Graph Convolution Network Based Classification of Subjects with Prefrontal Cortex Lesion via Information-Theoretic Brain Network Features

Sai Sanjay Balaji and Keshab K. Parhi

Department of Electrical & Computer Engineering, University of Minnesota, Twin Cities, , Minneapolis, 55455, MN, USA.

*Corresponding author(s). E-mail(s): parhi@umn.edu;
Contributing authors: balaj037@umn.edu;

Abstract

This paper investigates scalp electroencephalogram (EEG) data from 14 subjects with unilateral prefrontal cortex (pFC) lesions and 20 healthy controls during lateral visuospatial working memory (WM) tasks. The goal is to differentiate the brain networks involved in WM processing between these groups. The EEG recordings are transformed into graph signals, with proximity-weighted brain connectivity measures as edges and centrality measures as nodal features. Graph convolutional network (GCN) layers are used for feature representation, followed by a fully connected layer for classification. The GCN-based model effectively handles nine classification tasks, proving that graph-based network representation is versatile for describing brain interactions. The sparse MI-GCI-based graph model's accuracy effectively captures the functional segregation of distinct WM tasks. The classifier using *mutual information-guided Granger causality index (MI-GCI)* with 20% of top edges matched prior classification performance with 67% fewer parameters and 80% less graph density, identifying the correct class of all 34 subjects in group identification using leave-one-out cross-validation and two-thirds majority voting.

Keywords: Brain network, effective connectivity, graph convolution (GCN) networks, mutual information, prefrontal cortex (pFC) lesions

1 Introduction

Working memory (WM) is representative of the capability to actively store information in the human brain for a short duration [1]. The deterioration of WM can be used to diagnose neurological disorders affecting cognition, such as Alzheimer's and Parkinson's [2, 3]. Furthermore, the modification of WM through therapy has been shown to treat anxiety symptoms and post-traumatic stress disorder (PTSD) [4]. These factors underscore the requirement for understanding the dynamics in cognitive functions that facilitate WM. Understanding the neural pathways responsible for WM can ease its modification process, treat the symptoms of severe neurological disorders, and provide insight into human cognition.

The prefrontal cortex (pFC) of the human brain's frontal lobe, which has been instrumental in complex cognitive behavior, personality expression, decision-making, and moderating social behavior [5], also plays a vital role in the WM process. Prior studies using brain imaging techniques that employed different methods to test WM revealed a linear relationship between pFC activity and WM load [6, 7]. Despite the existence of such studies indicating the essential role of pFC in the WM process, it is surprising to observe patients successfully complete WM tasks despite suffering severe tissue damage (lesions) in the pFC cortical region. This discovery alludes to alternate neural pathways for WM that may not rely on pFC. The study in [8] supported this claim by showing that pFC activity is not always necessary in WM tasks. However, further work is required to describe the consequences of damage to pFC tissues on memory encoding and the reason behind successful memory encoding and retrieval despite such impairment.

Researchers are increasingly turning to advanced computational techniques, particularly machine learning (ML) and signal processing, to delve deeper into the neural mechanisms of WM. Traditional ML and deep learning (DL) models have become ubiquitous in analyzing brain electrophysiology datasets as they show robust predictions with high accuracy [9, 10]. Despite their utility, these traditional methods often disregard the underlying network-like structure of the brain [11]. A graphical representation is better suited to explain coherent electrophysiological activities across the different cortical regions that are spatially segregated but functionally connected [12]. The increased evidence towards brain connectivity-based analysis revealing functional differences of different cortical regions and their intercommunications induced a significant shift toward connectome-based analysis in EEG signals. *Connectivity* within the brain ranges from anatomical interconnection to functional communication between multiple brain regions. Brain connectivity is broadly categorized into three levels [13]. Anatomical or *Structural connectivity* is defined by the neural pathways between the two areas, which can be identified using noninvasive imaging techniques. *Functional connectivity* is defined by temporal correlation measures in the electrophysiological activity of neuron populations in two distinct regions. The *effective connectivity* measure aims to provide a sense of directional causality by quantifying the influence of activity in one neural system exerted directly or indirectly over another.

Several studies have explored brain network properties, highlighting the significance of functional connectivity during cognitive tasks [14]. Abnormal connectivity patterns have been linked to various neurological and psychiatric disorders [15].

Functional connectivity is crucial for characterizing brain network architecture and understanding small-world properties that support cognitive functions like memory consolidation and information integration [16]. It also illuminates dysfunction dynamics in conditions such as schizophrenia [17] and depression [18], offering clinical insights. However, methodological issues can affect data interpretation [19], including the assumption of connectome stationarity during EEG recordings and risks of false positives or negatives from tiny sample sizes or poor data selection. Defining connectome nodes and edges and managing artifacts like eye movements are further challenges. High-density EEG requirements and lengthy acquisition times limit connectome analyses in clinical or resource-constrained settings. Connectivity measure choice is crucial, with some measures affected by volume conduction, which can distort inter-regional activity dependencies. Techniques like source localization and signal decomposition help but add complexity.

Effective connectivity (EC), describing directional information flow within brain networks, is vital in cognitive neuroscience, revealing neural mechanisms behind perception, decision-making, and cognitive control [13]. This approach is popular in neuroscience for exploring causal brain region relationships, though capturing neuron activity with limited channels is debated [20]. EC measures excel in tasks like classifying brain states, pinpointing seizure foci, detecting neurological conditions, and distinguishing patients with prefrontal cortex lesions during working memory trials [21–25]. It involves deriving directional measures from the statistical interdependence of electrophysiological signal time-series data [26]. This manuscript's references to *causal networks* pertain to EC measures.

2 Related Work and Contributions of the Paper

EEG studies have long investigated the neural mechanisms behind WM processes, including encoding, maintenance, and retrieval, by analyzing brain activity across distinct phases. Traditional research has often focused on isolated EEG-derived measures such as spectral power and event-related potentials (ERPs) to identify how regions, such as the pFC, engage across WM tasks. However, such methods have faced limitations in capturing the complex, dynamic nature of WM [27, 28].

