EEG-Based Interpretable Motor Classification

Sergey Barabanoff, Beñat Froemming-Aldanondo, James Sargsyan, Raiza Soares

Abstract

Motor-imagery EEG decoding is a promising technique for brain–computer interfaces (BCIs) designed for quadriplegic individuals, amputees, and others with motor impairments. Existing assistive technologies, such as sip-and-puff wheelchairs and EMG-based prosthetics, remain limited in flexibility and performance. Recent advances in deep learning have significantly improved the decoding of motor imagery and movement-related neural signals. In this project, we implemented three deep learning architectures to perform multiple classification tasks using the PhysioNet EEG Motor Movement/Imagery Dataset. Our results show that the CNN–LSTM–Attention model achieved the highest performance, with an AUC of approximately 0.725. Additionally, post-hoc interpretability techniques were applied to identify the brain regions most strongly associated with each class, offering potential benefits for users with cross-modal motor-area remapping and for optimizing electrode placement in future BCI systems.

Introduction

Brain-computer interfaces have widespread clinical significance. For example, quadriplegics suffer from at least partial paralysis of all four limbs. Since they are unable to move on their own, they rely heavily on caretakers for nearly everything they do. Current devices to assist these individuals rely on things like speech or breath as inputs, but these are not ideal controls. Not all quadriplegics are able to speak, and having to carefully control one's breaths to use a device seems rather tiring. These approaches work for assistance devices, but they are not intuitive, not very ergonomic, and are very limited in their capabilities. For example, a mechanical prosthetic is somewhat useful, but it lacks the dexterity of a real hand because it is not possible to move it at will. Brain-computer interfaces (BCIs) are a more promising solution for these individuals because nearly anybody is able to use them, especially those with motor impairments. One specific example of BCI involves decoding motor imagery, that is, the thought of moving a limb. This sort of approach could restore disabled individuals' ability to act by thinking about performing an action, very similar to how abled individuals can move with a thought.

Motor imagery-based BCIs can be based on decoding imagined movements from electroencephalography (EEG) to control external devices like prosthetics and wheelchairs. EEG is a common choice for BCIs because it is non-invasive and relatively cheap. However, the clinical adoption of MI-BCIs using EEG remains limited due to challenges in decoding accuracy across users and a lack of interpretability in model decisions, alongside the fact that EEG is inherently noisy and does not generalize well between users. Clinicians require transparent insights into which brain regions and time periods contribute to each prediction if such models were to be adopted more widely for rehabilitation and diagnostic support.

This project aims to develop an interpretable deep learning model for classifying motor imagery EEG signals using the EEG Motor Movement/Imagery Dataset. Inspired by the work done by Li, Shi, and Li [1], we developed three model architectures. To enhance clinical interpretability, the project used tools such as Saliency Map and Grad-CAM visualizations to help identify which EEG channels and time segments are most influential in the model's decisions. The models were evaluated on their classification accuracy, generalization across subjects, and how interpretable the models' decisions were. Our results show that the CNN–LSTM–Attention model achieved the highest performance out of the 3 architectures, averaging an AUC of approximately 0.725 across all 7 classification tasks. Furthermore, there appear to be several significant differences in channels that correspond to classifying motor imagery versus movement. In particular, the motor and premotor cortices are important for classifying actual movement, but other areas like the visual cortex are helpful for classifying motor imagery.

These methods not only help demonstrate the feasibility of EEG for BCIs, but more importantly, elicit some of the important features that can guide clinicians and researchers in designing BCIs for individuals with motor impairments. BCIs could focus on recording from the brain areas that are more important for classifying motor imagery as opposed to movement, as these are the areas that disabled individuals are more likely to use when thinking about motion (motor brain areas may have undergone cortical remapping from not being used).

