Multimodal EEG and Spectrogram Model for Harmful Brain Activity Classification

Haoyang Ling (hyfrankl@umich.edu)
Sijun Tao (sijuntao@umich.edu)
Yifei Zhang (yifeizh@umich.edu)

April 24, 2024

1 Introduction

Electroencephalography (EEG) is vital for doctors to track brain activity. By putting electrodes on the scalp, EEG allows a non-invasive way to detect brain signals produced by the firing of neurons in the brain [1]. EEG is an effective tool to help identify harmful brain activities that can cause more damage if untreated [2]. This includes seizures, periodic discharges (generalized or one-sided), and rhythmic delta wave activity (generalized or one-sided).

However, the current method requires specialized neurologists to manually review EEG recordings, which is a very labor-intensive and time-consuming process. They have to stare at the squiggly lines on an EEG readout for hours looking for abnormalities. It is mentally taxing and prone to errors from fatigue and differences between experts' interpretations [3]. This bottleneck in EEG analysis is a major obstacle. It hinders efficient, reliable neurocritical care. Developing automated computer systems to identify these brain patterns from EEG data could revolutionize the process [4]. Automated EEG classification could detect issues more quickly, reduce the workload for neurologists, minimize human error, and ensure consistent analyses.

In the context of automated classification, both accuracy and interpretability are crucial. Since the results can directly impact a patient's health and treatment, high interpretability ensures that the findings are reasonable and trustworthy [5]. The ultimate goal of our project is to develop an automated process for EEG classification that achieves high accuracy while maintaining a high level of interpretability.

1.1 Related Works

For years, scientists have been creating automated systems to analyze EEGs and classify patterns of brain activity. Early methods took a traditional machine learning approach. First, they pulled out handcrafted statistical features from the EEG data. Then, they trained models like random forests or support vector machines on those features. The models aimed to identify patterns like periodic discharges or rhythmic delta waves [6, 7]. Studies using this feature engineering approach showed some success but also faced limitations. Extracting the right features requires expert neurological knowledge and extensive manual effort. And the handcrafted features may not fully capture the complexity of real EEG signals [8].

More recently, advances in deep learning have opened up new possibilities for automated EEG analysis. Deep neural networks like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can learn to identify patterns from raw EEG readings and spectrograms, which show frequency information over time [5, 8]. Several research teams have utilized these deep learning techniques to detect specific patterns such as seizures, periodic discharges, and delta activity [9, 10]. The deep neural models outperformed traditional machine learning methods by uncovering more detailed predictive features in the data. Even with these new deep-learning models, ambiguous EEG patterns pose a

significant challenge. Experts often disagree on how to classify them, blurring category boundaries. These "edge cases" and "proto-patterns" are major obstacles to reliable automated detection [11].

More innovation is needed to develop robust, reliable, and widely applicable EEG analysis systems. Such systems can significantly aid neurocritical care and reduce expert workloads.

2 Data

2.1 Data Overview

The data used in this project is sourced from Kaggle [12]. The original data we selected consists of three parts: EEG data, spectrograms, and metadata which is provided in a csv format.

a. EEG Data: EEG data captures voltage fluctuations resulting from ionic current flows within neurons of the brain. The EEG data used in this project consists of multiple overlapping samples, each uniquely identified by an EEG ID. These recordings were collected at a frequency of 200 samples per second. Each EEG recording covers a specific time duration, and metadata provides details on individual segments annotated by expert consensus.

b. Spectrograms: Spectrograms are visual representations derived from EEG data, illustrating changes in frequency and amplitude over time. In this project, spectrograms are constructed from assembled EEG data, comprising one or multiple EEG recordings. The spectrogram metadata includes detailed frequency information and recording regions of EEG electrodes, such as left lateral (LL), right lateral (RL), left parasagittal (LP), and right parasagittal (RP). Since our study mainly focuses on the slow wave, we use the Butterworth low-pass filter to only study the wave frequency within 20 Hz.

c. Metadata: This csv file contains essential metadata for the training set, linking EEG recordings to their respective spectrograms. Important fields include EEG and spectrogram IDs, patient IDs, and expert consensus labels for various brain activity classes.