Historically, studies have shown that the pFC plays a central role in WM, modulating neural activity during encoding, maintenance, and retrieval tasks. For example, early EEG and fMRI studies demonstrated distinct oscillatory patterns in the pFC associated with each WM phase. Encoding is often linked with increased theta power in frontal regions, while the maintenance phase is associated with alpha and theta oscillations in frontoparietal circuits [29, 30]. Retrieval is marked by frontal-midline theta and beta activity changes, highlighting pFC engagement in re-accessing stored information [31, 32]. Despite these insights, traditional EEG studies encounter limitations in isolating the phases and defining the extent of interaction between the pFC and other WM-related regions, such as the parietal cortex. Classical EEG metrics capture linear correlations but often need to improve their representation of the entire network dynamics necessary for complex WM tasks [32].

Recent advancements in ML techniques have improved the classification of WM phases from EEG data by extracting features like power spectral density and connectivity measures such as coherence and phase-locking value. These methods have shown promising performance, with accuracies typically ranging from 75% to 85%, depending on the dataset and feature set [33, 34]. However, relying on model-based static connectivity measures hinders the ability to capture the dynamic interactions of brain networks, and these approaches are often tailored to subject-specific analysis, which requires large, individual datasets to achieve accurate classifications.

Graph theory has emerged as a robust framework for modeling brain networks, capturing both intra- and inter-regional interactions. EEG-based studies have applied graph metrics, such as clustering coefficients and path lengths, to characterize WM networks across different phases. For instance, recent studies have used graph metrics to show how different EEG-derived networks represent encoding, maintenance, and retrieval in WM [35]. Graph-based convolutional networks (GCNs) [36] will be particularly useful in this context as they are equipped to handle the non-Euclidean nature of EEG connectivity data, allowing researchers to explore high-dimensional relationships among brain regions that classical methods could not capture. Studies incorporating GCNs have shown the potential to model phase-specific networks more precisely by leveraging temporal and spatial EEG features to detect phase transitions [37, 38]. This progression underscores a broader trend towards network-centric frameworks in neuroscience, where complex cognitive processes are understood through their distributed neural architectures rather than isolated regional activity alone.

This paper proposes a data-driven, model-free approach to investigate brain connectivity as a dynamic graph signal that overcomes previously discussed limitations. We investigate brain connectivity by employing measures based on mutual information to capture inter-channel interactions in scalp EEG recordings as graph signals. We focus on utilizing a graph neural network architecture to classify the brain network organization in two distinct populations: healthy controls and subjects with unilateral PFC lesions during working memory trials. The classification approach allows us to characterize the group-specific network dynamics during different phases of three distinct working memory tasks. This paper is an expanded version of [24], which showcases the potential of graph convolutional network (GCN) models in capturing the cognitive aspects of WM encoding in both healthy individuals and patients with pFC lesions. Specific contributions of the paper are outlined as follows:

- *Emphasis on mutual information*: This paper focuses on applying information theory, specifically guided by mutual information (MI), to obtain functional and effective connectivity measures.
- *Incorporation of spatial aspects in defining the graph adjacency matrix*: In the prior work [24], the adjacency matrix was constructed using the absolute value or change in connectivity measure, resulting in a highly dense matrix without further processing. In this paper, we build a sparse representation of the adjacency matrix. To achieve this, we consider the top k % of the overall connectivity features and scale them using a proximity matrix to attenuate long-distance connectivity features.

- *Differentiation between types of WM tasks*: Our previous work in [24] treated all kinds of WM tasks (identity, spatial, and temporal relation tasks) as a unified category. However, in this paper, we conduct a more detailed analysis by segregating the three types of WM trials and developing separate classifiers for each case.
- *Expansion of the classifier to incorporate different phases within a WM trial*: The previous study only considered the relative change in connectivity features from pretrial to encoding for the classifier [24]. In contrast, this paper includes all stages (pretrial, encoding, maintenance, and processing) in the analysis.
- *Model simplification*: The classifier in the previous work[24], consisted of 7378 trainable parameters, incorporated four centralities and five relative band power measures for node features, and did not utilize any technique for edge feature reduction. In this paper, we have taken multiple steps to simplify the GCN-based classifier by imposing sparsity and reducing the number of nodal features to achieve a performance comparable to the prior work.

3 Materials and Methods

Fig. 1 provides an overview of the algorithm utilized in this manuscript to classify patients with prefrontal cortex (PFC) lesions compared to healthy controls. The algorithm encompasses three principal steps, namely: extraction of brain connectivity measures from scalp EEG recordings, conversion of connectome features into graph signals, and classification of the resulting graph signals. The task and dataset used in this analysis are briefly reviewed before outlining the key steps of the algorithm.

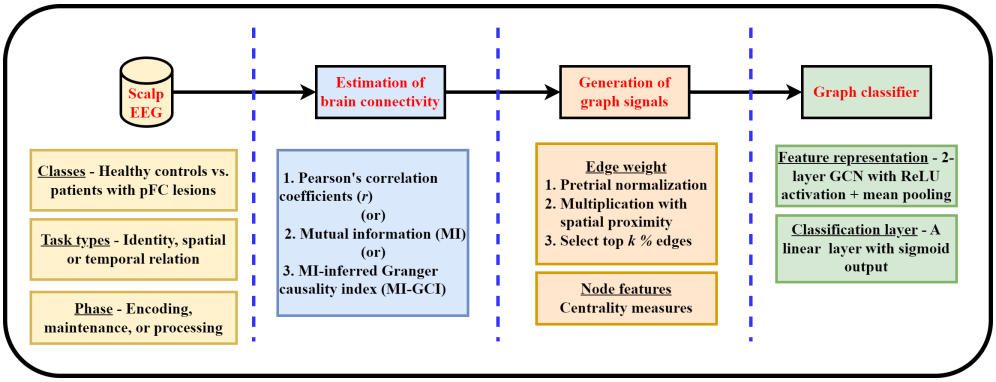

Fig. 1 Algorithm Overview for Classifying Patients with pFC Lesions from Healthy Controls: Estimation of Brain Connectivity from Scalp EEG, Generation of Graph Signals, and GCN-based Classification.

3.1 Assessment of working memory (WM)

The WM is tested using *lateralized visuospatial* working memory tasks. These tasks fall under two categories [8]:

(a) *Identity test* - Subjects are shown a pair of shapes and then asked to identify whether a given pair of shapes is the same as they had just observed.

(b) *Spatio-temporal relation test* - The subjects are initially shown a pair of shapes (similar to an identity test). The spatial aspect is examined by subjects, indicating the shape observed in the top/bottom, and the temporal element is reviewed by them, indicating the shape observed first/second.

