Harmful Brain Activity Classification

Keen Huei Liew KHL456@NYU.EDU
Deep Learning in Medicine
New York University
New York, NY, USA

Daria Chylak DUC9244@NYU.EDU
Deep Learning in Medicine
New York University
New York, NY, USA

Abstract

The Harmful Brain Activity Classification project aims to automate EEG-based classification of harmful brain activities, reducing reliance on manual interpretation which is time-consuming and prone to human error. This method promises more efficient diagnoses, cost reduction, and improved patient outcomes in neurology. We based our analysis and methodology on a Kaggle competition.

Keywords: EfficientNetB0, EEG

1 Introduction

The Harmful Brain Activity Classification project endeavors to revolutionize neurological diagnostics by leveraging EEG-based machine learning, thereby diminishing reliance on manual interpretation for quicker and more accurate diagnoses (Chung et al., 23). Through the utilization of EfficientNet, the model automates EEG analysis, thereby standardizing readings and potentially elevating diagnostic reliability.

The hypothesis in this classification project is that different types of harmful brain activity or other abnormal patterns may exhibit distinct EEG signatures. By training a deep learning model on EEG recordings, we can classify different types of harmful brain activity based on their unique EEG patterns. Specifically, we anticipate that our model will be able to differentiate subtle variations in EEG patterns associated with different neurological conditions.

Currently, EEG analysis requires extensive manual labor, where trained neurologists examine complex waveforms to identify pathological activity. This method is not only time-consuming but also prone to inconsistencies due to subjective interpretations by different experts. By deploying a model based on EfficientNet, which excels in handling intricate image and signal data, our approach automates the detection process, thereby standardizing the readings and potentially increasing the diagnostic reliability.

In literature, various machine learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), exhibit promise in seizure detection (Tsiouris et al., 2018; Thodoroff et al., 2016). The demonstrated potential of EfficientNet in PET scan analysis (Alsubai et al., 2024; Ding et al., 2018) suggests an avenue for improved EEG pattern detection. Personalized models (Shoeb Guttag, 2010) and transfer learning (Han et al., 2021) offer avenues for boosting accuracy while addressing challenges associated with data labeling. Ethical considerations, including patient data security, remain paramount (Nadhan et al., 2024), emphasizing the need for a balanced approach to achieve meaningful transformation in EEG analysis.

2 Data

This dataset was ready to use with no access restrictions. The dataset consisted of 17,089 EEGs recordings (50 seconds each) and 11,138 spectrograms (10 minutes each) taken from 1,950 patients. Experts reviewed 50 second long EEG samples along with matched spectrograms covering a 10-minute window centered at the same time, and labeled the central 10 seconds. To be more specific, with a center point called T, the EEG time window was 50 seconds long [T-25 : T+25] and the spectrogram time window was 600 seconds long [T-300 : T+300], and both the EEG and spectrogram had the same center point timestamp. The task was to predict the event occurring in the middle 10 seconds of both these time windows [T-5 : T+5].

3 Materials and Methods

We generated the following pipeline to detect and classify types of harmful brain activities with EEG data.

3.1 Data preprocessing

We began by importing dataset information from a CSV file and encoded our target column `('expert_consensus')` into numerical values. EEG data was then grouped by `eeg_id` to extract relevant details like spectrogram IDs, patient IDs, and expert consensus labels. Grouping at the `eeg_id` level encourages model generalization across different instances of brain activity. After obtaining this non-overlapping `eeg_id` data, we split the data into train (80%) and test (20%) sets. This split was done after grouping by `eeg_id` to ensure that each EEG recording was used exclusively in either the training or testing phase. Finally, we loaded EEGs and spectrograms stored as `.parquet` files. Visualization of spectrograms was performed to verify data accuracy.

