

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

Adrián Hodúr¹, Roman Rosipal¹, Nina Evetović¹, Zuzana Rošćáková¹, Jozef Jakubík¹

¹Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia

E-mail: adrian.hodur@savba.sk

ABSTRACT: Neural networks are increasingly applied in brain-computer interface research; however, their decision-making process is often considered difficult to interpret. This pilot study investigates whether a convolutional neural network can provide not only accurate but also neurophysiologically meaningful classification of electroencephalogram signals recorded during mirror box training. Two conditions, rest and left hand movement, were analyzed with a focus on interpretability. EEG signals from contralateral electrode C3 were transformed into time-frequency representations using the Continuous Wavelet Transform, and the resulting spectrograms were used as input to the convolutional neural network (CNN). To interpret the model's decisions, Gradient-weighted Class Activation Mapping and other methods were applied, enabling identification of discriminative time-frequency regions. By relating these patterns to established neuroscientific knowledge, we assess whether the model relies on physiologically meaningful features, providing insight into brain dynamics during mirror-induced motor activity and supporting interpretable BCI applications.

INTRODUCTION

Neural networks (NNs) have become a highly popular tool in artificial intelligence due to their ability to model nonlinear relationships and detect complex patterns in classification tasks. However, despite their strong performance, their internal parameters (optimized during the training process) are often difficult to interpret, particularly in deep and highly complex architectures. This limited interpretability presents a challenge in applications where understanding the decision-making process is essential.

Electroencephalogram (EEG) signals, which measure the electrical activity of the brain, are inherently complex and difficult to interpret. The combination of complex NN models and complex neurophysiological signals motivates our effort to develop a NN-based classification model with an emphasis on interpretability. In this study, we classify two labeled conditions recorded during mirror box training: **(1) rest**, during which the patient observes a fixation cross, and **(2) exercise**, during which the patient performs one of four movements with left hand.

To enhance the time-frequency interpretability of EEG

signals, we applied the Continuous Wavelet Transform (CWT) [1], which enables clear detection of temporal changes in frequency bands. This is particularly important because movement or motor imagery are associated with event-related desynchronization (ERD) and event-related synchronization (ERS). ERD corresponds to a decrease in power spectral density in the μ and β sensorimotor rhythms, whereas ERS represents an increase in these bands [2]. During movement, decreased power in the μ and β bands is typically observed, while rest is characterized by stronger μ activity.

For classification, CWT spectrograms were computed from EEG and these spectrograms were represented as images and used as input to a convolutional neural network (CNN) [3] adapted to our dataset and classification task.

To evaluate model interpretability, we employed explainable AI (XAI) methods, primarily Gradient-weighted Class Activation Mapping (Grad-CAM) [4], and for comparison also Local Interpretable Model-agnostic Explanations (LIME) [5] and Occlusion Sensitivity (OS) [6]. These methods highlight input regions that contribute most to classification decisions and were complemented by frequency-based analyses to assess whether the model focuses on neurophysiologically meaningful patterns.

Although the individual components of the proposed approach (CWT, CNN, and XAI methods) are well established, their combination is applied here within a neurophysiologically motivated framework with an emphasis on interpretability. This pilot study, conducted on a single subject, aims not only to distinguish rest and movement conditions but also to determine whether the model relies on physiologically meaningful features, such as frequency-specific synchronization and desynchronization in sensorimotor rhythms. By linking time-frequency representations with model explanations, the approach enables a transparent evaluation of consistency between deep learning decisions and established EEG knowledge in the context of mirror box training. Additionally, the proposed method was compared with the EEGNet architecture [7], which operates directly on raw EEG signals.

MATERIALS AND METHODS

Recording Setup:

EEG signals were recorded using a wireless g.Nautilus

PRO FLEXIBLE system (g.tec Medical Engineering GmbH, Austria), an FDA-cleared device designed for both clinical and research applications. The system was equipped with 28 Ag/AgCl wet electrodes arranged according to the international 10-20 system. The electrodes were integrated into a multi-channel EEG cap connected to a portable amplifier unit, enabling high-quality signal acquisition during experimental procedures. Signals were sampled at 250 Hz. During acquisition, A1 served as the reference electrode and AFz as the ground. For offline processing, the EEG data were re-referenced using a common average reference. Specifically, for each electrode, the mean signal across all recorded electrodes was computed and subtracted from the signal at that electrode, thereby reducing reference-related bias in subsequent analyses.