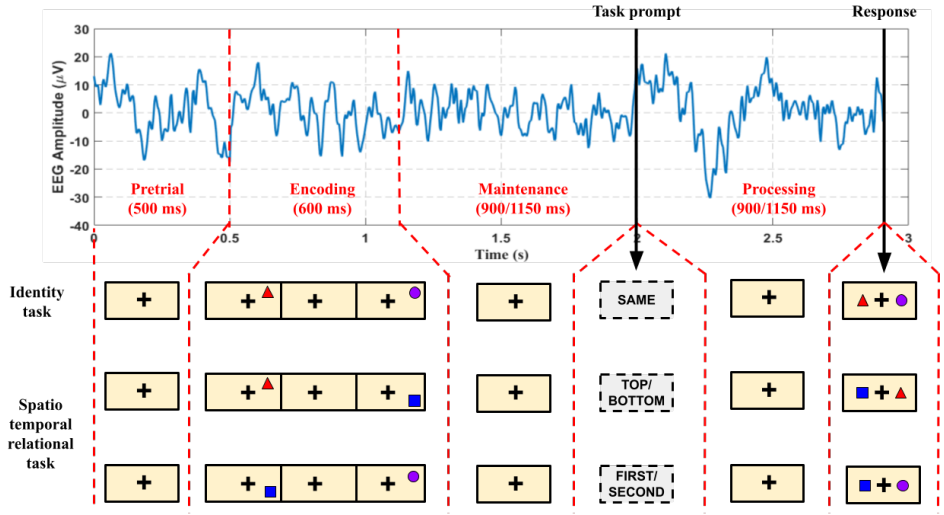

Fig. 2 Scalp EEG and screen display during the different phases of a lateralized visuospatial WM task.

Each WM trial can be divided into five phases, as illustrated in Fig. 2. Central fixation is shown to record the resting state EEG during the 2 s *pretrial* phase. After this, subjects are shown two common shapes in a top/bottom spatial orientation for 200 ms each sequentially with a 200 ms break in between, marking the *encoding* phase. A 900 ms or 1150 ms *maintenance* interval follows the encoding phase, where the subjects actively store the information shown during the encoding phase. This is followed by the *active processing* stage, where a text prompt appears for the same duration as the maintenance phase. Finally, the subjects indicated their *response*.

3.2 Dataset and preprocessing

The data set consists of scalp EEG recorded from 20 healthy human subjects (control) and 14 patients with unilateral pFC lesions performing multiple trials of the lateralized visuospatial working memory task [39]. The University of California, Berkeley, the Institutional Review Board, and the Regional Committee for Medical Research Ethics,

Region South approved the study protocol, and the study was carried out as per the Declaration of Helsinki, and all participants gave informed written consent. Using a 64 + 8 channel BioSemi ActiveTwo amplifier with Ag-AgCl pin-type active electrodes mounted on an elastic cap, the scalp EEG was recorded at 1024 samples/sec. Each participant completed 120-240 working memory task trials, with each trial having an equal probability of being an identity or relation type. The EEG signals from the 64 channels were recorded during the five phases.

The preprocessing of EEG data, as described in [8], is blinded to group membership. First, trials, where the gaze deviated to include the ipsilateral visual hemifield during stimulus presentation, were excluded to maintain data integrity and prevent eye movement artifacts. Following this, the raw data was referenced to the mean potential of two earlobe electrodes, down-sampled to 256 Hz, and filtered using 1-Hz high-pass and 70-Hz low-pass finite impulse response filters. Electromyography artifacts were automatically removed using the AAR external plug-in with default settings for a 30-second sliding window. The 60-Hz line noise harmonics were eliminated using a discrete Fourier transform. The continuous data was then epoched into 1000-ms buffers, with trials flagged based on eye gaze position excluded. A manual inspection was conducted to remove channels with abnormal signals, followed by independent components analysis to eliminate artifacts further. Rejected channels were replaced by interpolating the mean of neighboring channels. Finally, for patients with pFC lesions, a surface Laplacian filter was applied to refine connectivity estimates and minimize volume conduction. Channels were swapped across the midline in patients with right-hemisphere lesions to normalize them to the left hemisphere. This exact procedure was applied to 10 randomly selected control datasets to prevent potential confounding effects from inter-hemispheric variation. The recordings for each trial were then segregated into three multivariate time-series corresponding to the pretrial phase, encoding and maintenance phase, and active processing phase. Only successful WM trials are considered for our analysis. Table 1 summarizes the statistics of successful WM trials across the two groups for the three types of tasks.

Table 1 Statistics of different types of successful WM tasks among control and subjects with pFC lesions

Task type	Number of trials (mean ± std.)	
	Control (n=20)	pFC leisoned (n=14)
Identity	58.35 ± 11.8	49.2 ± 15.9
Spatial	63.0 ± 9.9	56.4 ± 16.7
Temporal	60.2 ± 12.3	53.6 ± 19.5

3.3 Brain connectivity measures

The two information-theoretic measures of brain connectivity examined in this manuscript are discussed below:

Table 2 Key Parameters and Steps Used in Estimating Different Connectivity Measures

S.No	EC measure	Assumptions, key parameters, and estimation method
1.	**Pearson's r and MI**	• The scalp EEG recording corresponding to each phase of a WM trial is split into smaller stationary windows of 200 ms duration with 20 ms step size. • The linear correlation coefficients and the bias-corrected GC normalized MI values are evaluated for every channel pair for each time window. The mean r and MI values across the windows are considered for the given phase of a given trial. • As the r and MI values are symmetric, i.e., $r_{xy} = r_{yx}$ and $I(\mathbf{X};\mathbf{Y}) = I(\mathbf{Y};\mathbf{X})$, the scalp EEG recorded from 64 channels result in $\binom{64}{2} = 2016$ functional connectivity measures in each case.
2.	**MI-GCI**	• The estimation of EC measures also involves the identification of optimal lag/Markovian parameters that characterize the number of past samples that influence the value of the present sample. For the analyzed dataset, the optimal lag value $\tau = 5$ and embedding dimension k=2 were identified using the average mutual information (AMI) function and false nearest neighbors (FNN) [40]. These were used to construct the embedded vectors for the two time-series before calculating the directed information between a channel pair[41]. • As the EC measure of MI-GCI is non-symmetric, the recordings from 64 EEG channels produce $2 \times \binom{64}{2} = 4032$ directional features. • For given channel-pair *'x'* and *'y'*, two directional measures/features are extracted (*'x'*→*'y'* and *'y'*→*'x'*). The difference between the two values provides the net directional flow of information between the two channels. To generalize, if $\mathbf{C}^{raw}$ represents the original 64×64 connectivity matrix, the simplified sparse representation, $\mathbf{C}$, is defined as $$c_{ij} = max\{c_{ij}^{raw} - c_{ji}^{raw}, 0\},\ i,j = 1,2,\ldots,64 \quad (1)$$

1. Mutual information (MI): MI quantifies the dependence between variables by measuring how much knowing one variable reduces uncertainty about the other [42]. It captures linear and nonlinear dependencies and is less affected by spatial blurring like volume conduction, making it more sensitive than correlation-based measures. However, MI assumes signals are Markov processes, necessitating more

data samples and intensive computation. Previous works have used MI to study brain region dependencies and functional connections, combining it with graph theory for brain network analysis [43], and integrating it with deep neural networks to improve cognitive workload prediction [44], highlighting its importance in understanding brain organization.

