Identification of Early EEG Biomarkers for Alzheimer's Disease Using Multi-channel Analysis and Interpretable Deep Learning

Nguyen Duc Anh – 22022661
Thai Nguyen Hoang Bach – 22022672
Thai Thi Thuy Linh – 22022631
Ha Kim Duong – 22022621
Nguyen Trong Huy – 22022545
Institute of Artificial Intelligence,
VNU University of Engineering and Technology,
Hanoi, Vietnam

Abstract—**In this project, we employ feature extraction methods to observe potential biomarkers of Alzheimer's Disease based on multi-channel EEG signal data. Subsequently, spectrogram images are generated and used as input for a deep learning model to classify two categories in the dataset. MCI and NORMAL.**

1. Introduce

1.1. What is the problem?

Identify and analyze spectral, temporal, and spatial features in multi-channel EEG signals that may serve as early biomarkers of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI).

Use these data to extract spectrogram images from each EEG segment, which serve as input to a deep learning model for binary classification (MCI and NORMAL)

Integrate interpretable deep learning methods with explainable AI techniques to identify and explain the most important EEG features that contribute to early detection.

1.2. What do the key concepts mean?

1) EEG data: EEG data records the brain's electrical activity through electrodes placed on the scalp. This is a non-invasive method commonly used in research.
2) Spectrogram: A spectrogram is a visual representation of a signal showing how its frequency spectrum changes over time.
3) Interpretable model: An interpretable model is a machine learning or artificial intelligence model whose decision-making or prediction process can be understood by humans
4) Acronyms:
   MCI: Mild Cognitive Impairment
   NORMAL: Cognitively normal individuals

1.3. What has been tried?

Currently, through research, it has been found that various methods have been used to process EEG signal data as input for machine learning models. In the paper "Deep Learning for Electroencephalogram (EEG) Classification Tasks: A Review"[**1**] by Alexander Craik, Yongtian He, and Jose L. Contreras-Vidal, the authors highlight three main approaches: using spectrogram images extracted from the signals, using power spectral density (PSD) features, and using raw EEG data.

In this study, we used spectrogram extraction as the input method for the model. According to several related research papers, the authors have explored various spectrogram extraction techniques. For example, a spectrogram image can be generated for each channel in the EEG data, and then the spectrograms from different channels within the same signal segment are stacked together to form a 3D spectrogram stack [**3**], which is used as input for models such as 3D CNNs. Alternatively, traditional 2D spectrograms [**2**] can be created by combining all channels into a single spectrogram image for each frequency band of a given signal segment. In this experiment, we conducted tests using 2 approaches: the spectrogram-based method, which involves using 2D images and stacked 3D spectrograms, and the raw data method for model training[**5**].

2. Methods

2.1. Dataset

In this study, we used the data set from https://misp.mui.ac.ir/en/eeg-data-0 consisting of 61 patients, including 29 in the MCI group and 32 in the NC group. Data were recorded based on the 10–20 International System using a 32-channel digital EEG device with a 256 Hz sampling rate (Galileo NT, EBneuro, Italy). Each recording includes 19 channels and lasts approximately 1800 seconds.

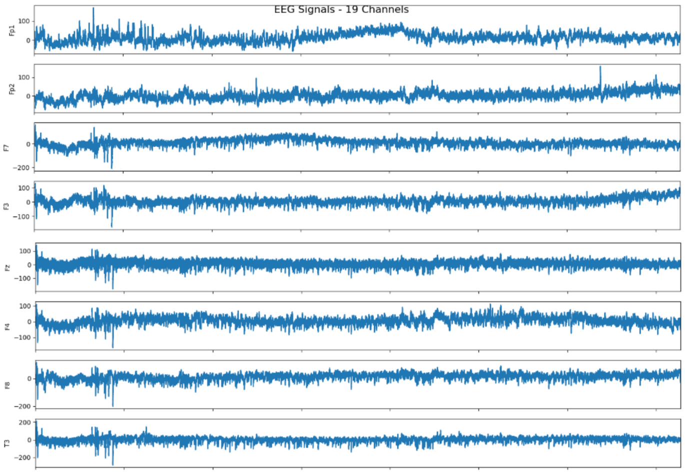

Fig. 1: The EEG waveform of one channel from a single data sample.

2.2. Data Preprocessing

After reading the raw data from the .edf format, the data is processed for noise removal using ICA and bandpass filtering. It is then segmented into epochs appropriate for each method. For the two spectrogram-based approaches, the data undergoes preprocessing followed by feature extraction. In contrast, the raw data is labeled and directly prepared for model input.

2.3. Feature Extraction

1) Time-domain Features

After segmenting EEG signals into epochs, we calculated statistical features in the time domain for each of the 19 EEG channels in each epoch. Specifically, the following five features were calculated:

- Mean: the average signal amplitude,
- Standard deviation (STD): the degree of signal dispersion,
- Peak: the maximum amplitude within the signal,
- Skewness: the asymmetry of the signal's amplitude distribution,
- Kurtosis: the sharpness or peakedness of the signal distribution.

