

Finding Neural Biomarkers for Motor Learning and Rehabilitation Using an Explainable Graph Neural Network

J. Han¹, Graduate Student Member, IEEE, A. Embs, Graduate Student Member, IEEE, F. Nardi², Graduate Student Member, IEEE, S. Haar³, and A. A. Faisal⁴, Member, IEEE

Abstract—Human motor learning is a neural process essential for acquiring new motor skills and adapting existing ones, which is fundamental to everyday activities. Neurological disorders such as Parkinson's Disease (PD) and stroke can significantly affect human motor functions. Identifying neural biomarkers for human motor learning is essential for advancing therapeutic strategies for such disorders. However, identifying specific neural biomarkers associated with motor learning has been challenging due to the complex nature of brain activity and the limitations of traditional data analysis techniques. In response to these challenges, we developed a novel Spatial Graph Neural Network (SGNN) model to predict motor learning outcomes from electroencephalogram (EEG) data using the spatial-temporal dynamics of brain activity. We used it to analyse EEG data collected during a visuomotor rotation (VMR) task designed to elicit distinct types of learning: error-based and reward-based. By doing so, we establish a controlled environment that allows for

precisely investigating neural signatures associated with these learning processes. To understand the features learned by the SGNN, we used a set of spatial, spectral, and temporal explainability methods to identify the brain regions and temporal dynamics crucial for learning. These approaches offer comprehensive insights into the neural biomarkers, aligning with current literature and ablation studies, and pave the way for applying this methodology to find biomarkers from various brain signals and tasks.

Index Terms—Brain-computer interface, EEG, human motor learning, graph neural network, explainable AI, XAI.

I. INTRODUCTION

HUMAN motor learning is a complex mechanism that involves acquiring, refining, and retaining skills that enable us to execute a wide range of movements, from basic to highly specialised tasks [1]. This process is facilitated by two primary mechanisms: error-based and reward-based learning. Error-based learning is characterised by continuous sensory-visual error signals, while reward-based learning relies on integrating feedback information about rewards to improve task performance [2]. In everyday life, a combination of these learning mechanisms occurs, making the isolation of unique neural biomarkers challenging [3]. This complexity is further compounded in patients with neural disorders such as Parkinson's Disease (PD), where dopamine deficiency can impair reward-based learning [4]. Consequently, investigating the distinct neural biomarkers of these learning types becomes crucial, especially in the context of PD, to develop more effective treatment strategies. Finding unique biomarkers of human motor learning could offer significant insights into motor learning mechanisms and their rehabilitation applications [5].

Building on the foundational understanding of motor learning mechanisms, previous research attempted to conduct experiments under controlled environments to isolate specific features of motor learning and study them individually [6], [7], [8], [9], [10], [11]. Functional brain imaging methods such as Functional magnetic resonance imaging (fMRI) [6] and Electroencephalography (EEG) [11], [12], [13] were used to record the neural activities in the brain during these experiments. Compared to fMRI, EEG can provide much higher

Received 20 April 2024; revised 2 November 2024; accepted 11 January 2025. Date of publication 16 January 2025; date of current version 28 January 2025. The work of F. Nardi was supported by U.K. Research and Innovation (UKRI) Centre for Doctoral Training in Artificial Intelligence (AI) for Healthcare under Grant EP/S023283/1. The work of S. Haar was supported in part by the Edmond and Lily Safran Fellowship Program and in part by UK Dementia Research Institute Care Research and Technology Centre. The work of A. A. Faisal was supported by UKRI Turing AI Fellowship under Grant EP/V025449/1. (A. Embs and J. Han are co-first authors.) (Corresponding author: J. Han.)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Imperial College London Research Ethics Committee (ICREC) under Application No. 21IC6754.

J. Han and A. Embs are with the Brain and Behavior Laboratory, Department of Computing, Imperial College London, SW7 2AZ London, U.K. (e-mail: j.han20@imperial.ac.uk; alexandra.embs21@imperial.ac.uk).

F. Nardi is with the UKRI Centre for Doctoral Training in AI for Healthcare and the Brain and Behavior Laboratory, Department of Computing, Imperial College London, SW7 2AZ London, U.K. (e-mail: f.nardi21@imperial.ac.uk).

S. Haar is with the Department of Brain Sciences and the UK Dementia Research Institute Care Research and Technology Centre, Imperial College London, W12 0BZ London, U.K. (e-mail: s.haar@imperial.ac.uk).

A. A. Faisal is with the Brain and Behavior Laboratory, the Department of Computing, and the Department of Bioengineering, Imperial College London, SW7 2AZ London, U.K., and also with the Chair for Digital Health and Data Science, University of Bayreuth, 95447 Bayreuth, Germany (e-mail: a.faisal@imperial.ac.uk).

Digital Object Identifier 10.1109/TNSRE.2025.3530110

temporal resolution, good portability, low acquisition cost, and a trade-off in signal-to-noise ratio (SNR). Considerable research efforts have been dedicated to understanding and identifying distinct neural signatures for motor learning from EEG signals. Specifically, Feedback-related Negativity (FRN) has been identified as a marker for reward-based learning, whereas P300 waves have been linked to general adaptation processes [9]. Furthermore, variations in post-movement beta rebound (PMBR) have been found to correlate with these learning types: an increase in PMBR is observed in error-based learning, while a decrease is noted in reward-based learning [13], [14].

Despite these advancements, the methodologies employed in these studies primarily rely on a statistical analysis of manually extracted features and focus narrowly on specific brain regions associated with movement and motor control. These approaches may overlook more complex yet potentially more informative features, which are often more challenging to discern. Consequently, the identified features might represent superficial markers, failing to capture the complex neural activities and high dimensional correlations between other cortical and subcortical areas. This limitation underscores the need for a novel method to understand the neural biomarkers of motor learning fully.

With the rapid development of deep learning algorithms, there has been a surge of research interest in developing advanced computational models for EEG classification [15], [16] and representation learning [17], [18]. Most of these EEG decoding methods have backbone models based on convolutional neural networks (CNN) [15], [16] and Recurrent Neural networks (RNN) [19], or Transformers [20], [21], which outperformed conventional classification models [22] and feature extraction methods [23]. However, these architectures process spatial information of the channels by stacking them in a grid, similar to how pixels are stacked together in an image [24]. These stacked representations of the EEG sensors do not account for spatial positioning, which neglects the functional neural connectivity within different brain areas [25]. Additionally, despite advances in explainable AI (XAI) methods, the black-box nature of these deep neural networks hinders the determination of what specific spatial relationships or features are important. These factors limit the ability to use these state-of-the-art models for identifying biomarkers during the human motor learning process since the process is heavily influenced by both functional connectivity and spatial connectivity [26], [27].

Inspired by the graph theory, Graph Neural Networks (GNN) have been increasingly used to learn complex features from EEG data for emotion recognition, motor imagery (MI), and neurological disorders [28]. GNNs can effectively leverage the spatial structure within EEG data to extract features, uncover patterns, and make predictions based on the complex interactions between different electrodes. Furthermore, when paired with specialized GNN XAI methods, they offer enhanced explainability regarding the features they learn due to their inherent nodes and edge structure. While GNNs have shown promise in standard BCI applications, their application in deciphering the physiological patterns of human movement

remains underexplored. We aim to harness the capabilities of GNN models to uncover neurologically significant biomarkers for human motor learning.

Existing work has explored various biomarkers associated with motor learning, such as FRN, P300, and PMBR, and the application of machine learning techniques, including CNNs, RNNs, and GNNs, to decode EEG signals. However, these studies often face challenges in fully capturing the complexity of neural dynamics and ensuring the interpretability of their findings. In our work to find motor learning biomarkers. Our contributions are listed below,

We conducted novel experiment setups to collect EEG signals during a specific learning type. This research collects high-density EEG data gathered during a VMR task specifically designed to provoke either error-based or reward-based learning. This approach provides a controlled environment to investigate the neural signatures of these distinct learning processes.

We developed an SGNN model to predict motor learning states and learn important features from EEG signals. This research designed and implemented a GNN model to learn neural activities from EEG data during error-based and reward-based learning. The model was used to perform two classification tasks: 1. classify EEG activity as occurring before learning or during reward-based learning and 2. classify EEG activity as occurring before learning or during error-based learning.

