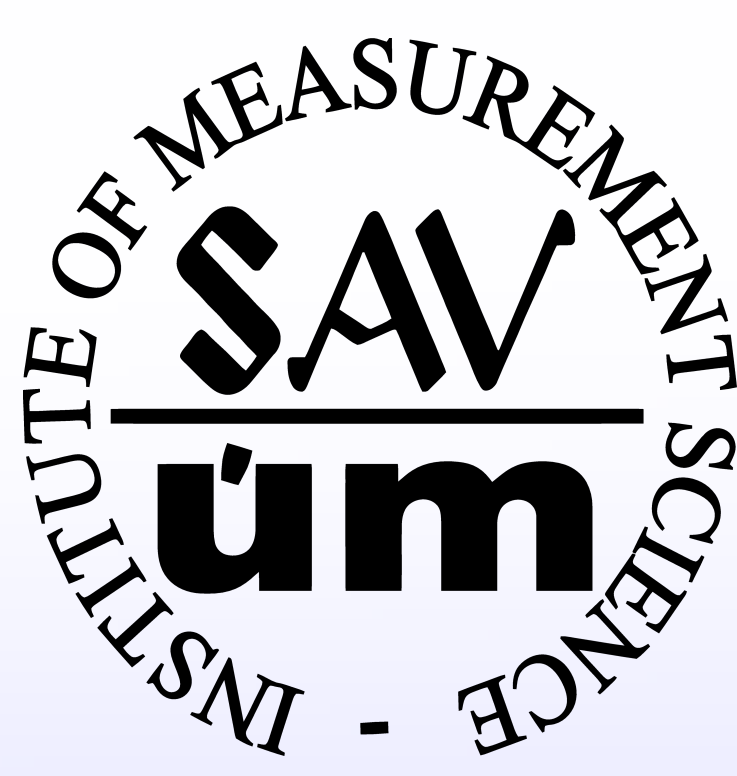


Interpretability of CNN-Based EEG Decoding During Mirror Box Training: A Pilot Study using CWT Spectrograms and Explainable AI

Adrián Hodúr, Roman Rosipal, Nina Evetović, Zuzana Rošťáková, Jozef Jakubík

adrian.hodur@savba.sk Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia



Introduction

Deep learning models achieve high classification accuracy in EEG decoding, but their decision-making process is often difficult to interpret. In neurorehabilitation, it is important to verify that the model relies on physiologically meaningful EEG features rather than dataset-specific patterns. Popular architectures such as EEGNet [1] operate directly on raw EEG signals and provide excellent classification performance. In this work, we investigate whether a CNN [2] based on Continuous Wavelet Transform (CWT) spectrograms [3] can achieve comparable performance while offering improved interpretability through Explainable AI (XAI) [4],[5]. By linking the model's decisions to established sensorimotor rhythms, we assess whether the learned features are neurophysiologically meaningful.

Data collection

- **Subject:** 1 participant (EEG - electrode C3)
- **Mirror box training:** 3 sessions
- **Classes:**
 - **Rest:** Looking at a black cross
 - **Movement:** 4 unique left-hand movements
- **Validation:** 4-fold cross-validation

