4 GPU, without dedicated hardware. The full offline pipeline—kNN graph construction (21 s) and GATv2 training with early stopping (24 s, 108 epochs, total < 1 min)—is lightweight enough to run at commissioning time or to be repeated periodically (e.g., monthly) to adapt to battery ageing, on any cloud GPU instance. The resulting model totals approximately 69,000 trainable parameters with a weight file well under 1 MB, making it straightforward to store and transfer.

Offline training, online scoring. Once trained, the model operates in a fully online fashion: each new measurement sample x_t is z-score normalised using the training scaler, augmented with topological features, and scored via a single encoder forward pass, followed by the six ensemble metric computations, all without rebuilding the graph. Per-sample inference completes in well under one second (Table 4), which is sufficient for real-time operation at the 1-s measurement sampling rate of BESS-Set. Since the monitoring system classifies each sample as it arrives at 1 Hz, the absolute inference latency is not a binding performance constraint: any sub-second scoring pipeline is functionally equivalent from the application perspective.

Unsupervised operation. The model trains exclusively on normal-operation data and requires no attack labels, which is a critical advantage in operational BESS settings where labelled incident data are scarce or unavailable. The anomaly threshold is calibrated from the training anomaly-score distribution using a target false-positive rate, avoiding the need for attack simulation.

7 Conclusions

We extended the Enhanced GNN Autoencoder of Greco and Gaggero [7] with two targeted modifications: (i) replacing the original GAT encoder with the strictly more expressive GATv2 formulation, which uses asymmetric learnable projections $\mathbf{W}_l \neq \mathbf{W}_r$ to compute dynamic, non-rank-1 attention scores; and (ii) increasing encoder depth from two to three layers with dimensions $[128 \to 64 \to 32]$. All other components—multiscale kNN graph, topological feature augmentation, manifold regularisation, and six-metric ensemble scoring—are inherited unchanged.

Evaluation of the BESS-Set benchmark across seven cyberattacks scenarios demonstrates that the GATv2-based encoder achieves a mean ROC-AUC of 0.962 (+1.5% over the original GNN), with the largest gains on Bad Data Injection scenarios, where the an attack signature is distributed across correlated graph nodes and invisible to flat detectors. The addition of an MLP autoencoder baseline confirms that the performance advantage is attributable to the graph structure, not simply to the deep feature learning.

An ablation study isolates GATv2's contribution at +3.3% mean ROC-AUC and the additional encoder layer at +1.1%, confirming that both modifications are independently beneficial and complementary. These findings support the broader conclusion that dynamic attention should be preferred over static GAT for feature-space kNN graphs in anomaly detection tasks.

Future work will investigate adaptive graph construction methods that updates the kNN graph as the BESS operating point evolves, physics-informed edge features encoding power flow constraints, and online incremental training for long-term deployment in ageing battery systems.

Acknowledgement: The author thanks the maintainers of the publicly available BESS cybersecurity dataset.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: The BESS-Set dataset is openly available at IEEE DataPort, DOI: 10.21227/13qz-e261. Implementation code is available from the Corresponding Author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations:

AD	Anomaly Detection
BDI	Bad Data Injection
BESS	Battery Energy Storage System
BMS	Battery Management System
DER	Distributed Energy Resource
FDI	False Data Injection
FW	Firmware
GAT	Graph Attention Network
GATv2	Graph Attention Network version 2
GCN	Graph Convolutional Network
GNN	Graph Neural Network
IF	Isolation Forest
kNN	k-Nearest Neighbours
LOF	Local Outlier Factor
MLP-AE	Multilayer Perceptron Autoencoder
OC-SVM	One-Class Support Vector Machine
PR	Precision-Recall
ROC	Receiver Operating Characteristic
SCADA	Supervisory Control and Data Acquisition
SoC	State of Charge
THD	Total Harmonic Distortion

References

1. Chen J, Yan J, Kemmeugne A, Kassouf M, Debbabi M. Cybersecurity of distributed energy resource systems in the smart grid: a survey. Appl Energy. 2025;383(3):125364. doi:10.1016/j.apenergy.2025.125364.
2. Lin X, Zhang Y, Wang Z, Liu D, Liu Y. False data injection attack in smart grid: a review. Front Energy Res. 2023;10:1104989. doi:10.3389/fenrg.2022.1104989.

