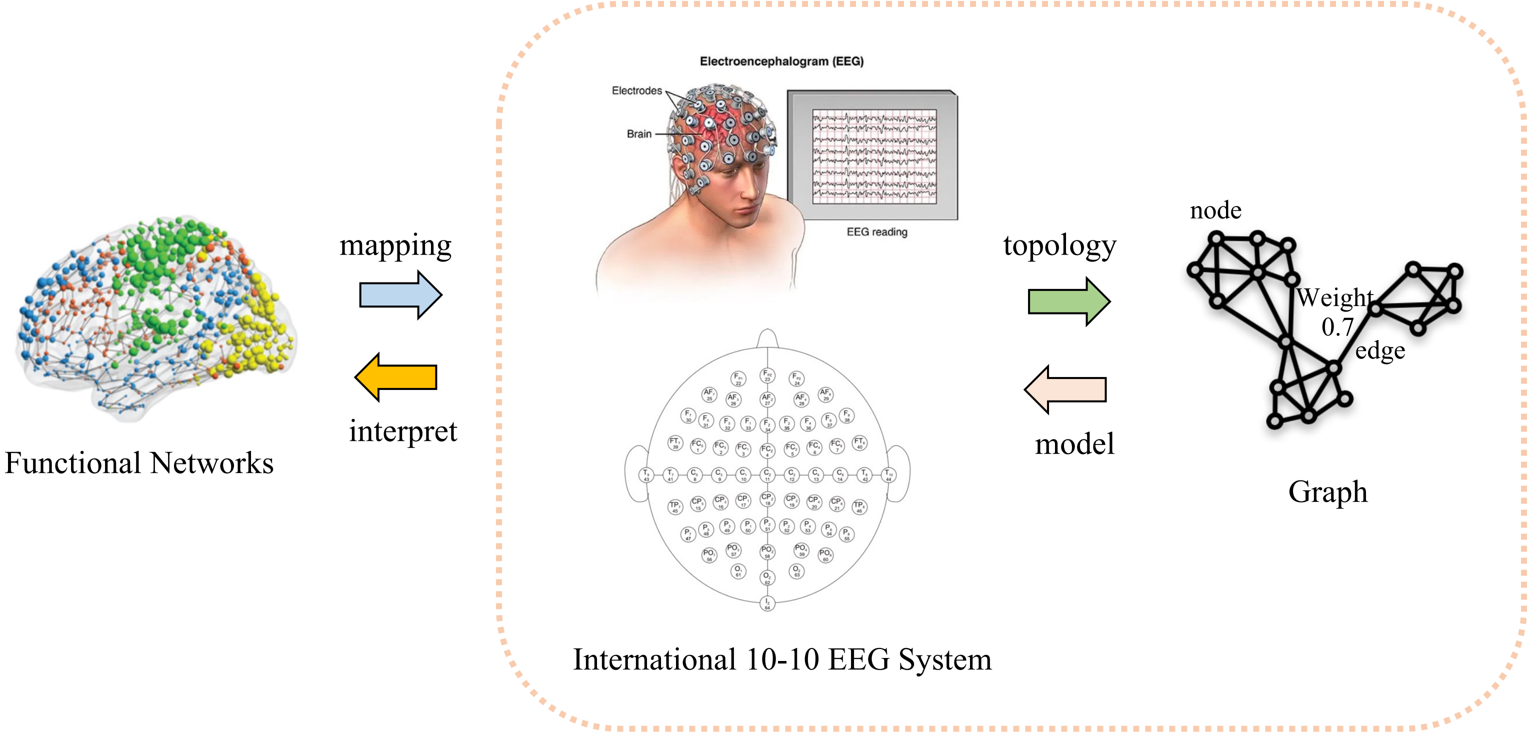
Deep Feature Mining via Attention-based BiLSTM-GCN for Human Motor Imagery Recognition

Shuyue Jia, Jan. 15th, 2023

Tasks

✓ Electroencephalogram (EEG) Tasks **Classification**

Control a wheelchair via EEG

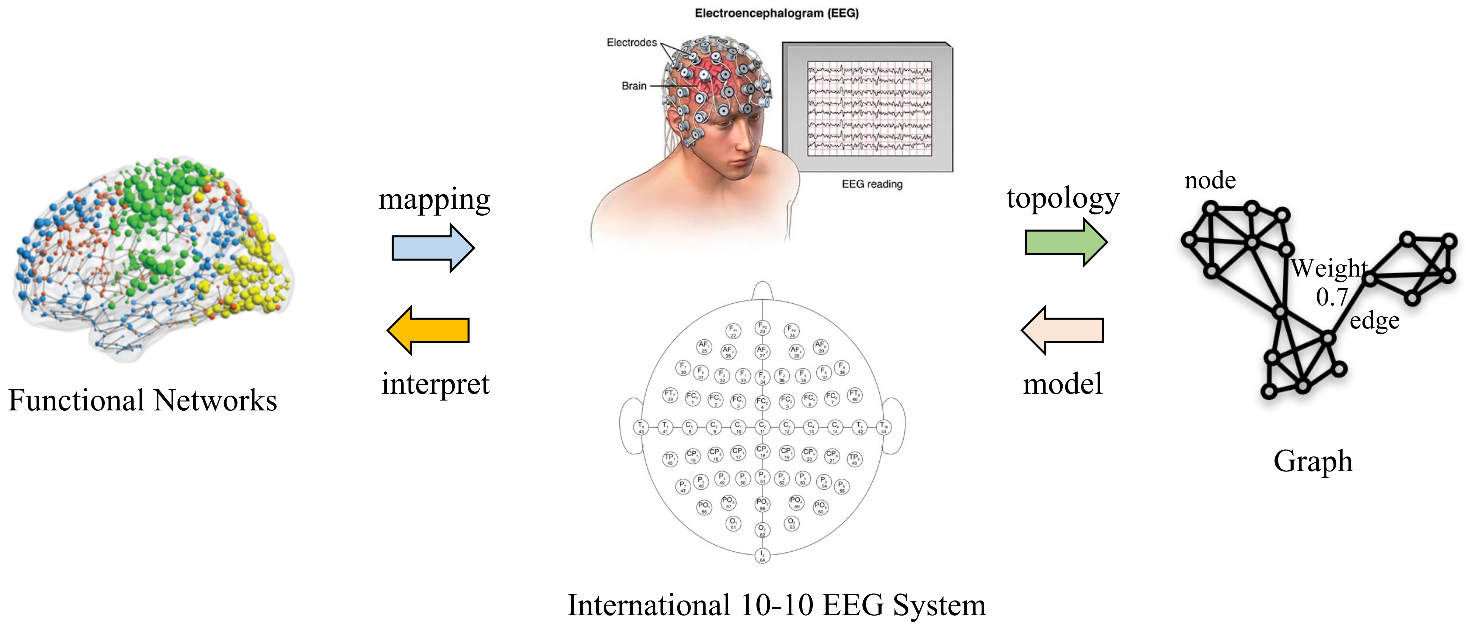


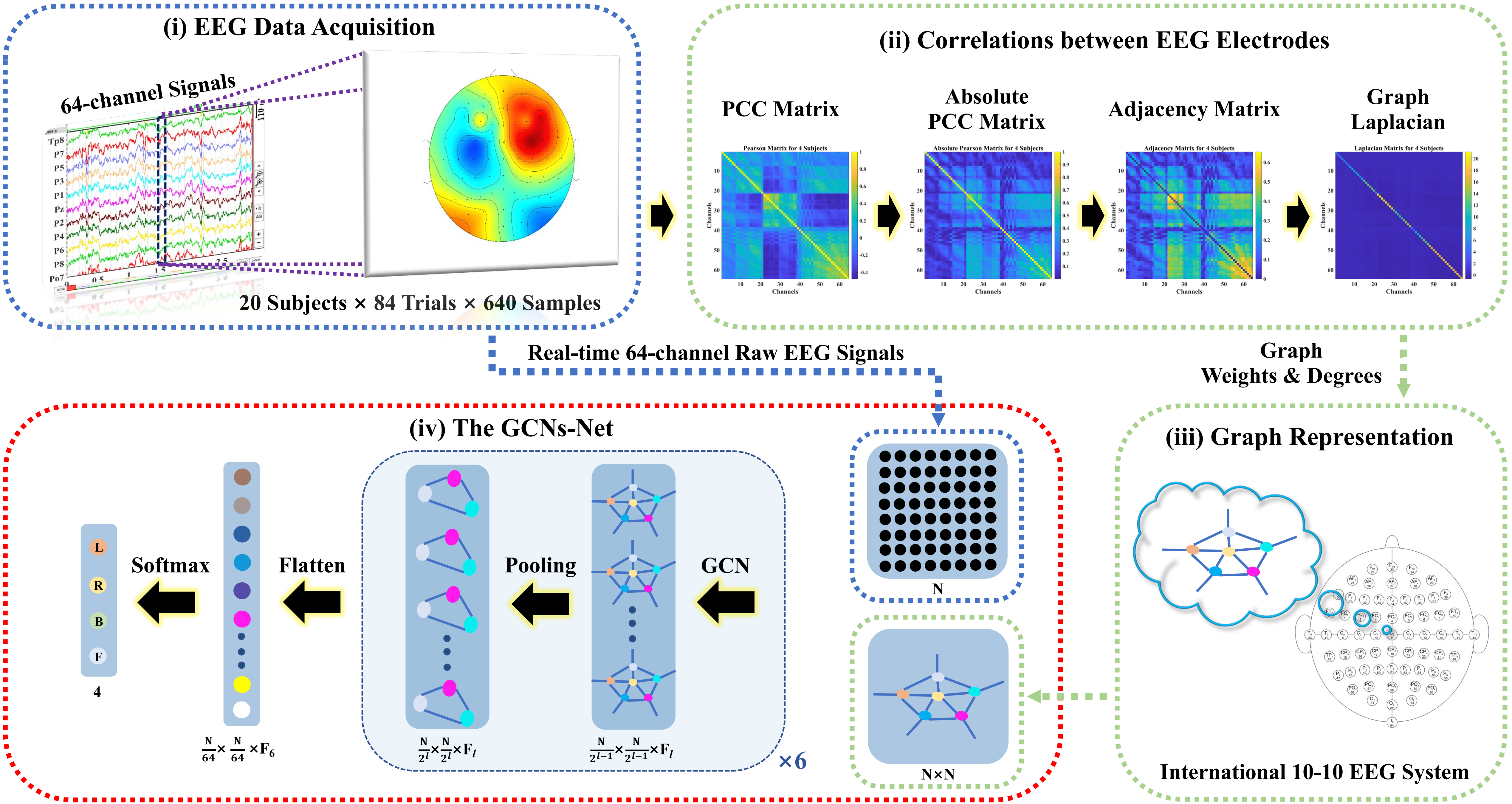
Interpret Functional Networks and better understand human brain

Image Credit: in the public domain.