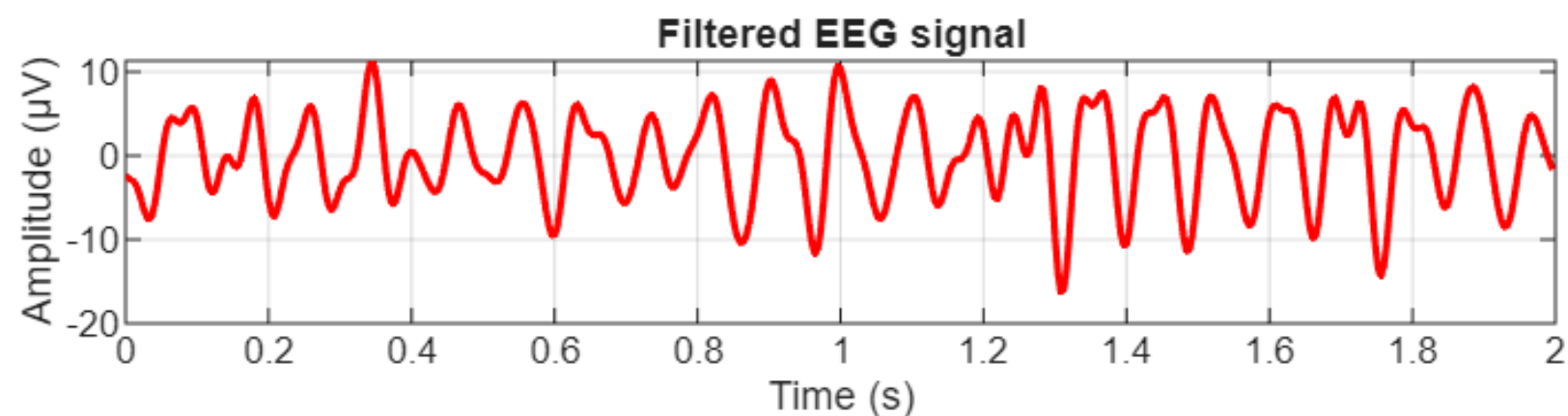
Preprocessing steps

- **Initial removal:** First 2s of each task period were excluded
- **Epoching:** 2s epochs, 80% overlap
- **Artifacts:** Contaminated epochs removed
- **Final dataset:** 2107 data inputs

Data inputs

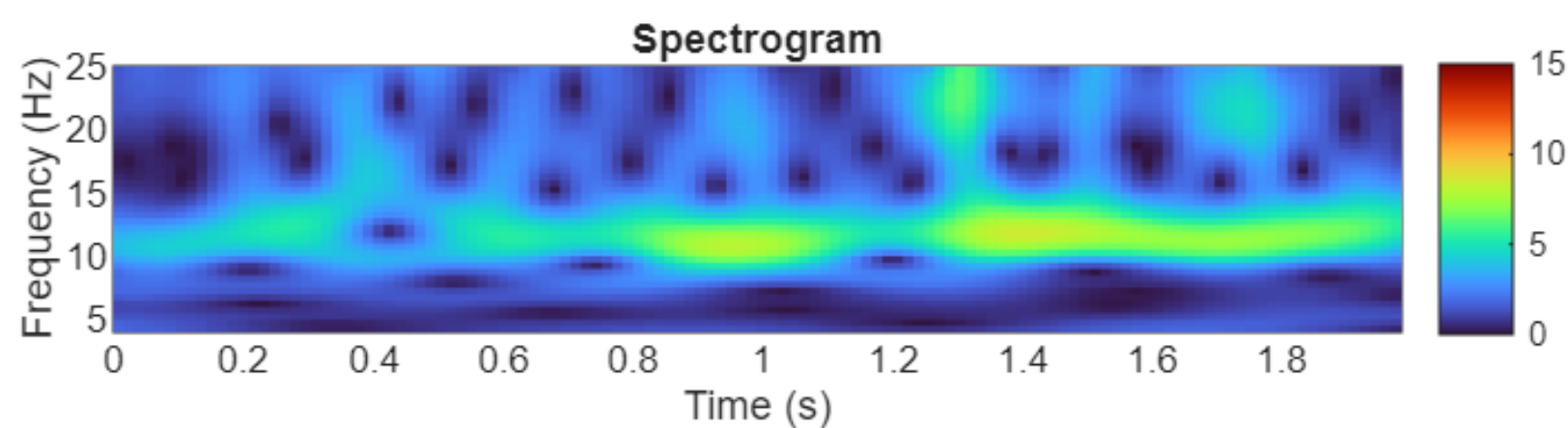
Filtered EEG signal

- **Band-pass filter:** 4-25 Hz
- **CNN input:** 1×500 vector



Spectrogram

- Continuous Wavelet Transform (CWT)
- **Frequency range:** 4-25 Hz
- **Frequency resolution:** 0.5 Hz
- **CNN input:** 43×125 matrix



Neurophysiological Validation

Independent validation of XAI findings

- IRASA: oscillatory activity estimation [6]
- PARAFAC: spatial/spectral decomposition [7]