2.2 Data Loader

The data loader module is crucial for efficiently loading and preparing EEG data for training and validation. It seamlessly handles both raw EEG signal data and their corresponding spectrograms, ensuring that the data is ready for the machine learning model.

The core functionality of the data loader is contained within a dataset class, which is responsible for indexing and retrieving data samples from the dataset. When an item is accessed from the dataset, the data loader performs the necessary preprocessing steps on the EEG signal data and spectrogram. The preprocessing of the EEG signal data involves several key steps. These include computing differences between specific channel pairs to capture relevant information, normalizing the data to ensure consistent scaling, applying a low-pass filter to remove high-frequency noise, and quantizing the data to reduce its size. These preprocessing techniques collectively contribute to improving the signal-to-noise ratio and preparing the data for input to the model. Similarly, the spectrogram data is processed to align with the model's input channels. This may involve repeating the spectrogram along the channel dimension or applying specific transformations to extract meaningful features.

By centralizing the data loading and preprocessing logic within the data loader module, it becomes more manageable to feed the EEG data to the model effectively. This modular approach enhances code reusability and maintainability, enabling efficient experimentation and iteration during the development process. It allows for a clear separation of concerns, with the data loader focusing on data preparation, while the model can concentrate on learning and making predictions based on the preprocessed data.

3 Methodology

3.1 Model

In this section, we would like to present the design and training framework of our multimodal model that leverages the information from waveforms and spectrograms. This design is because waveforms and spectrograms can provide both temporal and frequency resolution [13]. Waveforms are useful for detecting abrupt changes, while spectrograms highlight frequency patterns associated with brain activity. Based on this design decision, we chose the two most commonly used methods on waveforms and spectrograms: WaveNet and EfficientNet. In the following subsections, we would like to introduce these two baseline models.

3.1.1 Baseline Models

WaveNet: The original version is a deep learning architecture developed by Google DeepMind with valuable applications in EEG detection [14], which is designed for text-to-speech synthesis. However, many medical researchers have uncovered that WaveNet excels at extracting intricate patterns from waveform data, particularly in EEG detection [15, 16]. At its core, WaveNet is based on the PixelCNN architecture and employs dilated causal convolutions to handle long-range dependencies [17]. PixelCNN operates on the concept of autoregression, generating images pixel by pixel while utilizing masks to ensure each pixel only considers information from previously generated pixels, avoiding future information [17]. This stands in contrast to traditional CNNs, which process entire images at once. Besides, the idea of dilation facilitates the creation of filters with broader receptive fields, enabling the model to capture multi-scale features efficiently, shown in Fig. 1.

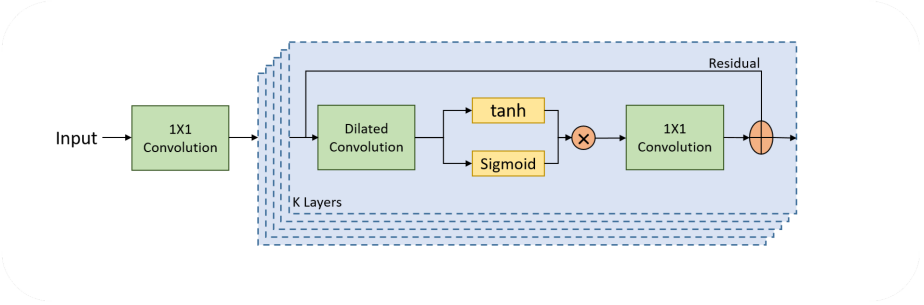


Figure 1: Block diagram of the WaveNet architecture [15]

This architecture design enables WaveNet to perceive relationships between distant elements within a sequence, which proves valuable in scenarios where abrupt changes occur at recurring intervals. Moreover, medical researchers often augment WaveNet with recurrent neural networks to enhance the extraction of temporal information, enabling the model to discern evolving patterns over time. In our project, while modifications to the kernel size in convolutional neural networks are made for performance enhancements, the core structure of the model remains largely unchanged.