Contributions:

Sergey: Preprocessing and model implementation/training/evaluation

Beñat: Model implementation and interpretability

James: Model implementation/training/evaluation

Raiza: Model implementation/training/evaluation and BrainBERT experiments

Background

Motor impairments can be caused by a variety of means. Individuals may develop quadriplegia after receiving a spinal cord injury, they might become amputees from losing a limb, or many other conditions. When these things happen, individuals lose some of their ability to interact freely with their environment. Amputees can no longer rely on the limb they lost and must instead use a prosthetic with limited capabilities; quadriplegics likely rely on a caretaker for all of their needs, since they are unable to move. BCIs have the potential to be very helpful for these individuals because it would allow them to interact directly with an external device without relying on their lost motor abilities.

Unfortunately, BCIs are not a trivial technology. They are difficult to design and implement for a variety of factors. First of all, the nervous system is extremely complex and constantly processing a lot of information, so extracting motor signals is not possible without recording lots of noise. Also, there are many ways to record brain activity, but these vary based on resolution and invasiveness. EEG could be very fruitful for this technology because it is not invasive, so it is cheaper and safer, but it has quite poor spatial resolution; this is detrimental because information is processed in the nervous system on very small scales. Therefore, to use EEG for a BCI, it is imperative that the correct features are extracted.

Decoding movement from EEG is a popular task because the motor cortex is large and well-defined in the brain, and motor tasks are very easy to perform. Activity in the motor cortex is also very strongly correlated to actual movement (as opposed to, say, language processing in language networks, which is extremely complex and distributed throughout the brain). Therefore, BCIs commonly target the motor cortex to decode movement.

Unfortunately, while the motor cortex is a prime target for decoding in abled individuals, it is unclear if it will be as promising for disabled individuals. This is because the brain has a natural tendency to repurpose unused brain areas. In the case of people with motor impairments, they are likely not making use of their motor cortex for movement, so it is probable that it will be remapped to other functional areas. Therefore, it will no longer be as fruitful for decoding motor intent.

Since the motor cortex could be less reliable for individuals with motor impairments, and since they cannot perform actual motions anyways, it is valuable to identify which features could be extracted from EEG that correlate to performing some sort of motor task. These features could be used to guide researchers in developing better BCIs.

Although there are existing motor imagery BCIs, there have not been approaches that use interpretability for models using preprocessed time-series signals as inputs. However, motor imagery BCIs have shown to be fairly successful. In a 2023 paper by Cajigas et al., a team of researchers developed a BCI based on upper limb motor imagery from a quadriplegic patient via ECoG (essentially invasive EEG).[7] Their personalized model achieved ~80% accuracy, but they did not explore the features responsible for this accuracy. Since they used motor imagery using ECoG of the motor cortex, it is possible that these results could have improved if they had placed the implant at a different location that may have not undergone remapping.

In a similar sense, decoding has been done for amputees. In a paper by Bruurmijn et al., researchers explored whether or not phantom limb movements could be decoded for motor imagery of moving the amputated limb.[8] However, instead of recording from the contralateral motor cortex (the side that would be responsible for movement of the amputated limb), they recorded from ipsilateral motor and somatosensory cortices. Using fMRI and a SVM, they were able to achieve above-chance accuracy. Since the ipsilateral hemisphere corresponds to movement of the non-amputated limb, this is further evidence that decoding should be done from brain areas that do not correspond to the area of motor impairment. Also, this paper did not explore relevant features, either.

Although BCIs exist to aid individuals with motor impairments, they rarely explore features that are important for neural decoding. Instead, BCIs are often designed around neuroscientific principles, but these may be inconsistent for disabled individuals compared to healthy baselines. Therefore, it is important that relevant features are uncovered for neural decoding of motor imagery for disabled individuals so that the design of BCIs can be guided in a more specific manner.

Dataset

The dataset contains over 1500 one- or two-minute EEG recordings, obtained from 109 subjects without motor impairments. The recordings were captured when the subjects participated in motor execution and imagery tasks. The features include EEG signals recorded from 64 scalp electrodes, each sampled at 160 samples per second as seen in Figure 1. The annotation (label) column includes one of three codes (T0, T1, or T2):

- **T0** corresponds to rest
- **T1** corresponds to onset of motion (real or imagined) of the left fist (in runs 3, 4, 7, 8, 11, and 12) and both fists (in runs 5, 6, 9, 10, 13, and 14).
- **T2** corresponds to onset of motion (real or imagined) of the right fist (in runs 3, 4, 7, 8, 11, and 12) and both feet (in runs 5, 6, 9, 10, 13, and 14).