2. MI-guided Granger causality index (MI-GCI): Granger causality (GC) is a statistical test for inferring causality between time-series and is extensively used in neuroscience to explore directional interactions between brain regions and oscillatory activities [45]. However, GC measures have limitations in capturing nonlinear dependencies and can be affected by assumptions of linearity and Gaussianity. To address these limitations, researchers have proposed MI-based approaches to estimate GC, leveraging the Kullback-Leibler divergence to assess predictability improvements while capturing nonlinear relationships and being tolerant to volume conduction effects [46].

Despite its linear limitation, Pearson's correlation (r) is a popular measure in brain network analysis, so it would be apt to include it and compare its performance against the two information-theoretic measures. Table 2 summarizes the assumptions, key parameters, and techniques used to estimate the connectivity measures

3.4 Graph modeling of brain connectivity:

A graph signal $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{A}\}$ is described by a set of nodes $\mathcal{V}$, a set of edges $\mathcal{E}$ and the adjacency matrix $\mathbf{A}$. The adjacency matrix is a square matrix of size $N \times N$, where N is the total number of nodes in the graph. The adjacency matrix is symmetric for an undirected graph, i.e., $A_{ij} = A_{ji}$. This subsection discusses the steps taken to generate a graph signal representative of a brain network from the connectivity measures discussed in section 3.3

3.4.1 Pretrial normalization of connectivity features

A group-level analysis involving numerous trials from different subjects suffers from inter-trial and inter-subject variances arising from EEG recordings and the subject's physiology [47]. To tackle this issue and to improve the inter-trial and inter-subject associativity of the brain connectivity measures, the change in the connectivity measure from the previous phase relative to the pretrial baseline is used as the modified connectivity feature ($\mathbf{C}$) for the graph modeling. The change in connectivity measure of directed information has proven in prior work to overcome inter-subject variances while classifying subjects during WM encoding in [24] and [25]. The modified connectivity feature can be mathematically represented as shown in equation (2), where the previous phase for encoding, maintenance, and processing phases are pretrial, encoding, and maintenance phases, respectively.

$$\mathbf{C} = \frac{|\mathbf{C}_{phase} - \mathbf{C}_{prev\ phase}|}{\mathbf{C}_{pretrial}} \tag{2}$$

3.4.2 Spatial and connectivity-based adjacency matrix

The connectivity matrix, denoted by $\mathbf{C}$, is a $N \times N$ matrix representing the functional or effective brain connectivity measure. Incorporating the connectivity matrix for generating the adjacency matrix ($\mathbf{A}$) is commonly used in graph-based brain network classification problems [48]. However, spatial connectivity plays a vital role in determining a suitable graphical model of brain connectivity. The actual flow of information between two cortical regions depends on the distance between them despite solid functional connectivity [49]. For this reason, the effect of distance while generating the adjacency matrix is incorporated using a proximity matrix. The steps to create the proximity matrix $\mathbf{P}$ are discussed below:

- First, the Euclidean distance between all pairs of regions is obtained to generate the $N \times N$ distance matrix $\mathbf{D}$ ($N = 64$ in this case).

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{3}$$

- The distance matrix is normalized to the range $[0, 1]$

$$\widetilde{d_{ij}} = \frac{d_{ij}}{\max\limits_{\{0 \leq i,j \leq N\}} d_{ij}} \tag{4}$$

- The proximity matrix ($\mathbf{P}$) is generated from the normalized distance matrix using an exponential function

$$p_{ij} = exp(-h\widetilde{d_{ij}}^2), \ h > 0 \tag{5}$$

  $h = 5$ is chosen to indicate minimal spatial proximity of $e^{-5} \approx 0.006$ between the regions with maximal spatial distance.

A sparse graph representation is obtained by retaining only the top k % of the 2016 edges. The value of k is varied from 10 to 30 in increments of 10 to identify the optimal sparsity that still results in superior prediction. Finally, the adjacency matrix is defined by the element-wise multiplication of the proximity matrix and connectivity matrix ($a_{ij} = c_{ij} \times p_{ij}$).

3.4.3 Network nodes and node features

The 64 scalp EEG channels are represented as graph nodes. The node features are derived from the topological graph measures of node centrality inferred from the different connectivity features during the WM trials' encoding, maintenance, and processing phases. The four centrality measures used in this study are discussed below:

- *Betweenness centrality* (BC) is a network analysis measure that quantifies a node's importance or centrality within a network based on its position in facilitating communication between other nodes. It estimates the extent to which a node lies on the shortest paths between pairs of other nodes in the network [50]. Nodes with high BC have a more significant influence on the flow of information

in the network. They act as critical intermediaries, facilitating communication and enabling efficient transfer of information between different parts of the network. Prior works have established BC as a reliable measure of a node's influence over information flow within a graph while identifying significant regions during cognitive tasks [51, 52].

- *Eigenvector centrality* considers the number of connections a node has and the importance of those connections. Nodes with high eigenvector centrality are not only well-connected but also connected to other highly influential nodes [53]. It is estimated iteratively based on the principle that a node's importance is proportional to the importance of its neighbors.
- *PageRank centrality* is calculated iteratively based on the concept that a node's importance depends on the significance of the nodes that link to it. It considers the incoming links and redistributes the importance scores among nodes. Initially developed by Google to assess the importance of a page on the web [54], PageRank has now found application in many graphical structures to assess the relative importance of a node.
- *Closeness centrality* quantifies how close a node is to other nodes regarding the shortest path distance. It captures how quickly information can spread from a node to other nodes in the network [55]. The closeness centrality of a node is calculated by taking the reciprocal of the average shortest path distance from that node to all other nodes in the network.