These features provide valuable information about the shape and variability of EEG signals over time. As each channel yields five features, a total of 95 features (19 channels × 5 features) were extracted per epoch.

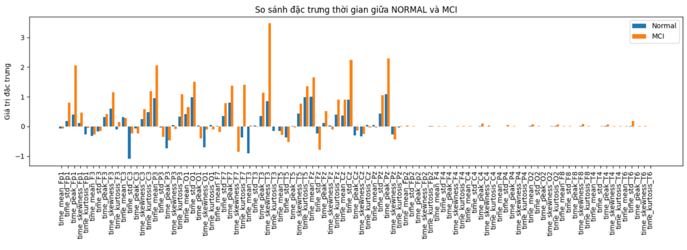

Fig. 2: The bar chart compares time-domain features across channels between the MCI and Normal groups.

2) Power Spectral Density (PSD) features.

For spectral feature extraction, we applied Welch's method to estimate the Power Spectral Density (PSD) of each EEG channel in every epoch. We focused on the average spectral power within four common EEG frequency bands:

- Delta (1–4 Hz)
- Theta (4–8 Hz)
- Alpha (8–13 Hz)
- Beta (13–30 Hz)

The process involved iterating through each epoch and then across all EEG channels. For each channel, Welch's method was used to compute the PSD, after which we averaged the power values within the frequency ranges mentioned above. This resulted in 4 features per channel, and hence 76 spectral features (19 channels × 4 bands) per epoch.

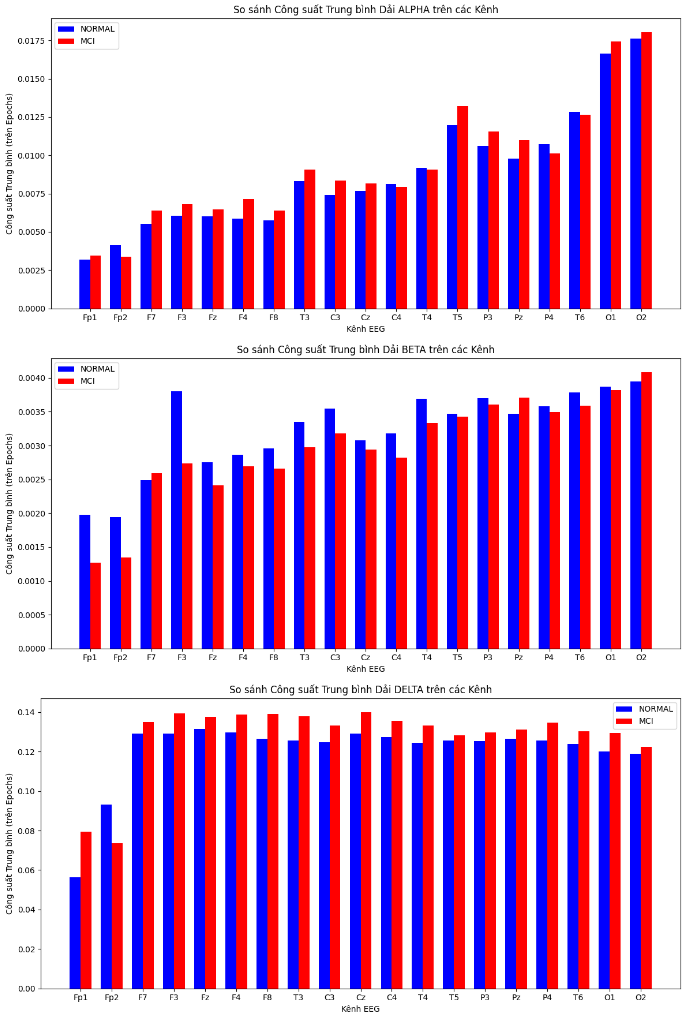

Fig. 3: A comparative plot of average spectral power between the two groups (MCI and Normal) across all EEG channels and frequency bands.

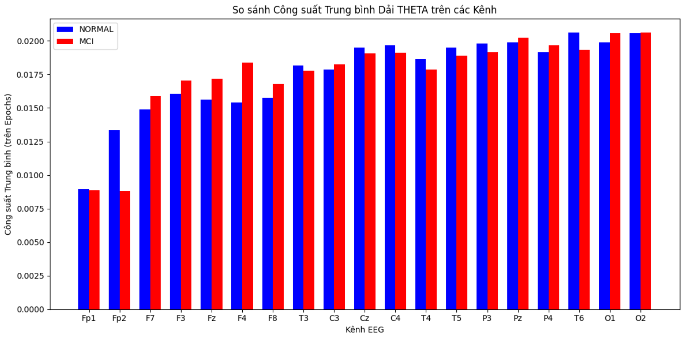

Fig. 4: A comparative plot of average spectral power between the two groups (MCI and Normal) across all EEG channels and frequency bands.