No special permission is required for access or use in research, provided you agree to the PhysioNet usage terms and cite the dataset accordingly.

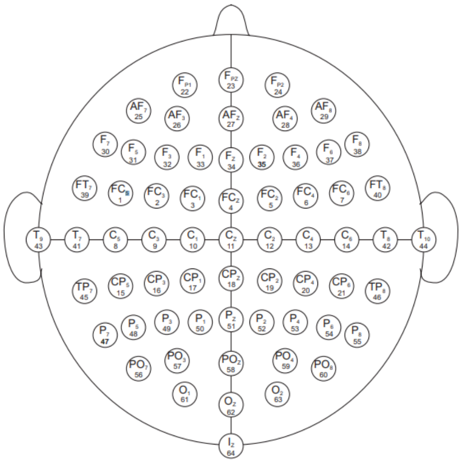


Figure 1. Scalp electrode names and locations.

Each subject completed a total of fourteen experimental runs. The first two runs served as one-minute baseline recordings, one with eyes open and one with eyes closed. The remaining twelve runs, each lasting two minutes, involved four different motor or motor imagery tasks, each repeated three times. In the first task, subjects physically opened and closed their left or right fist depending on whether a target appeared on the left or right side of the screen. In the second task, they imagined performing the same left or right fist movements without actual motion. The third task required subjects to open and close both fists when the target appeared at the top of the screen, or both feet when the target appeared at the bottom. Finally, in the fourth task, subjects imagined opening and closing both fists or both feet according to the target's position.

The recordings used 64 electrodes sampled at 160 Hz. Each of the 109 subjects had two 1 minute baseline recordings and twelve 2 minute recordings (three recordings for four tasks). Therefore, each subject has 26 minutes of EEG per channel, which is equal to 15,974,400 timepoints.

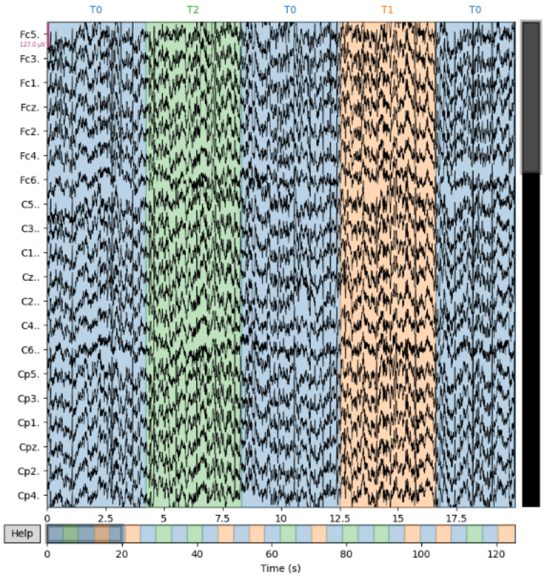


Figure 2. Patient 1, run 3, Task 1 (rest T0, open and close left T1 or right T2 fist).

All recordings were preprocessed. This was done by firstly filtering them between 1 and 30 Hz, and then filtering out line noise at multiples of 60 Hz. Recordings were then average referenced and split into 2 second epochs with 0.25 second overlaps. These two second epochs were ultimately the input data and they were labelled based on the task that the subject was performing. We initially experimented with 1-second windows, but performance was poor due to insufficient temporal information. With the final settings, each sample has shape (64, 320), corresponding to 64 EEG channels and 160 time points per 2-second window.

Methodology

We applied a 60/20/20 split for training, validation, and testing. To avoid data leakage, all samples from a given subject were assigned entirely to one split; for example, if subject S001 was placed in the test set, none of their epochs appeared in the training or validation sets.