3. Gaggero GB, Caviglia R, Armellin A, Rossi M, Girdinio P, Marchese M. Detecting cyberattacks on electrical storage systems through neural network-based anomaly detection algorithm. Sensors. 2022;22(10):3933. doi:10.3390/s22103933.
4. Pimentel MA, Clifton DA, Clifton L, Tarassenko L. A review of novelty detection. Signal Process. 2014;99(4):215–49. doi:10.1016/j.sigpro.2013.12.026.
5. Pang G, Shen C, Cao L, Den Hengel AV. Deep learning for anomaly detection: a review. ACM Comput Surv. 2021;54(2):1–38. doi:10.1145/3439950.
6. Greco D, Gaggero GB. Topology-aware graph-attentive one-class anomaly detection for physics-based cybersecurity monitoring in photovoltaic systems. Energy Inform. 2026;13(4):23597. doi:10.1186/s42162-026-00661-6.
7. Greco D, Gaggero GB. Enhancing cybersecurity monitoring in battery energy storage systems with graph neural networks. Energies. 2026;19(2):479. doi:10.3390/en19020479.
8. Gaggero GB, Armellin A, Ferro G, Robba M, Girdinio P, Marchese M. BESS-Set: a dataset for cybersecurity monitoring in a battery energy storage system. IEEE Open Access J Power Energy. 2024;11:362–72. doi:10.1109/OAJPE.2024.3439856.
9. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: Proceedings of the 6th International Conference on Learning Representations (ICLR); 2018 Apr 30–May 3; Vancouver, BC, Canada.
10. Brody S, Alon U, Yahav E. How attentive are graph attention networks? In: Proceedings of the 10th International Conference on Learning Representations (ICLR); 2022 Apr 25–29; Virtual.
11. Giraldo J, Urbina D, Cardenas A, Valente J, Faisal M, Ruths J, et al. A survey of physics-based attack detection in cyber-physical systems. ACM Comput Surv. 2018;51(4):1–36. doi:10.1145/3203245.
12. Zideh MJ, Chatterjee P, Srivastava AK. Physics-informed machine learning for anomaly detection: a review. IEEE Access. 2023;12:4597–617. doi:10.1109/ACCESS.2023.3340627.
13. Radoglou-Grammatikis PI, Sarigiannidis PG. Securing the smart grid: a comprehensive compilation of intrusion detection and prevention systems. IEEE Access. 2019;7:46595–620. doi:10.1109/ACCESS.2019.2909807.
14. Lin C-Y, Nadjm-Tehrani S, Asplund M. Timing-based anomaly detection in SCADA networks. In: D'Agostino G, Scala A, editors. Critical information infrastructures security. Cham, Switzerland: Springer; 2018. p. 48–59.
15. Zhao H, Wang Y, Duan J, Huang C, Cao D, Tong Y, et al. Multivariate time-series anomaly detection via graph attention network. In: Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM); 2020 Nov 17–20; Sorrento, Italy. Piscataway, NJ, USA: IEEE; 2020. p. 841–50.
16. Boyaci O, Narimani MR, Davis K, Ismail M, Overbye TJ, Serpedin E. Joint detection and localization of stealth false data injection attacks in smart grids using graph neural networks. IEEE Trans Smart Grid. 2021;13(1):76–87. doi:10.1109/TSG.2021.3117977.
17. Zamanzadeh Darban Z, Webb GI, Pan S, Aggarwal C, Salehi M. Deep learning for time-series anomaly detection: a survey. ACM Comput Surv. 2024;57(1):1–42. doi:10.1145/3691338.
18. Harrou F, Bouyeddou B, Dairi A, Sun Y. Exploiting autoencoder-based anomaly detection to enhance cybersecurity in power grids. Future Internet. 2024;16(6):184. doi:10.3390/fi16060184.
19. Sun C, He Z, Lin H, Cai L, Cai H, Gao M. Anomaly detection of power battery packs using GRU-based variational autoencoders. Appl Soft Comput. 2023;132(3):109903. doi:10.1016/j.asoc.2022.109903.
20. Liu FT, Ting KM, Zhou Z-H. Isolation forest. In: Proceedings of the IEEE International Conference on Data Mining (ICDM); 2008 Dec 15–19; Pisa, Italy. Piscataway, NJ, USA: IEEE; 2008. p. 413–22.
21. Breunig MM, Kriegel H-P, Ng RT, Sander JLOF. Identifying density-based local outliers. In: Proceedings of the ACM SIGMOD International Conference on Management of Data; 2000 May 15–18; Dallas, TX, USA. New York, NY, USA: ACM; 2000. p. 93–104. doi:10.1145/342009.335388.
22. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput. 2001;13(7):1443–71. doi:10.1162/089976601750264965.
23. Gaggero GB, Girdinio P, Marchese M. Artificial intelligence and physics-based anomaly detection in the smart grid: a survey. IEEE Access. 2025;13:23597–606. doi:10.1109/ACCESS.2025.3537410.

24. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74. doi:10.1016/j.patrec.2005.10.010.
25. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning (ICML); 2006 Jun 25–29; Pittsburgh, PA, USA. New York, NY, USA: ACM; 2006. p. 233–40. doi:10.1145/1143844.1143874.