- DELTA: The MCI group shows higher delta activity than the NORMAL group in almost all channels, especially prominent in the frontal region (F3, Fz, F4) and central region (C3, Cz), reflecting cognitive decline.
- THETA: The MCI group tends to have higher theta activity than the NORMAL group in many channels, particularly in frontal channels such as F3, F4, F7, Fz, and also temporal regions like T3, T5, indicating slower brain activity and memory impairment.
- ALPHA: The decrease in alpha activity is usually associated with cognitive decline, but here it is not very clear. It is possible that early-stage MCI does not strongly affect this band, or individuals in the MCI group still have good compensatory mechanisms.
- BETA: The MCI group tends to have lower beta activity than the NORMAL group in many channels, especially Fp1, Fp2, F3, T3, C3, Cz, reflecting reduced concentration, attention, and alertness.

3) Spectrogram image features

The spectrogram is generated using the Continuous Wavelet Transform (CWT) method with the Morlet wavelet and is applied iteratively across each channel with different frequency bands. This produces information-rich spectrograms, particularly useful for monitoring brain activity in specific frequency ranges. CWT preserves frequency, time, and phase information — making it extremely valuable in applications such as neurological diagnosis, brain-computer interfaces (BCI), or pathological classification (e.g , MCI, Alzheimer's disease).

- Spectrogram 2D: With this method, after preprocessing, each patient's data is divided into multiple segments, each approximately 4 seconds long and filtered to extract key frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–25 Hz). Then, for each frequency band within each segment, 19 spectrogram images are extracted corresponding to the 19 recorded EEG channels. These 19 channel-wise images are then combined to generate a single aggregated spectrogram image.

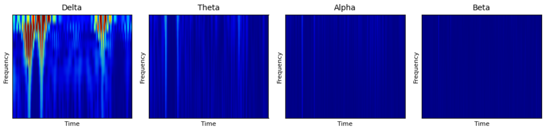

Fig. 5: Illustrative spectrogram images of a normal subject across four frequency bands.

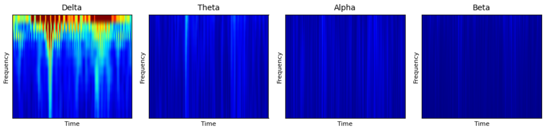

Fig. 6: Illustrative spectrogram images of a MCI subject across four frequency bands.

- Stack spectrogram: With this method, we still divide the data into epochs; however, each epoch lasts approximately 60 seconds due to computational resource limitations for each data segment of a patient. Nineteen 2D spectrogram images are generated and compressed into a single 3D block, which is then stored as a NumPy array (.npy).

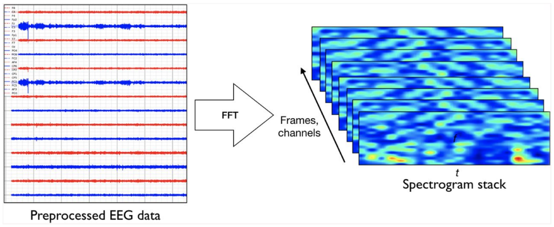

Fig. 7: Illustrative spectrogram images of a MCI subject across four frequency bands.

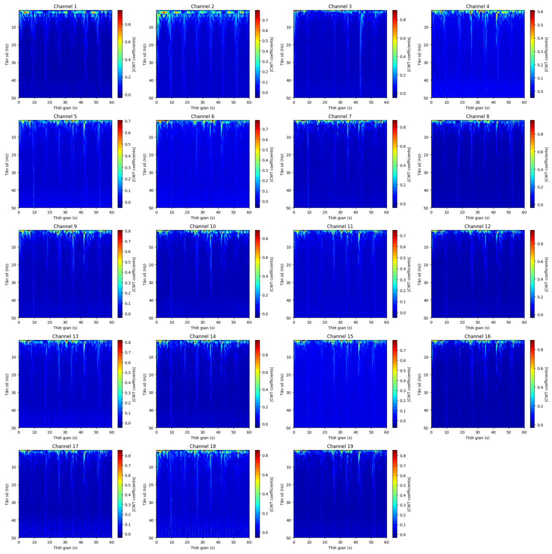

Fig. 8: 19 spectrogram images from 19 channels combined into a single stacked spectrogram.

2.4. Model Architecture

1) Spectrogram 2D

With 2D images, we use a conventional 2D CNN architecture for the images. In this application, we implement the ResNet18 architecture: ResNet18 is the lightest and shallowest variant in the Residual Network (ResNet) family, consisting of 18 weighted layers (layers with parameters).

Main structure: It includes 5 main blocks, each block containing 2 residual blocks (each residual block has 2 Conv2D layers).