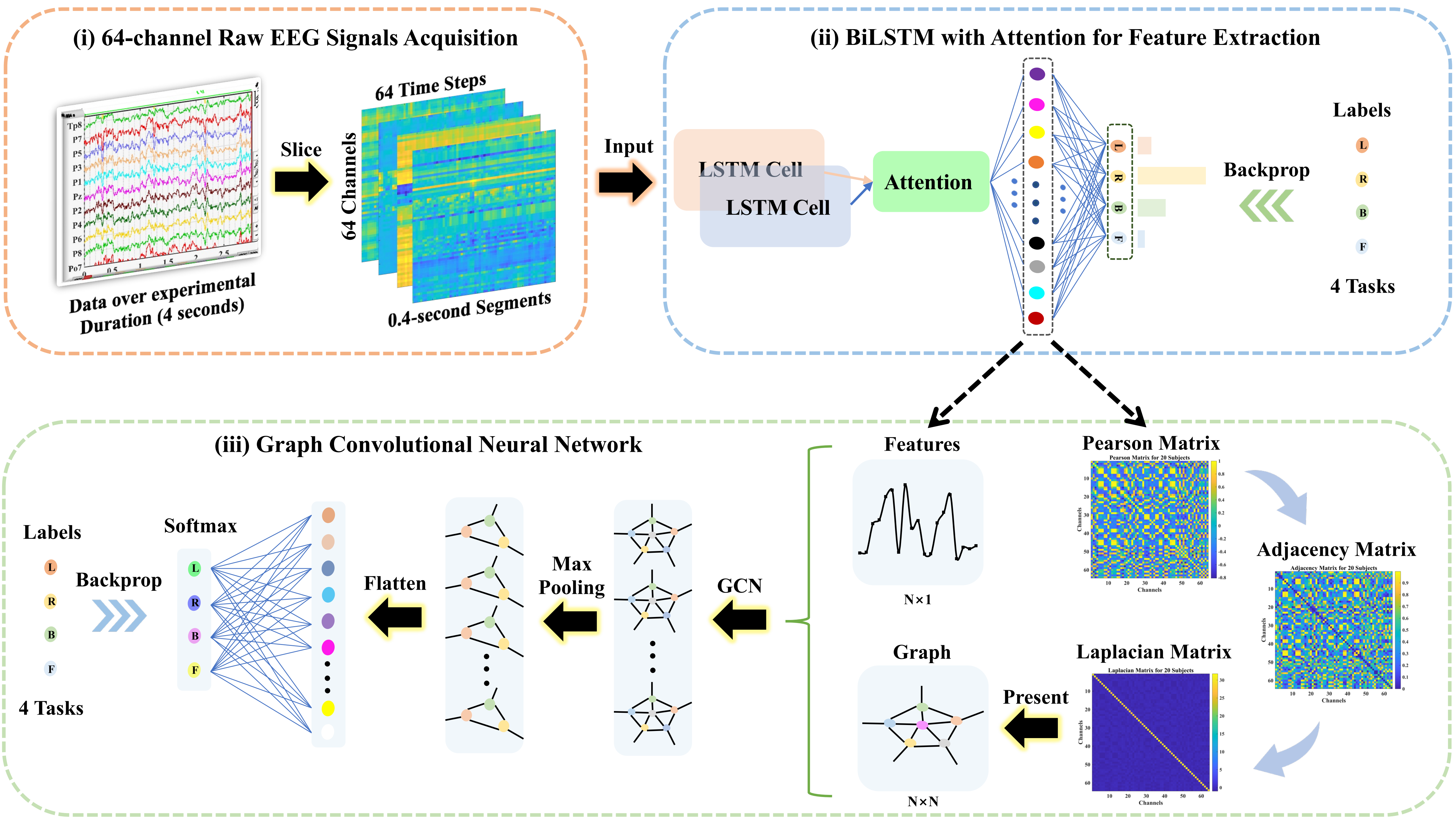
EEG Research Novelty

- ✓ [**Motivation**] Graph Modeling for EEG Electrodes System
- ✓ [**Method**] Graph Representation Learning of EEG Signals
- ✓ [**Motivation**] Spatial-Temporal Analysis of EEG Signals
- ✓ [**Method**] Deep Feature Mining of EEG Signals

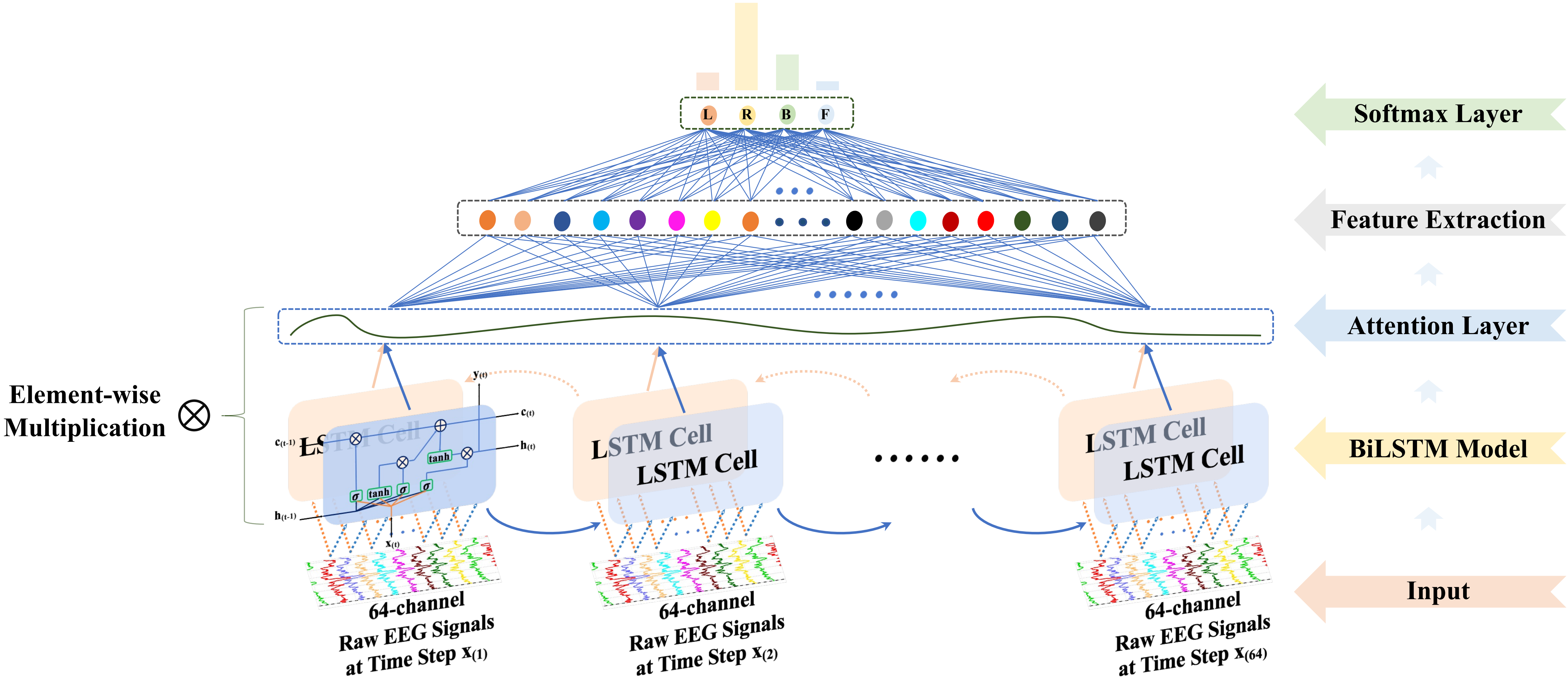




(i) EEG Data Acquisition
64-channel Signals
20 Subjects × 84 Trials × 640 Samples
(ii) Correlations between EEG Electrodes
PCC Matrix
Absolute PCC Matrix
Adjacency Matrix
Graph Laplacian
Pearson Matrix for 4 Subjects
Absolute Pearson Matrix for 4 Subjects
Adjacency Matrix for 4 Subjects
Laplacian Matrix for 4 Subjects
Channels
Real-time 64-channel Raw EEG Signals
Graph Weights & Degrees
(iii) Graph Representation
International 10-10 EEG System
(iv) The GCNs-Net
Softmax
Flatten
Pooling
GCN
L
R
B
F
4
$\frac{N}{64}\times\frac{N}{64}\times F_6$
$\frac{N}{2^l}\times\frac{N}{2^l}\times F_l$
$\frac{N}{2^{l-1}}\times\frac{N}{2^{l-1}}\times F_l$
×6
N
N×N


(i) 64-channel Raw EEG Signals Acquisition
Slice
64 Time Steps
64 Channels
Data over experimental Duration (4 seconds)
0.4-second Segments
Input
(ii) BiLSTM with Attention for Feature Extraction
LSTM Cell
LSTM Cell
Attention
Labels
Backprop
4 Tasks
(iii) Graph Convolutional Neural Network
Features
N×1
Pearson Matrix
Pearson Matrix for 20 Subjects
Adjacency Matrix
Adjacency Matrix for 20 Subjects
Laplacian Matrix
Laplacian Matrix for 20 Subjects
Channels
Graph
N×N
Present
GCN
Max Pooling
Flatten
Softmax
Labels
Backprop
4 Tasks

Attention-based Bidirectional Long Short-term Memory (Bi-LSTM)



EEG Deep Learning

GCNs-Net: A Graph Convolutional Neural Network Approach for Decoding Time-Resolved EEG Motor Imagery Signals

Yimin Hou [1], Shuyue Jia [1, 2], Xiangmin Lun [1], Ziqian Hao [3], Yan Shi [1],

Yang Li [4], Rui Zeng [5], and Jinglei Lv [5 *]

[1] School of Automation Engineering, Northeast Electric Power University

[2] Department of Computer Science, City University of Hong Kong

[3] Jinan Vocational College

[4] School of Electrical Engineering, Northeast Electric Power University

[5] School of Biomedical Engineering and Brain and Mind Center, The University of Sydney

EEG Deep Learning Library: https://github.com/SuperBruceJia/EEG-DL

Background

- **BCI**: establish connections between the brain and machines

  (1) **Acquire and analyze brain signals** while conducting **actual** or **imagery** tasks

  (2) **Control machines**

- **Significance**: help the disabled and understand the human brain
- **Types of BCI**:
  - Electroencephalography (EEG)
  - Magnetoencephalography (MEG)
  - Functional Magnetic Resonance Imaging (fMRI)
  - Invasive BCI Technologies (*e.g.*, Neuralink)

- **Reasons for using EEG for this project**:
  - Non-Invasiveness
  - High Temporal Resolution
  - Portability
  - Inexpensive Equipment

  A potential market

- **Specific Task**: EEG Motor Imagery (*e.g.*, control a wheelchair via imagery-based EEG signals)
- **Our Research**: develop EEG-based BCI technologies to improve current stroke rehabilitation strategies

Image Credit: in the public domain.

Key Points in dealing with EEG time series

- **Individual Variability → Lower Classification Accuracy**
  - ✓ Low SNR
  - ✓ Different brain electrical conductivity ← different anatomical structure of brain
  - ✓ Electrodes' positional error

Feature Extraction

EEG Electrodes' Structure Modeling

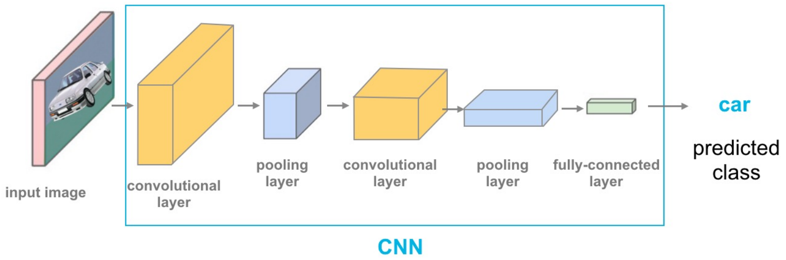
- **Slow Responding → Hard to develop Real-life Applications**
  - ✓ [**most literature**] Trial-level prediction (*e.g.*, 4 s)
  - ✓ Window/Slide-level prediction (*e.g.*, 0.4 s)
  - ✓ Time-resolved prediction (*e.g.*, 6.25 ms) (Our Work)

Time-resolved or Window-based Signal Sampling

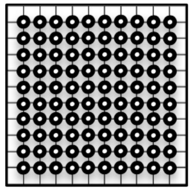
- **Lower Group-level Accuracy → Hard to develop Applications for a Group of People**
  - ✓ [**most literature**] Subject-level prediction (Our Work)
  - ✓ Group-level prediction (Our Work)

Motivation

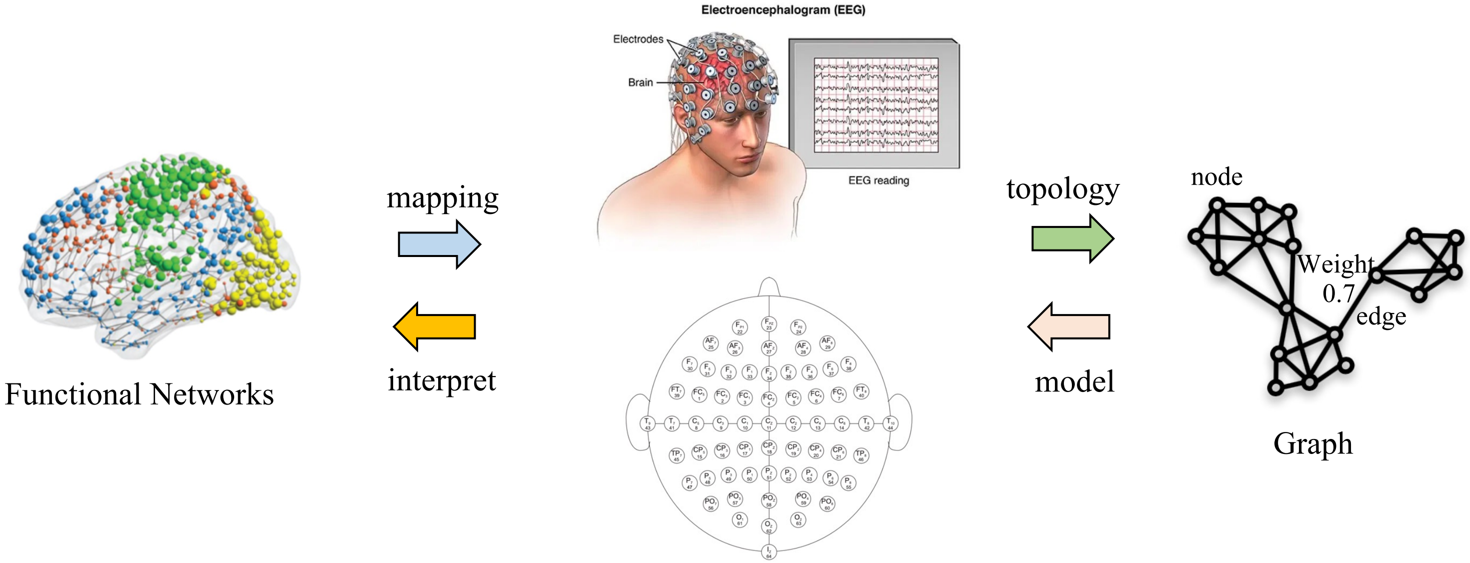
Convolutional Neural Networks:



- **Module: Convolution → Pooling → Fully-connected**
- **Modeling: Euclidean-Structured Data** (*e.g.*, Image, Speech, Natural Language)



- **Neuroscience** research has increasingly emphasized **Brain Network Dynamics**
  - Model **Functional Topological Connectivity** of EEG Electrodes → **Graph** (Non-Euclidean Structure)



International 10-10 EEG System

Our Question

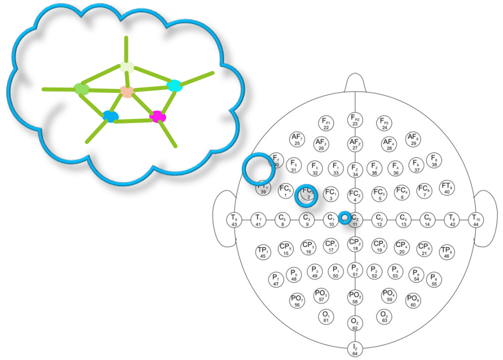
How to model the EEG System as a **Graph**?