References & Funding

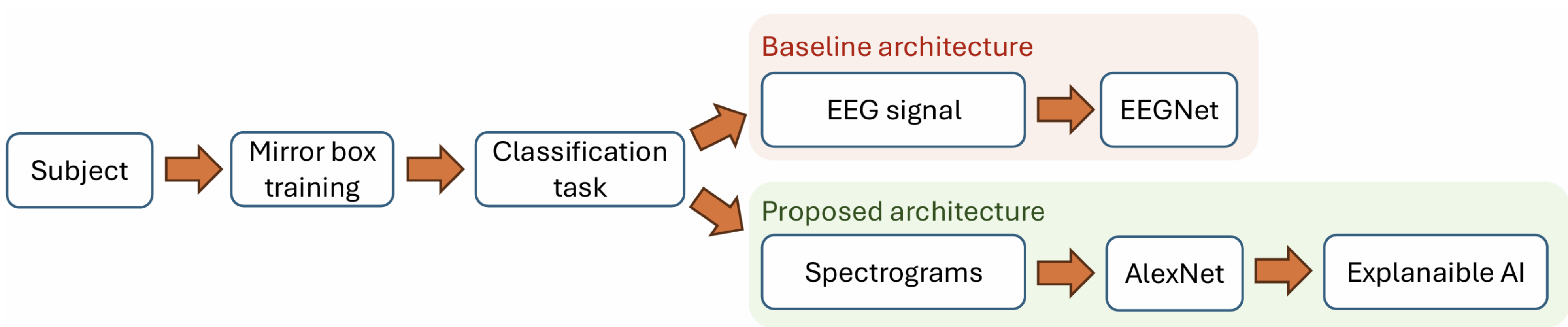
References

- [1] Lawhern et al., EEGNet: A Compact Convolutional Neural Network for EEG-Based Brain-Computer Interfaces, J. Neural Eng., 2018.
- [2] Alom et al., The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches, 2018.
- [3] Gao & Yan, Wavelets, Springer, 2011.
- [4] Selvaraju et al., Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization, ICCV, 2017.
- [5] Zeiler & Fergus, Visualizing and Understanding Convolutional Networks, ECCV, 2014.
- [6] Wen & Liu, Separating Fractal and Oscillatory Components in the Power Spectrum of Neurophysiological Signal, Brain Top., 2016.
- [7] Harshman & Lundy, PARAFAC: Parallel Factor Analysis, Comput. Stat. Data Anal., 1994.

Funding

This research was supported by the EU NextGenerationEU initiative through the Recovery and Resilience Plan for Slovakia under project No. 09I03-03-V04-00205, project No. 09I03-03-V04-00443, HORIZON-MSCA 2022-DN, 101118964-DONUT project, VEGA 2/0094/26, APVV-21-0105, APVV-25-0408.

Comparison of two approaches



Comparison of two CNN architectures

EEGNet Block	Component	Parameters
Convolution 1	Filters (size, number, stride, padding)	1x125, 16, 1x1, same
	Normalization/Activation	Batch/ELU
	Average pooling (size, stride, padding)	1x4, 1x4, same
	Dropout (rate)	0,25
Convolution 2	Filters (size, number, stride, padding)	1x32, 32, 1x1, same
	Normalization/Activation	Batch/ELU
	Average pooling (size, stride, padding)	1x8, 1x8, same
	Dropout (rate)	0,25
Final	Fully connected layer	2 (output)
	Soft max layer	

Left table-EEGNet. Right table-modified AlexNet. Both CNNs have a similar number of parameters.

AlexNet Block	Component	Parameters
Convolution 1	Filters (size, number, stride, padding)	11x11, 16, 1x1, same
	Normalization/Activation	Batch/ELU
	Max pooling (size, stride, padding)	1x5, 1x5, same
Convolution 2	Filters (size, number, stride, padding)	5x5, 16, 1x1, same
	Normalization/Activation	Batch/ELU
	Max pooling (size, stride, padding)	2x2, 2x2, same
Convolution 3	Filters (size, number, stride, padding)	3x3, 16, 1x1, same
	Normalization/Activation	Batch/ELU
	Max pooling (size, stride, padding)	2x2, 2x2, same
Final	Fully connected layer	16 (output)
	Activation	ELU
	Dropout (rate)	0.4
	Fully connected	2 (output)
	Soft max layer	

Results

1. Neurophysiological validation

IRASA: $\mu/\alpha/\beta$ -band event-related desynchronization ERD during movement (Fig.1)

PARAFAC: clear μ -peak (~ 11 Hz) around C3 electrode (Fig.2)