3.5 Graph classification model

3.5.1 Overview of GCN

Graph Convolutional Networks (GCNs) [36] have gained significant attention for their ability to process data structured as graphs. *Graph convolution* forms the core of GCNs, which extends the convolution operation from regular grids to irregular structures represented as graphs. The graph convolution operation involves aggregating information from neighboring nodes in a graph and updating the features of each node based on this aggregated information. This enables GCNs to capture local and global patterns in graph-structured data [56]. For a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{A}\}$, each node $\mathcal{V}_i$ is associated with a feature vector $\vec{x}_i$, representing the characteristics of that node. A GCN aims to learn node representations that encode both the local information from a node's own features and the global information from its neighboring nodes.

In brain connectivity analysis, GCNs can model functional or structural connectivity patterns between brain regions using EEG data, enabling the detection of abnormal connectivity associated with neurological disorders, identification of cognitive biomarkers, or decoding mental states from EEG signals [57–59]. By integrating the strengths of graph representation and convolutional operations, GCNs offer a promising framework for extracting meaningful information from complex brain networks represented by EEG data [60].

The graph neural network operator described in [61] exhibits greater expressive power than standard GNNs, including GCNs. It surpasses GNNs in terms of its ability

to distinguish non-isomorphic subgraphs and is capable of capturing and distinguishing a wider range of graph properties. Thus, the described operator was used in lieu of the standard graph convolutional layers in our analysis. The forward pass of the graph convolution using this operator consists of a combination of message passing, aggregation, and linear transformation operations and can be mathematically described as:

$$\mathbf{x}_i' = \mathbf{w}_0 + \mathbf{W}_1\mathbf{x}_i + \mathbf{W}_2 \sum_{j \in \mathcal{N}(i)} e_{ij}.\mathbf{x}_j \tag{6}$$

where $\mathbf{x}_i$ is the input feature vector of node i, $\mathbf{x}_i'$ is the convolutional output, $\mathcal{N}(i)$ represents the neighborhood of node i, e_{ij} denotes the weight for the edge to node i from its neighbor node j, $\mathbf{w}_0$ is the bias vector, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices. Any future reference to GCN in this manuscript refers to the above-mentioned graph operator [61] unless explicitly mentioned otherwise.

3.5.2 Classifier architecture

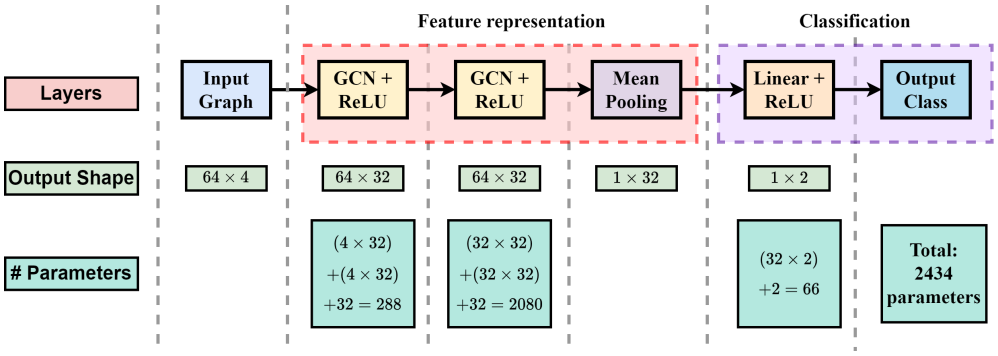

Fig. 3 Architecture of the GCN-based classifier showing the number of parameters for each layer: i) Two layers of GCN followed by mean pooling to learn the graph representation, ii) Fully connected layer for classification.

The architecture of the proposed GCN-based classifier is illustrated in Fig 3. The input to the classifier is the graph represented by its adjacency matrix and node features. Feature representation from the generated graph is performed using two layers of GCN, followed by a mean pooling layer that also acts as a regularizer. Each GCN layer consists of 64 nodes corresponding to the 64 channels with a rectified linear unit (ReLU) activation function and 32 output features. A ReLU-activated dense layer follows the GCN layers and forms the output layer. Other hyper-parameters, such as batch size, learning rate, number of epochs, and optimization algorithm, are tuned to obtain the best-performing model. Callback features are employed to reduce the learning rate and for early stopping by monitoring the training and validation loss.

The leave-one-subject-out cross-validation (LOOCV) technique is employed, with the trials corresponding to one subject held out for testing and the remainder used for

training. This is recursively repeated until all subjects are tested. The training data is further split into independent training and validation sets in the ratio 7:3. In each case, a single label is assigned to the test subject using two-thirds majority voting from the predicted labels of all the trials of the test subject. A majority failure is also considered a misclassification when calculating the overall performance.

4 Results

Table 3 summarizes the performance of the three brain connectivity features for various phases of the three types of WM trials. All the methods correctly identified the 20 control subjects in all iterations, resulting in a 100 ± 0 % specificity. The sensitivity, however, varied based on the chosen connectivity parameter. The primary observation is the superiority in classification by the MI-GCI compared to the two functional connectivity measures. MI-GCI not only consistently outperforms the other measures in terms of mean accuracy, but it is also robust across the iterations, as indicated by the significantly smaller standard deviation. Note that there are only 14 patients with pFC lesions. Thus, misclassifying even one subject results in a sensitivity drop of over 7%. Considering only the top $k = 20\%$ of the edges results in the best performance in 18 of the 27 cases shown in Table 3. For the other one-third of the cases, the difference in performance with $k = 20\%$ from the best performing model is less than 10% (except for the encoding phase of temporal task using r values). Based on this observation, the optimal value for k that results in a sparse graph signal without significantly compromising the classification performance is 20%.

5 Conclusion

This study demonstrates the effectiveness of information-theoretic measures in modeling brain connectivity for classifying WM stages among healthy controls and individuals with unilateral pFC lesions. EEG time series are transformed into graph signals using MI and MI-GCI measures, incorporating spatial information and retaining the top 20 % edges. Centrality measures are used as nodal features, and a 2-layer GCN classifier identified all 34 subjects with 100% accuracy across nine classification tasks using LOOCV and majority voting.