EfficientNet: It is a well-known convolutional neural network (CNN) architecture by Google Research [18], which aims for fewer parameters and computational resources without loss of prediction accuracy. Moreover, its model structure is scalable to achieve better performance. Due to these features, we choose it as our backbone model. The following picture shows the general structure of EfficientNet. The main component of EfficientNet is mobile inverted bottleneck convolution denoted as MBConv in the Fig. 2 and squeeze-and-excitation (SE) block. They respectively reduce the computational complexity and help extract channel-wise features.

Due to EfficientNet's efficiency and effectiveness, it is a powerful tool for EEG detection using spectrograms. In our project, we use the same structure; however, we stack the EEG spectrograms from different brain regions as different channels of the input data. It aligns the timeline of different channel information, which is more helpful than aggregating information in the same channel layer.

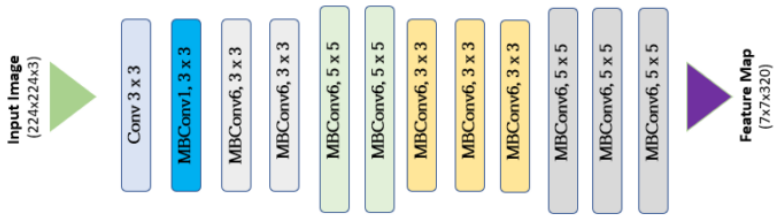


Figure 2: EfficientNet architecture [19]

3.1.2 Training Framework

The core of our training framework is to employ a multimodal fusion approach with a contrastive loss function. This approach differs from traditional contrastive learning, which typically aims to find similarities between pairs of data. In our case, we seek to augment the training data with information from both the waveform and spectrogram modalities.

The rationale behind this approach is that in the common setting of contrastive learning, the training process is often divided into two stages. First, the model is trained to find similarities between pairs of data, and then the learned features are used to train the final model. In contrast, our approach uses a one-stage learning process, which eliminates the need for this burdensome two-stage training.

In our framework, we use the downstream loss of each backbone model to supervise the feature extraction, allowing the models to learn more informative representations from the latent space. Additionally, we include the multimodal output as part of the overall loss function, as we believe that multimodal learning can help improve the model's performance.

To avoid the issue of multimodal collapse, where both backbone models extract similar information, we introduce a contrastive loss function for the positive pairs. This contrastive loss aims to lower the cosine similarity between the positive pairs, which in turn helps the linear dependence of the feature inputs to the multimodal component, ultimately improving the overall model performance. Moreover, we also introduce non-linearity to the classification layer with the activation function GeLU. In summary, the overall loss is defined as

$$\text{Loss} = \sum_{i=1}^{N} \alpha \text{L}(\text{P}_\text{w}(x_i^w), y_i) + \alpha \text{L}(\text{P}_\text{s}(x_i^s), y_i) + (1-2\alpha)\text{L}(\text{P}(x_i^w, x_i^s), y_i) + \beta \text{Sim}(x_i^w, x_i^s),$$

where N is the number of data samples we have; x_i^w and x_i^s are respectively features extracted from the i^{th} waveform data and i^{th} spectrogram; P_w, P_s, and P define the parameters in the classification layer respectively of the waveform backbone model, the spectrogram backbone model, and the multimodal model; L is the classification loss function; Sim define the similarity function; α and β are hyper-parameters in this equation. In this project, we use $\alpha = 0.25$ and $\beta = 0.5$.

In summary, the proposed model architecture shown in Fig. 3 leverages multimodal information for classification tasks. The input data, comprising raw waveform signals and spectrograms, are processed through separate feature extraction pathways. The waveform data is passed through a WaveNet-based feature extractor, while the spectrogram data is processed by an EfficientNet-based encoder. These modality-specific features are then fused to learn the extra information through the contrastive loss function. Features from separate backbone models as well as fused features are then separately passed through different classification layers to generate the aggregated final predictions. It aims to leverage the complementary information from both the waveform and spectrogram modalities to enhance the overall model performance.