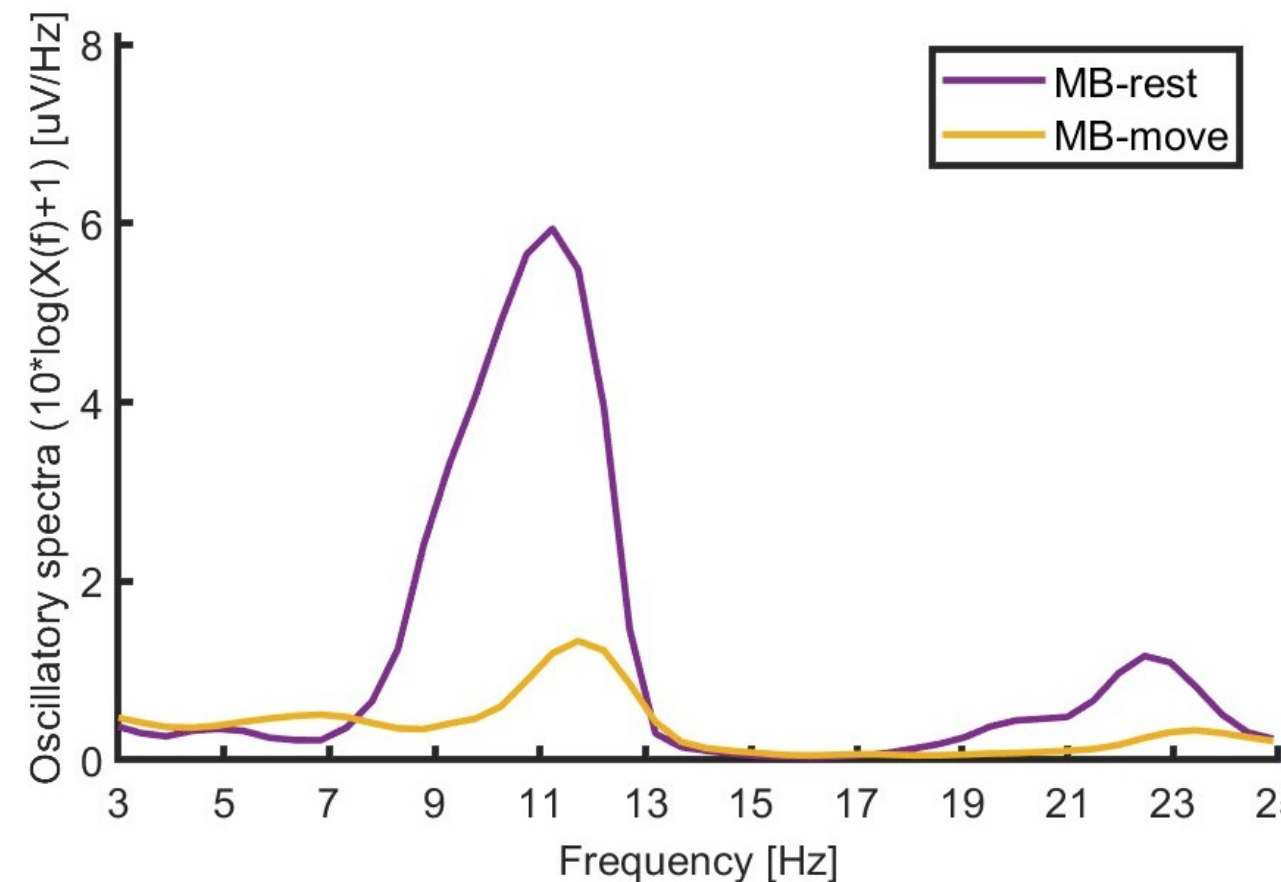


Figure 1: IRASA spectra at C3. Rest (dark red) vs movement (orange).