We deployed a set of spatial, spectral, and temporal explainability methods to find the important biomarkers for our GNN model. The GNN-SST-Explainer consists of multiple steps to determine different brain regions (spatial), different frequency bands (spectral), and time points (temporal) responsible for error-based and reward-based learning.

II. RELATED WORK

A. Human Motor Learning

Motor learning processes are essential for acquiring and refining new motor skills and involve a combination of adaptive and strategic mechanisms. These two mechanisms are known as *error-based learning*, which adjusts movements based on sensory feedback, and *reward-based learning*, driven by outcomes or reinforcement. Studies have found that error-based learning, associated with cerebellar activity, leads to stable task learning through continuous feedback [1], [29]. In contrast, reward-based learning, linked to basal ganglia activity, focuses on retention and involves more variable learning processes through reward signals [7], [30].

Most of the existing studies often isolate these learning mechanisms under lab conditions to identify specific biomarkers [6], [8]. Specifically, VMR tasks were commonly used to measure learning while inducing error-based or reward-based learning through the provision of specific feedback [7], [9]. Among the various biomarkers, feedback-related negativity (FRN) and P300 activity have been investigated in the recorded EEG signals between different learning types. FRN is characterised by a shift in event-related potentials (ERPs) 200-400 ms after feedback is provided to an individual [31]. Positive voltage potential shifts are

associated with a desirable reward outcome, and negative voltage potential shifts are associated with an undesirable reward outcome [9]. The P300 wave is another positive shift in ERPs that occurs 300-600 ms after feedback is provided to an individual [32]. This wave is hypothesised to aid in the processing of unexpected stimuli and adaptive responses. Ultimately, FRN was identified as a biomarker for reward-based learning and P300 as a general adaptation biomarker present in error-based and reward-based learning [9]. In addition to this, Post-Movement Beta Rebound (PMBR) has been investigated as a biomarker of motor learning [13], [14], [33]. PMBR consists of increased beta power in the sensory-motor cortex after movement termination [13]. Error-based learning has been found to be associated with an increase in PMBR over learning [14], [33], whereas reward-based learning has been found to be associated with a decrease in PMBR [34].

While these measures play an important role in the study of learning mechanisms and their underlying neural dynamics, manual feature extraction may overlook critical features and focus narrowly on known event-related patterns. Deep learning methods have the potential to automatically extract high-dimensional biomarkers and explore a broader range of brain regions, enabling a deeper understanding of the neural basis of motor learning.

B. EEG Signals Decoding

EEG analysis has significantly impacted clinical diagnostics and patient monitoring by identifying critical biomarkers within EEG signals. Traditionally, this process began with the extraction of relevant features, informed by expert knowledge, such as analysing different frequency bands [35], [36] or graph theoretical metrics to model the functional connectivities [37]. Subsequent steps involved conducting statistical analyses to identify correlations within the data, followed by employing straightforward machine learning models like logistic regression or decision trees for classifying neurological conditions, including Alzheimer's [35], [37] or Parkinson's disease [36]. While these conventional methods have demonstrated effectiveness in condition prediction, they heavily rely on manual feature extraction based on existing domain knowledge. This requirement for manual intervention presupposes certain assumptions about the data, potentially neglecting unrecognised yet pivotal features or complex relationships.

The evolution of deep learning has revolutionised this paradigm, enabling direct analysis of raw EEG data through an end-to-end approach without the necessity for manual feature extraction. These advanced deep learning models, based on Convolutional Neural Networks (CNN) [15], [16], Recurrent Neural Networks (RNN) [19] and Transformers [20], [21], have expanded EEG applications to fields such as emotion recognition [15] and sleep monitoring [17]. The superiority of deep learning models lies in their ability to discover high dimensional patterns in EEG data that were previously inaccessible to conventional classification [22] and feature extraction approaches [23]. This capability has led to enhanced accuracy in classification tasks and opened up new avenues for understanding brain activity in unprecedented detail.

However, the application of deep learning models in identifying biomarkers for clinical data or human behaviours remains less explored. While adept at handling structured pixel data, these deep learning models often overlook the intricate spatial relationships and neural connectivity patterns inherent in EEG signals, a gap noted by recent research in the field [24], [25]. Another significant challenge is the "black box" nature of these models, where the features or biomarkers learned may not be interpretable through neuroscience knowledge. Furthermore, the low data quality and variability across patients add to the difficulty of employing deep learning models for biomarker identification [38].

C. Graph Neural Networks

Leveraging the principles of graph theory, GNNs have become a pivotal tool in deep learning, significantly impacting the field of network neuroscience [39]. GNN enables the modelling of the brain as a graph, with nodes representing anatomical brain regions and edges representing morphological, functional, or structural connectivities. It can preserve the topological properties of graphs while capturing complex relationships and neural connectivity patterns across different brain regions [28], [39].

Several studies have successfully used GNNs for EEG decoding, particularly for classification tasks such as motor imagery [25], emotion [40], [41], [42] or disease diagnosis [43], [44]. Previous works such as [40] and [41] used frequency features instead of the raw signals as input, which requires the domain knowledge of important features for the training target. Yet, the area of biomarker identification, which focuses on developing GNN models that can identify discriminative biomarkers while classifying brain graphs without needing domain knowledge, remains less explored [39]. Few studies attempted to explain what biomarkers their GNN models used for EEG decoding [45], [46]. The challenge here lies in creating reproducible models, ensuring consistent identification of biomarkers across various datasets, and explanation-agnostic, which means they can interpret the reasoning behind their predictions [39]. Achieving reproducibility is crucial for establishing model reliability in network neuroscience while enhancing explainability would increase the model's trustworthiness, which is one of the most important aspects of healthcare models.

D. Explainability Methods

As discussed in the previous section, the explainability of deep learning models is crucial for gaining trust from the human user, especially for their application within the healthcare domain. The field of XAI attempts to make the operations of complex models transparent and understandable to humans. The most commonly used XAI methods are *LIME* [47] and *SHAP* [48].

While significant efforts have been directed towards developing explainability methods for conventional deep learning models, adapting these methods to GNNs faces unique challenges, particularly in explaining the graph structure.

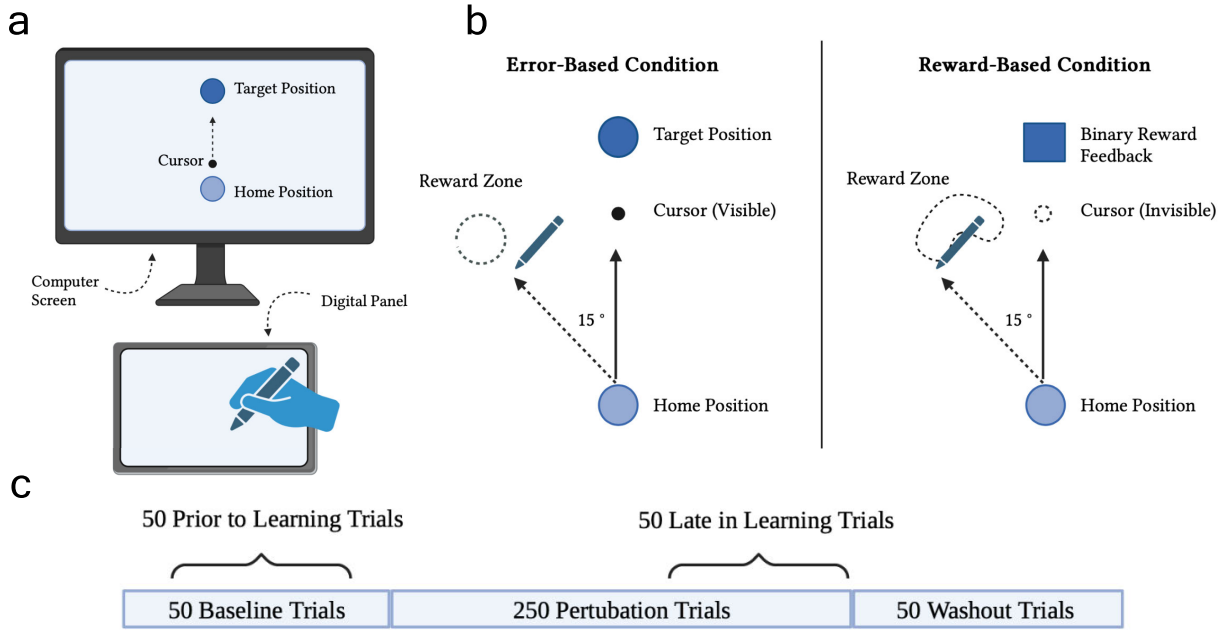


Fig. 1. (a) Illustration of the digital panel and computer screen set up during the VMR task. (b) The screen displays the error-based (left) and reward-based conditions (right). (c) An overview of the experiment trials.