3.2 Loss Function

In the aforementioned model, we employed Kullback-Leibler Divergence (KLD) as our selected loss function [20]. KLD is a measure used to quantify how one probability distribution diverges from an

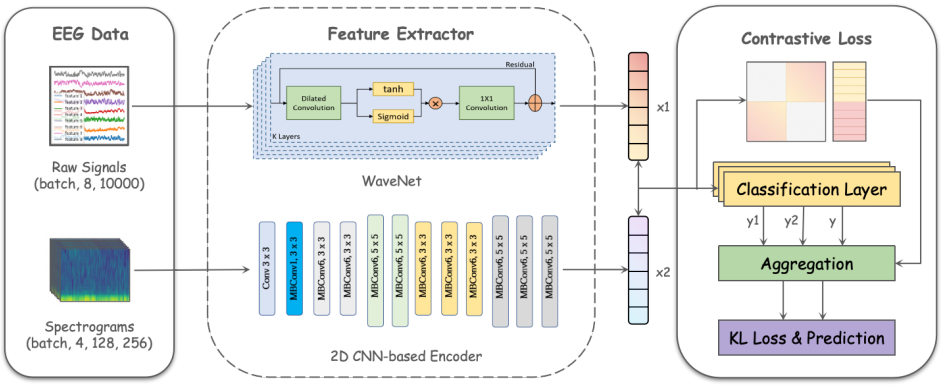


Figure 3: Proposed Multimodal Architecture

expected probability distribution. It measures the information lost when one distribution is used to approximate another. Since our project is a soft multi-class classification task that produces probabilistic predictions, KLD is well-suited for our needs. Based on this consideration, we selected KLD as our loss function, which is defined as:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

where $P(x)$ is the probabilistic predictions generated in each step in the training process and $Q(x)$ is the true possibilities of the classes.

3.3 Optimization & Evaluation Strategies

In order to facilitate the selection of hyperparameter settings that maximized our model effectiveness and to enable a thorough evaluation of the models, we utilized rigorous optimization and evaluation strategies.

Using grid search and random search techniques, hyperparameter optimization was carried out, examining various combinations of hyperparameters such as batch size and learning rate to find setups that optimize model performance. We leveraged validation metrics including accuracy, precision, recall, and F1-score to guide the optimization process and select the most effective hyperparameter settings.

To evaluate the model performance, we adopted a stratified k-fold cross-validation approach with a value of k equals to 5. This strategy was employed to minimize potential bias and variance in model assessment. It partitions the dataset into k subsets while preserving the percentage of samples for each class in each fold. By rotating the dataset partitions during training and validation, we assessed model generalization across diverse subsets of the data. Average metrics computed across folds provided reliable estimates of model performance.

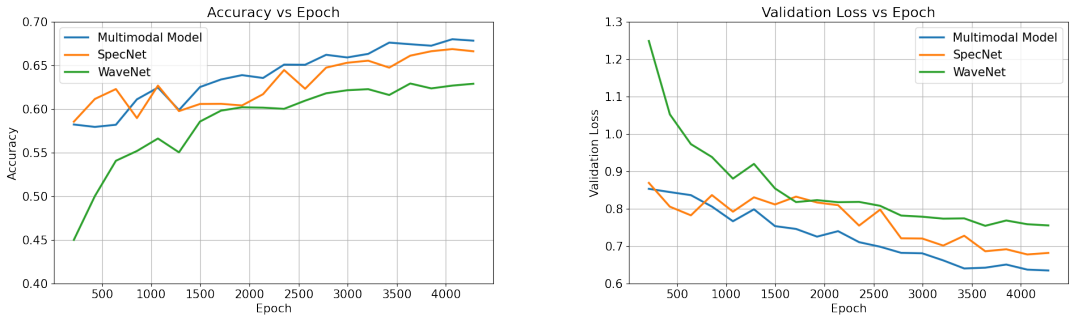
4 Results

In this project, we aim to conduct a thorough empirical evaluation of our proposed multimodal model and our baseline models (SpecNet and WaveNet). SpecNet stands for the model that operates on spectrograms using EfficientNet and WaveNet stands for the model that processes raw EEG signals using WaveNet. By comparing the performance of these models, we seek to determine whether multimodal learning offers benefits over unimodal approaches for our specific application.