In total: 17 Conv2D layers + 1 fully connected layer → 18 layers. This is a lightweight, fast, and accurate architecture often used in spectrogram image tasks to extract image features. It performs well with small to medium-sized datasets, which is common in EEG applications.

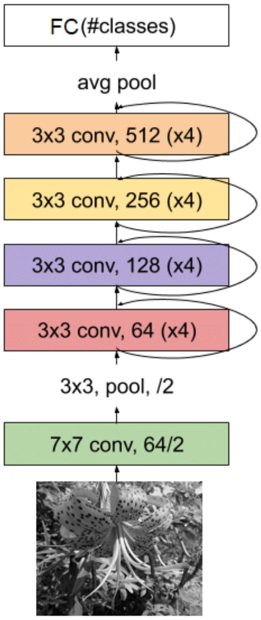

Fig. 9: resnet18 architecture

2) Stack spectrogram.

Since each data sample is represented as a 3D tensor, we employ the 3D variant of ResNet, namely r3d_18. This architecture is adapted to handle volumetric data and has been customized to support binary classification tasks. The input EEG tensor has the format [1, Channel, Depth/Time, Height, Width], for instance [16, 1, 19, 128, 128]. Similarly to ResNet2D, ResNet3D has a similar architecture but replaces the blocks with Conv3d, MaxPool3d, and BatchNorm3d layers.

3) Raw EEG data

With this raw data, we used a 1D CNN model. The architecture consists of multiple convolutional layers, activation functions, pooling layers, dropout for regularization, and a final fully connected layer for binary classification (outputting probabilities through a sigmoid function). The model takes as input a tensor of shape (batch_size, in_channels, sequence_length).

Layer	Output Shape	Operation
Input	[B, 1, 19, 112, 112]	Raw EEG stack
Conv3D	[B, 64, 19, 56, 56]	kernel=(3,7,7), stride=(1,2,2)
MaxPool3D	[B, 64, 19, 28, 28]	kernel=(1,3,3), stride=(1,2,2)
Layer1	[B, 64, 19, 28, 28]	BasicBlock × 2
Layer2	[B, 128, 10, 14, 14]	BasicBlock × 2, stride=2
Layer3	[B, 256, 5, 7, 7]	BasicBlock × 2, stride=2
Layer4	[B, 512, 3, 4, 4]	BasicBlock × 2, stride=2
AvgPool	[B, 512, 1, 1, 1]	AdaptiveAvgPool3d
FC + Sigmoid	[B, 1]	Binary Classification

TABLE I: The architecture of the r3d_18 model applied to EEG stack data

2.5. Training and Evaluation

1) Spectrogram-base

Both methods spectrogram-base are trained and evaluated using 80% of the data for training and 20% for testing, ensuring that the data is split by patient so that no patient appears in both sets to avoid data leakage. The models are trained for about 10 to 20 epochs, with accuracy monitored after each epoch.

For the 2D spectrogram: the model is trained for about 30 epochs, equivalent to 120 images per patient due to hardware limitations and a training time of about 1 hour.Epochs were randomly selected. The hyperparameters are:

learning rate: 1e-4

optimizer: Adam

loss function: CrossEntropyLoss

batch size: 32

The images are resized to 224x224, the standard input size for ResNet18.

For the stacked spectrograms: the model is trained on 30 stacks per patient, with training time and hyperparameters similar to those previously mentioned.

The model is evaluated using metrics such as test accuracy, sensitivity, specificity, and the ROC curve.

2) Raw EEG data

With this method, we selected 10 patients (5 MCI, 5 NORMAL) from the original dataset to be used later for testing. The data was separated into MCI and NORMAL groups from the beginning and was read and processed into 25-second epochs. It was then formatted as NumPy arrays with the shape (trials, channels, length).

Afterwards, the epochs from each patient were assigned corresponding labels: 0 for Normal and 1 for MCI. The arrays for MCI and Normal were then concatenated into a single array and grouped by subjects to ensure that each epoch belongs to one specific subject.

Then, Group K-Fold Cross-Validation with 5 folds was used for training. The folds were split based on subject groups, and for each fold, the data was divided into training and validation sets, ensuring that epochs from the same subject were not split across different sets.

Each fold was iterated through and trained using Binary Cross-Entropy Loss and the Adam optimizer with a learning rate of 0.001. The training process was monitored using accuracy on both the training and validation sets, and the best model for each fold was saved based on the highest validation accuracy.

3. Results

3.1. Spectrogram-base

Overall, the two spectrogram-based methods produce accuracy results that are not very good.