Surrogate methods like GraphLime [49] and PGM-Explainer [50] simplify GNN predictions using interpretable models. GraphLime, focusing on node-level explanations, assesses feature importance through a surrogate model but does not extend to graph-level classifications or edge information [49]. Conversely, PGM-Explainer, suitable for both node- and graph-level tasks, employ probabilistic models to indicate crucial components but also overlooks edge details [50].

Perturbation methods, including PGExplainer [51], and GNNExplainer [52], modify inputs to identify influential graph components. PGExplainer uses a perceptron for edge weight learning, optimizing model predictions through cross-entropy reduction. The goal of the GNNExplainer method is to identify a subgraph of the complete graph that maintains the same predictions and classification accuracy as the original graph [52]. This process identifies the nodes and edges that are most important to the classification decision and removes nodes and edges that are unnecessary to the classification decision. These approaches aim to enhance GNNs' transparency and interpretability, which is crucial for their application in complex domains. Due to the successes of these methods in other domains, they provide promising potential for explaining the high dimensional biomarkers identified by our GNN model during motor learning tasks.

III. MATERIALS AND METHODS

A. Data Collection

We conducted a novel experimental setup to collect EEG signals during specific learning types, isolating the error-based and reward-based learning mechanisms. The EEG data was recorded in a Faraday cage using a High-Density EEG system (GTEN 200) at a sampling rate of 1000 Hz from 20 healthy

participants (12 men and 8 women) during a Visuomotor Rotation (VMR) task. The task was designed to isolate and provoke either error-based or reward-based learning in a controlled environment. Each participant performed both tasks.

The VMR task setup (Figure 1 (a)) included a digital panel for interaction and a screen for visual feedback. Participants made reaching movements using an electronic pen on the digital panel, guiding a cursor on the screen from a starting point to a target. For the error-based condition, as shown in Figure 1 (b), visual feedback was provided by displaying the cursor, allowing participants to adjust their movements based on the visual discrepancy from the target. In contrast, the reward-based condition utilised binary feedback signals, with a smiley face indicating successful target acquisition or the target's disappearance signalling a miss. The reward zone during the rotation phase ranged between the median of the last 10 reach angles and the 15-degree perturbation angle [7].

As shown in Figure 1 (c), each task comprised 300 trials across three phases: baseline, rotation, and washout. The baseline phase did not involve any perturbation, allowing direct correspondence between the pen's location and the cursor's movement. The rotation phase introduced a 15-degree perturbation angle, challenging participants' motor adaptation skills. The final washout phase removed the perturbation to assess participants' ability to readjust to the original settings and measure the proportion of the retained perturbation.

The participants were divided into four groups to ensure a counterbalanced design by rotation direction (Clockwise or Counter Clockwise) and order of the tasks (error or reward first). This study was approved by The Imperial College London Research Ethics Committee (ICREC), reference number: 21IC6754 and informed consent was obtained from all participants.

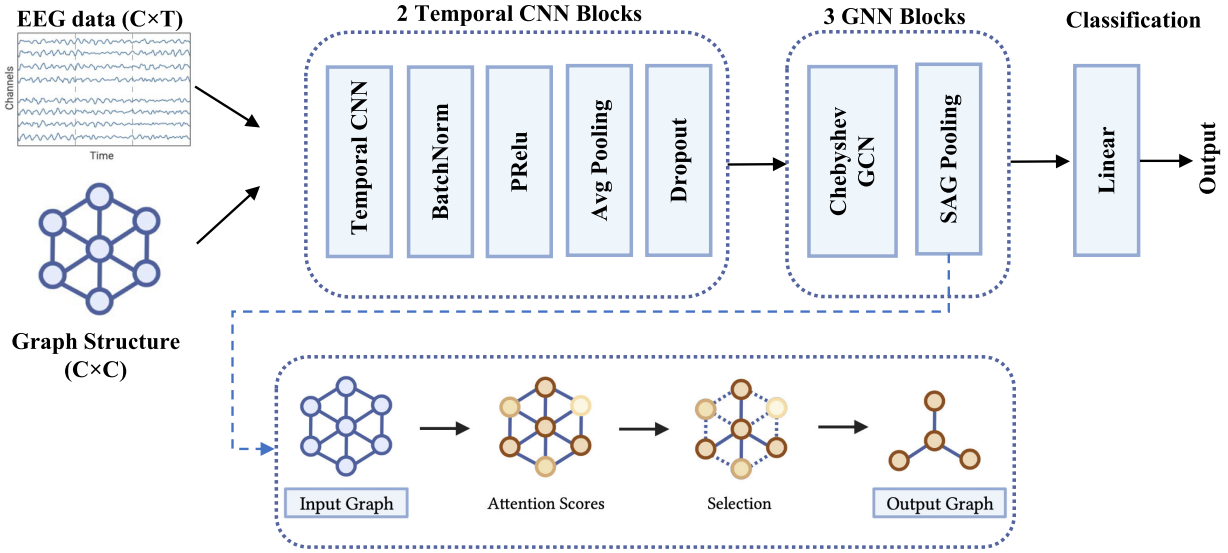


Fig. 2. An illustration of our proposed SGNN architecture and its SAGPooling process. The framework first uses 2 temporal CNN blocks to learn the node features from the EEG signal, then combines with the graph structure defined by an adjacency matrix to learn the spatial relationships using 3 GNN blocks.

B. Data Pre-Processing

Our High-Density EEG recordings were average referenced and high pass filtered using a 1 Hz finite impulse response (FIR) filter. A notch filter was applied at 50 Hz and its harmonics to address power line interference. We selected 168 sensors in the centre of the brain area out of the 256 available on the EEG cap, targeting the cranial region to minimise muscular noise from the neck for training.

The continuous EEG recordings were then segmented into 300 trials. We identified the duration of the longest trial to standardise the length of all trials, marking the start of each trial one second before the starting cue and extending to the end of the longest trial's duration. This approach ensured comprehensive coverage of neural activity during the planning, execution, and completion of each movement, facilitating uniform trial lengths suitable for analysis via convolutional and graph neural networks.

We selected relevant trials from the dataset for the classification task, discarding any trials where the go cue was not properly engaged. Specifically, we included the initial 50 trials of the baseline phase and the final 50 trials of the perturbation phase for model training (Figure 1 (c)). These two classes of trials represent EEG activity before and after the learning process, capturing the neural adaptations attributable to motor learning.

C. GNN Decoding Framework

The proposed Spatial-GNN (SGNN) framework as shown in figure 2 follows a similar design to our previous work where a multi-branch GNN was developed to aggregate multiple Motor Imagery (MI) EEG datasets with varying numbers of channels [25]. In this approach, we first represent our EEG data as brain graphs, where an adjacency matrix A of dimensions $C \times C$ denotes the graph's structure. The matrix's elements, $A_{i,j}$, represent the connection strength or edge

weight between nodes i and j . A stronger connection between two nodes results in a higher edge weight, while an absence of connection is indicated by $A_{i,j} = 0$.

To calculate the edge weights, we utilised the nearest neighbour method in a non-binary fashion [42]. This involved organizing the EEG channels into a grid matrix according to their positional coordinates, with each channel occupying its own grid cell. Channels C_i and C_r were considered neighbours if their grid cells fell within a specified radius, defined as 25% of the grid matrix's dimensions. All neighbour relationships were assigned a weight of $w = 0.5^{(d-1)}$, with distance being the number of grid cells between them. While this method emphasises the connectivity between closer sensors, it might overlook the brain's functional connectivity across different regions. Recognising the brain's intricate network of both local and global circuits [25], [42], we incorporated global edge connections to reflect these broader anatomical and functional linkages, critical for the brain's efficient activity transfer and synchronisation. Specifically, we established global connections involving central sensors between F3 and F4, and CP5 and CP6, respectively (see figure 3 (a)), capturing the comprehensive connectivity landscape. The resulting adjacency matrix with both local and global connections is illustrated in figure 3 (b).