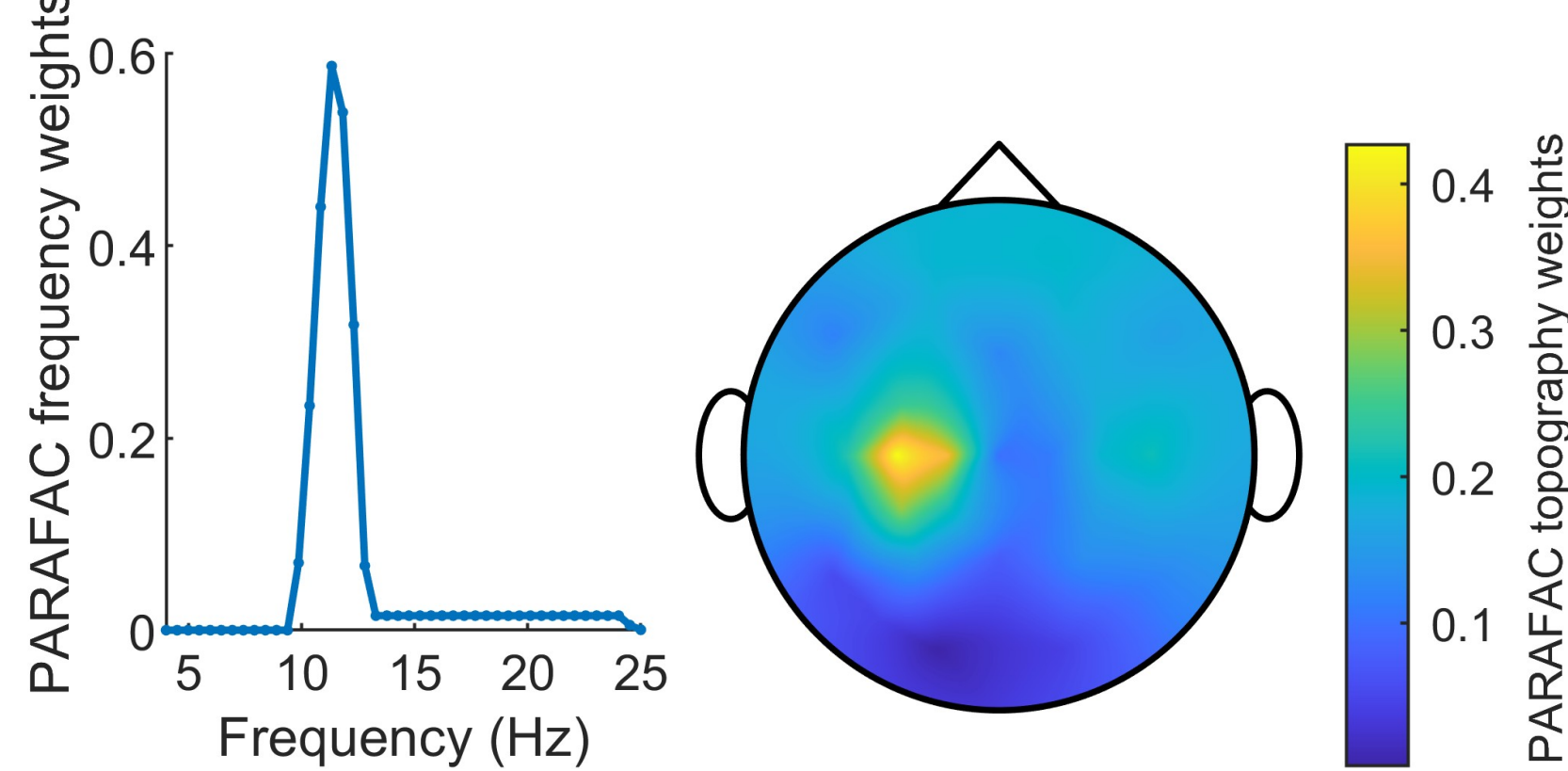


Figure 2: Left: Extracted μ component (Frequency profile) Right: Scalp weights (C3 electrode focus).