1) Spectrogram 2D

The accuracy is around 80% on the training set, but only fluctuates around 56% on the test set. Sensitivity (Recall): 0.50 Specificity: 0.60 F1 Score: 0.51 AUC: 0.55

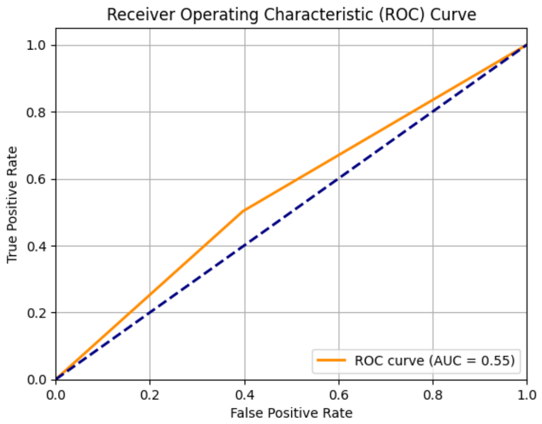

Fig. 10: ROC curve

Use CAM to explain how the model makes predictions.

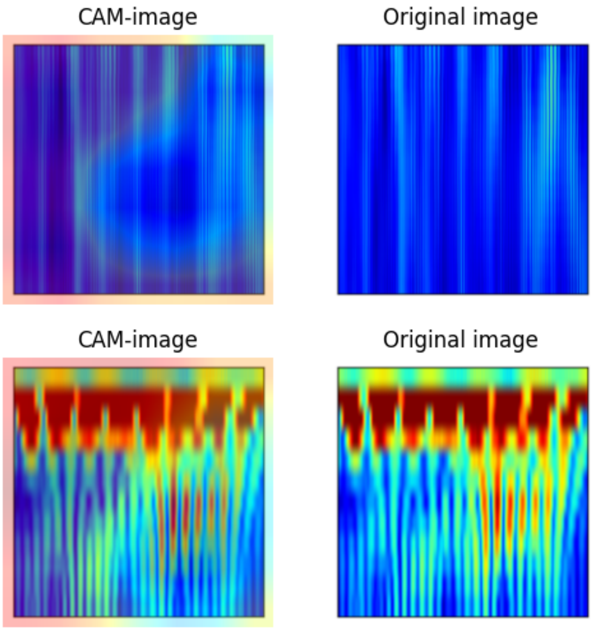

Fig. 11: CAM with 2D image

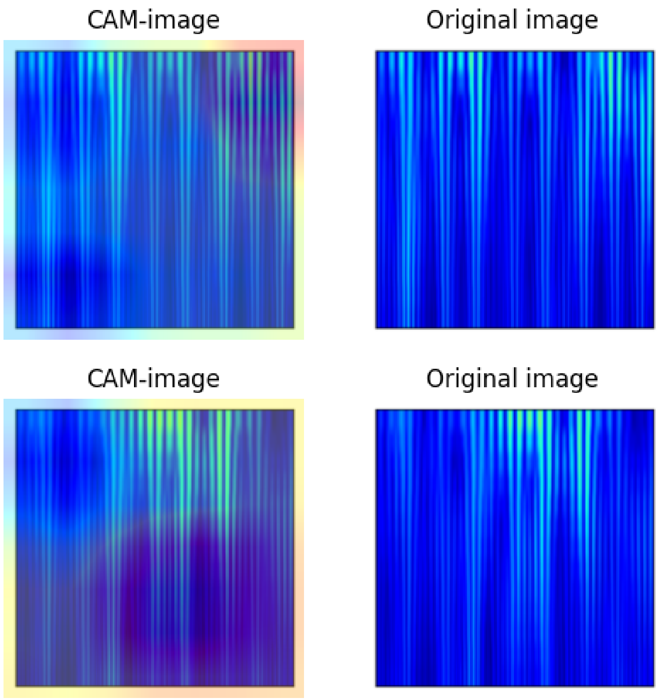

Fig. 12: ROC curve

2) Stack spectrogram

With the stacked spectrograms, the results improved slightly but not significantly. The highest accuracy (Acc) reached about 98% on the training set and 62% on the test set. On the test set, the loss and accuracy fluctuate greatly, indicating that the model has not truly learned well and shows signs of overfitting.

For the two spectrogram-based methods, the data is processed and then passed to the feature extraction stage. In contrast, the raw data is labeled and prepared directly for input into the model.

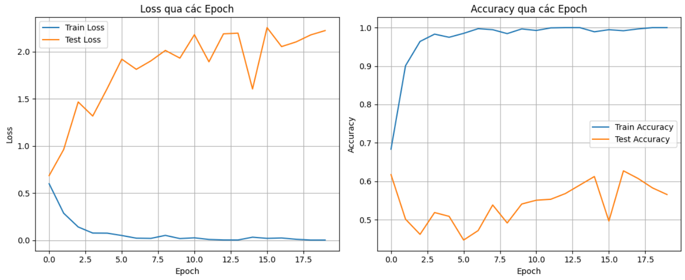

Fig. 13: Training and loss curves

3.2. Raw EEG data

With the raw data, the results obtained from the best-performing fold are slightly better, achieving around 70% accuracy on 10 test patients. The other metrics are as follows: Sensitivity (Recall): 0.6098 Specificity: 0.7720 F1-Score: 0.6647 ROC-AUC: 0.6925