Our Spatial-GNN (SGNN) framework, illustrated in figure 2, is designed to process EEG data through a graph-based approach, where each EEG sensor corresponds to a graph node. The overall architecture takes two inputs: the raw EEG signals as node features, with dimensions $C \times T$ (where C is the number of nodes or channels, and T signifies the time points), and the graph's structural information, as a $C \times C$ adjacency matrix representing the connectivity (edge weights) among the nodes.

The framework begins by performing two 1-D CNN convolutions along the time axis of the raw EEG signals to extract

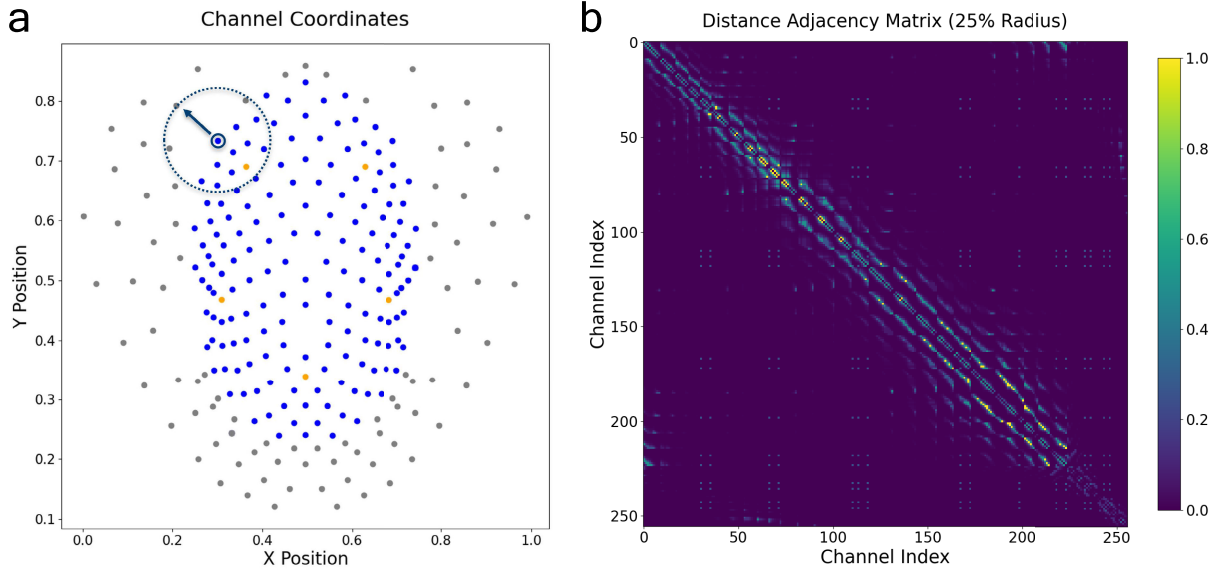


Fig. 3. (a) Illustration of EEG channel coordinates in 2D. Blue dots represent included EEG sensors, orange dots represent sensors with global connections, and grey dots represent non-central EEG sensors that were later excluded (b) The resulting adjacency matrix utilises the distance method with a 25% search radius.

the temporal features of the nodes. The temporal convolution blocks can learn frequency information in the EEG signal. Additionally, by performing temporal convolution prior to the graph convolution, the feature dimensions of the nodes are drastically reduced. This aids in preventing over-fitting which GNNs are prone to [25]. Batch normalisation, the PReLU activation function, average pooling, and dropout are implemented following each CNN convolution.

These newly extracted node features are then passed to the 3 GNN blocks with the adjacency matrix. Each of the GNN blocks involves a Chebyshev graph convolution to propagate information across neighbouring and functionally connected sensors followed by an additional self-attention graph pooling (SAGPooling) layer. Lastly, in the final portion of the network, the node features underwent two additional convolutional layers, which reduced the number of features per node to one and summarised these final features across all the nodes. SAGPooling leverages self-attention mechanisms to prioritise nodes based on their relevance. It employs a GATConv layer to calculate self-attention scores $Z \in \mathbb{R}^{C \times 1}$. A selection of top- k nodes C_{idx} is made based on these scores, retaining only the highest-ranking nodes for further processing. In our model, the nodes with the top 50% self-attention scores are retained after each convolutional layer, with updates to both node features and the adjacency matrix to reflect the pruned graph structure. Finally, the remaining latent features are flattened and fed into a Linear layer for binary classification between the EEG signals before and after motor learning. Overall, all blocks and layers minimise a cross-entropy loss.

where N is the number of samples.

D. GNN-SST-Explainer Pipeline

Taking inspiration from the literature [51], [52], we then explored a range of GNN XAI methods aiming to interpret

the Spatial, Spectral and Temporal (SST) characteristics of the biomarkers our SGNN learned during the training process.

For the spatial step of our XAI pipeline, we first used the SAGPooling node attention scores, Z , to visualise node importance. To obtain the aggregated node importance scores for a model, the Z scores for each SAGPooling layer were normalised between 0 and 1 to make the values comparable across layers. Nodes not selected during the pooling process were assigned a Z score of zero. The scores were then averaged across all three layers to obtain an overall importance value for each node. This process of normalisation and averaging was repeated to obtain node importance scores for each fold, each subject, and across subjects. Since attention scores are not fixed weights and instead depend on both learnable weights and the input data, we separately visualize the attention scores of the model for the EEG segments prior to learning, and after learning to observe if the network had learned to attend to different brain regions based on the learning type.

Inspired by [52], we used node and edge masking methods to explore the spatial characteristics of our biomarker. This involves freezing the weights of the trained SGNN model and applying a learnable node mask, $M_n \in \mathbb{R}^{C \times 1}$, to the raw data, $X \in \mathbb{R}^{C \times T}$, and an edge mask, $M_e \in \mathbb{R}^{C \times E}$, across the GNN's edges, where E denotes the number of edges and C is the number of nodes or channels. The sigmoid function is applied to each mask to encourage discrete binary masking of nodes and edges, thus refining the model's ability to highlight significant spatial features within the biomarker identification process. These learnable masks are then trained with a loss function made up of 5 terms: classification cross-entropy loss $\mathcal{L}_{CE}$, node mask size loss $\mathcal{L}_{NMS}$, edge mask size loss $\mathcal{L}_{EMS}$, node mask entropy loss $\mathcal{L}_{NME}$, and edge mask entropy loss $\mathcal{L}_{EME}$. The classification cross-entropy loss $\mathcal{L}_{CE}$ ensures that the classification accuracy of the original SGNN model is maintained when applying the masks, only non-essential nodes

and edges will be removed from the graph. The loss in node mask size $\mathcal{L}_{\text{NMS}}$, the loss in edge mask size $\mathcal{L}_{\text{EMS}}$ are the sum of all elements in the masks:

$$\mathcal{L}_{\text{NMS}} = \beta \sum_{i=1}^C M_{n_i}, \quad \mathcal{L}_{\text{EMS}} = \alpha \sum_{i=1}^E M_{e_i} \quad (1)$$

where $\beta = 1.0$ is the node size coefficient and $\alpha = 0.005$ is the edge size coefficient. The purpose of these loss terms is to penalise extraneous nodes and edges, thus reducing the complete graph as much as possible. Finally, the mean element-wise entropy of the node and edge masks are calculated using:

$$\begin{aligned} \mathcal{L}_{\text{NME}} &= -\frac{\gamma}{C} \sum_{i=1}^C (M_{n_i} \log(M_{n_i}) + (1 - M_{n_i}) \log(1 - M_{n_i})) \\ \mathcal{L}_{\text{EME}} &= -\frac{\lambda}{E} \sum_{i=1}^E (M_{e_i} \log(M_{e_i}) + (1 - M_{e_i}) \log(1 - M_{e_i})) \end{aligned} \quad (2)$$

where $\gamma = 0.1$ is the node entropy loss coefficient and $\lambda = 1.0$ is the edge entropy loss coefficient. The entropy loss terms assist the mask elements in being discrete values of zero and one. Given this, the entire loss function can be summarised as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{NMS}} + \mathcal{L}_{\text{EMS}} + \mathcal{L}_{\text{NME}} + \mathcal{L}_{\text{EME}}$. The importance score of a node or an edge is determined by the corresponding value of its mask after optimisation since a higher mask value indicates higher importance for the model to minimize the classification loss. However, continuous or “soft” masks can lead to the ‘introduce evidence’ issue, where non-binary values inject noise into the input graph, complicating the explainability process [52]. In contrast, binary or “hard” masks enhance intuitiveness and interpretability. To address this, the method combines the use of the sigmoid activation function with element-wise entropy in the loss function to promote mask discretisation. Specifically, the hard node mask technique calculates gradients for each mask element after the initial backward pass in the learnable mask training. Elements with non-zero gradients are set to one in the hard mask, effectively filtering out nodes that do not impact the classification outcome through an element-wise multiplication with the soft mask. This strategic application ensures only influential nodes are retained, simplifying the explainability process.