How can we process EEG Signals via **Graph Representation Learning**?

Image Credit: The PhysioNet Dataset and the Functional Network Image is in the public domain.

Can we **directly apply convolutions on graphs**?

- Traditional CNN **cannot** directly process graph signals
  - **Graph is irregular** (*i.e.*, unordered and vary in size)
  - Convolution **cannot** keep **Translation Invariance** on non-Euclidean signals
- **Graph Convolutional Neural Networks** (GCN)
  - Directly process **non-Euclidean graph-structured signals**
  - Consider **relational properties** (*e.g.*, correlations) between nodes



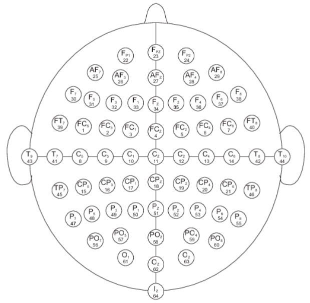
→ Model **Functional Topological Relationships** among EEG electrodes

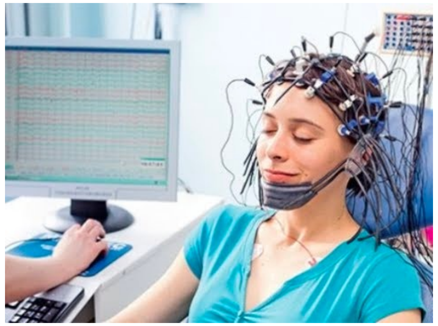
→ Analyze and interpret **Brain Network Dynamics**

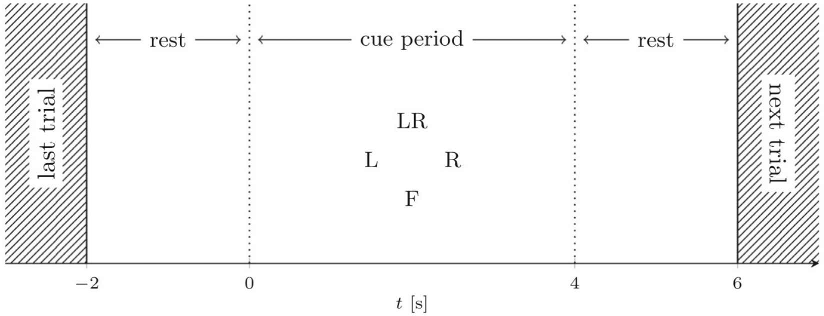
Benchmark Dataset

- The PhysioNet Dataset (EEG Motor Movement/Imagery Dataset)
- International 10-10 EEG System → **64 electrodes**

  (excluding electrodes Nz, F9, F10, FT9, FT10, A1, A2, TP9, TP10, P9, and P10)



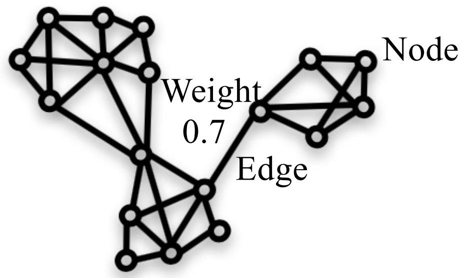




- **109 subjects (the largest number of participants in the field of EEG Motor Imagery)**
- Task: **4-class EEG Motor Imagery Classification**
  - ✓ Imagining (Task 1) left fist, (Task 2) right fist, (Task 3) both fists, (Task 4) both feet
- Each subject → 3 runs, 7 trials, 4 classes → 84 trials in total
- Each trial → 4 seconds experimental duration, 160 Hz Sampling Rate → 640 Time Points
- We apply the **Time-resolved Sampling Method**
  - ✓ Total samples per subject: 3 runs × 7 trials × 4 classes × 4 seconds × 160 Hz = 53,760 samples
  - ✓ Experimental Setting: 90% as the training set and the left 10% as the test set

Image Credit: The PhysioNet Dataset and the middle image is in the public domain.

Preliminary: Graph Representation



Definition: An Undirected and Weighted Graph with N nodes: $\mathbf{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{A}\}$
- $\mathbf{V}$: nodes (vertices), $|\mathbf{V}| = N$
- $\mathbf{E}$: edges (links) that connect nodes
- $\mathbf{A}$: weights (correlations) between nodes

Nodes Correlations: Pearson Matrix $\mathbf{P} \in \mathrm{R}^{N \times N}$ (denotes as PCC matrix)
- Measure the linear correlations between node $\mathbf{x}$ and node $\mathbf{y}$
- μ is the mean, σ is the standard deviation, and $P_{x,y}$ is the Pearson Correlation Coefficient between node $\mathbf{x}$ and node $\mathbf{y}$

$$P_{x,y} = \frac{\mathrm{E}((\mathbf{x} - \mu_x)(\mathbf{y} - \mu_y))}{\sigma_x \sigma_y}$$

- Absolute Pearson Matrix: $|\mathbf{P}| \in \mathrm{R}^{N \times N}$ and $|P_{ij}| \in [0, 1] \rightarrow$ Note: In this work, we only consider **scale**.

Graph Weights: Adjacency Matrix $\mathbf{A} = |\mathbf{P}| - \mathbf{I} \in \mathrm{R}^{N \times N}$, where $\mathbf{I}$ is an Identity Matrix

Graph Degrees: Degree Matrix $\mathbf{D} \in \mathrm{R}^{N \times N}$

$$D_{ii} = \sum_{j=1}^{N} A_{ij}$$

Graph Representation: Combinatorial Laplacian $\mathbf{L} \in \mathrm{R}^{N \times N}$

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

Normalized:

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{\frac{1}{2}}$$

Preliminary:

Spectral Theorem for Graph Laplacian $\mathbf{L}$

$$\mathbf{L} = \mathbf{U\Lambda U}^{\mathrm{T}}$$

$$\mathbf{LU} = \mathbf{\Lambda U}$$

– **U**: Fourier basis → **real** and **orthonormal** eigenvectors of **L**

– **Λ**: Fourier modes → the diagonal is the **ordered** and **real nonnegative** eigenvalues of **L**

Graph Fourier Transforms of Signal f

can be seen as the $\mathrm{e}^{-j\omega \mathrm{t}}$ in Fourier Transforms

$$\mathrm{F}[f(\boldsymbol{\lambda})] = \hat{f}(\boldsymbol{\lambda}) = \sum_{i=1}^{\mathrm{n}} f(i) \times U(i)$$

$$\hat{f}(\boldsymbol{\lambda}) = \mathbf{U}^{\mathrm{T}} f \Leftrightarrow f = \mathbf{U}\hat{f}(\boldsymbol{\lambda})$$

$\hat{f}(\boldsymbol{\lambda})$ is the projection value of the Fourier basis **U**

Preliminary: Graph Convolution via Graph Fourier Transform

Notation:

Signal f

Signal h

F: Fourier Transforms

F^{-1}: Inverse Fourier Transforms

$\hat{f}(w)$: F(f)

$\hat{h}(w)$: F(h)

Note: Fourier Transforms of **Convolution in the spatial domain**

$\Leftrightarrow$

Point-wise Multiplication of two Fourier transformed signals

$$\mathrm{F}\left((f * h)_{\mathbf{G}}\right) = \hat{f}(w) \times \hat{h}(w)$$

Convolution $(f * h)_{\mathbf{G}} = \mathrm{F}^{-1}(\hat{f}(w) \times \hat{h}(w))$

$$\hat{f}(\boldsymbol{\lambda}) = \mathbf{U}^{\mathrm{T}} f$$

Hadamard Product (Element-wise Multiplication)

$$(f * h)_{\mathbf{G}} = \mathrm{F}^{-1}\left(\left(\mathbf{U}^{\mathrm{T}} f\right) \odot \left(\mathbf{U}^{\mathrm{T}} h\right)\right)$$

$$f = \mathbf{U}\hat{f}(\boldsymbol{\lambda})$$

$$(f * h)_{\mathbf{G}} = \mathbf{U}\left(\left(\mathbf{U}^{\mathrm{T}} f\right) \odot \left(\mathbf{U}^{\mathrm{T}} h\right)\right)$$

$$(f * h)_{\mathbf{G}} = \underbrace{\mathbf{U}}_{[\mathrm{n} \times \mathrm{n}]} \underbrace{\mathrm{diag}[\hat{h}(\lambda_1), \hat{h}(\lambda_2), \ldots, \hat{h}(\lambda_{\mathrm{n}})]}_{[\mathrm{n} \times \mathrm{n}]} \underbrace{\mathbf{U}^{\mathrm{T}}}_{[\mathrm{n} \times \mathrm{n}]} \underbrace{f}_{[\mathrm{n} \times \mathrm{d}]}$$

Source: https://en.wikipedia.org/wiki/Convolution_theorem

Graph Convolution

$$(f * h)_{\mathbf{G}} = \mathbf{U}\, \boxed{\mathrm{diag}[\hat{h}(\lambda_1), \hat{h}(\lambda_2), \ldots, \hat{h}(\lambda_n)]}\, \mathbf{U}^{\mathrm{T}} f$$

Activation Function

$$\mathcal{Y} = \sigma(\mathbf{U}\mathbf{g}_{\boldsymbol{\theta}}\mathbf{U}^{\mathrm{T}}\boldsymbol{\chi})$$

K^{th} Polynomial Function

$$\mathcal{Y} = \sigma(\mathbf{U}\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\Lambda})\mathbf{U}^{\mathrm{T}}\boldsymbol{\chi})$$
$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Approximate

$$\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\Lambda}) = \sum_{k=0}^{K} \theta_k \boldsymbol{\Lambda}^k$$

$$\mathcal{Y} = \sigma\left(\mathbf{U}\sum_{k=0}^{K} \theta_k \boldsymbol{\Lambda}^k\, \mathbf{U}^{\mathrm{T}}\boldsymbol{\chi}\right) = \sigma\left(\sum_{k=0}^{K} \theta_k\, (\mathbf{U}\boldsymbol{\Lambda}^k\mathbf{U}^{\mathrm{T}})\boldsymbol{\chi}\right) = \sigma\left(\sum_{k=0}^{K} \theta_k\, (\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}})^k\boldsymbol{\chi}\right) = \sigma\left(\sum_{k=0}^{K} \theta_k\, \mathbf{L}^k\boldsymbol{\chi}\right)$$

$$\boxed{\mathcal{Y} = \sigma\left(\sum_{k=0}^{K} \theta_k\, \mathbf{L}^k\boldsymbol{\chi}\right)}$$

Credit: Defferrard *et al.*, Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering, In NeurIPS 2016.