3.2 CustomDataset and DataLoader

The `CustomDataset` class loaded and preprocessed the data for training, validation, or testing the neural network model. It ensured correct formatting and transformation before feeding data into the model. The class processed EEGs and EEG spectrograms into an 8-channel image (128, 256, 8). This consolidation enabled the model to learn from multiple data modalities simultaneously and simplified the model architecture. Finally, a `DataLoader` was defined to load data from `CustomDataset` for visualization of post-processed spectrogram images.

3.3 EfficientNetB0 model

The training process utilized EfficientNet, known for its efficiency and effectiveness in handling complex image data. `CustomModel`, utilizing the `timm` library, constructed the model architecture (specified as EfficientNet in the earlier `config` class), with pretrained weights. A fully connected linear layer, `classifier`, was defined for class prediction.

Additionally, we reshaped the input tensor from `(128, 256, 8)` to `(512, 512, 3)` to suit the model. This method separated the input tensor into spectrograms (first 4 channels) and EEG signals (last 4 channels), concatenated them along the channel dimension and where the tensor is duplicated to create three identical channels, and permuted the dimensions to match the model's input expectations.

3.4 Training loop

We initiated folds with `GroupKFold`, iterating over each epoch and fold. Each epoch traverses the entire training dataset, while each fold represents a distinct split for cross-validation, enhancing model robustness and generalization to new data. During training, data batches were loaded using `train_loader`, and each batch was passed through the model to obtain predictions. The `CrossEntropyLoss` function computed the loss between predictions and actual labels, suitable for multi-class classification. Within the training loop, we calculated training loss and accuracy for the current epoch and fold, providing insights into model learning and updating parameters to minimize loss. Additionally, a `ReduceLROnPlateau` scheduler adjusted the learning rate as training progressed, ensuring model convergence to an optimal solution.

4 Results

The overall performance metrics of our classification is as follows:

`Accuracy: 0.4847, Precision: 0.7490, Recall: 0.2412, F1 Score: 0.2279`

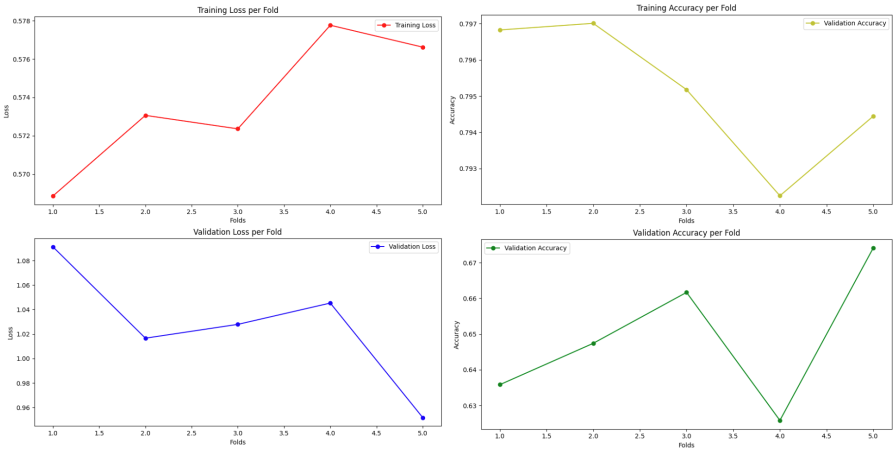
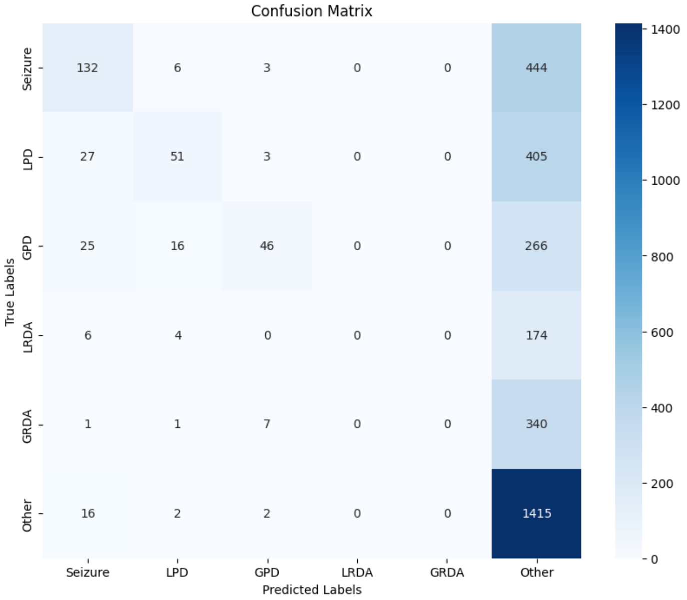