Raw signal preprocessing:

Raw EEG recordings were first converted to the European Data Format (EDF) and subsequently preprocessed using the BrainVision Analyzer software (version 2.3.03). The preprocessing workflow consisted of several steps aimed at improving signal quality and minimizing artifact contamination. We used a semi-automatic threshold-based artifact detection procedure implemented to identify segments contaminated by high-amplitude or non-physiological activity. All automatically marked segments were visually inspected to ensure accurate classification. Electromyographic (EMG) artifacts were also identified prior to exporting the cleaned dataset for further processing.

Dataset:

EEG data were recorded from a single participant during 3 mirror box (MB) training sessions. Continuous EEG signals were acquired throughout both resting and movement conditions. During the movement periods, each task with left hand was repeated multiple times. Each MB session included four distinct exercises with self-paced pauses between repetitions:

- **Exercise 1: Elbow-Supported Hand Lift** - The participant lifted the hand while keeping the forearm in contact with the table.
- **Exercise 2: Palm Rotation** - The participant rotated the palm upward and downward while the forearm remained on the table.
- **Exercise 3: Ball Roll** - The participant rolled a ball on the table using the palm, lifting the fingers as the ball reached the base of the palm.
- **Exercise 4: Bottle Grasp and Release** - The participant grasped a bottle, released it, and returned the hand to the table.

All movements were performed with the left hand while observing its mirror reflection, creating a visual illusion of movement. This mirror-induced illusion is known to modulate activity in the contralateral sensorimotor cortex [8]. For this reason, EEG signals from electrode **C3** were selected for further analysis.

EEG Preprocessing:

The preprocessing pipeline consisted of five steps: **i)**

Epoching: Continuous EEG signals were segmented into 2-second epochs using a sliding-window approach with 80% overlap between consecutive segments. **ii) Frequency selection:** Only frequency components within the 4-25 Hz range (0.5 ~ Hz resolution) were retained for further analysis. **iii) Time-frequency transformation:** Each epoch was transformed using the CWT (see next section with CWT). From each resulting spectrogram, 125 temporal samples corresponding to the 2-second interval were extracted, forming the input matrices for the CNN. **iv) Initial segment removal:** The first two seconds of each task period were excluded to avoid contamination from head movements and task transition effects. **v) Artifact handling:** All spectrograms containing artifacts were excluded from further analysis. After preprocessing, 2107 spectrogram images with size 43×125 (frequency $\times$ time) were obtained from movement and rest conditions. For validation, data from one specific movement task and its corresponding post-movement rest period were reserved. The remaining movement and rest data were used for CNN training.

CWT for creating input data matrices:

We generated 2-second-long CWT spectrograms within the 4-25 Hz frequency range using the CWT. The CWT provides a complex time-frequency representation of a signal by projecting it onto scaled and translated versions of a mother wavelet. The analytic Gabor-Morlet wavelet [9] was used as mother wavelet. This wavelet provides near-optimal joint time-frequency localization due to its Gaussian envelope and is well suited for oscillatory, non-stationary signals such as EEG. For further analysis, the magnitude of the CWT representing the local signal energy was used. Because lower frequencies require a longer temporal context for accurate magnitude estimation, an additional 100 time points were temporarily appended before and after each 2-second epoch during the CWT computation. These additional samples were used solely to improve the estimation of low-frequency components at the window boundaries and were discarded after the transformation. Consequently, only the original 2-second long spectrograms were retained for further analysis.