4.1 Accuracy and Validation Loss

Fig.4a and Fig.4b shows the validation accuracy and validation loss logged during training for each model. We can see from the figures that our proposed multimodal model achieves the highest accuracy, reaching nearly 70% by the end of training, substantially outperforming both SpecNet and WaveNet.

Also, our proposed model attains the lowest validation loss, decreasing to around 0.7 in the final epochs, while SpecNet and WaveNet exhibit higher losses.



(a) Accuracy v.s. Epoch

(b) Validation Loss v.s. Epoch

Figure 4: Accuracy and Validation Loss

These results provide strong evidence that integrating information from multiple modalities enables the multimodal model to make more accurate predictions and reduce errors on unseen data compared to the unimodal baselines.

4.2 Radar plot Comparison

Fig. 5 shows the accuracy score of different models on six categories. As illustrated in Fig. 5, our multimodal model consistently outperformed the two baseline models across all categories, demonstrating superior accuracy and robust performance. Notably, our model achieved exceptional accuracy levels, exceeding 80% in the *GPD* and *Other* categories, and surpassing 60% in *LPD*, *GRDA*, and *Seizure* categories.

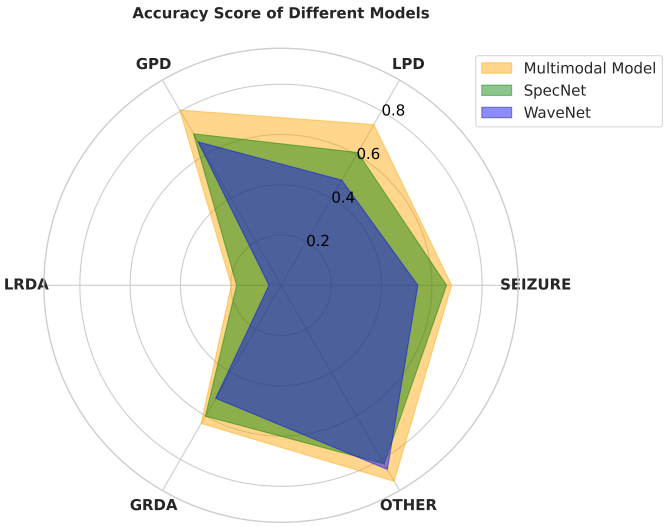


Figure 5: Accuracy Score of Different Models on Six Categories

These results underscore the effectiveness of our model in accurately classifying diverse types of harmful brain activity based on EEG signals and spectrograms. The consistent performance across multiple categories highlights the versatility and potential of our model for real-world applications in neurocritical care and EEG analysis. These findings suggest promising prospects for improving diagnostic accuracy and patient outcomes in clinical settings.

4.3 Confusion Matrices

To gain deeper insight into the prediction patterns of our model, we generated confusion matrix considering only out-of-fold (OOF) samples with labels above 0.5. This focus on confident instances allows us to assess the models' performance on the most important cases. This can also reduce the impact of the examples that even human experts can't really agree on.

In our case, it's a multi-class setting where the performance on each class can vary significantly. Therefore, it's crucial to evaluate confusion matrices. As shown in the plot (Fig. 6), our model achieves strong performance on most classes, with the deepest color along the diagonals. However, for the "LRDA" class, the model's performance is lower compared to other classes, despite getting most of the results correct. This can be attributed to the original data distribution, where only 5% of the data belongs to the "LRDA" class. It's also worth highlighting the few misclassifications and the low number of false positives for the "Other" class, which demonstrates the model's ability to provide reliable predictions instead of overly categorizing cases into the "Other" class.

These results highlight the potential of the multimodal model for accurate and trustworthy EEG classification in real-world applications.