Figure 1: Performance metrics from the model training process across different folds.



	Seizure	LPD	GPD	LRDA	GRDA	Other
Seizure	132	6	3	0	0	444
LPD	27	51	3	0	0	405
GPD	25	16	46	0	0	266
LRDA	6	4	0	0	0	174
GRDA	1	1	7	0	0	340
Other	16	2	2	0	0	1415

Figure 2: Confusion matrix visualizes the performance of our classification model across various categories of harmful brain activities, illustrating both correct predictions (diagonal) and misclassifications (off-diagonal).

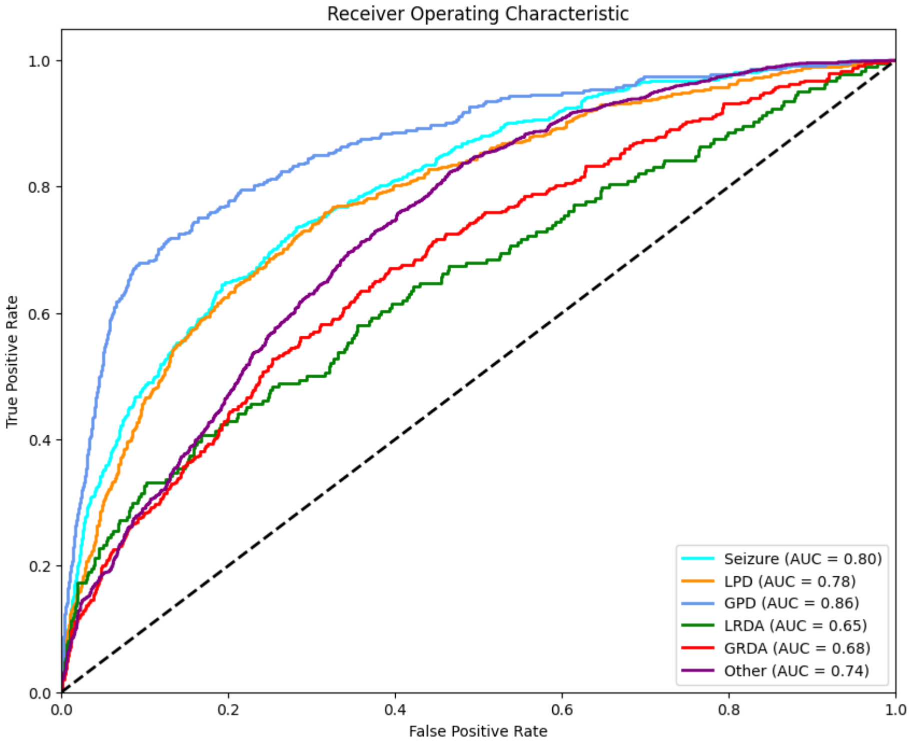


Figure 3: Receiver Operating Characteristic (ROC) curve illustrates the performance of our model in distinguishing between different types of harmful brain activities.

5 Discussion

5.1 Model evaluation

5.1.1 Training and validation curves

The training loss decreases consistently across epochs in each fold which is typical and desired during training, whereas the validation loss exhibits some fluctuations across different folds, suggesting that the model might be sensitive to the specific subsets of data used in each fold. The training accuracy consistently outperforms validation accuracy, which might suggest overfitting to the training data.