Convolutional Neural Network Architecture:

Because the size of our dataset was limited, we designed a simple CNN with a small number of convolutional layers. The architecture was inspired by the well-known AlexNet model [10]. The network was employed to classify CWT-derived EEG spectrograms. The model consists of an input layer followed by convolutional blocks and fully connected layers. The number of trainable parameters was reduced to mitigate overfitting. Each convolutional layer is followed by batch normalization (BN) and the exponential linear unit (ELU) activation function. The combination of BN and ELU activation has been shown to improve performance in deep CNN models applied to EEG data [11]. The architecture of the convolutional blocks after the input layer is summarized in Table 1.

After the convolutional blocks, the feature maps were

Table 1: Overview of the first three convolutional blocks of the proposed CNN. K - filter size, N - number of filters, S - stride, P - padding

Block	Component	Parameters
Conv1	Filters (K/N/S/P)	11×11, 16, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Max 1×5, 1×5, same
Conv2	Filters (K/N/S/P)	5×5, 16, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Max 2×2, 2×2, same
Conv3	Filters (K/N/S/P)	3×3, 16, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Max 2×2, 2×2, same

flattened and passed to a fully connected layer with 16 units, followed by ELU activation and dropout (rate = 0.4) to further reduce overfitting. The final fully connected layer contained 2 neurons, and a softmax function was applied to produce class probabilities.

For comparison, we also evaluated the proposed approach using an EEGNet-based architecture [7] operating directly on raw EEG signals. The input data were band-pass filtered in the 4-25 Hz range and centered around zero mean. The network was adapted to the input size of 1×500 , corresponding to a 2-second EEG segment sampled at 250 Hz. To ensure a fair comparison, the architecture was adjusted while preserving the core design principles of EEGNet. The detailed configuration of the adapted network is summarized in Table 2.

Table 2: Overview of the blocks of EEGNet. K - filter size, N - number of filters, S - stride, P - padding

Block	Component	Parameters
Conv1	Filters (K/N/S/P)	1×125, 16, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Avg, 1×4, 1×4 same
	Dropout (rate)	0,25
Conv2	Filters (K/N/S/P)	1×32, 32, 1×1, same
	Normalization / Activation	BN / ELU
	Pooling (K/S/P)	Avg, 1×8, 1×8, same
	Dropout (rate)	0,25
Final	Full. connected (output)	2
	Soft max classification	

XAI:

To interpret the decisions of the CNN, Grad-CAM [4] was employed. Grad-CAM provides a visual explanation by highlighting the regions of the input that contribute most strongly to a specific class prediction. For comparison, we also used (OS) [6] and LIME [5] methods.

Training options:

For data preprocessing and CNN training, we used MATLAB R2025b. The network was trained using the cross-entropy loss function with class weight correction to account for the imbalance between rest and movement data. The following training options were applied for effective learning. Training was performed using the Adam optimizer [12] with a mini-batch size of 32 and an initial learning rate of 3×10^{-4} . A piecewise learning-rate schedule with periodic decay was employed to ensure

stable convergence during later epochs. L2 regularization with $\lambda = 1 \times 10^{-4}$ was applied to reduce overfitting. Early stopping was implemented based on validation performance, with a patience of 8 and validation checks every 50 iterations. These hyperparameters were selected based on preliminary experiments, which indicated that this configuration provided a good balance between convergence speed, model stability, and generalization performance.

Additional Supplementary Methods:

To assess the neurophysiological validity of the CNN-based analysis we applied traditional EEG analysis methods, including the IRASA method [13] and PARAFAC decomposition [14], to estimate the spatial and spectral characteristics of the EEG signal.

RESULTS

EEG spatial and frequency analysis:

IRASA was applied to separate oscillatory and aperiodic components of the amplitude spectrum at electrode C3 during resting and movement conditions. Baseline recordings included one minute of eyes-closed (EC) and one minute of eyes-open (EO) rest. As shown in Figure 1 (left), a broader oscillatory peak within 8-12 Hz was present at C3 in both conditions. To assess visual alpha activity, we additionally analyzed electrode O1 (Figure 1, right). A pronounced peak around 9 Hz was observed, particularly during the eyes-closed condition, corresponding to occipital α rhythm. Given that occipital alpha can spatially spread toward parietal and central regions due to large-scale thalamo-cortical synchronization [15], this may partly explain the broader and less sharply defined peak at C3 during EC, reflecting overlap between sensorimotor μ and occipital α activity.