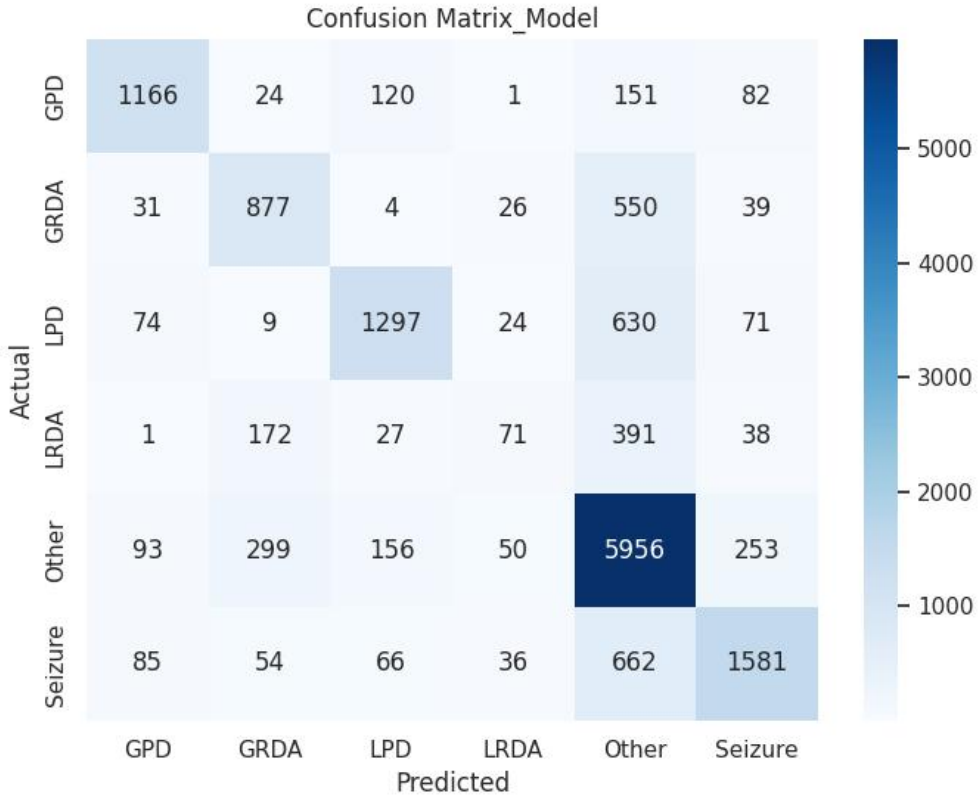
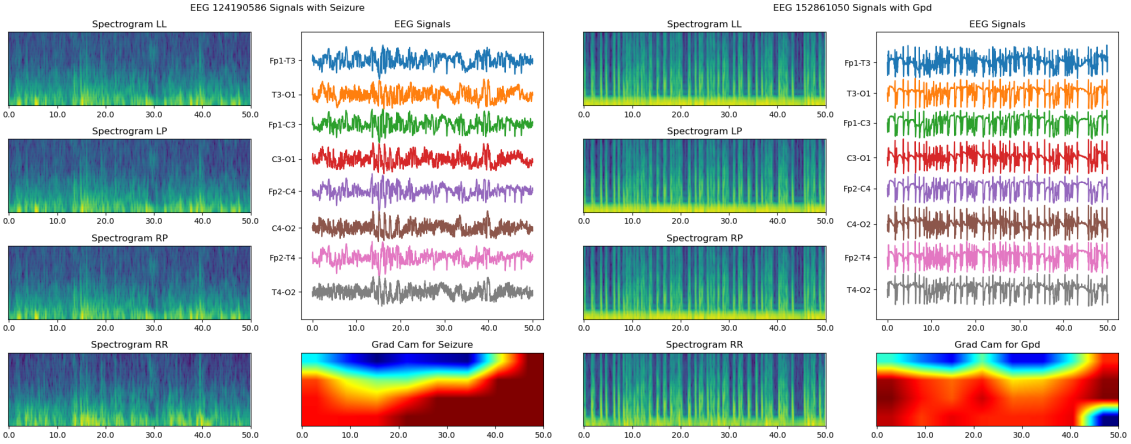