Label	Train	Validation	Test
Rest	20972	7490	7420
Move both fists	2177	777	784
Move both feet	2233	798	791
Move right fist	2198	770	770
Move left fist	2212	805	805
Imagine moving both fists	2205	791	763
Imagine moving both feet	2205	784	742
Imagine moving right fist	2198	777	770
Imagine moving left fist	2212	798	805

Table 1: Number of appearances for each task between the three splits.

As expected, the resting class is substantially larger than any of the movement or imagery classes. One open question is how to best define “rest.” The dataset contains two baseline recordings, one with eyes open and one with eyes closed, along with resting intervals that occur between tasks. Currently, all of these states are grouped together as a single rest category to encourage model generalization. However, it remains unclear whether this approach is ideal; restricting rest to only the inter-task periods and excluding baselines may produce cleaner or more meaningful distinctions.

For modeling, our baseline approach was to reproduce the architecture presented by Li et al., which integrates convolutional layers, a bidirectional LSTM, and an attention mechanism for decoding motor imagery. Their model applies one-dimensional convolutions to each epoch, uses the LSTM to capture temporal dependencies, and incorporates attention for final classification. Given our large number of samples, we hypothesized that the presence of noise and artifacts might actually enhance generalization.

We evaluated multiple classification tasks: Rest vs. Move vs. Imagine; Rest vs. Move; Rest vs. Imagine; Move Left vs. Right; Imagine Left vs. Right; and the full 9-class problem. In addition to implementing the CNN → LSTM → Attention model, we developed two related architectures: a CNN → Transformer model and a CNN → LSTM → Transformer model. All architectures shared the same initial component, a per-channel convolutional feature extractor followed by a global convolutional layer to aggregate cross-channel information, before their respective sequence modeling and attention/transformer modules. Hyperparameters were tuned empirically by evaluating validation performance. All training was done using L4 GPUs on Google Colab with high-RAM settings.

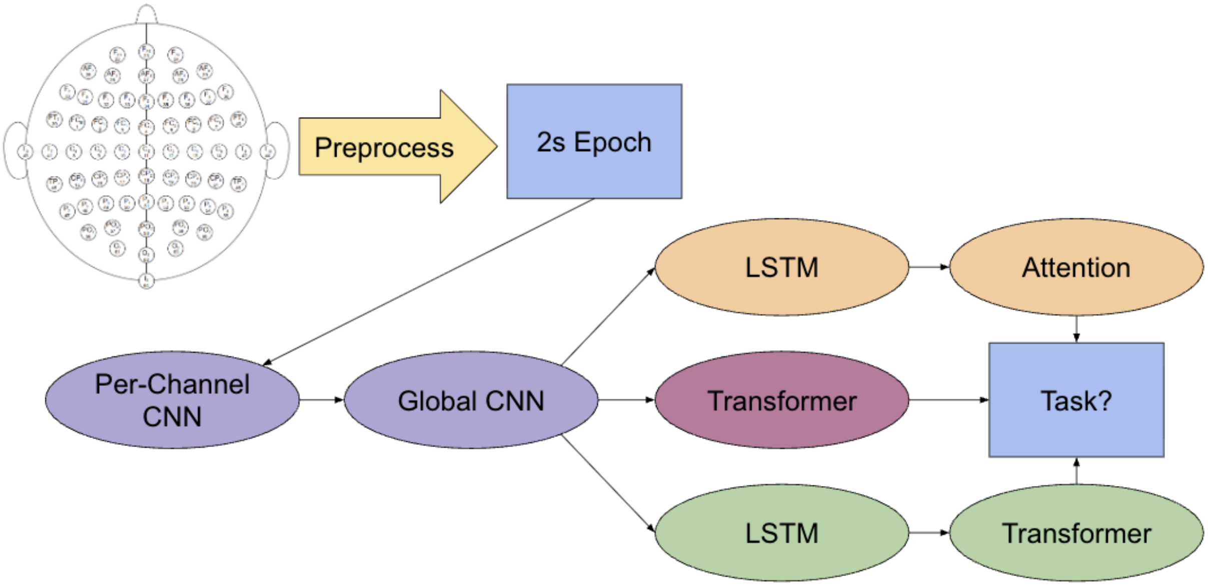


Figure 3. Prediction pipeline. Note that each of the three architectures uses a per-channel convolutional layer followed by a global convolutional layer, but then these features are fed into an LSTM followed by attention, a transformer, or an LSTM followed by a transformer.