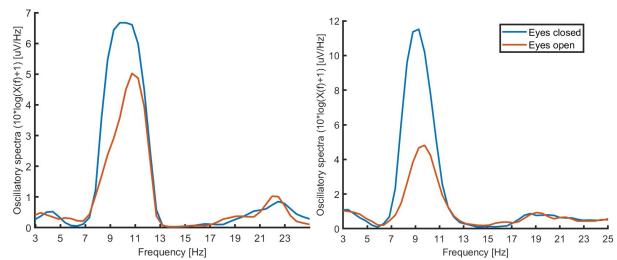


Figure 1: Oscillatory part of the amplitude spectrum as obtained by IRASA method/decomposition. **Left:** Electrode C3. An oscillatory peak is visible between 8 and 12 Hz. EC - blue curve, EO - orange curve. **Right:** Electrode O1. A dominant oscillatory peak is visible around 9 Hz, corresponding to occipital alpha activity. EC - blue curve, EO - orange curve.

During MB training (Figure 2), a clear oscillatory peak around 11 Hz was observed at C3 during rest, corresponding to the μ rhythm. This peak was markedly reduced during left-hand movement, indicating ERD. A comparable desynchronization pattern was also present at O1, reflecting suppression of visual α activity during movement.

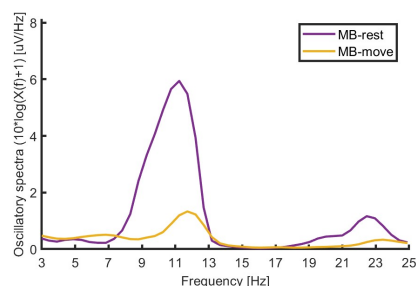


Figure 2: IRASA oscillatory spectra at electrode C3 during MB training. The dark red curve represents the resting/pause condition, while the orange curve corresponds to left-hand movement. A clear reduction of μ amplitude is observed during movement, indicating event-related desynchronization.

PARAFAC analysis further confirmed the presence of a task-related μ component. As illustrated in Figure 3, the extracted component showed a peak frequency of approximately 11.2 Hz and was localized over the left sensorimotor cortex (C3 region).

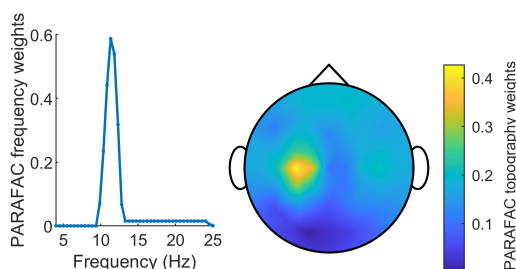


Figure 3: **Left:** Frequency profile of the extracted PARAFAC μ component with a peak at approximately 11.2 Hz. **Right:** Corresponding scalp spatial weights showing predominant activation over the left sensorimotor cortex (C3 region).

CNN, XAI:

After data preprocessing, the CNN model was trained. The training process is illustrated in Figure 4. The blue curve represents the training accuracy over time, while the orange curve indicates the validation accuracy (counted every 50th iteration). As described previously, one of the four exercises, together with its immediately following rest period, was always used as the validation dataset. Therefore, the training procedure was repeated four times, each time using a different exercise-rest pair for validation.

Across the validation runs, the average classification accuracy (rest vs exercise) was approximately 89% for the proposed CNN (AlexNet-inspired) with CWT-spectrogram input and 90% for the EEGNet-based model with raw EEG input. These results indicate comparable performance between the two approaches.

In the next step, the Grad-CAM, LIME and OS method was applied to the trained model using samples from the validation dataset. As an example, two spectrograms were selected. The first was selected from exercise 1 and the second from the rest period after it. In Figure 5, the left panel shows a 2-second spectrogram segment

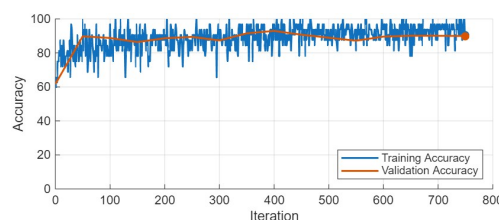


Figure 4: CNN model training process showing training and validation accuracy.

recorded during the exercise 1 and rest period. This samples were correctly classified as the exercise/rest class by the trained AlexNet inspired CNN model. The middle panel shows the Grad-CAM map, and the right panel shows the OS map. Both highlight the time-frequency regions that most influenced the model's decision.