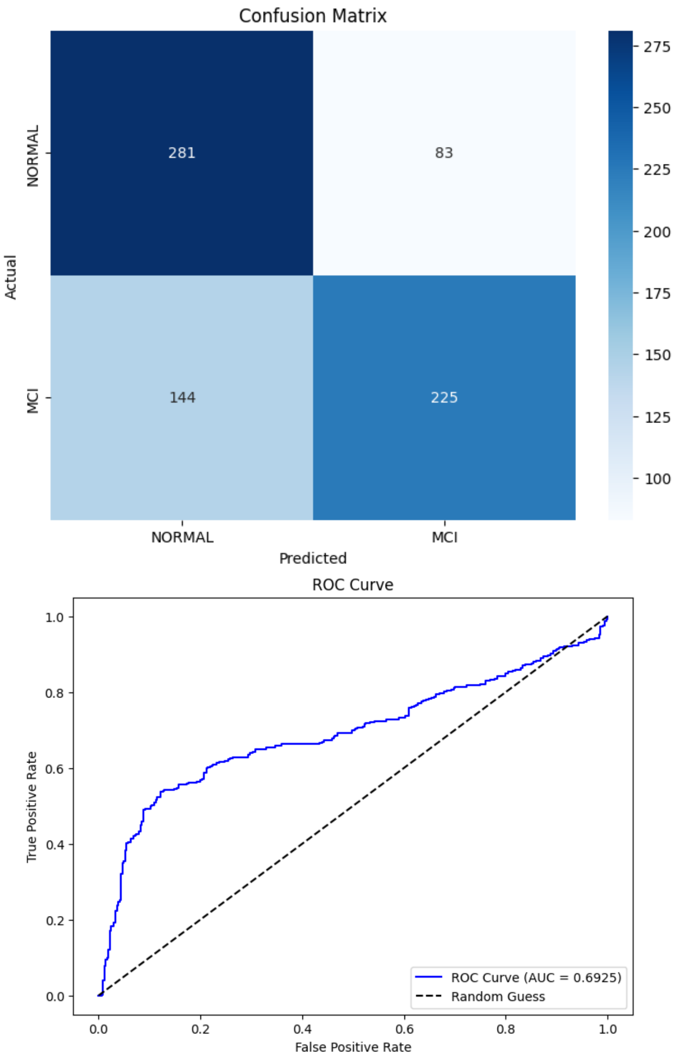

Fig. 14: Training and loss curves

We employ Grad-CAM to enhance interpretability in analyzing True Positive, False Positive, True Negative, and False Negative cases.