This paper introduces several advancements over prior work [24]. It incorporates spatial considerations by selecting the top k% of connectivity features and scaling them based on proximity, achieving a graph that is 80 % more sparse compared to the dense matrix in [24] without affecting the classification performance. It differentiates between identity, spatial, and temporal relation tasks for more detailed analysis and develops separate classifiers for each WM task type. Additionally, the classifier now includes all phases of WM trials and achieves 100% accuracy with a reduced parameter count of 2434, a 67% decrease compared to [24]. Table 4 summarizes these improvements.

The study's limitations include the small sample size, which may affect generalizability despite pretrial normalization and LOOCV efforts. Segmenting EEG recordings into smaller time windows and assessing mean performance addressed the need for stationarity assumptions in neural signals. Pretrial normalization mitigated inter-trial

Table 3 Performance of GCN-based classifier after retaining the top k % of the graph edges for $k = 10, 20, 30$: **Sensitivity reported as % of the 14 patients with pFC lesions identified correctly by the algorithm based on two-thirds majority voting.** All models achieve 100% specificity, i.e., the control subjects in all cases are identified correctly.

WM type	Phase	Connectivity	Sensitivity (mean ± std. %)		
			k = 10%	k = 20%	k = 30%
Identity	Encoding	R	73.8 ± 10.1	**78.6 ± 13.5**	76.2 ± 16.8
		MI	61.9 ± 37.0	**69.0 ± 26.9**	57.1 ± 26.9
		MI-GCI	**100.0 ± 0.0**	92.9 ± 6.7	97.6 ± 3.4
	Maintenance	R	83.3 ± 16.8	**95.2 ± 6.7**	83.3 ± 10.1
		MI	83.3 ± 16.8	**95.2 ± 6.7**	83.3 ± 20.2
		MI-GCI	95.2 ± 3.4	**100.0 ± 0.0**	97.6 ± 3.4
	Processing	R	95.2 ± 6.7	**95.2 ± 3.4**	85.7 ± 10.1
		MI	85.7 ± 16.8	90.5 ± 10.1	**95.2 ± 6.7**
		MI-GCI	**100.0 ± 0.0**	**100.0 ± 0.0**	95.2 ± 3.4
Spatial	Encoding	R	71.4 ± 10.1	**81.0 ± 10.1**	78.6 ± 13.5
		MI	97.6 ± 3.4	**100.0 ± 0.0**	95.2 ± 6.7
		MI-GCI	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**
	Maintenance	R	**90.5 ± 3.4**	81.0 ± 10.1	85.7 ± 13.5
		MI	90.5 ± 10.1	88.1 ± 13.5	**97.6 ± 3.4**
		MI-GCI	95.2 ± 3.4	**100.0 ± 0.0**	95.2 ± 3.4
	Processing	R	76.2 ± 16.8	**90.5 ± 10.1**	88.1 ± 13.5
		MI	78.6 ± 16.8	81.0 ± 10.1	**92.9 ± 6.7**
		MI-GCI	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**
Temporal	Encoding	R	**90.5 ± 10.1**	73.8 ± 16.8	88.1 ± 13.5
		MI	76.2 ± 20.2	**88.1 ± 13.5**	71.4 ± 23.6
		MI-GCI	95.2 ± 6.7	**100.0 ± 0.0**	92.9 ± 6.7
	Maintenance	R	95.2 ± 3.4	**100.0 ± 0.0**	97.6 ± 3.4
		MI	85.7 ± 16.8	85.7 ± 10.1	**88.1 ± 13.5**
		MI-GCI	95.2 ± 3.4	**100.0 ± 0.0**	**100.0 ± 0.0**
	Processing	R	83.3 ± 16.8	88.1 ± 13.5	**97.6 ± 3.4**
		MI	88.1 ± 13.5	88.1 ± 6.7	**90.5 ± 6.7**
		MI-GCI	**100.0 ± 0.0**	**100.0 ± 0.0**	97.6 ± 3.4

Table 4 Comparative summary of prior work and proposed work on GCN-based classifier during WM tasks between healthy controls and subjects with pFC lesions. Both methods identified accurately all 34 subjects using LOOCV and two-thirds majority voting.

	Prior work [24]	**Proposed work**
Effective connectivity (EC) measure	Directed information (DI)	Mutual information-guided Granger causality index (MI-GCI)
Graph adjacency matrix	Only considers the connectivity estimate between all channel pairs	The connectivity estimate is scaled based on spatial proximity
Nodal features	Four centralities and five relative band power features (total 9)	Four centrality measures
Distinction between different types of WM tasks	No distinction	Separate models developed for identity, spatial, and temporal relational tasks
WM phase analyzed	Encoding phase	Encoding, maintenance, and processing phases
Graph density	Dense full connected graph signal	Sparse graph signals with top 20% edges
# of parameters used for the GCN classifier	7378	2434 (reduction of $\sim$ 67%)

and inter-subject variances, but significant inter-subject variance remains challenging. Selecting an appropriate lag for MI-GCI is crucial, as incorrect lag selection can impact the accuracy of connectivity measurement.

Future research should explore the impacts of pFC lesions on connectivity and cognitive processes using larger, more diverse cohorts and multimodal data, such as combining EEG with functional MRI or diffusion tensor imaging. These steps will validate findings, enhance the generalizability of the classification framework, and advance our understanding of WM neural mechanisms.

Declarations

Acknowledgements. This work was supported in part by the National Science Foundation under grant number CCF-1954749.

Conflict of Interest. The authors have no conflicts of interest or competing interests to declare that are relevant to the content of this article.

Data availability. The data used for analysis is publicly available at the Collaborative Research in Computational Neuroscience (CRCNS) repository [39] (https://crcns.org/data-sets/pfc/pfc-5) and can be downloaded by creating a CRCNS.org account.