Graph Convolution

Convolution:
Weighted Sum

Weight Sharing

Node Aggregation

K is Filter Size

$$\mathcal{Y} = \sigma\left(\sum_{k=0}^{K} \theta_k \mathbf{L}^k \chi\right)$$

No need for Fourier Transform

Beauty is in Simplicity

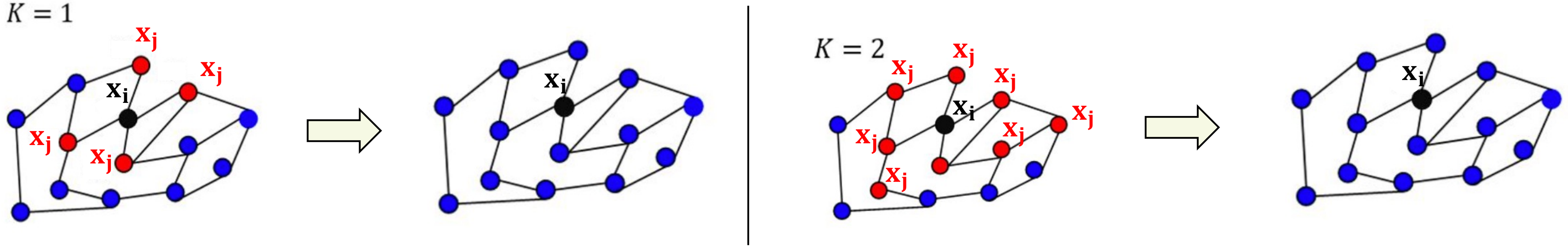
GCN Key Idea: Use "**edge information**" to **aggregate** "**node information**" to generate a new "**node representation**"

Laplace Operator

Local connectivity

$$\mathbf{x_{new}} \leftarrow \mathbf{L}\mathbf{x_i} = \sum_j A_{ij}(\mathbf{x_i} - \mathbf{x_j})$$

Localize in Space

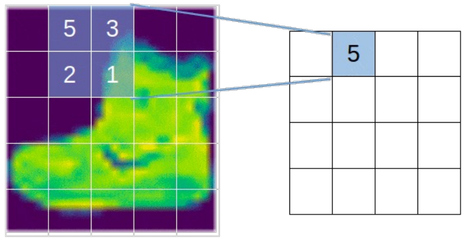


Pros:

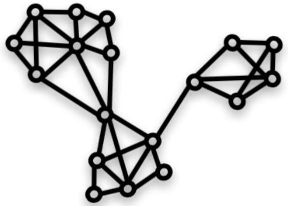
1. No need for Spectral Decomposition of **L**
2. Less number of parameters (decrease model complexity) → $K \ll N$

Cons: Need to compute $\mathbf{L}^k$

Pooling on Graphs (Graph Coarsening)



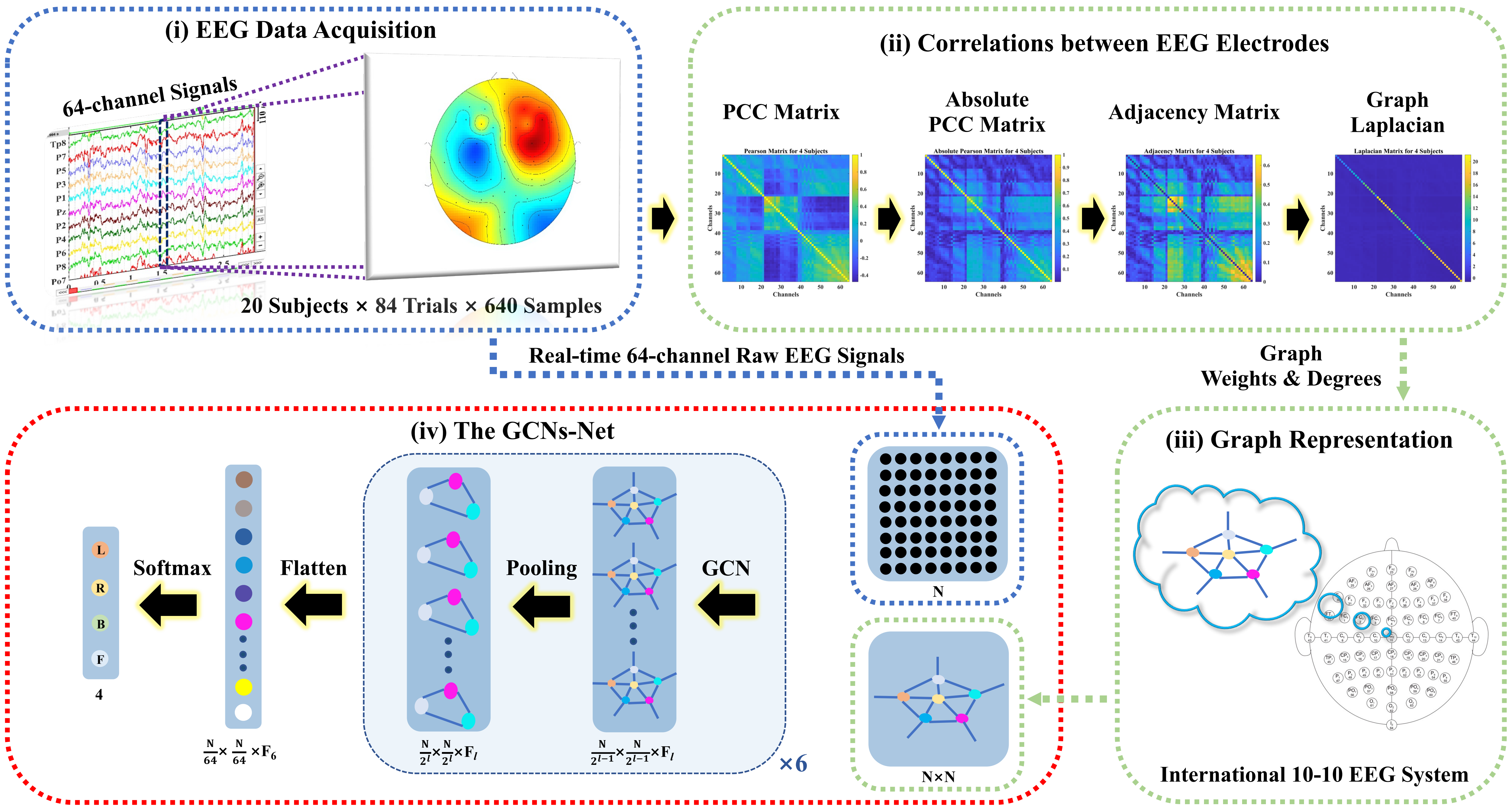
- Traditional CNN doesn't need to consider neighbors after convolutions
  - [**Euclidean Structure**] The output Feature Maps are "regular"
  - The neighbor is "meaningful"
- GCNs need to consider neighbors after convolutions
  - [**Non-Euclidean Structure**] The output **graphs' nodes are not arranged in any meaningful way**
  - Use **Graclus Multilevel Clustering Algorithm** to find "meaningful" neighbors
  - Minimize the ***Local Normalized Cut*** (a cluster grouping method)



$$-W_{ij}(\frac{1}{d_i}+\frac{1}{d_j})$$

  - i and j denote node i and node j
  - W_{ij} is the **learned weight** between node i and node j

Image Credit: in the public domain.


(i) EEG Data Acquisition
64-channel Signals
20 Subjects × 84 Trials × 640 Samples
(ii) Correlations between EEG Electrodes
PCC Matrix
Absolute PCC Matrix
Adjacency Matrix
Graph Laplacian
Pearson Matrix for 4 Subjects
Absolute Pearson Matrix for 4 Subjects
Adjacency Matrix for 4 Subjects
Laplacian Matrix for 4 Subjects
Channels
Real-time 64-channel Raw EEG Signals
Graph Weights & Degrees
(iii) Graph Representation
International 10-10 EEG System
(iv) The GCNs-Net
Softmax
Flatten
Pooling
GCN
L
R
B
F
4
$\frac{N}{64}\times\frac{N}{64}\times F_6$
$\frac{N}{2^l}\times\frac{N}{2^l}\times F_l$
$\frac{N}{2^{l-1}}\times\frac{N}{2^{l-1}}\times F_l$
×6
N
N×N

Correlation among EEG electrodes
Two Subjects: Subject 10 and 5

Problem: Individual Variability

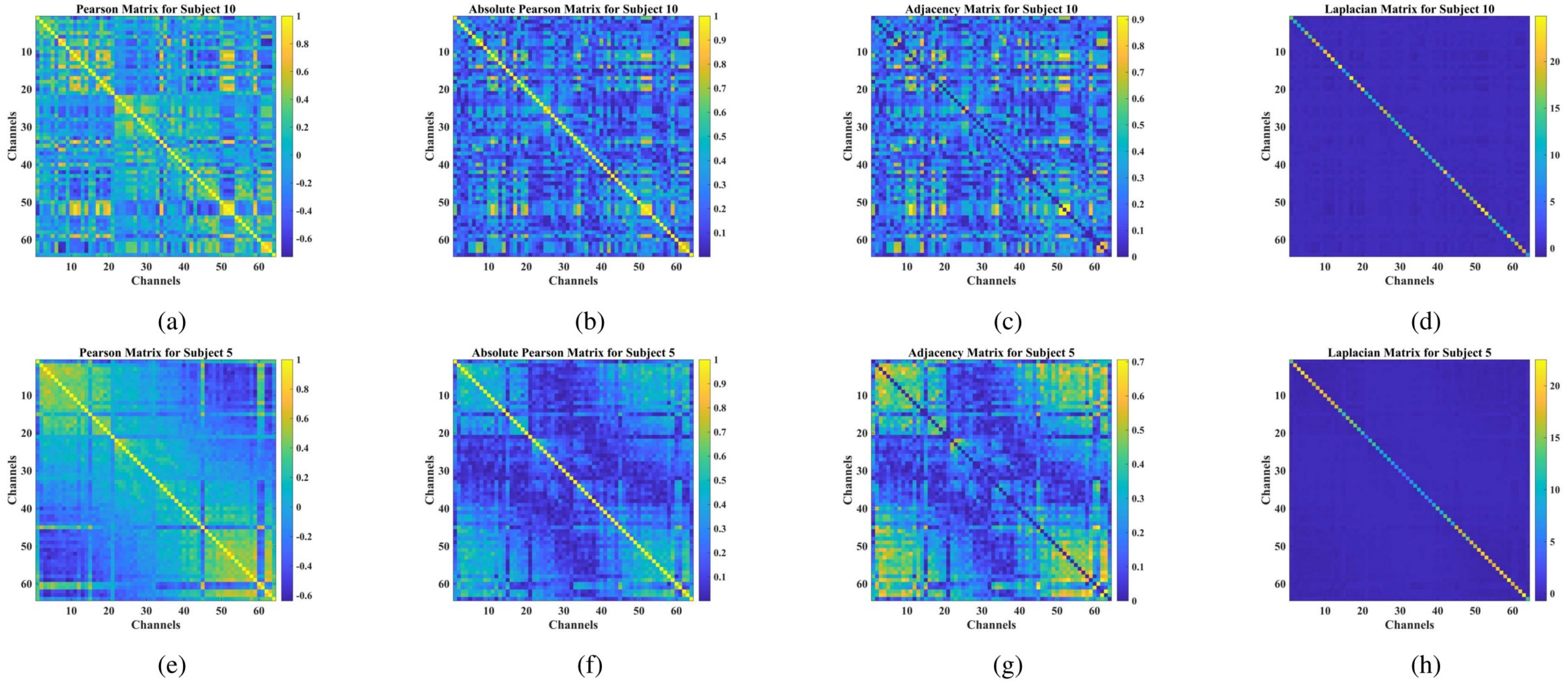


Fig. 6. PCC matrix, absolute PCC matrix, adjacency matrix, and graph Laplacian for Subjects 10 and 5 from the PhysioNet dataset. (a) PCC matrix for Subject 10. (b) Absolute PCC matrix for Subject 10. (c) Adjacency matrix for Subject 10. (d) Graph Laplacian for Subject 10. (e) PCC matrix for Subject 5. (f) Absolute PCC matrix for Subject 5. (g) Adjacency matrix for Subject 5. (h) Graph Laplacian for Subject 5.