Additional Grad-CAM, LIME and OS maps were generated for multiple spectrograms from all validation datasets. The highlighted regions were often consistent across different samples, with only minor variations. Similar patterns were observed for each validation fold, further supporting the stability of the learned features.

DISCUSSION

The proposed approach achieved classification performance comparable to an EEGNet-based model operating on raw EEG signals. However, a key advantage of the presented method lies in its interpretability. By combining time-frequency representations with XAI techniques, it is possible to directly assess whether the model relies on neurophysiologically meaningful features, providing additional insight beyond classification accuracy.

XAI visualizations provide insight into which spectral regions were most relevant for the CNN classification. In Figure 5 bottom-left panel, a spectrogram segment from the beginning of the exercise period was correctly classified as movement. Since this segment was extracted from the onset of the exercise block, it is reasonable to assume that the subject was actively performing the movement during this interval. Compared to the magnitude observed during the rest condition (top-left panel in Figure 5), this spectrogram shows a clear reduction in μ and β band activity after approximately 0.5s. This pattern is consistent with ERD in the sensorimotor rhythm during motor execution. The Grad-CAM map (Figure 5, bottom-middle) highlights broad regions within these frequency bands, indicating that suppression of μ and β activity played a substantial role in the movement classification. Moreover, occipital α activity is suppressed during the movement, reflecting visual processing of the mirrored hand, and likely contributed alongside μ and β ERD to the classification. These findings are consistent with the IRASA results, which showed reduced oscillatory amplitude in both the μ and occipital α bands during movement. Similarly, the PARAFAC decomposition identified a μ component with a peak frequency of 11.2 Hz, whose amplitude decreased during motor activ-

