Spatio-Temporal Graph Modeling for EEG-Based Seizure Detection: A Comparative Study with Temporal Baselines

Martin Le Bras Ludek Cizinsky Jean Siffert Golnoosh Abdollahinejad

École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

***Abstract*—Epileptic seizure detection from electroencephalography (EEG) is a challenging task due to the high dimensionality, noise, and non-stationary nature of the signals. While conventional deep learning models typically process EEG channels independently, graph-based methods offer a principled way to incorporate spatial relationships between electrodes. In this study, we compare graph-based and non-graph-based approaches using EEG data collected under the 10–20 system. Through extensive experiments, we identify a graph-based model that outperforms the best non-graph baseline by 25 percentage points in macro-F1 score on the test set. Moreover, the graph-based model exhibits significantly lower generalization gap, indicating improved robustness and reduced overfitting.**

I. Introduction

Epilepsy, a neurological disorder affecting nearly 50 million people worldwide [18], is characterized by recurrent seizures resulting from abnormal neuronal activity in the brain [4]. Accurate and timely seizure detection plays a crucial role in improving the quality of life for patients, particularly for those who are not responsive to medication or surgical treatment.

Electroencephalography (EEG) is the primary non-invasive tool for monitoring brain activity in patients with epilepsy. It captures voltage signals from multiple electrodes placed on the scalp, producing high-dimensional and noisy time-series data [14]. Traditional approaches for seizure detection typically rely on signal processing techniques or deep learning models tailored to time-series data, such as recurrent neural networks [6], [7] and 1D convolutional networks [12]. However, these models often treat EEG channels independently, ignoring the spatial relationships between electrodes that reflect the underlying brain connectivity.

To address the limitations of conventional EEG processing, this study investigates graph-based representations in which each electrode is modeled as a node and edges encode spatial or functional relationships. This formulation enables the use of Graph Neural Networks (GNNs) to jointly capture spatial and temporal dynamics. A central question we explore is whether graph-based methods can outperform traditional non-graph approaches for seizure detection.

In addition, we assess whether improving temporal feature extraction enhances downstream performance, and whether neural architectures can surpass classical non-neural baselines for time-series processing. We further investigate the benefits of combining time-window-based aggregation with graph-based modeling, compared to standard temporal modules such as 1D CNNs. Finally, we examine the role of domain-specific spatial priors—such as inter-electrode distances - in shaping graph topology and influencing model performance. Through these investigations, our goal is to evaluate the practical advantages and limitations of graph-based modeling for EEG data and to provide insights into how spatial priors and architectural design choices impact seizure detection accuracy.

II. Methodology

Our method comprises four primary components: temporal segmentation via sliding windows, a distance-based adjacency matrix for spatial modeling, a Graph Convolutional Network (GCN), and a Long Short-Term Memory (LSTM) network enhanced with attention. Figure 1 provides an overview of the architecture along with key signal dimensions. We first segment EEG signals temporally and independently encode spatial relationships within each segment using graph-based methods. Subsequently, we employ the LSTM-attention module to model temporal dependencies across segments, thereby yielding the final prediction.

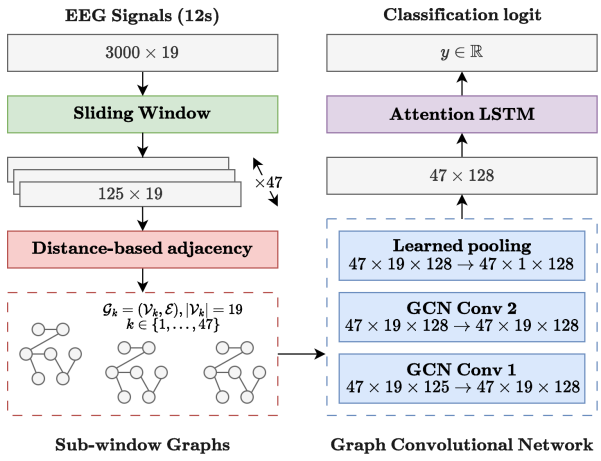

Fig. 1: Diagram of seizure classification model

A. *Sliding window processing*

To address the high dimensionality and non-stationarity of EEG signals, we segment each recording into overlapping sliding windows along the temporal axis. Using a window size of $W = 125$ and a stride of $S = 62$, each sample is divided into 47 sub-windows, which are processed independently.