Correlation among EEG electrodes 20 Subjects and 100 Subjects

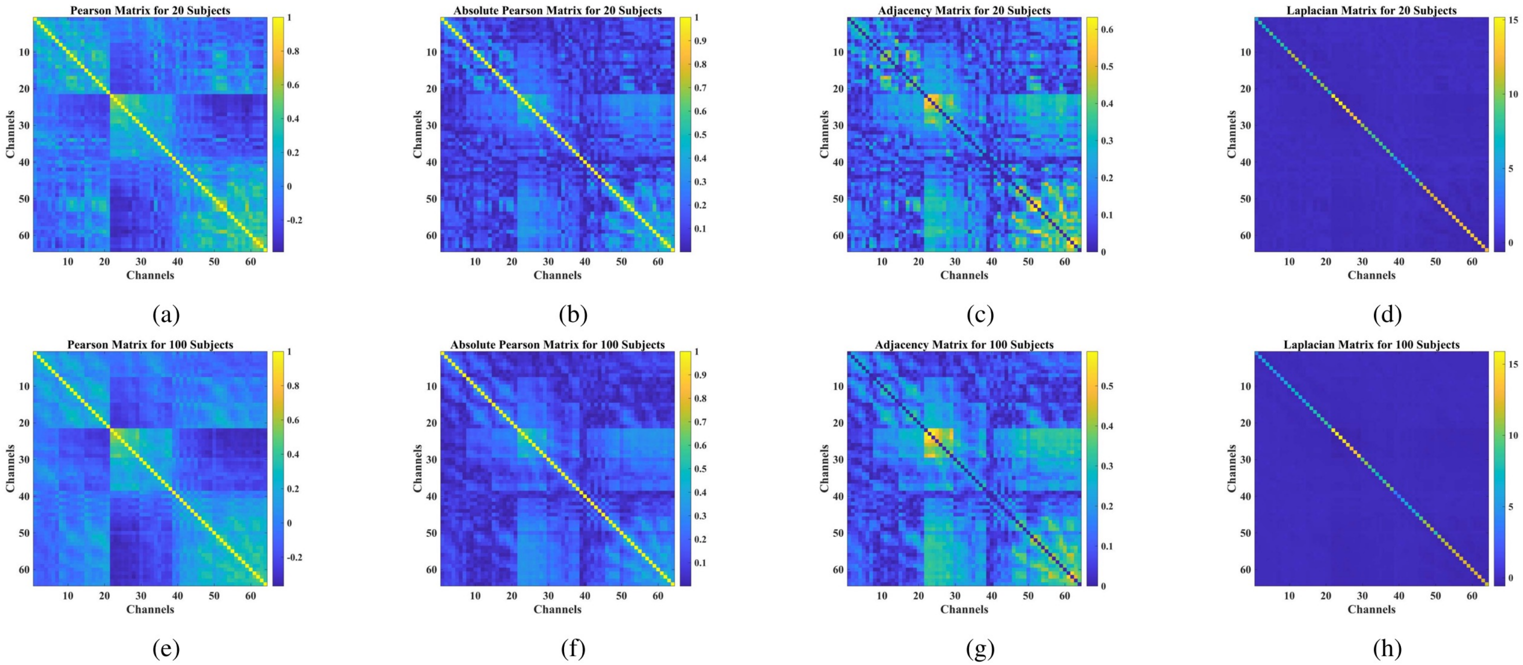


Fig. 2. PCC matrix, absolute PCC matrix, adjacency matrix, and graph Laplacian for 20 and 100 subjects, respectively, from the PhysioNet dataset. (a) PCC matrix for 20 subjects. (b) Absolute PCC matrix for 20 subjects. (c) Adjacency matrix for 20 subjects. (d) Graph Laplacian for 20 subjects. (e) PCC matrix for 100 subjects. (f) Absolute PCC matrix for 100 subjects. (g) Adjacency matrix for 100 subjects. (h) Graph Laplacian for 100 subjects.

Increasing ***the number of subjects*** alleviates ***individual variability***

Model Design for 64-electrode EEG System

TABLE I

IMPLEMENTATION DETAILS OF THE PROPOSED GCNS-NET ON THE PHYSIONET DATASET

Layer	Type	Maps	Size	Edges	Polynomial Order	Pooling Size	Activation	Weights	Bias
Softmax	Fully-connected	–	O	–	–	–	Softmax	$\frac{N}{64} \times \frac{N}{64} \times F_6 \times O$	O
Flatten	Flatten	–	$\frac{N}{64} \times \frac{N}{64} \times F_6$	–	–	–	–	–	–
P6	Max-pooling	F_6	$\frac{N}{32}$	$\sum_{i=1}^{\frac{N}{32}-1} i$	–	2	–	–	–
C6	Convolution	F_6	$\frac{N}{32}$	$\sum_{i=1}^{\frac{N}{32}-1} i$	K	–	Softplus	$F_5 \times F_6 \times K$	$\frac{N}{32} \times F_6$
P5	Max-pooling	F_5	$\frac{N}{16}$	$\sum_{i=1}^{\frac{N}{16}-1} i$	–	2	–	–	–
C5	Convolution	F_5	$\frac{N}{16}$	$\sum_{i=1}^{\frac{N}{16}-1} i$	K	–	Softplus	$F_4 \times F_5 \times K$	$\frac{N}{16} \times F_5$
P4	Max-pooling	F_4	$\frac{N}{8}$	$\sum_{i=1}^{\frac{N}{8}-1} i$	–	2	–	–	–
C4	Convolution	F_4	$\frac{N}{8}$	$\sum_{i=1}^{\frac{N}{8}-1} i$	K	–	Softplus	$F_3 \times F_4 \times K$	$\frac{N}{8} \times F_4$
P3	Max-pooling	F_3	$\frac{N}{4}$	$\sum_{i=1}^{\frac{N}{4}-1} i$	–	2	–	–	–
C3	Convolution	F_3	$\frac{N}{4}$	$\sum_{i=1}^{\frac{N}{4}-1} i$	K	–	Softplus	$F_2 \times F_3 \times K$	$\frac{N}{4} \times F_3$
P2	Max-pooling	F_2	$\frac{N}{2}$	$\sum_{i=1}^{\frac{N}{2}-1} i$	–	2	–	–	–
C2	Convolution	F_2	$\frac{N}{2}$	$\sum_{i=1}^{\frac{N}{2}-1} i$	K	–	Softplus	$F_1 \times F_2 \times K$	$\frac{N}{2} \times F_2$
P1	Max-pooling	F_1	N	$\sum_{i=1}^{N-1} i$	–	2	–	–	–
C1	Convolution	F_1	N	$\sum_{i=1}^{N-1} i$	K	–	Softplus	$1 \times F_1 \times K$	$N \times F_1$
Input	Input	1	N	$\sum_{i=1}^{N-1} i$	–	–	–	–	–

Model Optimization

- **Ablation Study**: **Optimal Model Structure** (64-electrode EEG System)
  - C6-P6-K2 with [16, 32, 64, 128, 256, 512] filters
- **Gradient Iterative Solver**: **Adam Optimizer** with **Stochastic Gradient Descent** (SGD) algorithm
  - Learning Rate: 0.01
  - Batch Size: 1,024
- **Activation Function**: **Softplus** (Smooth Rectified Linear Unit)

$$\mathrm{F}(\mathbf{x}) = \log(1 + \mathrm{e}^{\mathbf{x}})$$

- **Model Output**: **Softmax**: **y** are labels, $\hat{\mathbf{y}}$ are the final output tasks

$$\widehat{y}_i = \operatorname{argmax}\left(\frac{\mathrm{e}^{y_\mathrm{i}}}{\sum_{\mathrm{i}=1}^{4} \mathrm{e}^{y_\mathrm{i}}}\right)$$

- **Loss Function**: **Cross-entropy Loss** with **L2 regularization**

$$\mathrm{Loss} = -\sum_{\mathrm{i}=1}^{4} y_i \log(\widehat{y}_i) + \lambda\left(\sum_{\mathrm{j}=1}^{\mathrm{n}} w_\mathrm{j}^2 + b_\mathrm{j}^2\right)$$

$\lambda = 1 \times 10^{-6}$ is the coefficient of the L2 regularization.

Ablation Study

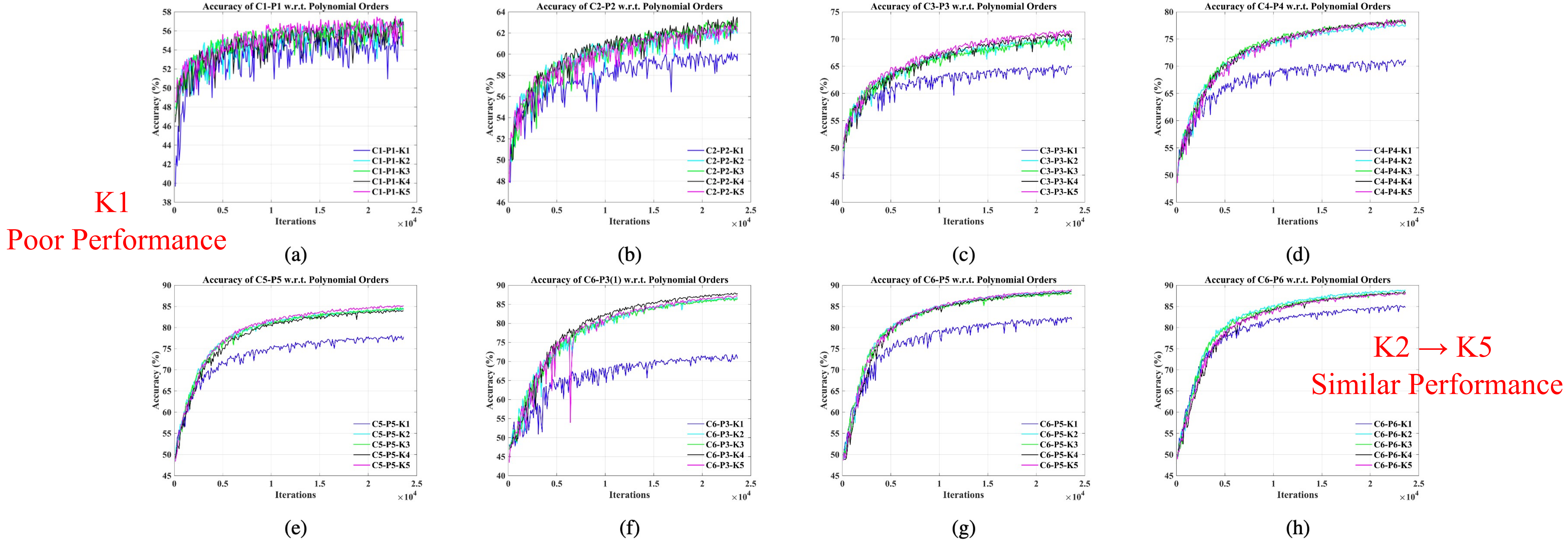


Fig. 3. Accuracy of some selected models regarding different polynomial approximation order. The models are selected from Table II. (a) Accuracy of the model C1-P1 (model 1). (b) Accuracy of the model C2-P2 (model 3). (c) Accuracy of the model C3-P3 (model 6). (d) Accuracy of the model C4-P4 (model 10). (e) Accuracy of the model C5-P5 (model 14). (f) Accuracy of the model C6-P3 (model 16). (g) Accuracy of the model C6-P5 (model 19). (h) Accuracy of the model C6-P6 (model 20).

Experimental Results
Groupwise Prediction and Subject-specific Adaptation

TABLE IV

PERFORMANCE COMPARISONS ON THE PHYSIONET DATASET

Related Work	Max. Accuracy	Avg. Accuracy	p-value	Level	Approach	Num. of Subjects
Dose *et al.* (2018) [22]	–	58.58%	–	Group	CNNs	105
	80.38%	68.51%	< 0.05	Subject		1
Ma *et al.* (2018) [60]	82.65%	68.20%	–	Group	RNNs	12
Hou *et al.* (2020) [20]	94.50%	–	–	Group	ESI-CNNs	10
	96.00%	–	> 0.05	Subject		1
Hou *et al.* (2022) [34]	94.64%	–	–	Group	BiLSTM-GCN	20
	98.81%	95.48%	> 0.05	Subject		1
Jia *et al.* (2022) [40]	94.16%	93.78%	–	Group	Graph ResNet	20
	98.08%	94.18%	> 0.05	Subject		1
Author	**89.39%**	**88.57%**	–	**Group**	**GCNs-Net**	**20**
	88.14%	–				**100**
	98.72%	**93.06%**		**Subject**		**1**

Note: ***p*-value < 0.05** → Statistically Significant Difference

Takeaways and Future Work

✓ **Graph Representation**

Graph Representation Learning to deeply extract **Network Patterns of Brain Dynamics** for EEG classification.

✓ **Model Converge**

Converge for both **Personalized and Groupwise Predictions**, indicating that the GCNs-Net is able to build a generalized representation of EEG time-series **against both Personalized and Groupwise Variations**.

✓ **Future Work**

Model EEG signals as **Dynamic Graphs** and process them via **Dynamic Graph Representation Learning**.