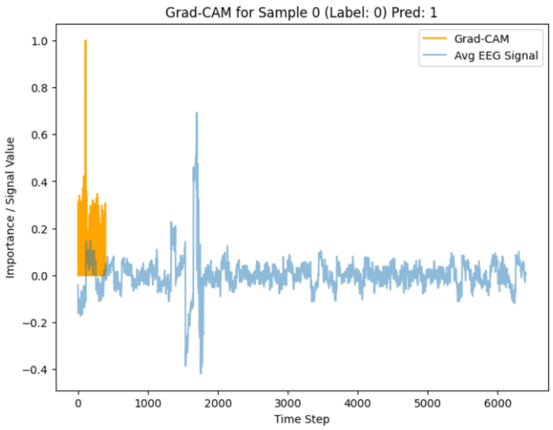

Fig. 15: Grad-CAM for some cases

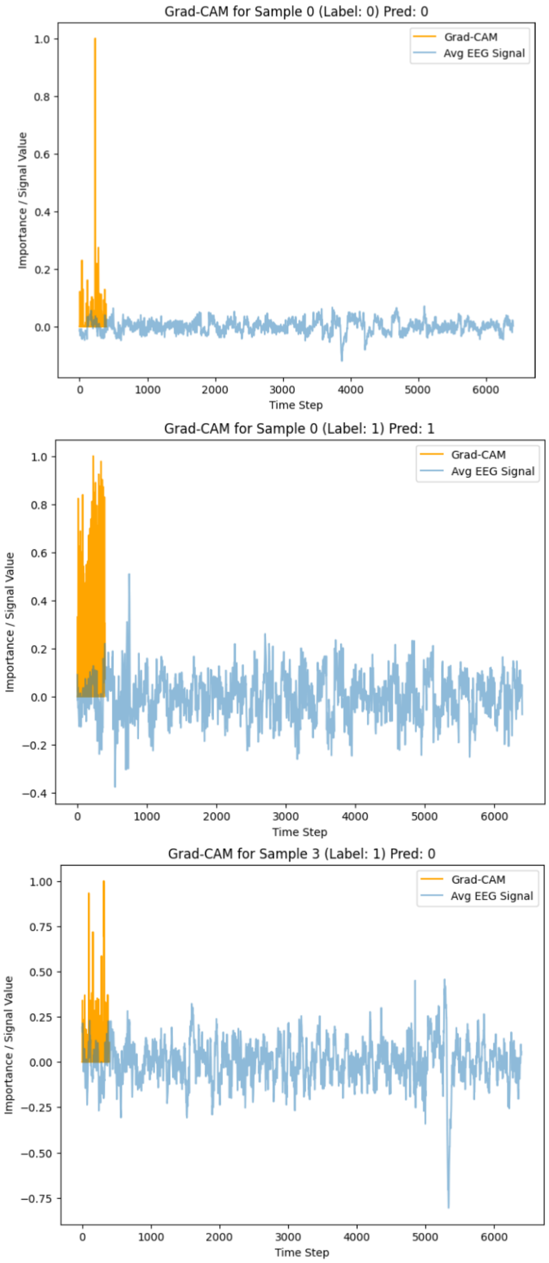

Fig. 16: Grad-CAM for some cases

4. Discussion

4.1. Interpretation of results

1) Spectrogram-base

The suboptimal results might be due to the data not being processed properly or not being transformed into spectrograms, or possibly due to poor feature representations between data samples. Splitting the data into many segments leads to a situation where a single patient has multiple data points in the dataset. For example, with each epoch lasting 60 seconds, a patient could have around 30 epochs, and each epoch contains data from 19 EEG channels. This results in a subject having an excessive number of data points, potentially introducing redundancy and noise that make it more difficult for the model to generalize. Additionally, feeding individual epochs one by one into the model for label prediction may not be the most appropriate approach. Having the model learn and predict based on each separate epoch might lead to biased or misleading results. Therefore, it is necessary to explore methods that can more clearly differentiate between patients with the disease, and to ensure that data between patients remains more independent.

The same issue arises with 2D images. According to previous studies, researchers often split the data into segments of about 4 seconds each. Thus, a 1800-second recording would be divided into 450 small segments, and each small segment corresponds to 4 spectrogram images. This leads to a single patient having an excessive number of images, which can result in irrelevant redundancy. Handling 1,800 images per patient exceeds our computational capacity. Also, we are unsure which methods can help identify segments that are truly relevant for classification. As previously mentioned, random selection might result in inconsistency and errors.

2) Raw EEG data

Training based on raw data ensures that the original properties of the EEG signals are preserved, maintaining their primitive nature. As a result, the performance is slightly higher compared to the other two approaches. Feeding the raw data directly into the model while grouping appropriately helps retain the essential characteristics and reduces data redundancy. However, the classification accuracy remains around 70%, which is still relatively low for medical-related tasks and needs improvement. This may be attributed to architectural limitations as well as the scarcity of data.

In some cross-validation folds, we observed that the model performed well in detecting the NORMAL class, but struggled with MCI classification. This could be attributed to the subtle signal differences between the two groups, making it challenging for the model to learn distinctive patterns. Moreover, with only 61 patients in the dataset, the generalization ability of the model remains limited emphasizing the need for larger and more diverse clinical datasets in future studies.

5. Conclusion

Through the process of researching and implementing a total of three different approaches for processing, analyzing, and extracting features from EEG data, we have gained a more comprehensive understanding of how artificial intelligence and advanced technologies can be applied in the healthcare domain—specifically in the task of classifying MCI and NORMAL groups based on EEG signals. This journey has equipped us with practical knowledge about handling EEG data, from preprocessing and feature extraction to using it as input for modern deep learning architectures. Moreover, we have developed a clearer grasp of the data preparation and training workflows that are crucial when working with sensitive and complex medical data.

Although the current results are still relatively limited, through the practical process of handling and analyzing EEG data—a type of biomedical signal that is both complex and highly variable—we have gained many valuable insights. These experiences deepened our understanding of preprocessing, feature extraction, and the unique challenges of applying deep learning models to healthcare data.

Looking forward, we identify several promising directions for improvement: Selecting more informative EEG segments, rather than relying on fully random sampling; Applying data augmentation techniques to improve the model's generalization capability; Exploring more advanced deep learning architectures that are better suited for modeling the spatiotemporal characteristics of EEG signals.