B. *Distance-based adjacency*

Graphs from sub-window EEG signals are constructed by generating a binary adjacency matrix $\mathbf{A} \in \mathbb{R}^{19\times19}$. EEG

signals were recorded using the 10–20 system [11], in which electrodes are positioned on the scalp based on standardized anatomical locations corresponding to the underlying brain cortex areas. Leveraging this structured placement, we construct a distance-based adjacency matrix to encode spatial relationships between electrodes [8]. This adjacency matrix is shared across all patients, promoting a patient-independent representation that enhances generalization and reduces the risk of overfitting to individual-specific spatial patterns. The adjacency matrix is created by applying a distance-based mask to the pairwise electrode distances. More formally,

$$\mathbf{A}_{i,j} = \begin{cases} 1 & d(i,j) \leq D \\ 0 & \text{otherwise} \end{cases} \qquad D = 1.5 \tag{1}$$

where $d(i,j)$ is the distance between electrodes i and j.

C. Graph Convolutional Network

To leverage the spatial structure of EEG signals captured via the 10–20 electrode placement system, we employ a two-layer GCN [9]. Each layer performs localized message passing, aggregating and transforming node features using their spatial neighbors. Formally, for each node $i \in \mathcal{V}$, $|\mathcal{V}| = 19$, the operation is defined as:

$$\mathbf{x}'_i = \mathbf{\Theta}^\top \sum_{j \in \mathcal{N}_i \cup i} \frac{ej,i}{\sqrt{\hat{d}_j \hat{d}_i}} \mathbf{x}_j, \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^{F_{\text{in}}}$ and $\mathbf{x}'_i \in \mathbb{R}^{F_{\text{out}}}$ are the input and output features, respectively. The first GCN layer operates on window-level graphs with $F_{\text{in}} = 125$ and $F_{\text{out}} = 128$; the second layer maintains $F_{\text{in}} = F_{\text{out}} = 128$. Node embeddings $\mathbf{x}_i \in \mathbb{R}^{128}$ are pooled via a learned linear projection, producing a fixed-length graph embedding $g_k \in \mathbb{R}^{128}$.

D. LSTM with Attention

The graph embeddings $g_k,\ k \in \{1, \ldots, 47\}$ produced by the GCN form a temporally ordered sequence. To capture the sequential dependencies, we employ a unidirectional Long Short-Term Memory (LSTM) network enhanced with an attention mechanism. This module models temporal dependencies in a causal manner, processing the sequence left to right. Given an input sequence $\{x_1, x_2, \ldots, x_T\}$, the LSTM outputs a hidden state h_t at each time step t, which is subsequently refined via attention to emphasize relevant temporal features.

This yields a sequence of contextualized hidden states $\{h_1, h_2, \ldots, h_T\}$ that encode information from the entire input. To highlight the most informative segments, we apply an attention mechanism over the LSTM outputs. For each time step t, a score α_t is computed to indicate its relevance to the classification objective. The attention weights and resulting context vector s are given by:

$$u_t = \tanh(W_a h_t + b_a), \tag{3}$$

$$\alpha_t = \frac{\exp(u_t^\top u_w)}{\sum_{k=1}^{T} \exp(u_k^\top u_w)}, \tag{4}$$

$$s = \sum_{t=1}^{T} \alpha_t h_t, \tag{5}$$

where $W_a \in \mathbb{R}^{d \times d}$, $b_a \in \mathbb{R}^d$, and $u_w \in \mathbb{R}^d$ are learnable parameters. The context vector s aggregates temporal features with learned importance weights, enabling the model to attend to discriminative patterns while retaining global sequence context. This vector is subsequently passed through a linear layer to produce the final prediction.

III. Results

In this section, we present the sequence of experiments that informed the development of our final method. Starting from simple baseline models, we progressively introduce architectural components based on empirical findings to incrementally improve performance.

Model	F1 (Macro)	F1$^+$	F1$^-$
LSTM	**0.8800**	**0.8026**	**0.9575**
1D Convolution	0.8452	0.7445	0.9472
Transformer Encoder	0.8633	0.7724	0.9543
GCN	0.8001	0.6667	0.9336
GAT	0.7546	0.5818	0.9273

TABLE I: Baseline models performance on validation set.