To better understand how each model arrived at its predictions, we applied post-hoc interpretability techniques tailored to the different architectural components. For the convolutional layers, we used Grad-CAM to highlight the temporal features most influential to the CNN, while for the recurrent and transformer-based components, we generated gradient-based saliency maps to measure the sensitivity of the output to each input channel and time point. To ensure fair and unbiased interpretation, rather than cherry-picking visually appealing examples, we computed label-wise averaged Grad-CAM and saliency maps across all samples of a class before visualization. These aggregated saliency values were first overlaid onto the EEG waveforms, allowing us to inspect the raw signals alongside the regions where the model consistently focused its attention. This helped identify meaningful neural patterns and detect potential issues, such as dependence on noise or a single electrode. For spatial interpretability, we then averaged saliency over time for each channel and projected the results onto EEG topographic maps. Topomaps proved especially informative because they preserve the spatial organization of the scalp, enabling direct comparison between model attention and known neurophysiological motor regions, and providing insight into whether the model's focus was anatomically plausible or indicative of artifacts or dataset biases.

Another approach that we attempted was to use an existing pre-trained model like BrainBert. The BrainBert pipeline required specialized preprocessing that transformed raw EEG recordings in EDF format into spectrograms for model training. Each epoch of EEG data was converted using a Short-Time Fourier Transform (STFT) with configurable parameters for window length and overlap, then clipped to a fixed number of frequency bins to standardize input dimensions. We found that the model underperformed on the dataset since BrainBert was trained on iEEG data and processes only a single channel per forward pass. Multi-channel epochs were collapsed into a single averaged waveform to work with BrainBert, thus losing spatial information.

Evaluation metrics included several approaches. To evaluate each model, the test set was used to calculate a balanced accuracy, AUC, and an F1-Score. This would help measure the model's robustness to different classes and holistically demonstrate how well the model performed for classification. Interpretability methods were visualized as average activations across the entire test set so that we could inspect for unique activations across classes.