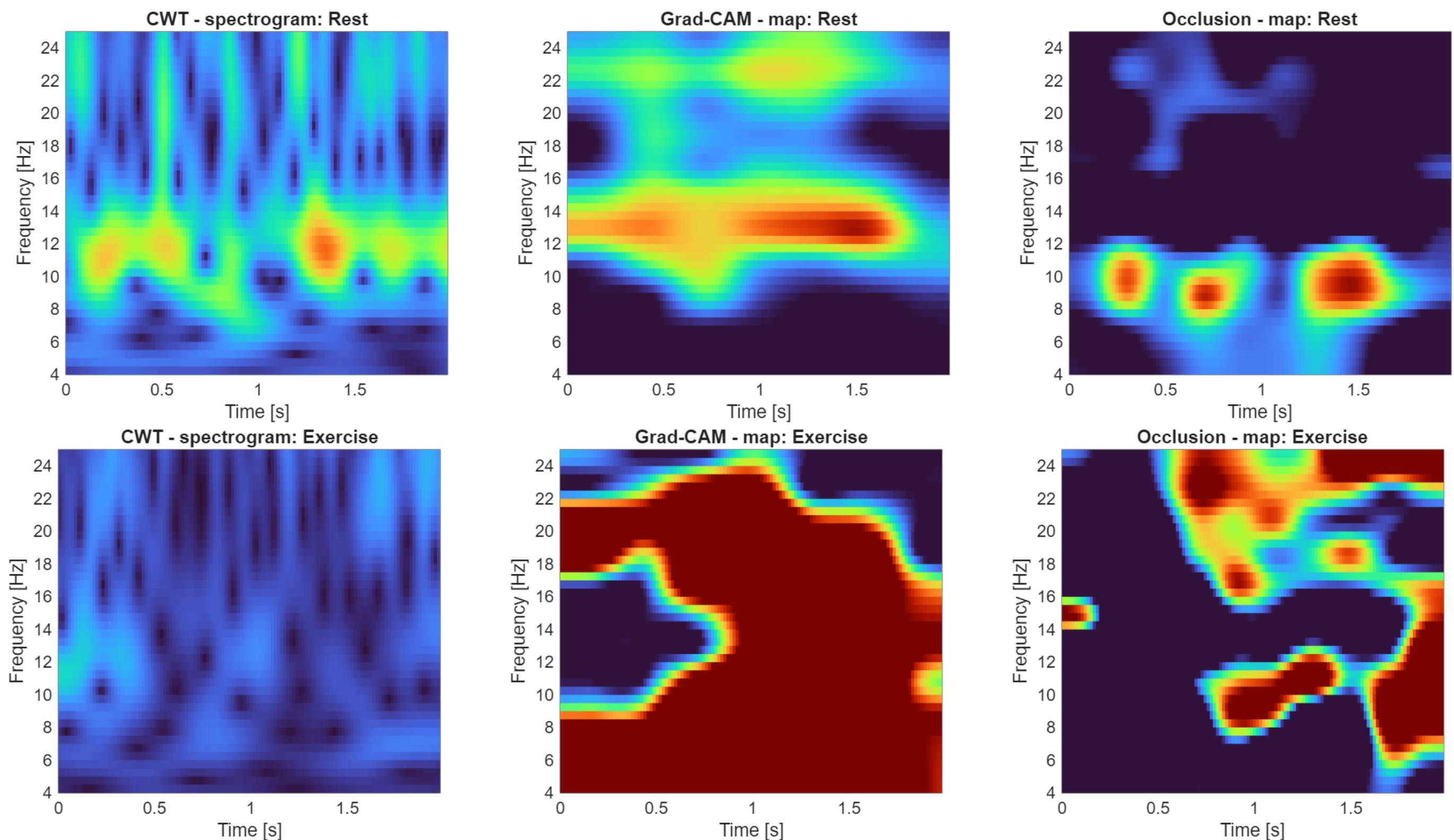
2. Classification performance

The proposed CWT-based CNN achieved classification performance comparable to EEGNet.

CNN Model	Input	Average accuracy
EEGNet	Filtered EEG signal	0,90
AlexNet	CWT spectrogram	0,89

3. Explainability

XAI (Grad-CAM and Occlusion Sensitivity) highlighted $\mu/\alpha/\beta$ frequency bands associated with movement-related ERD and enhanced $\mu/\alpha/\beta$ activity during rest, demonstrating that the CNN relied on physiologically meaningful EEG patterns. **Left panel:** Correctly classified spectrograms, Rest (top), Movement (bottom). **Middle panel:** Grad-CAM map. **Right panel:** Occlusion Sensitivity map.



Conclusion

- XAI highlighted physiologically meaningful EEG patterns, including $\mu/\alpha/\beta$ ERD during movement and increased $\mu/\alpha/\beta$ activity during rest.
- XAI maps showed consistent patterns across validation samples, indicating stable model behavior.
- XAI explained misclassifications by revealing movement-like patterns during rest and pause-like EEG patterns during movement.
- IRASA and PARAFAC confirmed the extracted neurophysiological patterns.

CWT-based CNN achieves comparable accuracy to EEGNet, with improved explainability.