To assess the significance of individual frequency bands on classification accuracy, we implemented spectral ablation studies. These studies apply either bandpass or bandstop filters of specific frequency bands onto the test data and test the trained models on these processed data to understand the impact of each frequency band on model performance. The bandstop spectral ablation study involves isolating a specific frequency band from the signal and removing all other frequencies outside this range. Whereas the bandpass spectral ablation study involves removing a specific frequency band from the signal while retaining the rest of the signal intact.

Finally, for temporal explainability, we further executed temporal ablation studies by applying either band-pass or band-stop filtering to distinct temporal segments of the test

data, specifically the pre-movement, movement, and post-movement segments. Subsequently, the trained models were evaluated on this altered data to assess the impact on classification performance. However, the resolution of this temporal ablation approach is limited and does not offer a continuous perspective on temporal significance. Therefore we extend a similar concept of [52] to the temporal domain for continuous evaluation of temporal importance. The goal is to identify a subset of temporal features, X_s , that maintains the same classification accuracy as the full set of features, X . In this instance a feature mask, $F \in \mathbb{R}^{1 \times T}$ was applied across each node feature vector. This allows for the identification of temporal time points that are most essential to classification decisions and removes those that are least essential. The loss function is composed of the prediction cross-entropy loss $\mathcal{L}_{\text{CE}}$, the feature mask size loss $\mathcal{L}_{\text{FMS}}$, the feature mask entropy loss $\mathcal{L}_{\text{FME}}$ similar to the spatial explainability method.

$$\begin{aligned} \mathcal{L}_{\text{FMS}} &= \beta \sum_{t=1}^T F_t \\ \mathcal{L}_{\text{FME}} &= -\frac{\gamma}{T} \sum_{t=1}^T (F_t \log F_t + (1 - F_t) \log(1 - F_t)) \end{aligned} \quad (3)$$

where $\beta = 1.0$ and $\gamma = 0.1$. The importance score of a temporal segment is determined by the value of its mask after the optimisation.

IV. EXPERIMENTS AND RESULTS

A. Classification Accuracy

We evaluated the classification performance of our proposed SGNN model in comparison with state-of-the-art EEG decoding models, including CNN-based models such as EEG-Net [15], DeepConvNet [16], and ShallowConvNet [16] and GNN-based models such as EEG Graph Convolutional Neural Network (EEG-GCNN) [40] and self-attention dynamic graph neural networks (SDGNN) [41]. These models were selected for their demonstrated effectiveness in EEG classification and the clarity of implementation details. For implementation, our model and the CNN-based models used the EEG signal as input whereas the EEG-GCNN models used the power density of 6 different frequency bands (delta, theta, alpha, lower beta, higher beta and gamma) as input node features and the SDGNN model takes the EEG spectrogram as the node features for input. All models were implemented using PyTorch. We calculated the mean accuracy and F1-scores of within-subject five-fold cross-validation to assess the model’s performance of classifying EEG signals before and after learning. A Wilcoxon signed-rank test was utilised to determine statistical significance due to the non-normal data distribution and the paired nature of the samples. Figure 4 displays the obtained accuracy and F1-scores with their standard deviation.

As seen in Figure 4 (a), our SGNN model, which uses CNN layers to automatically extract relevant features, achieved higher classification performance compared to other GNN-based methods. Its performance was comparable to EEGNet and ShallowConvNet, both of which have significantly fewer

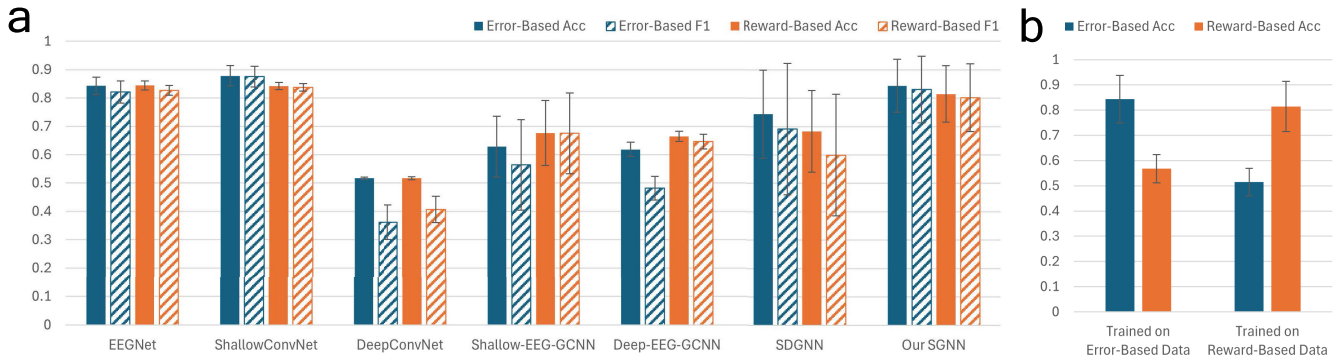


Fig. 4. (a) Average classification accuracies (Acc) and F1-scores (F1) of the baseline models compared to our proposed model. (b) The average test accuracy of our models when trained with data of the opposite condition.

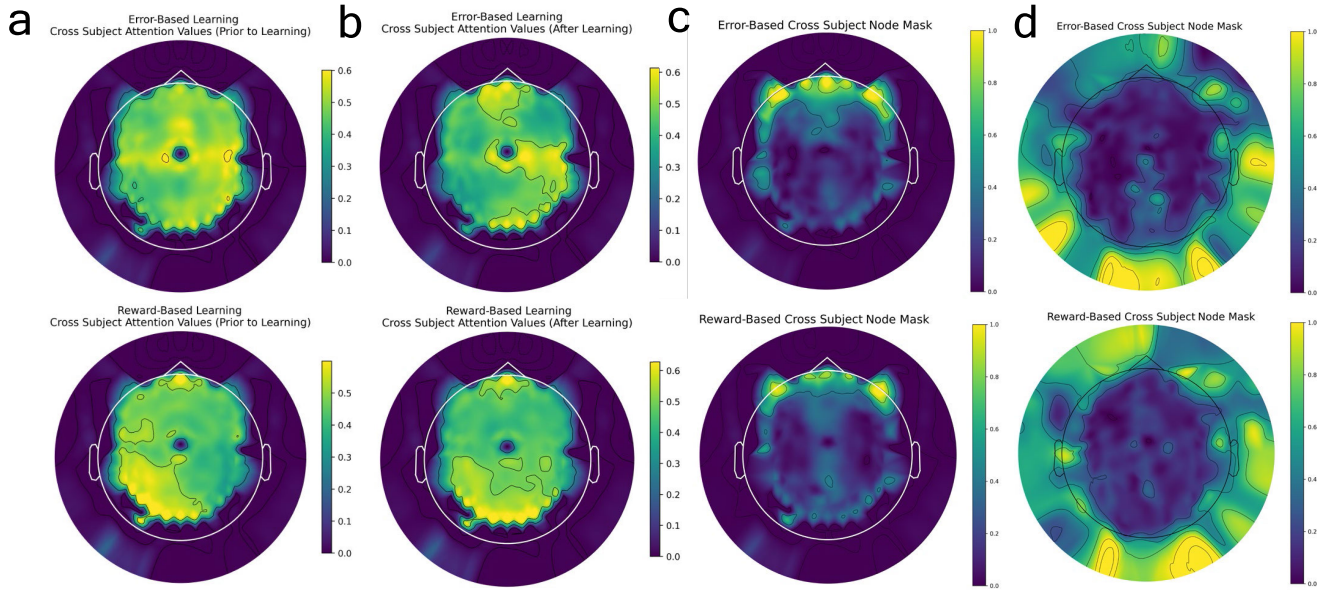


Fig. 5. Spatial explainability visualisation (a) The attention values of the SAGPool layer of data before learning (b) The attention values of the SAGPool layer of data after learning (c) The importance score of different channels when 168 channels in the centre were selected for training. (d) The importance score of different channels when all 256 channels were used for training.

parameters, underscoring the robustness of our model on small datasets.