Results

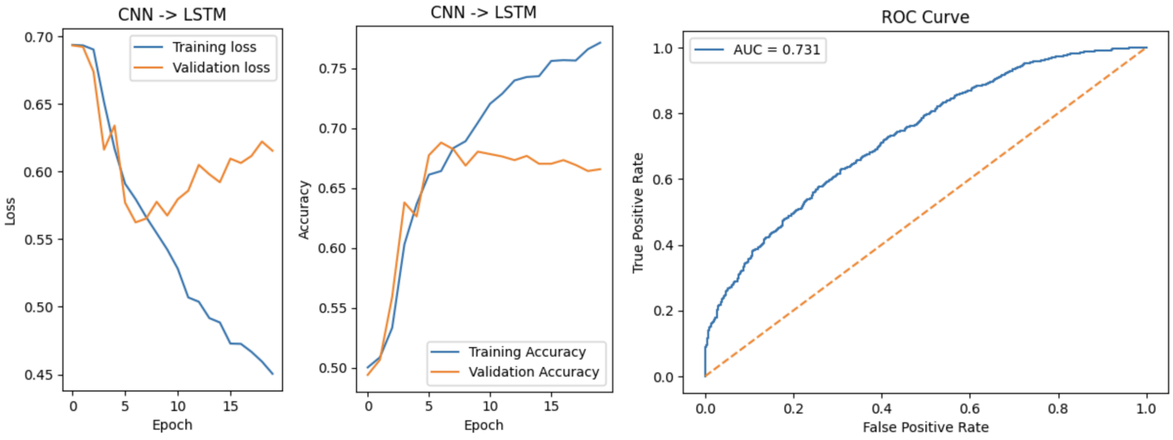
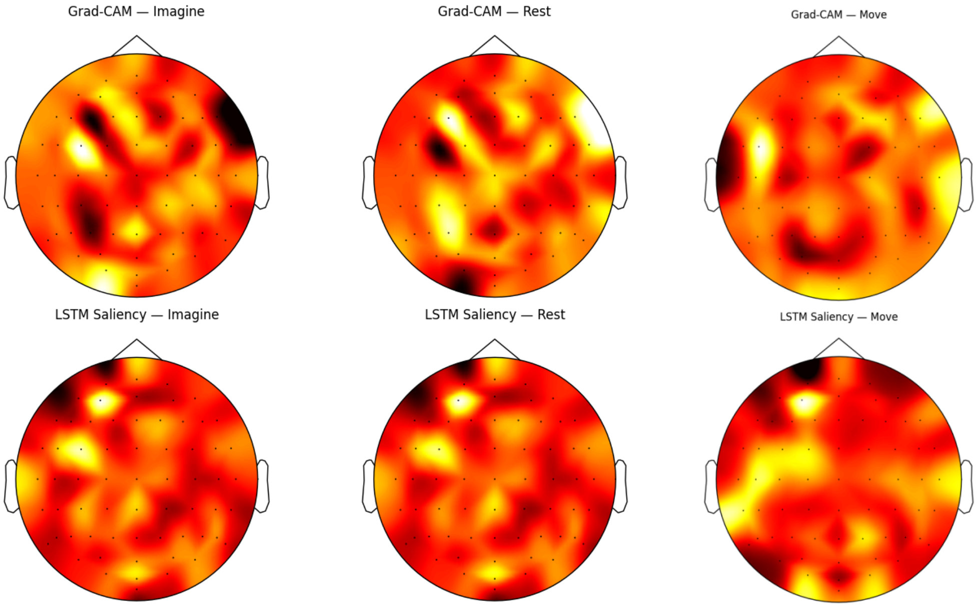


Figure 4. Training and evaluation plots for the CNN → LSTM → Attention architecture for predicting motor imagery of left versus right hand movements.

	Balanced Accuracy	AUC	F1 Score	Balanced Accuracy	AUC	F1 Score	Balanced Accuracy	AUC	F1 Score
Rest/Move/ Imagine	**0.465**	**0.691**	**0.462**	0.437	0.661	0.434	0.448	0.676	0.443
Move/Imagine	0.574	**0.622**	0.565	0.573	0.615	**0.569**	**0.580**	0.621	0.565
Rest/Move	0.629	**0.753**	0.636	0.615	0.684	0.612	**0.651**	0.749	**0.655**
Rest/Imagine	0.619	**0.705**	0.624	0.622	0.687	0.625	**0.637**	0.703	**0.641**
Move Right/Left	**0.665**	**0.758**	**0.655**	0.631	0.699	0.630	0.645	0.729	0.642
Imagine Right/Left	**0.653**	**0.731**	0.647	0.632	0.687	0.631	0.649	0.729	**0.648**
All Tasks	**0.210**	**0.722**	**0.219**	0.137	0.684	0.118	0.162	0.708	0.163

Table 2. Evaluation results for each of the three models (color-coded as in Figure 3) for different classification tasks.