Figure 6: Confusion Matrix

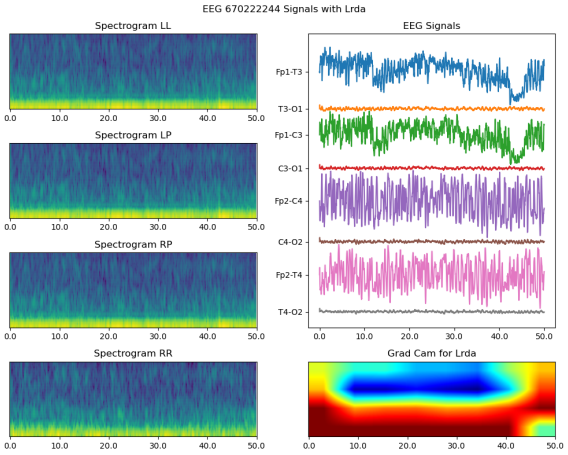
4.4 Interpretability

As we've emphasized before, interpretability is crucial for the project since it'll be used on patients. Here we employed Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM is a technique that highlights the regions of an input (in this case, the spectrogram and EEG signals) that are most important for the model's prediction of a particular class [21]. This allows us to understand which patterns and features the model is focusing on when making its predictions.

As shown in Fig.7a, Fig.7b, and Fig.7c, we applied Grad-CAM to visualize the model's attention for different classes.



(a) Seizure (b) GPD



(c) LRDA

Figure 7: Grad-CAM for different classes

For the Seizure class, the Grad-CAM heat map highlights sharp spikes and high-amplitude patterns in specific channels of the EEG signal, along with corresponding high-energy regions in the spectrogram. For the "GPD" class, the model focuses on the diffuse spikes in the EEG signals and the corresponding energy patterns in the spectrograms. And for the "LRDA" class, the Grad-CAM heat map shows the model attending to low-amplitude, diffuse activity across multiple EEG channels (which shows rhythmic) and lower energy levels across the spectrograms. These patterns are consistent with the clinical description of the activities, which proves that our model is learning to focus on clinically relevant patterns and features specific to each condition [22, 23].

This interpretability can facilitate the adoption of our model as a decision support tool, as clinicians can verify that the model's predictions are based on clinically relevant patterns and features, increasing their confidence in the system's outputs.

5 Discussion

The proposed multimodal architecture for EEG-based classification shows considerable promise in enhancing performance. However, it also presents notable limitations and ethical considerations that demand thoughtful examination.

5.1 Limitations of the Multimodal Design

One challenge is the risk of "multimodal collapse", where the feature extractors for waveform and spectrogram modalities may converge, potentially undermining the benefits of multimodal fusion. Although the current design employs a contrastive loss to mitigate this issue, additional research and theoretical exploration are needed to ensure the robustness of this approach. People can explore more on the choice of losses rather than cosine similarity.

Moreover, while WaveNet and EfficientNet serve as well-established baseline models, their selection might constrain the architecture's capacity to capture nuanced EEG data patterns. Exploring alternative feature extraction architectures or crafting custom models tailored to EEG signal characteristics could enhance performance. Transformer structure might be one potential direction of further explorations.

Additionally, as in the previous section, we have observed model prefers to predict others rather than LRDA symptoms. The under-represent class LRDA causes the model biasedness. It warns us to do more feature engineering and advanced sampling strategies to handle the data imbalance.

5.2 Ethical Considerations

The use of automated EEG analysis systems gives rise to some ethical concerns that necessitate careful consideration. EEG data harbors highly sensitive information on a patient's brain activity, carrying substantial privacy implications. Thus, data security and explicit patient consent protocols are required to ethically handle such sensitive information. We may also need to be aware of

Furthermore, excessive reliance on automated tools by healthcare practitioners might diminish critical thinking and clinical judgment. While the model offers valuable insights, it should serve as a decision support tool, with clinicians retaining ultimate authority and responsibility for patient diagnosis and treatment. Establishing appropriate training and guidelines for system use is crucial to mitigate the risks of overconfidence or over-reliance on automated output.