Furthermore, apart from evaluating the GNN architecture's performance relative to the CNN architectures, an additional analysis was undertaken. This involved testing the performance of the GNN models trained with each subject's error-based or reward-based data with test data from their opposite condition. This experiment was conducted to determine if the models learned the same or differing neural signatures. Figure 4 (b) shows that, while the models performed very well with the test data under the same condition, the cross-condition performances are significantly lower. This indicated a significant domain shift between the data collected under the error-based learning task and the reward-based learning task of the same subject. Therefore, this finding suggests that the models learned distinct neural activity biomarkers from error-based and reward-based learning data, and the learning types have different neural circuitry mechanisms.

B. Explainability Analysis

The attention visualisation method across our networks reveals a broad engagement of brain regions, with some emphasis on the prefrontal, sensorimotor and visual cortex, as shown in Figure 5 (a) and (b). However, this pattern does not consistently emerge across all spatial maps for both error-based and reward-based models in the datasets.

In our Node Masking approach our model identified the bilateral prefrontal cortex in both reward-based and error-based tasks as shown in Figure 5 (c). Furthermore, we also investigated the importance of the nodes when the models are trained on all 256 sensors in the dataset. Figure 5 (d) revealed that the model relied on muscle fatigue signals from the neck for the classification instead of the functional brain area. Although the model could perform the prediction of learning, it did not learn any useful biomarkers. This result demonstrated the growing importance of explainability in AI, particularly in healthcare applications.

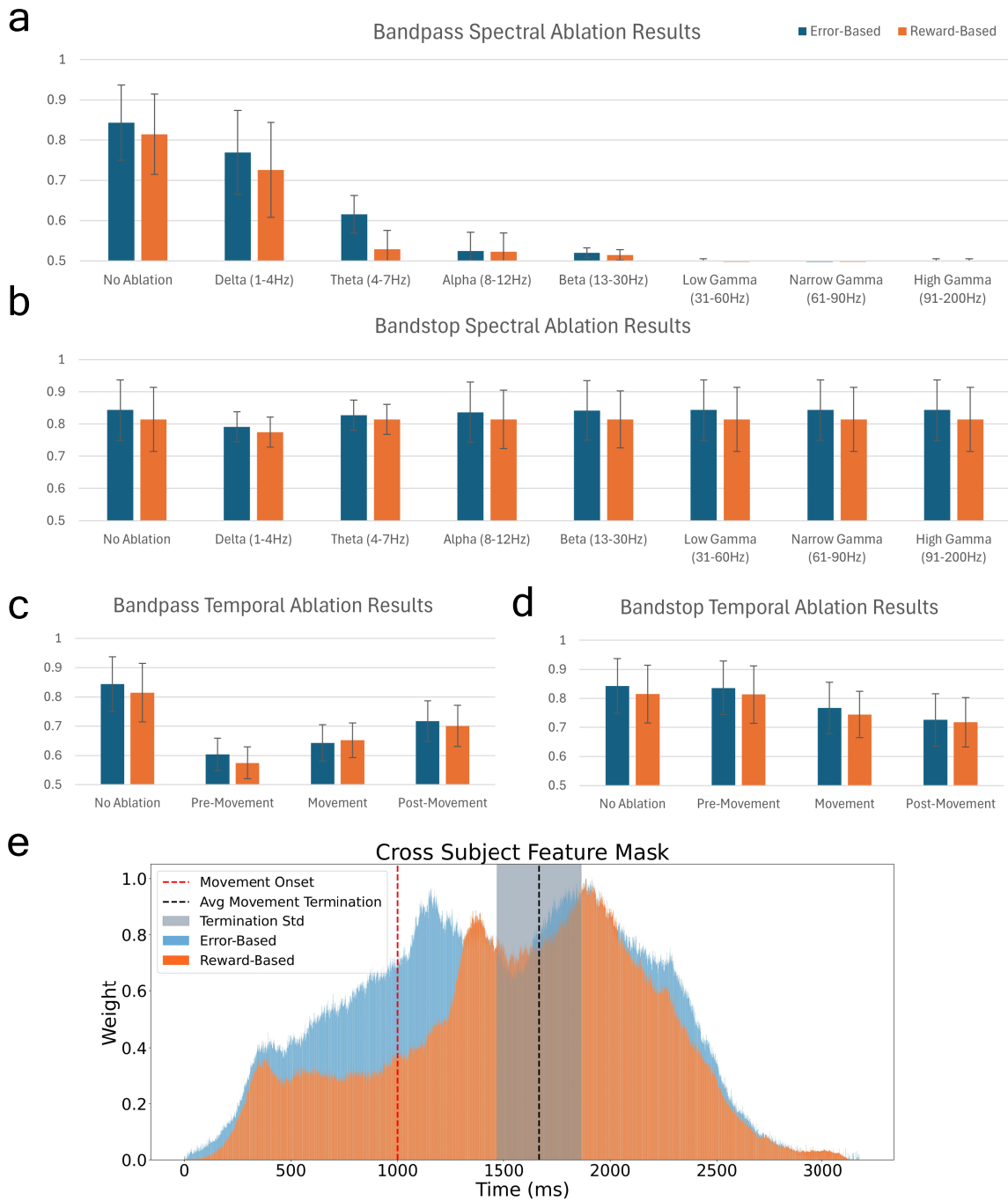


Fig. 6. Spectral and temporal explainability visualization. (a) Bandpass spectral ablation results. (b) Bandstop spectral ablation result. (c) Bandpass temporal ablation results. (d) Bandstop temporal ablation result. (e) Temporal masking explainability result.

The results of our bandpass and bandstop spectral ablation studies are shown in 6 (a) and (b), respectively. The bandpass spectral ablation results demonstrated the significance of the Delta band as the most significant biomarker. Having the Delta band in the EEG can still enable our model to achieve over 70% mean accuracy.

The temporal bandpass and bandstop ablation results are shown in Figure 6 (c) & (d). The three main temporal periods investigated were pre-movement, movement, and post-movement. Both ablation studies showed the post-movement period as most vital to learning classification.

Furthermore, the resulting temporal importance score averaged across all subjects determined from the temporal masking

method is displayed in Figure 6 (e). These results also confirm that the GNN model relies most heavily on the during-movement and post-movement temporal periods to make the classification decision for error- and reward-based tasks. In particular, the period of 300 ms following the termination of movement is considered most vital.

V. DISCUSSION

The classification results demonstrate that our proposed SGNN model outperforms the two GNN baseline models, in terms of both accuracy and F1-score for error-based and reward-based motor learning conditions. The superior performance of our model can be attributed to its end-to-end learning

capability, where CNN layers automatically learn important features from EEG data, as opposed to the manually extracted frequency features used in EEG-GCNN and SDGNN. This approach enables our model to better capture the complex spatial-temporal relationships inherent in motor learning.

Furthermore, while deeper models like DeepConvNet and Deep-EEG-GCNN suffered from overfitting due to the small training sample size, our model's architecture strikes a balance between complexity and robustness, allowing it to achieve high performance without sacrificing explainability.

The spatial explainability results show the prefrontal, sensorimotor and visual cortex as the most important areas for the model to classify the EEG data before and after learning the visuomotor rotation. The divergences in the attention scores before and after motor learning, outline a significant difference between the areas of interest, which are given higher attention values by the network. The shift in the error-based task from the whole brain to the prefrontal and parietal cortex represents perfectly the learning of the visuomotor rotation, as they are related to sensory-motor adaptation [53] and integration, as well as to reaching and grasping movements [54]. On the other hand, the reward-based learning attention values before and after learning show a major importance of the parietal cortex, related to the task—reward associations [55], coherently with the nature of the task and the feedback provided.