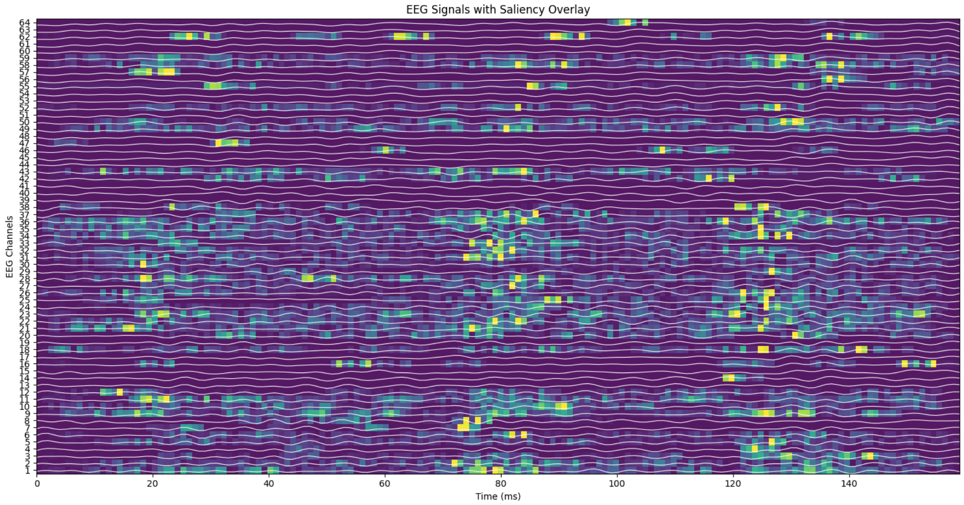


Figure 5. Saliency map overlaid on EEG waveforms and corresponding brain topography. Note: The topomaps are averaged across all samples, not cherry-picked.

Discussion

The goal of this project was to develop and evaluate interpretable deep learning models for decoding motor imagery and motor execution from EEG signals. In particular, we aimed to balance classification performance with interpretability, since it is a key requirement for clinical adoption of BCI systems. Using the PhysioNet EEG Motor Movement/Imagery dataset, we implemented and compared three architectures that combine convolutional layers with temporal modeling components, including LSTMs, attention mechanisms, and transformers. We analyzed their behavior using post-hoc interpretability techniques.

Across the evaluated tasks, we noticed that CNN-LSTM-based architectures consistently outperformed the transformer-only variant. The strongest results were observed for binary classification tasks such as rest versus movement and left versus right limb discrimination, where AUC values reached as high as 0.72-0.75. The results indicate that while the models are not yet suitable for clinical practice, they are able to capture meaningful structure in the EEG signals. We noticed a significant performance drop (0.210) for the full 9-class classification task, indicating the difficulty of fine-grained motor decoding under class imbalance and high inter-subject variability. Moreover, the weaker performance of transformer-based models hints that the convolutional features extracted from raw EEG may not provide sufficient information for effective attention-based modeling without additional feature engineering. From Figure 4, it is clear that the models were able to extract some features from the signals, but then quickly began to overfit. This is due to EEG's inability to generalize well across subjects: the model learned the training subjects, but struggled to predict from validation subjects.

Post-hoc interpretability analysis provided essential insights into how the models made their predictions. Grad-CAM and saliency maps revealed different activation patterns for motor execution and motor imagery tasks. Movement-related predictions tended to emphasize lateral temporal and frontal channels associated with motor regions, while motor imagery showed more medial and occipital activation (at the visual cortex). Therefore, it can be concluded that patients may rely on more visual imagery when imagining movement, rather than directly activating motor cortex regions; therefore, it is possible to classify motor imagery without the motor cortex, which is an important implication for disabled individuals. These indicate that the models are learning partially task-relevant features rather than exploitation noise or single-channel artifacts. This could also be valuable insight into the features that might be used by patients with motor impairments, whose motor cortices are likely less active.

The study has several limitations. All EEG recordings were collected from healthy individuals rather than patients with motor impairments, limiting conclusions about how the model would perform in real clinical practice with quadriplegics and amputees who have likely undergone cortical remapping. Moreover, the lack of data, the noise and variability of EEG signals across subjects, and class imbalance between rest and task conditions constrained overall classification accuracy. Finally, the lack of existing ready-to-run codes on the same topic to compare results adds difficulties in assessing how well the model has performed compared to other models for the same task.