References

[1] Baddeley, A.: Working memory. Science **255**(5044), 556–559 (1992)

[2] Morris, R.G.: Working memory in Alzheimer-type dementia. Neuropsychology **8**(4), 544 (1994)

[3] Kensinger, E.A., Shearer, D.K., Locascio, J.J., Growdon, J.H., Corkin, S.: Working memory in mild Alzheimer's disease and early Parkinson's disease. Neuropsychology **17**(2), 230 (2003)

[4] Andrade, J., Kavanagh, D., Baddeley, A.: Eye-movements and visual imagery: A working memory approach to the treatment of post-traumatic stress disorder. British journal of clinical psychology **36**(2), 209–223 (1997)

[5] Euston, D.R., Gruber, A.J., McNaughton, B.L.: The role of medial prefrontal cortex in memory and decision making. Neuron **76**(6), 1057–1070 (2012)

[6] Cohen, J.D., Forman, S.D., Braver, T.S., Casey, B., Servan-Schreiber, D., Noll, D.C.: Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI. Human brain mapping **1**(4), 293–304 (1994)

[7] Funahashi, S.: Prefrontal cortex and working memory processes. Neuroscience **139**(1), 251–261 (2006)

[8] Johnson, E.L., Dewar, C.D., Solbakk, A.-K., Endestad, T., Meling, T.R., Knight, R.T.: Bidirectional frontoparietal oscillatory systems support working memory. Current Biology **27**(12), 1829–1835 (2017)

[9] Sanei, S., Chambers, J.A.: EEG Signal Processing. John Wiley & Sons, ??? (2013)

[10] Hosseini, M.-P., Hosseini, A., Ahi, K.: A review on machine learning for eeg signal processing in bioengineering. IEEE reviews in biomedical engineering **14**, 204–218 (2020)

[11] Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)

[12] Li, Z., Hwang, K., Li, K., Wu, J., Ji, T.: Graph-generative neural network for EEG-based epileptic seizure detection via discovery of dynamic brain functional connectivity. Scientific Reports **12**(1), 18998 (2022)

[13] Friston, K.J.: Functional and effective connectivity in neuroimaging: a synthesis. Human brain mapping **2**(1-2), 56–78 (1994)

[14] Bressler, S.L., Tognoli, E.: Operational principles of neurocognitive networks. International journal of psychophysiology **60**(2), 139–148 (2006)

[15] Bassett, D.S., Sporns, O.: Network neuroscience. Nature neuroscience **20**(3), 353–364 (2017)

[16] Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of

structural and functional systems. Nature reviews neuroscience **10**(3), 186–198 (2009)

[17] Friston, K.J., Frith, C.D., *et al.*: Schizophrenia: a disconnection syndrome. Clin Neurosci **3**(2), 89–97 (1995)

[18] Greicius, M.D., Flores, B.H., Menon, V., Glover, G.H., Solvason, H.B., Kenna, H., Reiss, A.L., Schatzberg, A.F.: Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. Biological psychiatry **62**(5), 429–437 (2007)

[19] Van Diessen, E., Numan, T., Van Dellen, E., Van Der Kooi, A., Boersma, M., Hofman, D., Van Lutterveld, R., Van Dijk, B., Van Straaten, E., Hillebrand, A., *et al.*: Opportunities and methodological challenges in EEG and MEG resting state functional brain network research. Clinical Neurophysiology **126**(8), 1468–1481 (2015)

[20] Mehler, D.M.A., Kording, K.P.: The lure of misleading causal statements in functional connectivity research. arXiv preprint arXiv:1812.03363 (2018)

[21] Avvaru, S., Peled, N., Provenza, N.R., Widge, A.S., Parhi, K.K.: Region-Level Functional and Effective Network Analysis of Human Brain During Cognitive Task Engagement. IEEE Transactions on Neural Systems and Rehabilitation Engineering **29**, 1651–1660 (2021)

[22] Balaji, S.S., Parhi, K.K.: Seizure onset zone (SOZ) identification using effective brain connectivity of epileptogenic networks. Journal of Neural Engineering (2024)

[23] Avvaru, S., Parhi, K.K.: Effective Brain Connectivity Extraction by Frequency-Domain Convergent Cross-Mapping (FDCCM) and its Application in Parkinson's Disease Classification. IEEE Transactions on Biomedical Engineering (2023)

[24] Balaji, S.S., Parhi, K.K.: Classifying Subjects with PFC Lesions from Healthy Controls during Working Memory Encoding via Graph Convolutional Networks. In: 2023 11th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 1–4 (2023)

[25] Balaji, S.S., Parhi, K.K.: Classifying Patients with pFC Lesions from Healthy Controls Using Directed Information Based Effective Brain Connectivity Measured from the Encoding Phase of Working Memory Task. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2023)

[26] Balaji, S.S., Parhi, K.K.: Seizure Onset Zone Identification From iEEG: A Review. IEEE Access **10**, 62535–62547 (2022)

[27] Smith, E.E., Jonides, J.: Neuroimaging analyses of human working memory.

Proceedings of the National Academy of Sciences **95**(20), 12061–12068 (1998)

[28] Rypma, B., D'Esposito, M.: The roles of prefrontal brain regions in components of working memory: effects of memory load and individual differences. Proceedings of the National Academy of Sciences **96**(11), 6558–6563 (1999)

[29] Jensen, O., Tesche, C.D.: Frontal theta activity in humans increases with memory load in a working memory task. European journal of Neuroscience **15**(8), 1395–1399 (2002)

[30] Xu, T., Stephane, M., Parhi, K.K.: Abnormal Neural Oscillations in Schizophrenia Assessed by Spectral Power Ratio of MEG During Word Processing. IEEE Transactions on Neural Systems and Rehabilitation Engineering **24**(11), 1148–1158 (2016)

[31] Brookes, M.J., Wood, J.R., Stevenson, C.M., Zumer, J.M., White, T.P., Liddle, P.F., Morris, P.G.: Changes in brain network activity during working memory tasks: a magnetoencephalography study. Neuroimage **55**(4), 1804–1815 (2011)

[32] Haque, Z.Z., Samandra, R., Mansouri, F.A.: Neural substrate and underlying mechanisms of working memory: insights from brain stimulation studies. Journal of Neurophysiology **125**(6), 2038–2053 (2021)

[33] Wu, Y., Qian, H., Yang, X., Chu, H., Yan, C., Gong, X.: Classification of EEG signals during Working- Memory Maintenance based on Phase Space Reconstruction of Empirical Mode Decomposition. In: 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 675–680 (2020)

[34] Xi, J., Huang, X.-L., Dang, X.-Y., Ge, B.-B., Chen, Y., Ge, Y.: Classification for Memory Activities: Experiments and EEG Analysis Based on Networks Constructed via Phase-Locking Value. Computational and mathematical methods in medicine **2022**(1), 3878771 (2022)

[35] Toppi, J., Astolfi, L., Risetti, M., Anzolin, A., Kober, S.E., Wood, G., Mattia, D.: Different topological properties of EEG-derived networks describe working memory phases as revealed by graph theoretical analysis. Frontiers in Human Neuroscience **11**, 637 (2018)