6. Scientific References

References

[1] Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). *Deep learning for electroencephalogram (EEG) classification tasks: a review*. Journal of Neural Engineering, 16(3), 031001. https://doi.org/10.1088/1741-2552/ab0ab5

[2] B. R. Nayana, M. N. Pavithra, S. Chaitra, T. N. Bhuvana Mohini, Thompson Stephan, Vijay Mohan, and Neha Agarwal. *EEG-based neurodegenerative disease diagnosis: comparative analysis of conventional methods and deep learning models*. Scientific Reports, 15(1):15950, 2025. https://www.nature.com/articles/s41598-025-00292-z

[3] Giulio Ruffini, David Ibañez, Marta Castellano, Laura Dubreuil-Vall, Aureli Soria-Frisch, Ron Postuma, Jean-François Gagnon, and Jacques Montplaisir. *Deep Learning With EEG Spectrograms in Rapid Eye Movement Behavior Disorder*. Frontiers in Neurology, 10:806, 2019. https://www.frontiersin.org/articles/10.3389/fneur.2019.00806/full

[4] Elham S. Salama, Reda A. El-Khoribi, Mahmoud E. Shoman, and Mohamed A. Wahby Shalaby. *EEG-Based Emotion Recognition using 3D Convolutional Neural Networks*. International Journal of Advanced Computer Science and Applications, 9(8), 329–337, 2018. https://thesai.org/Downloads/Volume9No8/Paper_43-EEG_based_Emotion_Recognition.pdf

[5] Shu Lih Oh ,Jahmunah Vicnesh ,Edward J Ciaccio ,Rajamanickam Yuvaraj andU Rajendra Acharya . *Deep Convolutional Neural Network Model for Automated Diagnosis of Schizophrenia Using EEG Signals*. https://www.mdpi.com/2076-3417/9/14/2870

7. Contributions

- Nguyen Duc Anh
  - Extracted stacked spectrogram data and spectrogram images.
  - Trained and evaluated deep learning models combined with interpretability using three different approaches (Raw data, Stacks spectrogram, 2D Spectrogram).
  - Studied related research on EEG data processing and training in deep learning models.
  - Compiled and wrote the final report of results.
  - Explored concepts, terminology, tools, and techniques related to EEG data processing and analysis. Code:
  - Trained models based on stacked spectrogram and extracted stacked spectrogram data and spectrogram images from raw data *code link*
  - Trained model based on raw EEG data *code link*
  - Trained model based on spectrogram images *code link*
- Thai Nguyen Hoang Bach

  - Extracting and Comparing Temporal and Spectral Features of MCI and NORMAL subjects
  - Analyzed temporal features by computing mean values and visualizing both temporal and spectral characteristics; extracted spectrogram images.
  - Studied research papers and created presentation slides.
  - Read and reviewed research papers related to the topic.
  - Provided comments on the research results, identified challenges, and proposed future improvements.
  - *code link*
- Thai Thi Thuy Linh
  - Preprocessed and denoised EEG data.
  - Analyzed and interpreted features related to the Alpha, Beta, Delta, and Theta frequency bands.
  - Compared raw EEG signals over time between two subject groups.
  - Compared average power across frequency bands and channels.
  - Studied research on the effects of various EEG frequency bands.
  - Explored terminology, concepts, tools, and techniques for EEG data processing and analysis.
  - *code link*
- Ha Kim Duong
  - Researched and reviewed studies related to the topic.
  - Studied various training methods.
  - Provided insights into research results, identified challenges, and suggested directions for future improvements.
  - Trained EEG data using deep learning models.
- Nguyen Trong Huy
  - Research on terminology, concepts, tools, and techniques for processing and analyzing EEG data
  - Explore training methods
  - Explore feature extraction methods from EEG data

To improve EEG signal classification performance in detecting mild cognitive impairment (MCI), several coordinated development directions should be pursued. First, the issue of data redundancy caused by splitting EEG signals into numerous short segments can be addressed by **selectively identifying segments that contain meaningful discriminative information**. Techniques such as attention-based selection, unsupervised learning, or feature-based scoring (e.g., using entropy, energy, or spectral features) can help extract only the most informative segments, reducing both noise and computational load—especially when each patient can generate thousands of spectrograms or short windows.

Second, instead of feeding individual segments (epochs or images) into the model, future work should **develop architectures capable of learning global patient-level representations**. For instance, multiple consecutive epochs can be grouped and passed through RNNs, BiLSTMs, or transformers to capture temporal relationships. Techniques such as attention pooling or multi-instance learning can help the model focus on important segments without needing segment-level labels.

Third, improving generalizability requires **expanding and diversifying the dataset**. It is crucial to collect EEG datafrom various clinical sites, including patients at different stages of disease progression and across demographic factors (e.g., age, gender), to enhance the representativenessof the model. Moreover, data augmentation techniques adapted to physiological signals - such as jittering, time warping, frequency masking, or contrastive learning - can enrich the training data and improve robustness.

Finally, **designing EEG-specific deep learning architectures** is essential. Models like EEGNet, DeepConvNet, or custom transformer variants with multi-channel attention mechanisms can better leverage the unique spatial-temporal properties of multi-channel EEG data. Importantly, **patient-wise evaluation methods** such as leave-one-subject-out cross-validation should be used to avoid data leakage and ensure the model's ability to generalize to unseen patients. This is especially critical for real-world clinical deploymen, where the system must reliably process entirely new data.