A. Experimental Setup

We allocate 90% of the dataset (11,693 samples) for training and the remaining 10% (1,300 samples) for validation. All models are trained for 1,000 epochs, with the first 50 epochs designated for learning rate warm-up. To enhance training stability, we apply L2 gradient clipping with a maximum norm of 1.0. Models are optimized using the binary cross-entropy loss. We report the F1 score for each class (negative/positive) as well as the macro-average. All experiments are conducted on a single NVIDIA V100 GPU.

B. Baseline Approaches

To evaluate the relative performance of graph-based and non-graph-based methods, we establish a set of baseline models. Prior to training, we apply the Fast Fourier Transform (FFT) to project EEG signals from the temporal to the spectral domain, thereby reducing the input dimensionality from $\mathbb{R}^{3000 \times 19}$ to $\mathbb{R}^{384 \times 19}$. The resulting spectral features are then processed using either a *temporal* or a *graph-based* model. Temporal models interpret the frequency dimension as a sequence. We consider three architectures: LSTM network [5], 1D convolutional layers, and a Transformer Encoder [16].

For graph-based models, we construct a binary adjacency matrix by thresholding the pairwise distance matrix between scalp electrodes, as described previously. The resulting graph $\mathcal{G}$ comprises $|\mathcal{V}| = 19$ nodes. We implement two message-passing architectures: GCN [9] and a Graph Attention Network (GAT) [17].

Table I summarizes the baseline results. In this initial setting, non-graph-based models outperform their graph-based

Processing	Model	F1 (Macro)	F1$^+$	F1$^-$
Decimate ($Q = 30$)	LSTM	0.7755	0.6275	0.9235
Decimate ($Q = 5$)	GCN	0.6731	0.4651	0.8811
Window ($N = 12$)	LSTM	0.7956	0.6606	0.9305
Window ($N = 12$)	GCN	0.6623	0.4327	0.8919
DCT ($K = 600$)	LSTM	0.7900	0.6452	**0.9349**
DCT ($K = 100$)	GCN	0.6293	0.3846	0.8740
Wavelet ($L = 4$)	LSTM	**0.8021**	**0.6726**	0.9317
Wavelet ($L = 4$)	GCN	0.6600	0.4354	0.8847

(a) *Non-neural* approaches for time compression.

Processing	Model	F1 (Macro)	F1$^+$	F1$^-$
1D CNN (Large)	LSTM	**0.8605**	**0.7724**	**0.9486**
1D CNN (Large)	GCN	0.8103	0.6770	0.9435
PatchTST (NO, $L = 10$)	LSTM	0.8017	0.6712	0.9323
PatchTST (NO, $L = 10$)	GCN	0.6610	0.4300	0.8920
PatchTST (O, $L = 15$)	LSTM	0.7958	0.6589	0.9328
PatchTST (O, $L = 15$)	GCN	0.6600	0.4354	0.8847

(b) *Neural* approaches for time compression (NO: non-overlap; O: overlap)

WSize	WStride	NW	F1 (Macro)	F1$^+$	F1$^-$
125	62	48	0.8052	0.6716	**0.9389**
125	100	30	**0.8085**	**0.6842**	0.9328
125	125	24	0.8000	0.6723	0.9277
250	125	24	0.8007	0.6697	0.9318
250	200	15	0.7887	0.6503	0.9270
250	250	12	0.7746	0.6291	0.9201

(c) Window based approach. Comparison of different window sizes and strides configurations (NW: number of windows).

Model	F1 (Macro)	F1$^+$	F1$^-$
CONV1D	0.7913	0.6440	0.9387
Transformer Encoder	0.7765	0.6256	0.9275
LSTM Only	0.8052	0.6716	0.9389
LSTM + Attention	0.7900	0.6497	0.9304
Learned Pool + LSTM + Attention	**0.8473**	**0.7454**	**0.9493**

(d) Window based approach. Comparison of different classifier configurations.

TABLE II: Experiments focusing on time compression via (non)-neural approaches and window based processing.

counterparts, with the LSTM achieving the highest overall performance. While both model types perform well on the negative class, graph-based models exhibit significantly lower F1 scores for the positive class. We hypothesize that the dimensionality reduction induced by FFT may discard spatial information critical for effective graph-based learning.