The node masking method identified the bilateral prefrontal cortex as the most important area for both reward-based and error-based tasks. This finding aligns with the known involvement of the prefrontal cortex in higher-order cognitive functions such as learning, memory, planning, and problem-solving, as well as associative learning, where associations between events and stimuli are learned [56]. Both reward and error-based learning, subsets of associative learning that link actions to outcomes, likely engage the prefrontal cortex's association formulation activity, aiding in signal classification related to learning phases. Furthermore, the prefrontal cortex's connection to working memory, essential for retaining information for problem-solving and adjusting movements based on motor error information, underscores its relevance to motor learning [57], [58]. Additionally, spatial mapping highlighted the contralateral sensorimotor cortex, though less prominently, supporting the integrated role of cognitive processing and movement execution in motor learning, especially given the task's performance with the right hand, suggesting left cortex sensorimotor activity [59].

The spectral ablation results show the Delta and Theta bands as surprisingly fundamental for the accuracy of the model. Delta activity is traditionally linked to memory consolidation [60] and attention [61], as well as its oscillations and their relation with the Beta band shape motor behaviours and sensorimotor processes [62]. The Theta band is related to the coordination and timing of neural activity, which is essential for encoding and integrating new motor experiences.

In contrast, the importance of the Beta band, traditionally and strongly linked to motor learning, seems not to be used as the main driving factor by the model. In particular, we would have expected it to be more important due to the Post-Movement Beta Rebound (PMBR), an increase in

beta power following the termination of movement, found to be different across the different learning mechanisms [13], [14], [34]. However, the PMBR could be relevant to the temporal ablation and masking, as the post-movement window is of main importance for the network predictions and this peak in the Beta could occur up to 1000ms after the movement onset [63], which is exactly corresponding to the second peak visible in figure 6 (e).

Our temporal explainability methods identified the post-movement period, especially 300 ms after the movement termination to be the most important temporal segment for classification. This period likely corresponds to error detection and feedback processing which are crucial for motor learning. It involves the assessment of discrepancies between the intended and actual movement outcomes, enabling the fine-tuning of future motor actions.

Looking further at the importance of temporal aspects of the signals, the difference in the peak around the movement onset between error- and reward-based tasks could be explained by the timing of the reward provision, as in the error task it is immediately visible, while in the reinforcement-based task, the participants can retrieve the trial outcome just at the end of it.

VI. CONCLUSION

In this study, we developed an SGNN model and applied an explainability pipeline to identify biomarkers in EEG data related to motor learning. Our SGNN model achieved comparable performance to conventional CNN-based EEG classification models and outperformed other GNN-based methods by effectively leveraging the graph-like structure of EEG data and employing temporal CNN layers to automatically extract important features from the EEG signals. This highlights the model's ability to maintain high classification accuracy on small training datasets while providing enhanced interpretability for biomarker identification.

Furthermore, we introduced the GNN-SST-Explainer pipeline, incorporating spatial, spectral, and temporal analysis to enhance model interpretability. Our spatial analysis identified the prefrontal and sensorimotor cortices as key areas for classification decisions before and after learning. Spectral analysis revealed the model's emphasis on lower frequencies, especially the Delta band, while temporal analysis identified the post-movement phase, particularly 300 ms and 1000 ms following movement termination, as biomarkers for motor learning. In addition, our proposed explainability method has led to identifying theta band activity in the post-movement prefrontal cortex as a novel biomarker for error- and reward-based learning.

Our research addresses a gap in behavioural neuroscience by employing deep learning methods, especially GNN, to extract explainable biomarkers from EEG signals. The implications of our proposed method extend well into the healthcare domains, facilitating the adoption of deep learning and data-driven strategies for biomarker identification across various biosignals linked to neurological diseases. Our research has potential applications in personalised rehabilitation programs and the development of neurofeedback systems, offering new avenues for targeted interventions in motor learning and recovery.

This advancement sets the stage for innovative interdisciplinary research and cultivates new collaborative prospects between computer scientists and medical professionals, aiming to create trustworthy AIs for rehabilitation and healthcare.

ACKNOWLEDGMENT

(S. Haar and A. A. Faisal are co-senior authors.)