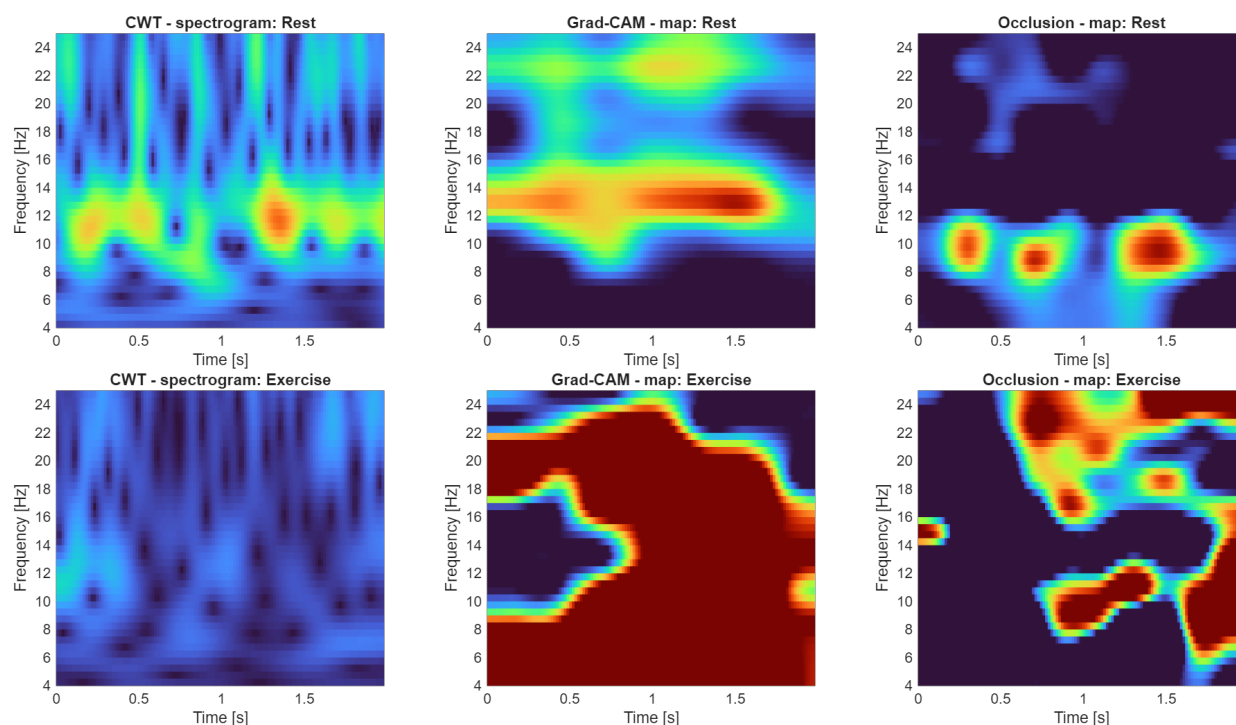


Figure 5: Correctly classified spectrograms. **Left panel:** Original input spectrogram labeled as Rest (top), Exercise (bottom) (color scale (CS): 0-15). **Middle panel:** Grad-CAM activation map highlighting the most influential regions (CS: 0-0.3). **Right panel:** Occlusion Sensitivity (OS) map showing how masking different time-frequency regions affects the model's output (CS: 0-0.01).

ity. Together, these complementary analyses support the physiological plausibility of the CNN decision process. In addition, lower activity in the θ band can be observed during movement compared to rest, and corresponding regions are partially highlighted by the Grad-CAM map. This may reflect reduced low-frequency activity associated with relaxed states or increased task engagement during movement, although this effect is less pronounced than the changes observed in the μ and β bands.

In contrast, Figure 5 (top-left panel) shows a rest-period spectrogram characterized by increased magnitude in the 9-15 Hz range. Based on the IRASA and PARAFAC findings, this activity likely reflects a combination of occipital (visual) α and sensorimotor μ rhythms. During the rest condition, the subject fixated on a visual cross and remained in a relaxed state, which may have contributed to enhanced posterior alpha activity. The corresponding XAI-maps (Figure 5, top-middle panel and top-right panel) emphasizes this frequency range, suggesting that elevated μ and α rhythms was a key discriminative feature for the rest class, with a slight increase in β activity also contributing to the classification.

Despite generally good classification performance, several misclassifications were observed. Analysis of incorrectly classified samples revealed that some rest-period spectrograms with weak sensorimotor μ , visual α , and β activity were classified as exercise. XAI maps visualizations indicated that regions with reduced magnitude in these frequency bands contributed to the incorrect predictions, suggesting that the model relied mainly on changes within the sensorimotor frequency range. One possible

explanation is variability in the subject's behavioral state during rest. Although instructed to fixate on a cross, fluctuations in attention or cognitive engagement may have modulated occipital α activity. In particular, α power over occipital regions has been shown to decrease during externally directed visual attention [16].

Another factor may be the structure of the exercise task itself. The subject performed repetitive movements separated by short pauses, and spectrogram segments extracted during these pauses could exhibit increased μ or α magnitude resembling rest-like patterns. An example is shown in Figure 6, where a spectrogram from the exercise period displays elevated magnitude between 8 and 15 Hz. This likely corresponds to a short pause between repetitions during which partial resynchronization of the sensorimotor rhythm occurred, potentially contributing to incorrect classifications.

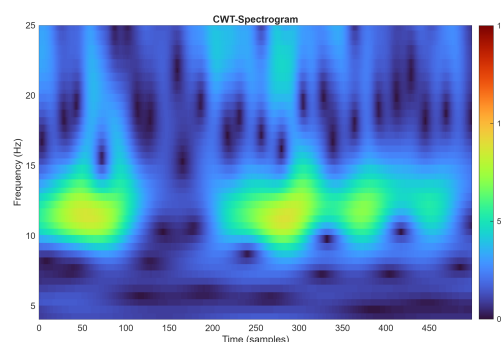


Figure 6: 2-second long spectrogram from exercise period.