C. Beyond FFT: Alternative Temporal Compression Methods

While the Fast Fourier Transform (FFT) provides a straightforward and effective spectral representation for compressing the temporal dimension of EEG signals, we hypothesize that alternative methods may better capture task-relevant features. Accordingly, we investigate two complementary categories of approaches: classical non-neural techniques and neural network-based methods.

1) Non-neural Approaches: We consider classical signal processing methods for temporal compression: **Simple decimation** applies an anti-aliasing filter followed by downsampling by retaining every q-th sample, effectively preserving low-frequency content; **Sliding window mean** partitions the signal into non-overlapping windows of fixed length and averages within each window to smooth the signal; **Discrete Cosine Transform (DCT)** decomposes the signal into cosine basis functions and retains the first K coefficients corresponding to low frequencies [2]; and the **1D Discrete Wavelet Transform** captures coarse signal structure by preserving top-level approximation coefficients.

2) Neural Approaches: We evaluate neural architectures designed for temporal encoding: **1D Convolution** employs convolutional layers along the temporal dimension of each EEG channel to learn local patterns and reduce dimensionality via striding and pooling; while the **Patch Time Series Transformer (PatchTST)** partitions the EEG sequence into fixed-size patches and leverages a Transformer encoder to operate on patches rather than individual timesteps, thereby reducing sequence length [13].

Table IIa presents the results for non-neural temporal compression methods. Among them, the wavelet transform yields the highest macro F1 score, yet underperforms the best model from the previous section by 8 percentage points (pp). As expected, the top-performing neural method-a 1D CNN signal processor-outperforms all non-neural alternatives by a margin of 6 pp (Table IIb). However, it still lags behind the FFT-based baseline by 2 pp. Notably, the transformer-based signal processor (PatchTST) performs significantly worse than other neural models. We hypothesize that this underperformance stems from the limited size of the training set, as transformer architectures are known to require large datasets to generalize effectively.

D. Aggregating time series using window based approach

The previous approaches directly compress the time series, potentially discarding important information depending on the hidden dimension. Instead of reducing the time dimension upfront, we split each EEG signal into multiple overlapping sub-windows, preserving the temporal resolution. Each sub-window is then independently processed using a graph-based model (GCN or GAT) to extract spatial embeddings. Finally, the sequence of graph embeddings is passed through a temporal model (LSTM, 1D Convolution, etc.) to perform the final classification. This two-stage architecture allows us to better balance temporal resolution and spatial structure, potentially improving performance by retaining more information throughout the pipeline.

Table IIc presents an ablation study on different window and stride configurations. We evaluate two window sizes-125

Adjacency Type	F1 (Macro)	F1$^+$	F1$^-$
Distance ($D = 1.0$)	0.7819	**0.6414**	0.9224
Distance ($D = 1.5$)	**0.7825**	0.6410	**0.9239**
Correlation ($C = 0.5$)	0.7770	0.6336	0.9204
Learned	0.7649	0.6124	0.9174

(a) Ablation on various adjacency matrices.

Model	F1 (Macro)	F1$^+$	F1$^-$	Test F1 (Macro)
LSTM Baseline	0.8350	0.7331	0.9369	0.5668
LSTM	**0.7956**	**0.6596**	**0.9316**	0.7967
LSTM+Attn	0.7904	0.6524	0.9284	**0.8184**
CONV1D	0.7717	0.6329	0.9106	0.7943
TENCODER	0.7562	0.6025	0.9099	0.7896

(b) Cross validation results (mean across 2 folds). All methods use learned pooling (except from the baseline LSTM).

TABLE III: Comparison of different graph building methods and cross-validation results.

and 250-corresponding to 0.5s and 1.0s, respectively. For each, we vary the stride to control the degree of overlap between windows, ranging from no overlap to full overlap. On average, smaller window sizes yield better performance, supporting our earlier hypothesis that reduced temporal compression improves downstream model accuracy. Interestingly, for the 0.5s window, performance remains largely invariant to the stride size. We hypothesize that this is due to the downstream classifier jointly processing multiple windows, thereby diminishing the benefit of introducing overlap at the preprocessing stage. Since a larger stride results in fewer windows, an intriguing direction for future work is to further reduce window size while using a stride equal to the window length, eliminating overlap entirely.