REFERENCES

- [1] A. S. Therrien and A. L. Wong, "Mechanisms of human motor learning do not function independently," *Frontiers Human Neurosci.*, vol. 15, Jan. 2022, Art. no. 785992.
- [2] D. M. Wolpert, J. Diedrichsen, and J. R. Flanagan, "Principles of sensorimotor learning," *Nature Rev. Neurosci.*, vol. 12, no. 12, pp. 739–751, Dec. 2011.
- [3] A. M. Haith and J. W. Krakauer, "Model-based and model-free mechanisms of human motor learning," in *Progress in Motor Control: Neural, Computational and Dynamic Approaches*. Cham, Switzerland: Springer, 2013, pp. 1–21.
- [4] N. C. van Wouwe, K. R. Ridderinkhof, G. P. H. Band, W. P. M. van den Wildenberg, and S. A. Wylie, "Dose dependent dopaminergic modulation of reward-based learning in Parkinson's disease," *Neuropsychologia*, vol. 50, no. 5, pp. 583–591, Apr. 2012.
- [5] A. Nieuwboer, L. Rochester, L. Müncks, and S. P. Swinnen, "Motor learning in Parkinson's disease: Limitations and potential for rehabilitation," *Parkinsonism Rel. Disorders*, vol. 15, pp. S53–S58, Dec. 2009.
- [6] J. Diedrichsen, Y. Hashambhoy, T. Rane, and R. Shadmehr, "Neural correlates of reach errors," *J. Neurosci.*, vol. 25, no. 43, pp. 9919–9931, Oct. 2005.
- [7] A. S. Therrien, D. M. Wolpert, and A. J. Bastian, "Effective reinforcement learning following cerebellar damage requires a balance between exploration and motor noise," *Brain*, vol. 139, no. 1, pp. 101–114, Jan. 2016.
- [8] S. Uehara, F. Mawase, and P. Celnik, "Learning similar actions by reinforcement or sensory-prediction errors rely on distinct physiological mechanisms," *Cerebral Cortex*, vol. 28, no. 10, pp. 3478–3490, Oct. 2018.
- [9] D. J. Palidis, J. G. A. Cashaback, and P. L. Gribble, "Neural signatures of reward and sensory error feedback processing in motor learning," *J. Neurophysiology*, vol. 121, no. 4, pp. 1561–1574, Apr. 2019.
- [10] F. Nardi, S. Haar, and A. A. Faisal, "Bill-EVR: An embodied virtual reality framework for reward-and-error-based motor rehab-learning," in *Proc. Int. Conf. Rehabil. Robot. (ICORR)*, Sep. 2023, pp. 1–6.
- [11] F. Nardi, A. A. Faisal, and S. Haar, "Motor learning mechanisms are not modified by feedback manipulations in a real-world task," *bioRxiv*, pp. 1–16, Apr. 2024.
- [12] O. E. Krigolson, D. Cheng, and G. Binsted, "The role of visual processing in motor learning and control: Insights from electroencephalography," *Vis. Res.*, vol. 110, pp. 277–285, May 2015.
- [13] S. Haar and A. A. Faisal, "Brain activity reveals multiple motor-learning mechanisms in a real-world task," *Frontiers Human Neurosci.*, vol. 14, p. 354, Sep. 2020.
- [14] H. Tan, N. Jenkinson, and P. Brown, "Dynamic neural correlates of motor error monitoring and adaptation during trial-to-trial learning," *J. Neurosci.*, vol. 34, no. 16, pp. 5678–5688, Apr. 2014.
- [15] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.
- [16] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.
- [17] H. Banville, I. Albuquerque, A. Hyvärinen, G. Moffat, D.-A. Engemann, and A. Gramfort, "Self-supervised representation learning from electroencephalography signals," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.
- [18] J. Han, X. Gu, G.-Z. Yang, and B. Lo, "Noise-factorized disentangled representation learning for generalizable motor imagery EEG classification," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 2, pp. 765–776, Feb. 2024.
- [19] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang, "LSTM-based EEG classification in motor imagery tasks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 11, pp. 2086–2095, Nov. 2018.
- [20] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers Human Neurosci.*, vol. 15, pp. 1–15, Jun. 2021.
- [21] C. Yang, M. Westover, and J. Sun, "Biot: Biosignal transformer for cross-data learning in the wild," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–21.
- [22] K. Keng Ang, Z. Yang Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 2390–2397.
- [23] A. Subasi and M. Ismail Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, Dec. 2010.
- [24] A. Demir, T. Koike-Akino, Y. Wang, M. Haruna, and D. Erdogmus, "EEG-GNN: Graph neural networks for classification of electroencephalogram (EEG) signals," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1061–1067.
- [25] J. Han, X. Wei, and A. A. Faisal, "EEG decoding for datasets with heterogeneous electrode configurations using transfer learning graph neural networks," *J. Neural Eng.*, vol. 20, no. 6, Dec. 2023, Art. no. 066027.
- [26] J. Wu, R. Srinivasan, A. Kaur, and S. C. Cramer, "Resting-state cortical connectivity predicts motor skill acquisition," *NeuroImage*, vol. 91, pp. 84–90, May 2014.
- [27] M. H. Heitger, R. Ronsse, T. Dholander, P. Dupont, K. Caeyenberghs, and S. P. Swinnen, "Motor learning-induced changes in functional brain connectivity as revealed by means of graph-theoretical network analysis," *NeuroImage*, vol. 61, no. 3, pp. 633–650, Jul. 2012.
- [28] D. Klepl, M. Wu, and F. He, "Graph neural network-based EEG classification: A survey," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 493–503, 2024.
- [29] P. Thoma, C. Bellebaum, B. Koch, M. Schwarz, and I. Daum, "The cerebellum is involved in reward-based reversal learning," *Cerebellum*, vol. 7, no. 3, pp. 433–443, Sep. 2008.
- [30] K. Doya, "Complementary roles of basal ganglia and cerebellum in learning and motor control," *Current Opinion Neurobiol.*, vol. 10, no. 6, pp. 732–739, Dec. 2000.
- [31] J. van der Helden, M. A. S. Boksem, and J. H. G. Blom, "The importance of failure: Feedback-related negativity predicts motor learning efficiency," *Cerebral Cortex*, vol. 20, no. 7, pp. 1596–1603, Jul. 2010.
- [32] N. Zhong et al., "Smaller feedback-related negativity (FRN) reflects the risky decision-making deficits of methamphetamine dependent individuals," *Frontiers Psychiatry*, vol. 11, p. 320, Apr. 2020.
- [33] H. Tan, C. Wade, and P. Brown, "Post-movement beta activity in sensorimotor cortex indexes confidence in the estimations from internal models," *J. Neurosci.*, vol. 36, no. 5, pp. 1516–1528, Feb. 2016.
- [34] J. Kolasinski, E. L. Hinson, A. P. Divanbeighi Zand, A. Rizov, U. E. Emir, and C. J. Stagg, "The dynamics of cortical GABA in human motor learning," *J. Physiol.*, vol. 597, no. 1, pp. 271–282, Jan. 2019.
- [35] S. Simpraga et al., "EEG machine learning for accurate detection of cholinergic intervention and Alzheimer's disease," *Sci. Rep.*, vol. 7, no. 1, p. 5775, Jul. 2017.
- [36] M. I. Vanegas, M. F. Ghilardi, S. P. Kelly, and A. Blangero, "Machine learning for EEG-based biomarkers in Parkinson's disease," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 2661–2665.
- [37] M. Jalili, "Graph theoretical analysis of Alzheimer's disease: Discrimination of AD patients from healthy subjects," *Inf. Sci.*, vol. 384, pp. 145–156, Apr. 2017.
- [38] X. Gu et al., "Beyond supervised learning for pervasive healthcare," *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 42–62, 2024.
- [39] A. Bessadok, M. A. Mahjoub, and I. Rekik, "Graph neural networks in network neuroscience," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5833–5848, Sep. 2022.
- [40] N. Wagh and Y. Varatharajah, "EEG-GCNN: Augmenting electroencephalogram-based neurological disease diagnosis using a domain-guided graph convolutional neural network," in *Proc. Mach. Learn. Health*, Jan. 2020, pp. 367–378.
- [41] C. Li et al., "EEG emotion recognition based on self-attention dynamic graph neural networks," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 292–296.
- [42] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul. 2022.
- [43] M. Jia et al., "Efficient graph convolutional networks for seizure prediction using scalp EEG," *Frontiers Neurosci.*, vol. 16, Aug. 2022, Art. no. 967116.

- [44] K. Raeisi, M. Khazaei, P. Croce, G. Tamburro, S. Comani, and F. Zappasodi, "A graph convolutional neural network for the automated detection of seizures in the neonatal EEG," *Comput. Methods Programs Biomed.*, vol. 222, Jul. 2022, Art. no. 106950.
- [45] D. Klepl, F. He, M. Wu, D. J. Blackburn, and P. Sarrigiannis, "Adaptive gated graph convolutional network for explainable diagnosis of alzheimer's disease using EEG data," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3978–3987, 2023.
- [46] M. Zhdanov, S. Steinmann, and N. Hoffmann, "Investigating brain connectivity with graph neural networks and GNNExplainer," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 5155–5161.
- [47] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. discovery data mining*, 2016, pp. 1135–1144.
- [48] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017, pp. 1–10.
- [49] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "GraphLIME: Local interpretable model explanations for graph neural networks," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6968–6972, Jul. 2022.
- [50] M. N. Vu and M. T. Thai, "PGM-explainer: Probabilistic graphical model explanations for graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 12225–12235.
- [51] D. Luo et al., "Parameterized explainer for graph neural network," in *Proc. NIPS*, vol. 33, 2020, pp. 19620–19631.
- [52] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
- [53] J. A. Taylor and R. B. Ivry, "Cerebellar and prefrontal cortex contributions to adaptation, strategies, and reinforcement learning," *Prog. Brain Res.*, vol. 210, pp. 217–253, Jan. 2014.
- [54] R. A. Andersen and C. A. Buneo, "Intentional maps in posterior parietal cortex," *Annu. Rev. Neurosci.*, vol. 25, no. 1, pp. 189–220, Mar. 2002.
- [55] D. Wisniewski, C. Reverberi, I. Momennejad, T. Kahnt, and J.-D. Haynes, "The role of the parietal cortex in the representation of task-reward associations," *J. Neurosci.*, vol. 35, no. 36, pp. 12355–12365, Sep. 2015.
- [56] D. R. Euston, A. J. Gruber, and B. L. McNaughton, "The role of medial prefrontal cortex in memory and decision making," *Neuron*, vol. 76, no. 6, pp. 1057–1070, Dec. 2012.
- [57] M. D'Esposito and B. R. Postle, "The cognitive neuroscience of working memory," *Annu. Rev. Psychol.*, vol. 66, pp. 115–142, Sep. 2015.
- [58] R. D. Seidler, J. Bo, and J. A. Anguera, "Neurocognitive contributions to motor skill learning: The role of working memory," *J. Motor Behav.*, vol. 44, no. 6, pp. 445–453, Nov. 2012.
- [59] J. E. Downey et al., "The motor cortex has independent representations for ipsilateral and contralateral arm movements but correlated representations for grasping," *Cerebral Cortex*, vol. 30, no. 10, pp. 5400–5409, Sep. 2020.
- [60] J. Kim, T. Gulati, and K. Ganguly, "Competing roles of slow oscillations and delta waves in memory consolidation versus forgetting," *Cell*, vol. 179, no. 2, pp. 514–526, Oct. 2019.
- [61] P. Lakatos, G. Karmos, A. D. Mehta, I. Ulbert, and C. E. Schroeder, "Entrainment of neuronal oscillations as a mechanism of attentional selection," *Science*, vol. 320, no. 5872, pp. 110–113, 2008.
- [62] B. Morillon, L. H. Arnal, C. E. Schroeder, and A. Keitel, "Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception," *Neurosci. Biobehavioral Rev.*, vol. 107, pp. 136–142, Dec. 2019.
- [63] M. T. Jurkiewicz, W. C. Gaetz, A. C. Bostan, and D. Cheyne, "Post-movement beta rebound is generated in motor cortex: Evidence from neuromagnetic recordings," *NeuroImage*, vol. 32, no. 3, pp. 1281–1289, 2006.