5.3 Conclusion

In conclusion, the proposed multimodal architecture demonstrates promising results in EEG-based classification. However, addressing the identified limitations and ensuring the ethical deployment of such technologies in neurocritical care requires further research and collaboration. Ongoing engagement with stakeholders, including clinicians, researchers, and ethicists, is essential to navigate these challenges and fully realize the innovation's potential while upholding ethical principles and safeguarding patient well-being.

References

[1] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[2] L. J. Hirsch and R. P. Brenner, *Atlas of EEG in critical care*. John Wiley & Sons, 2011.

[3] J. J. Halford, D. Shiau, J. Desrochers, B. Kolls, B. Dean, C. Waters, N. Azar, K. Haas, E. Kutluay, G. Martz, *et al.*, "Inter-rater agreement on identification of electrographic seizures and periodic discharges in icu eeg recordings," *Clinical Neurophysiology*, vol. 126, no. 9, pp. 1661–1669, 2015.

[4] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[5] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.

[6] A. Subasi and M. I. Gursoy, "Eeg signal classification using pca, ica, lda and support vector machines," *Expert systems with applications*, vol. 37, no. 12, pp. 8659–8666, 2010.

[7] R. Sharma and R. B. Pachori, "Classification of epileptic seizures in eeg signals based on phase space representation of intrinsic mode functions," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1106–1117, 2015.

[8] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[9] A. Antoniades, L. Spyrou, C. C. Took, and S. Sanei, "Deep learning for epileptic intracranial eeg data," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2016.

[10] R. Hussein, H. Palangi, R. K. Ward, and Z. J. Wang, "Optimized deep neural network architecture for robust detection of epileptic seizures using eeg signals," *Clinical Neurophysiology*, vol. 130, no. 1, pp. 25–37, 2019.

[11] A. A. Ein Shoka, M. M. Dessouky, A. El-Sayed, and E. E.-D. Hemdan, "Eeg seizure detection: concepts, techniques, challenges, and future trends," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42021–42051, 2023.

[12] Kaggle, "Hms - harmful brain activity classification," 2024.

[13] M. C. Ng, J. Jing, and M. B. Westover, "A primer on eeg spectrograms," *Journal of Clinical Neurophysiology*, vol. 39, no. 3, pp. 177–183, 2022.

[14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[15] H. Albaqami, G. M. Hassan, and A. Datta, "Automatic detection of abnormal eeg signals using wavenet and lstm," *Sensors*, vol. 23, no. 13, p. 5960, 2023.

[16] H. Pankka, J. Lehtinen, R. Ilmoniemi, and T. Roine, "Forecasting eeg time series with wavenet," *bioRxiv*, pp. 2024–01, 2024.

[17] A. van den Oord, N. Kalchbrenner, L. Espeholt, k. kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[18] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and

R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 09–15 Jun 2019.

[19] W. ML, "Efficientnet and its performance comparison with other transfer learning networks," 2023.

[20] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–317, IEEE, 2007.

[21] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016.

[22] N. Kane, J. Acharya, S. Beniczky, L. Caboclo, S. Finnigan, P. W. Kaplan, H. Shibasaki, R. Pressler, and M. J. van Putten, "A revised glossary of terms most commonly used by clinical electroencephalographers and updated proposal for the report format of the eeg findings. revision 2017," *Clinical neurophysiology practice*, vol. 2, p. 170, 2017.

[23] J. Claassen, N. Jette, F. Chum, R. Green, M. Schmidt, H. Choi, J. Jirsch, J. Frontera, E. S. Connolly, R. Emerson, *et al.*, "Electrographic seizures and periodic discharges after intracerebral hemorrhage," *Neurology*, vol. 69, no. 13, pp. 1356–1365, 2007.

Appendix

Please refer to our GitHub repository for detailed information about codes: Our GitHub repository.