Table IId reports the results of an ablation study on the choice of classifier, i.e., the component that aggregates the sequence of window-level graph representations produced by the GCN module into a single prediction. The best performance is achieved by an LSTM-based classifier enhanced with two key modifications. First, it employs a trainable pooling function that maps the node embeddings of each graph (corresponding to individual windows) into a single graph-level embedding. Second, it incorporates an attention mechanism, as detailed in Section II. Notably, the results indicate that the learnable pooling function is the primary driver of performance gains-substantially outperforming the default mean pooling strategy.

E. Building Graphs with EEG Spatial Priors

While previous methods primarily focused on modeling the temporal dimension, we now investigate approaches that incorporate spatial priors inherent to the EEG 10–20 electrode placement system. Specifically, we implement three strategies for constructing the graph structure. The first is the distance-based method described in Section II. The remaining two are defined as follows:

Correlation-based. A binary adjacency matrix $\mathbf{A}$ is constructed by thresholding the Pearson correlation coefficient $C(i, j)$ between time series of nodes i and j:

$$\mathbf{A}_{i,j} = \begin{cases} 1 & C(i,j) \geq T \\ 0 & \text{otherwise} \end{cases} \qquad T = 0.5 \tag{6}$$

Learned. The adjacency matrix $\mathbf{A}$ is treated as a trainable parameter and learned via backpropagation during training.

Table IIIa compares these approaches using the best-performing window-based architecture from Table IId. We evaluate the distance-based graph with two thresholds ($D = 1.0$ and $D = 1.5$), as well as the correlation-based and learned alternatives. Results indicate that the distance-based method with threshold $D = 1.5$ achieves the highest performance. We hypothesize that the fixed spatial prior encoded by this method provides a more generalizable inductive bias, particularly across different patients and recording sessions. Future work may explore combining distance- and correlation-based graphs, as proposed in [8], to further enhance performance.

F. Best Model and Investigation of the Performance Gap

Despite the strong performance of the LSTM baseline on the validation set, its significantly weaker results on the hidden test set (Kaggle) prompted further investigation. To obtain a more robust performance estimate, we conducted 2-fold cross-validation for the top-performing window slicing based models, as shown in Table IIIb. While the non-graph LSTM baseline still outperforms our window-based graph approaches on average in cross-validation, the latter consistently yield better results on the test set. In addition, validation/test set performance variance is minimal unlike for the non-graph baseline. Notably, our best overall model, described in Section II, achieves a macro-F1 score of 0.81 on the test set.

To explain the observed gap between validation and test performance, we hypothesized potential patient-level data leakage,i.e., training and validation sets containing recordings from the same individuals. However, this was ruled out after verifying that only a single patient had samples in both sets, contributing less than 1% of the validation data. We also ruled out performance underestimation on the minority class, which represents 20% of the validation split (260 samples).

IV. Conclusion

While non-graph models achieve competitive results for seizure prediction, they exhibit limited generalization-demonstrated by strong validation performance but significant drops on the test set. In contrast, graph-based methods provide a more robust framework by jointly capturing spatial and temporal dependencies. Our best-performing model combines a spatio-temporal graph architecture with learned pooling and a distance-based adjacency matrix. It converges in fewer epochs and generalizes better to unseen data, offering both improved performance and greater efficiency. These results underscore the potential of graph-based methods for scalable and reliable EEG-based seizure prediction.

References

[1] Z. Al Sahili and M. Awad. Spatio-temporal graph neural networks: A survey, 2023. arXiv preprint arXiv:2301.10569v2.

[2] D. Birvinskas, V. Jusas, I. Martišius, and R. Damaševičius. Fast dct algorithms for eeg data compression in embedded systems. *Computer Science and Information Systems*, 12:49–62, 01 2015.

[3] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[4] R. S. Fisher, W. van Emde Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel. Epileptic seizures and epilepsy: Definitions proposed by the international league against epilepsy (ilae) and the international bureau for epilepsy (ibe). *Epilepsia*, 46(4):470–472, 2005.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[6] X. Hu, S. Yuan, F. Xu, Y. Leng, K. Yuan, and Q. Yuan. Scalp eeg classification using deep bi-lstm network for seizure detection. *Computers in Biology and Medicine*, 124:103919, 2020.

[7] R. Hussein, H. Palangi, R. Ward, and Z. J. Wang. Epileptic seizure detection: A deep learning approach, 2018.

[8] J. Jia, B. Zhang, H. Lv, Z. Xu, S. Hu, and H. Li. Cr-gcn: Channel-relationships-based graph convolutional network for eeg emotion recognition. *Brain Sciences*, 12(8), 2022.