[36] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

[37] Gao, Y., Fu, X., Ouyang, T., Wang, Y.: EEG-GCN: Spatio-Temporal and Self-Adaptive Graph Convolutional Networks for Single and Multi-View EEG-Based Emotion Recognition. IEEE Signal Processing Letters **29**, 1574–1578 (2022)

[38] Zhang, Y., Tetrel, L., Thirion, B., Bellec, P.: Functional annotation of human

cognitive states using deep graph convolution. NeuroImage **231**, 117847 (2021)

[39] Johnson, E.L.: 64-channel human scalp EEG from 14 unilateral PFC patients and 20 healthy controls performing a lateralized visuospatial working memory task. CRCNS.org (2017). https://doi.org/10.6080/K0ZC811B

[40] Wallot, S., Mønster, D.: Calculation of average mutual information (AMI) and false-nearest neighbors (FNN) for the estimation of embedding parameters of multidimensional time series in matlab. Frontiers in psychology **9**, 1679 (2018)

[41] Quinn, C.J., Coleman, T.P., Kiyavash, N., Hatsopoulos, N.G.: Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. Journal of computational neuroscience **30**, 17–44 (2011)

[42] Cover, T.M., Thomas, J.A.: Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, USA (2006)

[43] Meunier, D., Lambiotte, R., Bullmore, E.T.: Modular and hierarchically modular organization of brain networks. Frontiers in neuroscience **4**, 200 (2010)

[44] Gupta, A., Siddhad, G., Pandey, V., Roy, P.P., Kim, B.-G.: Subject-specific cognitive workload classification using EEG-based functional connectivity and deep learning. Sensors **21**(20), 6710 (2021)

[45] Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, 424–438 (1969)

[46] Brovelli, A., Chicharro, D., Badier, J.-M., Wang, H., Jirsa, V.: Characterization of cortical networks and corticocortical functional connectivity mediating arbitrary visuomotor mapping. Journal of Neuroscience **35**(37), 12643–12658 (2015)

[47] Huster, R.J., Plis, S.M., Calhoun, V.D.: Group-level component analyses of EEG: validation and evaluation. Frontiers in neuroscience **9**, 254 (2015)

[48] Farahani, F.V., Karwowski, W., Lighthall, N.R.: Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. frontiers in Neuroscience **13**, 585 (2019)

[49] Mheich, A., Wendling, F., Hassan, M.: Brain network similarity: methods and applications. Network Neuroscience **4**(3), 507–527 (2020)

[50] White, D.R., Borgatti, S.P.: Betweenness centrality measures for directed graphs. Social networks **16**(4), 335–346 (1994)

[51] Rubinov, M., Sporns, O.: Complex network measures of brain connectivity: uses and interpretations. Neuroimage **52**(3), 1059–1069 (2010)

[52] Makarov, V.V., Zhuravlev, M.O., Runnova, A.E., Protasov, P., Maksimenko, V.A., Frolov, N.S., Pisarchik, A.N., Hramov, A.E.: Betweenness centrality in multiplex brain network during mental task evaluation. Physical Review E **98**(6), 062413 (2018)

[53] Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of mathematical sociology **2**(1), 113–120 (1972)

[54] Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems **30**(1-7), 107–117 (1998)

[55] Bavelas, A.: Communication patterns in task-oriented groups. The journal of the acoustical society of America **22**(6), 725–730 (1950)

[56] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A Comprehensive Survey on Graph Neural Networks. IEEE Transactions on Neural Networks and Learning Systems **32**(1), 4–24 (2021)

[57] Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D.: Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. Medical image analysis **48**, 117–130 (2018)

[58] Liu, S., Liang, B., Wang, S., Li, B., Pan, L., Wang, S.-H.: NF-GAT: A Node Feature-Based Graph Attention Network for ASD Classification. IEEE Open Journal of Engineering in Medicine and Biology **5**, 428–433 (2024)

[59] Feng, L., Cheng, C., Zhao, M., Deng, H., Zhang, Y.: EEG-based emotion recognition using spatial-temporal graph convolutional LSTM with attention mechanism. IEEE Journal of Biomedical and Health Informatics **26**(11), 5406–5417 (2022)

[60] Li, Y.: A Survey of EEG Analysis based on Graph Neural Network. In: 2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT), pp. 151–155 (2021). IEEE

[61] Morris, C., Ritzert, M., Fey, M., Hamilton, W.L., Lenssen, J.E., Rattan, G., Grohe, M.: Weisfeiler and Leman go neural: Higher-order graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4602–4609 (2019)

Sai Sanjay Balaji received the B.E. on Electronics and Instrumentation Engineering from the Anna University, Chennai, in 2015, the M.S. degree on Electrical and Computer Engineering from the University of Minnesota, Minneapolis, MN, USA in 2021. He is currently pursuing a Ph.D. degree in Electrical Engineering at the University of Minnesota. His research interests include biomedical signal processing and causal graph analysis of brain networks using machine learning techniques.

Keshab K. Parhi received the B.Tech. degree from the Indian Institute of Technology (IIT), Kharagpur, in 1982, the M.S.E.E. degree from the University of Pennsylvania, Philadelphia, in 1984, and the Ph.D. degree from the University of California, Berkeley, in 1988. He has been with the University of Minnesota, Minneapolis, since 1988, where he is currently the Erwin A. Kelen Chair and a Distinguished McKnight University Professor with the Department of Electrical and Computer Engineering. He has published over 725 papers, is the inventor of 36 patents, and has authored the textbook VLSI Digital Signal Processing Systems (Wiley, 1999). His current research interests include VLSI architecture design of machine learning and signal processing systems, hardware security, and data-driven neuroengineering and neuroscience. He is a fellow of IEEE, American Association for the Advancement of Science (AAAS), the Association for Computing Machinery (ACM), American Institute of Medical and Biological Engineering (AIMBE), and the National Academy of Inventors (NAI). He was a recipient of numerous awards, including the 2017 Mac Van Valkenburg Award and the 2012 Charles A. Desoer Technical Achievement Award from the IEEE Circuits and Systems Society, the 2003 IEEE Kiyo Tomiyasu Technical Field Award, and the Golden Jubilee Medal from the IEEE Circuits and Systems Society in 2000. He served as the Editor-in-Chief for IEEE Transactions on Circuits and Systems— Part I: Regular Papers from 2004 to 2005. He currently serves as the Editor-in-Chief for the IEEE Circuits and Systems Magazine. Since 1993, he has been an Associate Editor of the Springer Journal for Signal Processing Systems.