Deep Feature Mining via Attention-based BiLSTM-GCN for Human Motor Imagery Recognition

Yimin Hou [1], Shuyue Jia [1, 2 *], Xiangmin Lun [1], Shu Zhang [3], Tao Chen [1], Fang Wang [1], and Jinglei Lv [4]

[1] School of Automation Engineering, Northeast Electric Power University

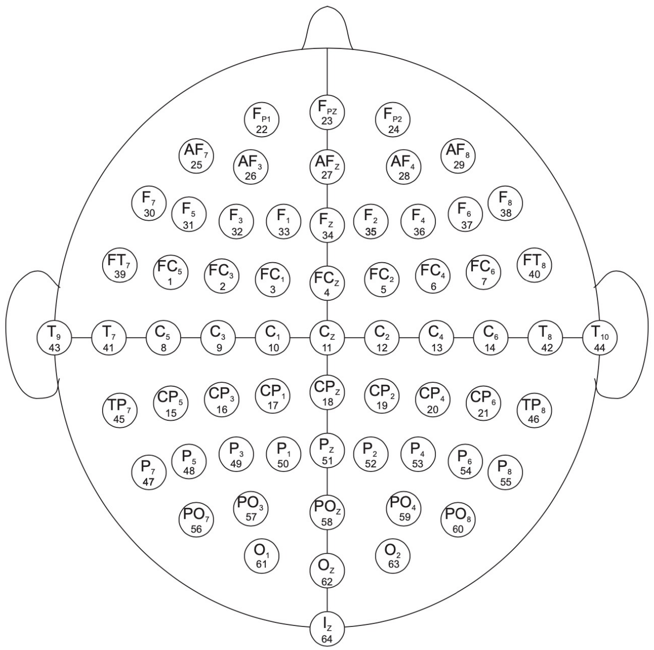
[2] Department of Computer Science, City University of Hong Kong

[3] School of Computer Science, Northwestern Polytechnical University

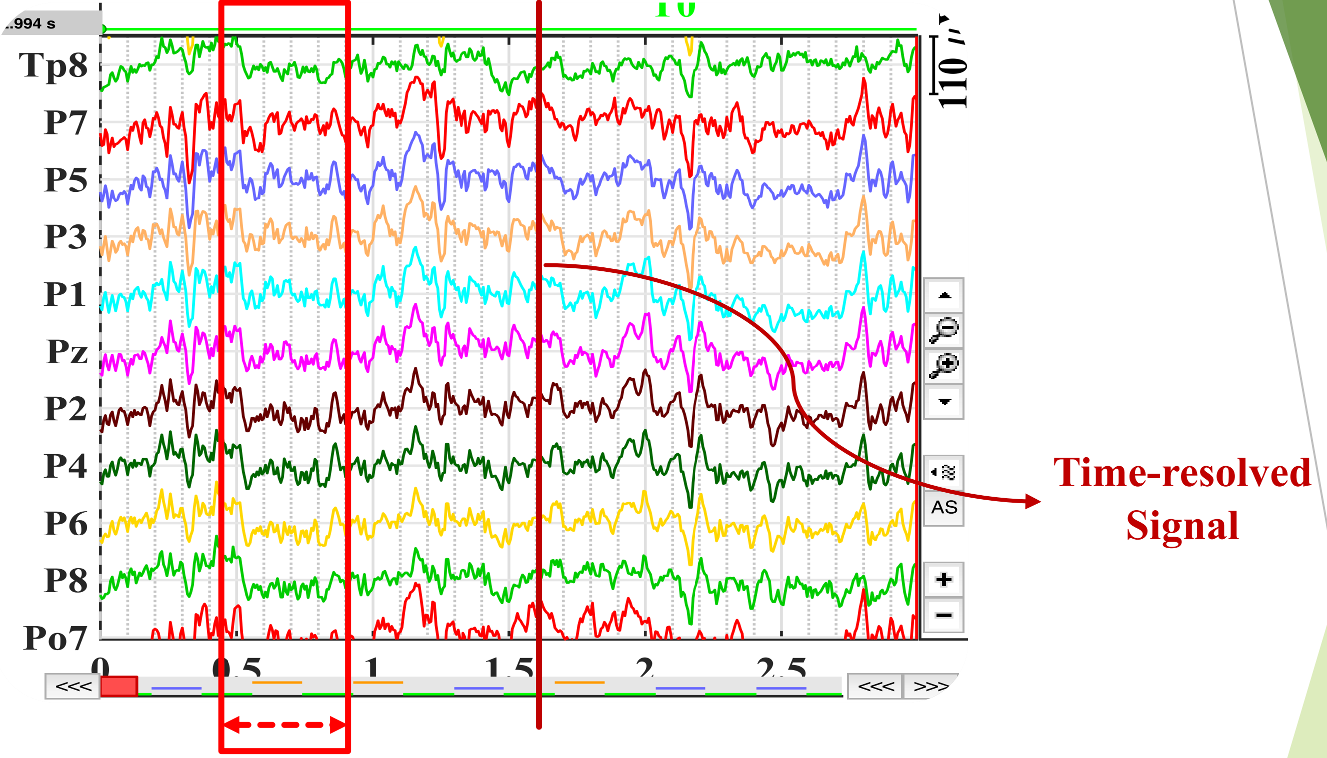
[4] School of Biomedical Engineering and Brain and Mind Center, The University of Sydney

EEG Deep Learning Library: https://github.com/SuperBruceJia/EEG-DL

One Problem of the GCNs-Net



Spatial information



Temporal information

✓ GCNs-Net is based on **Time-resolved Signal** → doesn't consider **Temporal Information**

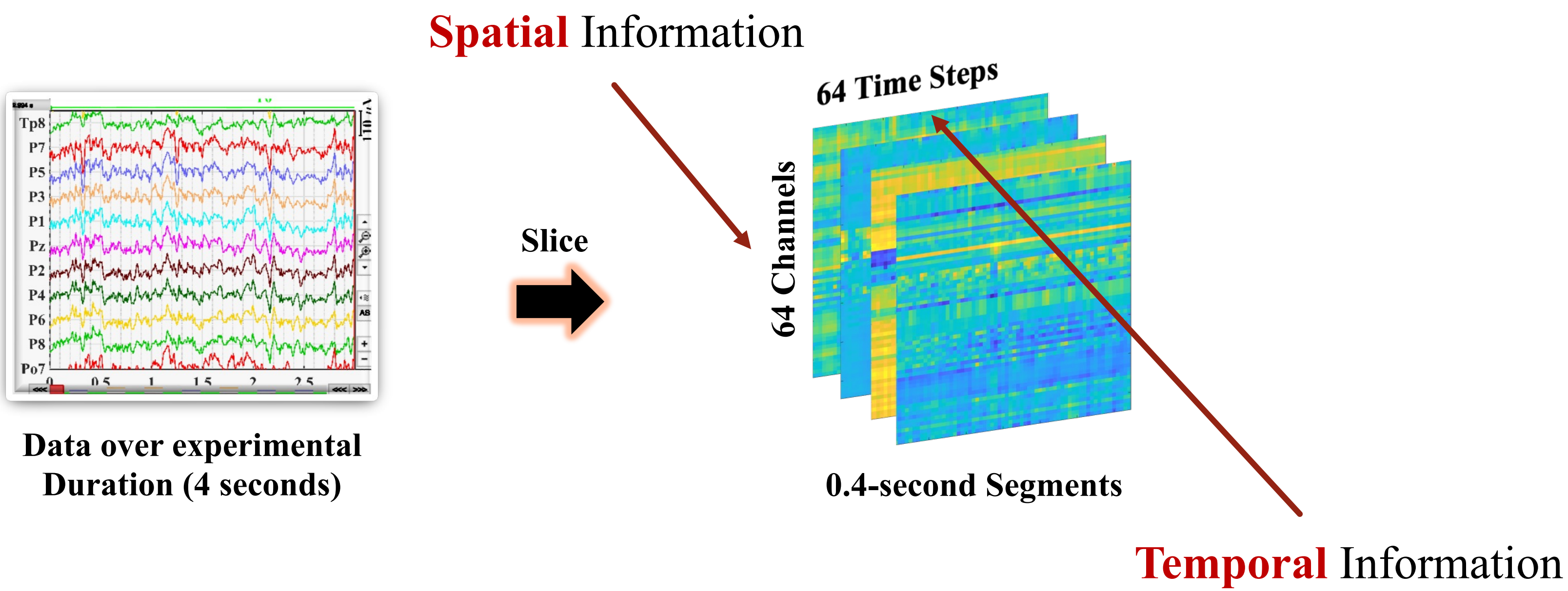
Motivation:

✓ **[Spatial-Temporal Analysis] Consider Temporal and Spatial Information from EEG signals**

✓ **[Responsive] Maintain High Responding Time**

Image Credit: The PhysioNet Dataset.

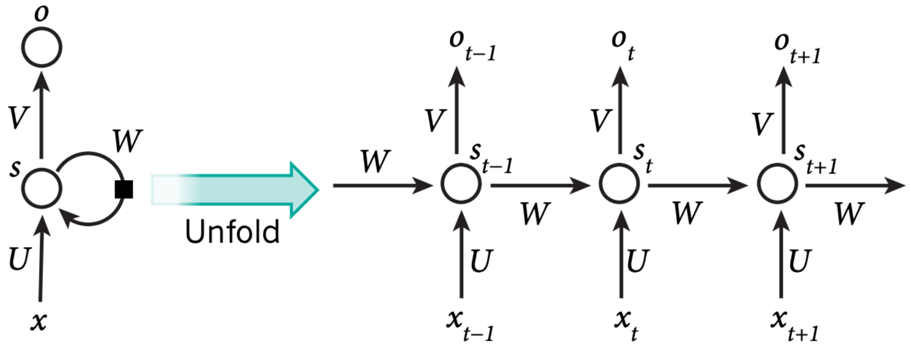
64-channel Raw EEG Signals Acquisition



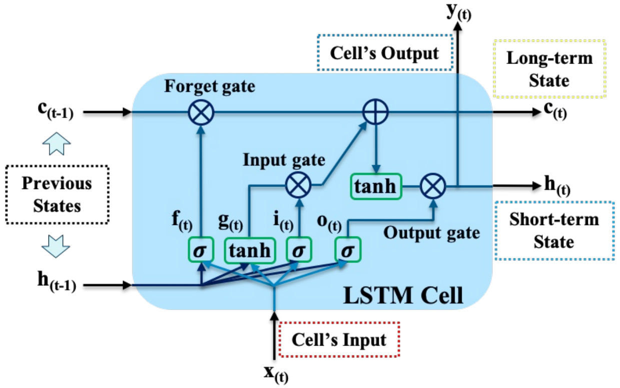
- ✓ 4-s Signals (experimental duration): **0.4-s segments** over time
- ✓ Each Segment: **64 channels × 64 time steps**
- ✓ Pre-processed Data: **Temporal** Information + **Spatial** Information

Temporal Information Extraction

unrolling the network through time



Recurrent Neural Network (RNN)

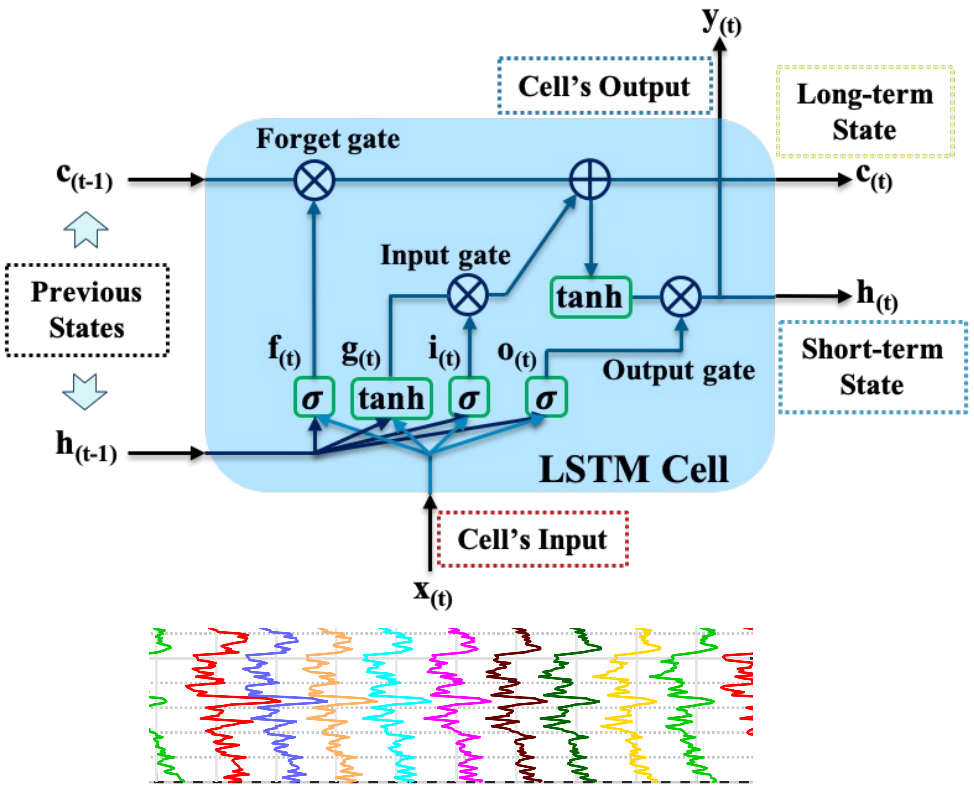


Long Short-Term Memory (LSTM)

- ✓ Designed for **order-mattered sequential data**, *e.g.*, time series
- ✓ The learned features at *time step t* are affected by $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ → **continuly learn** from time series
- ✓ LSTM: better capture **long-range sequence dependencies**
- ✓ Gated Recurrent Units (GRU): **lightweight** architecture with comparable performance

Image Credit: LeCun *et al.*, Deep Learning, In Nature.

Long Short-term Memory (LSTM)



64-channel Raw EEG Signals at Time Step $x_{(t)}$

- ✓ Capture **Long-range Dependencies** by the long-term state path $\mathbf{c}_{t-1} \rightarrow \mathbf{c}_t$
- ✓ **Input Gate**: store $\mathbf{x}_t$ and control $\mathbf{c}_t$'s input
- ✓ **Forget Gate**: control $\mathbf{c}_{t-1}$
- ✓ **Output Gate**: control $\mathbf{c}_t$'s output
  → short-term state $\mathbf{h}_t$ (**Cell's Output**)
- ✓ More parameters to store information
- ✓ Bidirectional:
  (1) $\mathbf{x}_1 \rightarrow \mathbf{x}_t$
  (2) $\mathbf{x}_t \rightarrow \mathbf{x}_1$

Attention Mechanism

- ✓ Signals or Outputs

<u>**Equally**</u> **treated/contributed**

vs.