[9] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks, 2017.

[10] S. Kiranyaz, T. Ince, O. Abdeljaber, O. Avci, M. Gabbouj, and D. J. Inman. 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151:107398, 2021.

[11] G. H. Klem, H. Lüders, H. H. Jasper, and C. E. Elger. The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology. *Electroencephalography and clinical neurophysiology. Supplement*, 52:3–6, 1999.

[12] S. Mallick and V. Baths. Novel deep learning framework for detection of epileptic seizures using eeg signals. *Frontiers in Computational Neuroscience*, Volume 18 - 2024, 2024.

[13] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[14] D. L. Schomer and F. H. Lopes da Silva. *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford University Press, 11 2017.

[15] Y. Tang, Q. Wu, H. Mao, and L. Guo. Epileptic seizure detection based on path signature and bi-lstm network with attention mechanism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:304–313, 2024.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[17] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[18] World Health Organization, International League Against Epilepsy, and International Bureau for Epilepsy. Epilepsy: A public health imperative. Technical report, World Health Organization, Geneva, 2019.

V. Appendix

A. Data Preparation

We utilize a subset of the Temple University Hospital (TUH) EEG Corpus, one of the largest publicly available EEG datasets. Our data comprises recordings from 75 patients, each potentially containing multiple sessions. The dataset is partitioned into 50 patients for training and 25 for validation. Each session consists of multivariate time series sampled at 250 Hz, capturing electrical potentials (in mV) from 19 scalp electrodes positioned according to the international 10-20 system. We segment each session into non-overlapping 12-second windows, labeling each segment as either seizure or regular brain activity.

B. Related Work

Seizure detection from EEG signals has been approached using various deep learning models. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [5], have been widely used due to their effectiveness in modeling sequential dependencies. One-dimensional Convolutional Neural Networks (1D CNNs) [10] are also common, offering localized temporal feature extraction with fewer parameters.

Transformer-based architectures [16] have become a strong alternative in time-series modeling due to their ability to capture long-range dependencies through self-attention. PatchTST [13], a recent model built on Transformers, improves long-term forecasting performance by using patch representations for more stable attention learning in time-series data.

Incorporating signal descriptors with neural models has also proven effective. For example, combining path-signature features with Bi-LSTM and attention mechanisms [15] has led to state-of-the-art performance in seizure detection.

Graph-based approaches have been increasingly adopted to capture spatial dependencies in multivariate signals such as EEG. Graph Convolutional Networks (GCNs) [3], [9] extend convolution operations to non-Euclidean graph domains by leveraging the spectral properties of the graph Laplacian. These methods, initially developed for general tasks such as node classification and link prediction, have since been integrated into spatio-temporal GNN frameworks to model spatial structures dynamically.

Graph Attention Networks (GATs) [17] introduced attention mechanisms into graph learning by enabling the model to learn the importance of neighboring nodes in a data-driven manner. Originally applied to static graphs, GATs have become a key component in spatio-temporal models by enabling more flexible and context-aware spatial modeling.

Recent surveys [1] emphasize that spatio-temporal GNNs typically combine a graph-based spatial module (e.g., GCN, GAT, GIN) with a temporal module (e.g., LSTM, 1D CNN, Transformer). These hybrid frameworks have been successfully applied in applications such as traffic forecasting, brain connectivity analysis, and pandemic prediction.

Our work builds upon these developments by comparing non-graph and graph-based methods for seizure detection while exploring architectural improvements in temporal processing and graph design informed by domain-specific priors.

C. Training efficiency of window-based approaches

Although our spatio-temporal graph models contain substantially more parameters (approximately 450,000) than LSTM counterparts (approximately 90,000), they converge significantly faster-typically within 45 epochs versus 950 for LSTMs. This efficiency likely stems from the structured representation afforded by subwindow decomposition and graph-based processing. Consequently, despite their larger size, these models achieve superior training efficiency in both time and computational resources.

D. Future work

Future work may improve model expressiveness by refining the graph structure and node representations. Specifically, employing continuous adjacency weights in place of binary connections would enable the model to capture nuanced degrees of electrode connectivity. Furthermore, integrating explicit positional information into node features could provide crucial spatial context, thereby enhancing the model's capacity to learn meaningful patterns from EEG data.