CONCLUSION

In this study, we classified CWT-based spectrograms derived from rest and exercise tasks during MB training using a CNN. The model achieved good classification accuracy across multiple validation runs, demonstrating stable performance in distinguishing between movement and rest conditions. Notably, the obtained results were comparable to those achieved by an EEGNet-based model operating directly on raw EEG signals.

Importantly, the proposed approach provides an additional advantage in terms of interpretability. The use of time-frequency representations combined with XAI enables direct assessment of whether the model relies on physiologically meaningful features. The visualizations indicated that the model's decisions were largely consistent with established neurophysiological principles. Movement-related segments were primarily associated with reduced power in the visual α frequency band and with μ and β frequency bands, corresponding to ERD, while rest segments were characterized by increased activity in the α range. These findings were further supported by IRASA and PARAFAC analyses, which independently confirmed modulation of the sensorimotor and occipital rhythms.

Future work will focus on integrating EMG recordings to precisely determine the timing of motor execution. Accurate identification of movement onset and offset would allow removal of spectrogram segments containing transitional pauses between repetitions, improving classification performance and interpretability. In addition, future studies will explore CNN architectures with interpretable kernels, enabling analysis of learned filters in relation to neurophysiological patterns.

Furthermore, combining EMG timing information with EEG analysis would enable a more detailed investigation of ERD dynamics relative to movement onset, providing deeper insight into the relationship between EEG spectral characteristics and CNN-based classification.

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 (Z.R.) and project No. 09I03-03-V04-00443 (R.R. and A.H.). Additional funding was provided by the European Commission through the HORIZON-MSCA 2022-DN, 101118964-DONUT project. This work was also supported by the Slovak Research and Development Agency under grant APVV-21-0105 (TInVR) and VEGA 2/0094/26.

REFERENCES

[1] Gao RX, Yan R. Wavelets. Springer US: Boston, MA (2011).

- [2] Pfurtscheller G, Lopes Da Silva F. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*. 1999;110(11):1842–1857.
- [3] O'Shea K, Nash R. An introduction to convolutional neural networks. Dec. 2, 2015. arXiv: 1511.08458 [cs]. Accessed: Feb. 26, 2026. [Online]. Available: <http://arxiv.org/abs/1511.08458>.
- [4] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017, 618–626.
- [5] Ribeiro MT, Singh S, Guestrin C. "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, 1135–1144.
- [6] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV). Springer, 2014, 818–833.
- [7] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*. 2018;15(5):056013.
- [8] Garry MI, Loftus A, Summers JJ. Mirror, mirror on the wall: Viewing a mirror reflection of unilateral hand movements facilitates ipsilateral m1 excitability. *Experimental Brain Research*. 2005;163(1):118–122.
- [9] Russell B, Han J. Jean morlet and the continuous wavelet transform. *CREWES Res. Rep*. 2016;28:115.
- [10] Alom MZ et al. The history began from alexnet: A comprehensive survey on deep learning approaches. 2018. arXiv: 1803.01164 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1803.01164>.
- [11] Schirrneister RT et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*. 2017;38(11):5391–5420.
- [12] Kingma DP, Ba J. Adam: A method for stochastic optimization. Jan. 30, 2017. arXiv: 1412.6980 [cs]. Accessed: Feb. 26, 2026. [Online]. Available: <http://arxiv.org/abs/1412.6980>.
- [13] Wen H, Liu Z. Separating fractal and oscillatory components in the power spectrum of neurophysiological signal. *Brain Topography*. 2016;29(1):13–26.
- [14] Rošťáková Z, Rosipal R, Seifpour S, Trejo LJ. A comparison of non-negative Tucker decomposition and parallel factor analysis for identification and measurement of human EEG rhythms. *Measurement Science Review*. 2020;20(3):126–138.
- [15] Klimesch W. EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews*. 1999;29(2):169–195.
- [16] Magosso E, colleagues. EEG alpha power is modulated by attentional changes during cognitive tasks and virtual reality immersion. *Frontiers in Human Neuroscience*. 2019;13:256.