<u>**Differently**</u> **treated/contributed** <u>**with preference/importance**</u>

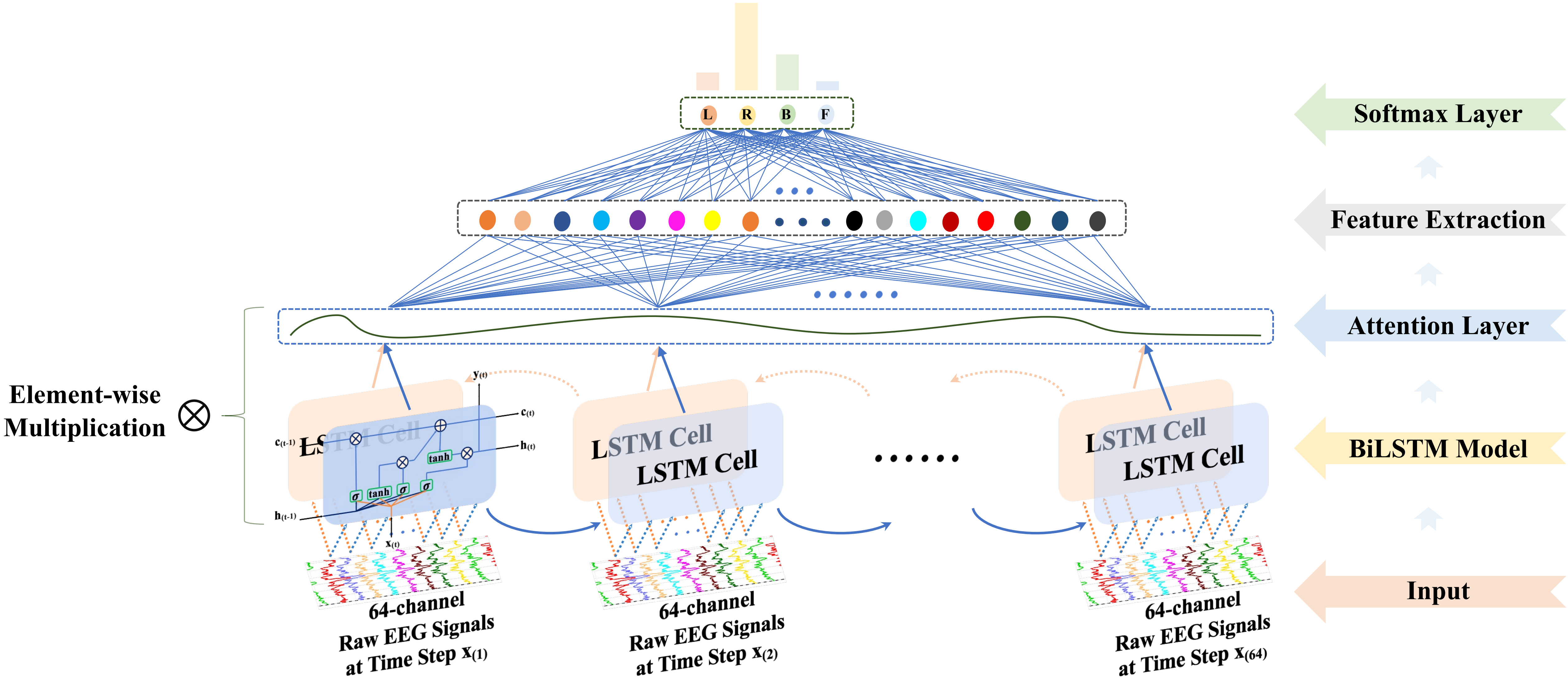
FC Layer $\mathbf{U}_t = \tanh(\mathbf{W}_w \mathbf{y}_t + \mathbf{b}_w)$

Attentional Weights $\boldsymbol{\alpha}_t = \dfrac{\exp(\mathbf{U}_t^T \mathbf{W}_U)}{\sum_t \exp(\mathbf{U}_t^T \mathbf{W}_U)}$

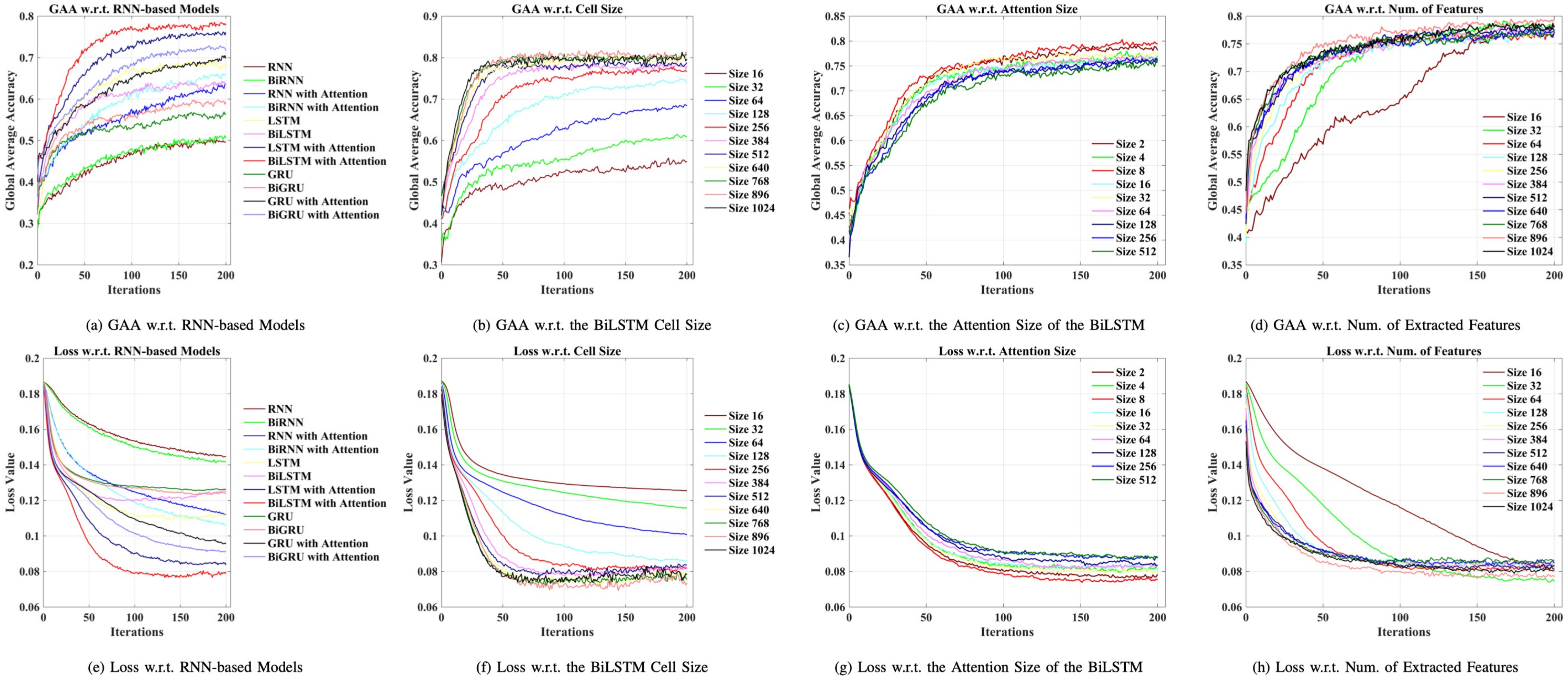
Weighted Sum $\hat{\mathbf{U}}_t = \sum_t \boldsymbol{\alpha}_t \mathbf{y}_t$

Image Credit: Prof. Bolei Zhou.

Attention-based Bidirectional Long Short-term Memory (Bi-LSTM)



Model Design Ablation Study



(a) GAA w.r.t. RNN-based Models

(b) GAA w.r.t. the BiLSTM Cell Size

(c) GAA w.r.t. the Attention Size of the BiLSTM

(d) GAA w.r.t. Num. of Extracted Features

(e) Loss w.r.t. RNN-based Models

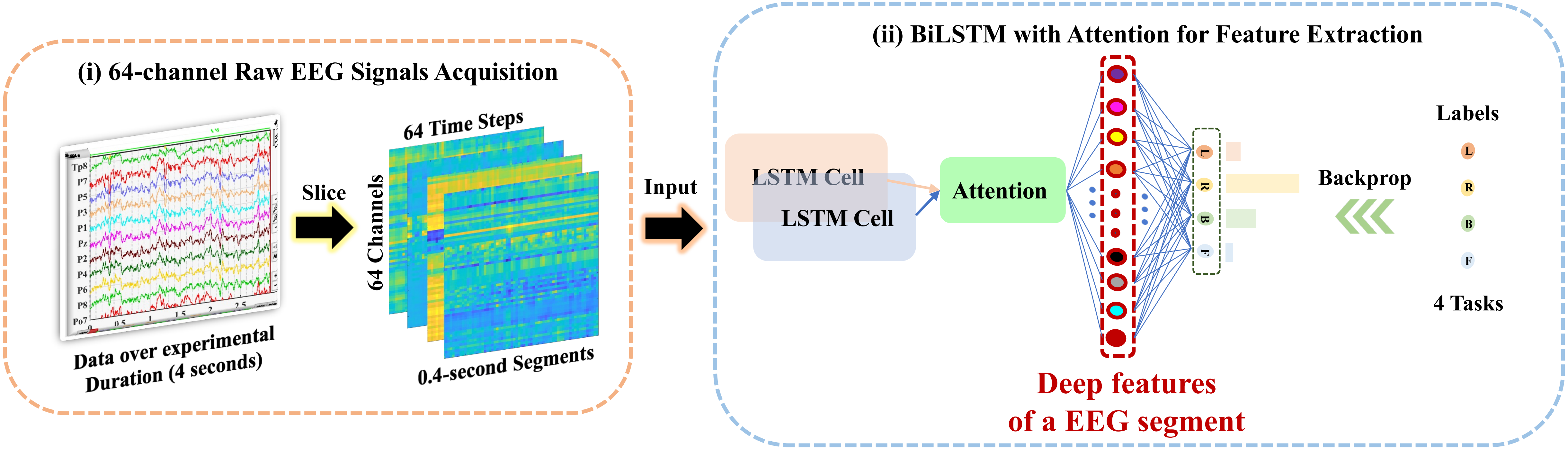
(f) Loss w.r.t. the BiLSTM Cell Size

(g) Loss w.r.t. the Attention Size of the BiLSTM

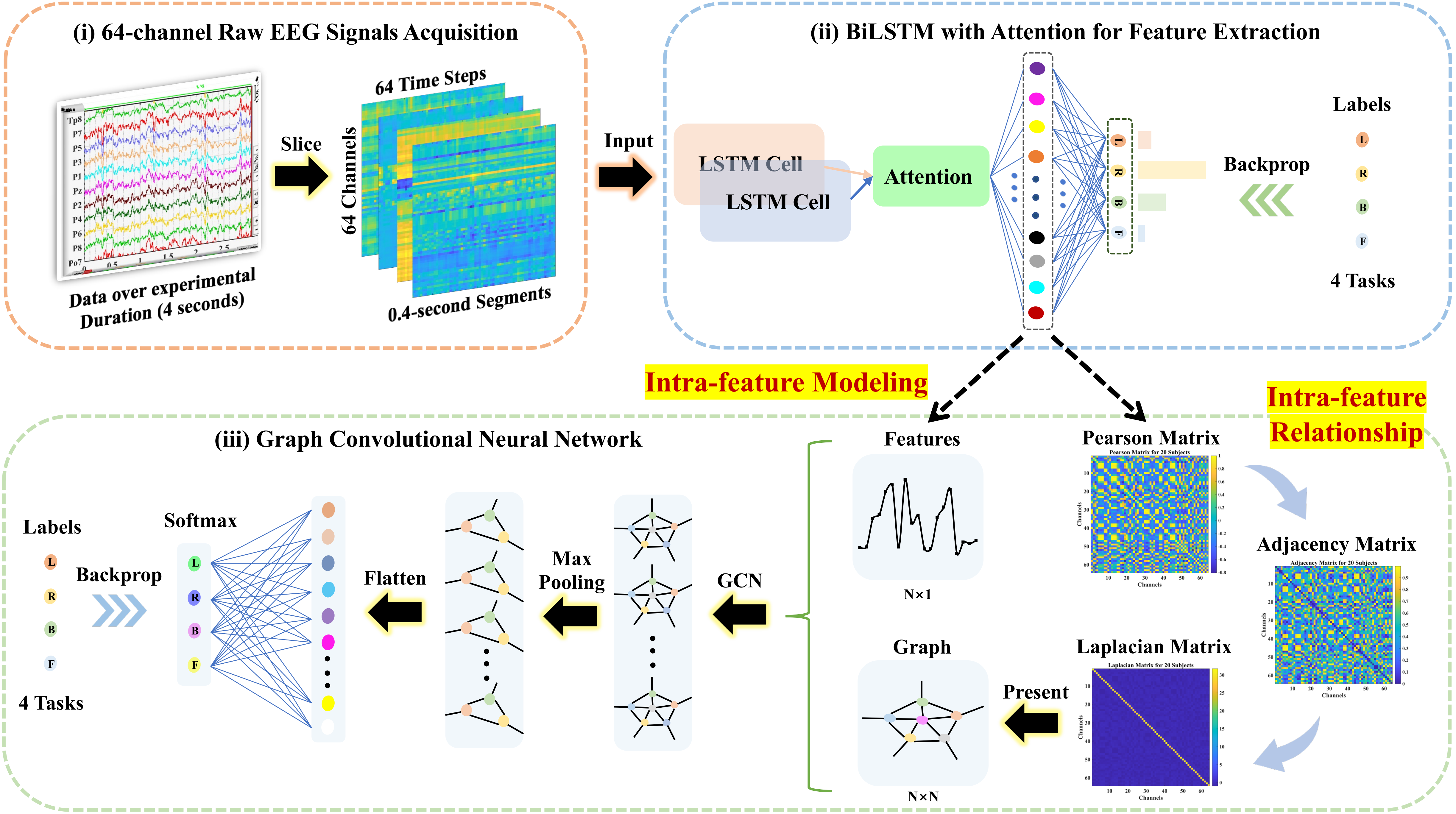
(h) Loss w.r.t. Num. of Extracted Features