Future work can have several directions. Model performance could be improved by incorporating more advanced preprocessing or using some publicly available model pipelines. Existing foundational models such as BrainWave can be leveraged and fine-tuned to compare model performance. Exploring additional interpretability techniques, such as integrated gradients or SHAP-based methods, could provide more robust explanations. Testing the models on datasets collected from individuals with motor impairments would be critical for evaluating clinical relevance. Finally, developing lightweight models capable of real-time inference would be an important step toward practical BCI deployment. Another approach that would likely improve results drastically would be to design personalized models based on data from individuals, rather than populations; this would help overcome the lack of generalizability that EEG suffers from. The PhysioNet EEG Motor Movement/Imagery dataset used in this project is publicly available and can be accessed through PhysioNet under standard usage terms. The link to the dataset can be found in [5]. All preprocessing, training, evaluation and interpretability analyses were implemented using open-source Python libraries, such as MNE and PyTorch. The code for this project is not publicly available, but it is available in the final project submission.

Conclusion

In this project, we studied the problem of decoding motor imagery and movement from EEG signals with a focus on both classification performance and interpretability. Our approach used deep learning models that combine convolutional layers with temporal modeling components, including LSTMs and attention or transformer mechanisms, applied to the PhysioNet EEG

Motor Movement/Imagery dataset. We evaluated multiple classification tasks ranging from simple rest versus movement to more fine-grained limb-specific motor imagery tasks, while enforcing subject-exclusive data splits to test generalization. Among the tested architectures, the CNN–LSTM–Attention model achieved the strongest overall performance, reaching AUC values of approximately 0.72 across several tasks.

In addition to classification results, post-hoc interpretability methods such as saliency maps and Grad-CAM revealed meaningful differences between movement and motor imagery. Movement showed stronger activation in lateral motor regions, and imagery exhibited more medial and occipital patterns. The results suggest that the models are not relying purely on noise and instead capture physiologically relevant features. This work can become a stepping stone to improve current BCI systems by increasing transparency and helping clinicians understand which brain regions contribute to predictions, potentially supporting personalized electrode placement and more reliable control interfaces. The unique features of motor imagery may also be valuable for patients with motor impairments or altered cortical organization.

To bring this work closer to clinical practice, future work needs to validate the models on data from individuals with motor impairments, improve robustness to noise and subject variability, and integrate real-time result output capabilities. Additional clinical testing and collaboration with healthcare professionals would be necessary to ensure safety, reliability, and usability in real-world settings.

References

1. Li, J., Shi, W., & Li, Y. (2024). An effective classification approach for EEG-based motor imagery tasks combined with attention mechanisms. Cognitive Neurodynamics, 18, 2689–2707.
2. Shuqfa, Z., Lakas, A., & Belkacem, A. N. (2024). Curation of Physionet EEG Motor Movement/Imagery Dataset for decoding and classification. Data in Brief.
3. Li, J., et al. (2023). Comparative study of EEG motor imagery classification based on DSCNN + ELM. Journal of Neuroscience Methods.
4. Lionakis, E., Karampidis, K., & Papadourakis, G. (2023). Current trends, challenges, and future research directions of hybrid and deep learning techniques for motor imagery brain–computer interface. Multimodal Technologies and Interaction, 7(10), 95.
5. Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., & Wolpaw, J. R. (2004). BCI2000: A general-purpose brain–computer interface (BCI) system. IEEE Transactions on Biomedical Engineering, 51(6), 1034–1043. https://physionet.org/content/eegmmidb/1.0.0/
6. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23), e215–e220. RRID:SCR_007345.

7. Cajigas, I., Davis, K. C., Prins, N. W., Gallo, S., Naeem, J. A., Fisher, L., Ivan, M. E., Prasad, A., & Jagid, J. R. (2023). Brain–computer interface control of stepping from invasive electrocorticography upper-limb motor imagery in a patient with quadriplegia. Frontiers in Human Neuroscience, 16.
8. Bruurmijn, L. C. M., Raemaekers, M., Branco, M. P., Vansteensel, M. J., & Ramsey, N. F. (2021). Decoding attempted phantom hand movements from ipsilateral sensorimotor areas after amputation. Journal of Neural Engineering, 18.