Fig. 3: Models and Hyperparameters Comparison w.r.t. the RNN-based Methods for Feature Extraction

Topological Structure of Features

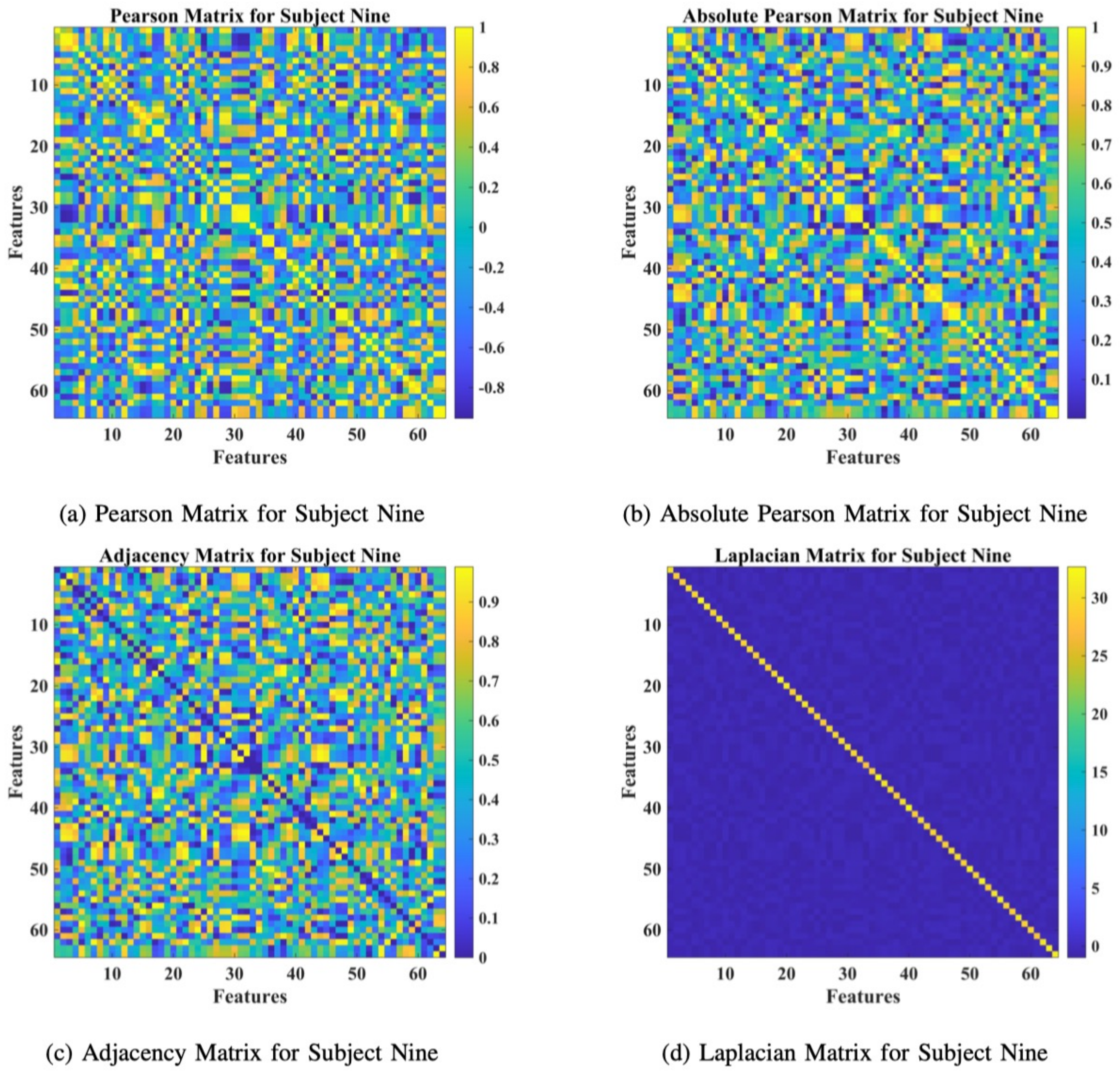


✓ **Deep Feature Mining → Intra-feature Relationship → Intra-feature Modeling**



(i) 64-channel Raw EEG Signals Acquisition
Tp8
P7
P5
P3
P1
Pz
P2
P4
P6
P8
Po7
Data over experimental Duration (4 seconds)
Slice
64 Time Steps
64 Channels
0.4-second Segments
Input
(ii) BiLSTM with Attention for Feature Extraction
LSTM Cell
LSTM Cell
Attention
L
R
B
F
Backprop
Labels
4 Tasks
Intra-feature Modeling
Intra-feature Relationship
(iii) Graph Convolutional Neural Network
Features
N×1
Pearson Matrix
Pearson Matrix for 20 Subjects
Channels
Adjacency Matrix
Adjacency Matrix for 20 Subjects
Graph
N×N
Present
Laplacian Matrix
Laplacian Matrix for 20 Subjects
GCN
Max Pooling
Flatten
Softmax
Backprop
Labels
4 Tasks

Topological Structure of Features



(a) Pearson Matrix for Subject Nine

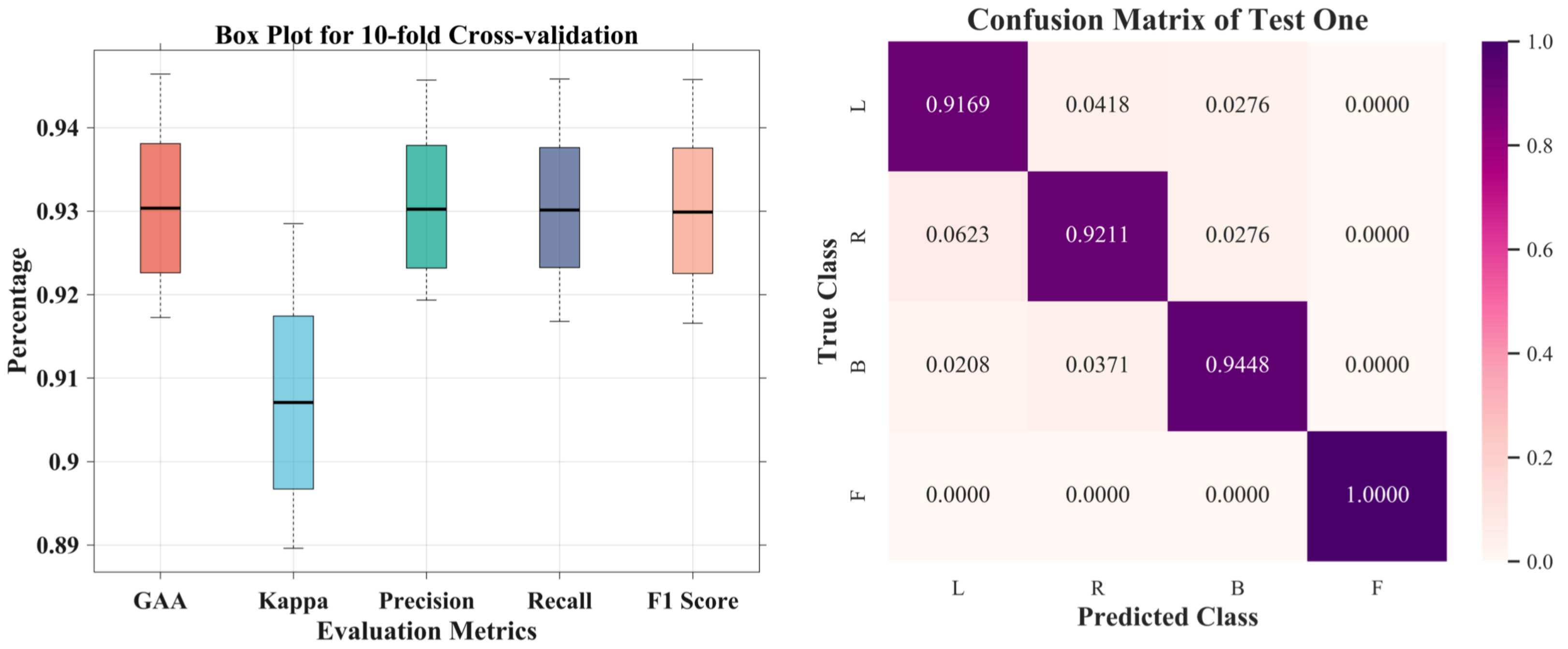
(b) Absolute Pearson Matrix for Subject Nine

(c) Adjacency Matrix for Subject Nine

(d) Laplacian Matrix for Subject Nine

Fig. 4: The Pearson, Absolute Pearson, Adjacency, and Laplacian Matrices for Subject Nine.

Experimental Results - Groupwise Prediction



(a) Box Plot for 10-fold cross validation

(b) Confusion Matrix of Test One

Fig. 5: Box plot and confusion matrix for 10-fold cross validation.

Note:
(1) Box Plot (Maximum Score, Upper Quartile, Median, Lower Quartile, and Minimum Score)
(2) Confusion Matrix: TP, TN, FP, and FN

Experimental Results - Subject-Specific Adaptation

TABLE II: Subject-level Evaluation

No. of Subject	GAA	Kappa	Precision	Recall	F1 Score
1	94.05%	92.06%	94.20%	94.32%	94.16%
2	96.43%	95.19%	96.06%	96.06%	96.06%
3	97.62%	96.79%	97.33%	97.08%	97.18%
4	90.48%	87.34%	91.30%	91.11%	90.42%
5	95.24%	93.61%	95.96%	95.06%	95.38%
6	94.05%	92.02%	93.40%	94.96%	93.66%
7	98.81%	98.40%	98.81%	99.07%	98.92%
8	95.24%	93.60%	95.39%	95.04%	95.19%
9	98.81%	98.39%	99.11%	98.68%	98.87%
10	94.05%	91.98%	93.39%	94.70%	93.61%
Average	**95.48%**	**93.94%**	**95.50%**	**95.61%**	**95.35%**

TABLE III: Current studies comparison on subject-level prediction

Related Work	Max. GAA	Approach	Database
Ortiz-Echeverri *et al.* (2019)	94.66%	Sorted-fast ICA-CWT + CNNs	BCI Competition IV-a Dataset
Sadiq *et al.* (2019)	95.20%	EWT + LS-SVM	
Taran *et al.* (2018)	96.89%	TQWT + LS-SVM	
Zhang *et al.* (2019)	83.00%	CNNs-LSTM	BCI Competition IV-2a Dataset
Ji *et al.* (2019)	95.10%	SVM	
Amin *et al.* (2019)	95.40%	MCNNs	
Dose *et al.* (2018)	68.51%	CNNs	Physionet Database
Hou *et al.* (2019)	96.00%	ESI + CNNs	
This work	**98.81%**	**Attention-based BiLSTM-GCN**	

Takeaways and Future Work

✓ Spatial-Temporal Analysis

(1) Converge to both **Subject-level and Groupwise Predictions** and handle **Individual Variability**.

(2) The 0.4-s sample size **Time-Resolved Solution** toward fast response.

✓ Deep Feature Mining

(1) ↑ **Highest Accuracy**

(2) Advance **Clinical Translation** of EEG-based BCI technology to meet diverse demands, such as those of paralyzed patients.

✓ Future Work

Long-range Dependencies among intra-subject or inter-subject EEG signals can be modeled via **Non-local Modeling**, **Self-attention Mechanism**, **Transformer**, and **AI foundation Models**.

Thank you!

Any question?

Acknowledgment: Dr. Zhijian Hou and Dr. Dingquan Li.