5.1.2 Overall performance metrics

The overall performance metrics of our classification model, displaying an accuracy of 48.47%, precision of 74.90%, recall of 24.12%, and an F1 score of 22.79% indicate a need for improvements in model sensitivity and predictive balance. While the model is reasonably precise in the labels it predicts to be positive, it fails to identify a significant portion of actual positive cases (low recall). The F1 score, which balances precision and recall, is notably low, indicating that the model is not effectively capturing the positive class in the dataset.

5.1.3 Confusion matrix

The confusion matrix reveals some specific strengths and weaknesses. For the Seizure class, there is high accuracy in predicting seizures but some misclassification with LPD and other classes. For LPD, there is significant confusion with Seizure and GPD classes. For GPD, there is decent identification but some confusion with LPD. For LRDA and GRDA, hese classes show considerable misclassification, often predicted as Other. For the Other class, there is a very high true positive rate, suggesting that the model effectively identifies this class but could potentially be overfitting to features specific to this class or underfitting other classes.

5.1.4 ROC Curve and AUC

The ROC curves for each class display varying degrees of discriminative ability, where Seizure (AUC = 0.80) and GPD (AUC = 0.86) have better performance in distinguishing these conditions from others. Other classes including LPD, LRDA, GRDA, and Other have lower AUC values (between 0.65 and 0.78), indicating weaker performance in classification.

5.2 Model evaluation discussion

The results from our model reveal both promising capabilities and notable limitations. The performance metrics across different folds demonstrate that while the model achieves a reasonable level of accuracy in some instances, there are fluctuations across different validation folds, as shown by the oscillating validation loss and accuracy. This variability suggests that our model's performance is sensitive to the data's partitioning, indicating potential overfitting to certain features in the training data or inherent complexities within the EEG patterns that are not consistently recognized by the model. Particularly telling is the confusion matrix, which provides a granular view of the model's classification ability across different types of brain activity. The high degree of accuracy in detecting 'Other' activities suggests that the model is effective at identifying general, non-specific brain activity patterns. However, the low accuracy for more clinically significant categories such as 'LRDA' and 'GRDA' and the substantial misclassification between 'Seizure', 'LPD', and 'GPD' categories highlight critical weaknesses. These errors could potentially lead to misdiagnoses in a clinical setting, emphasizing the need for refined training approaches, possibly incorporating more nuanced features or a more balanced dataset that better represents these rarer categories. The Receiver Operating Characteristic (ROC) curves provide an optimistic view, indicating decent model performance in distinguishing between classes. However, the practical application of such models requires not just high individual class performance but a robust capability to differentiate between all possible conditions accurately. The insights gained from this project underline the importance of continuous model evaluation and iteration. Future research should focus on enhancing data preprocessing, feature engineering, and exploring advanced model architectures or ensemble methods that might improve the generalization and reliability of the model across all categories of brain activity.

To improve model performance, several strategies can be employed:

- Hyperparameter optimization: Fine-tune settings like batch size, learning rate, epochs, and folds to enhance learning without overfitting.
- Enhanced validation and regularization: Implement stricter cross-validation and regularization methods to mitigate overfitting.
- Ensemble methods: Use ensemble techniques to gather predictions from several different models - an combination can improve accuracy and prediction stability.

These adjustments aim to refine training methods, better capture complex EEG patterns, and ensure more balanced representation across various brain activity categories.

6 Team Work

Daria Chylak and Keen Huei Liew made significant and equal contributions to their project. Daria handled data preprocessing and was in charge of presentations and the final paper. Her development and fine-tuning of preprocessing algorithms significantly improved the data quality for the models. Keen focused on the model's definition and evaluation, designing the neural network architecture for optimal performance and conducting detailed analyses to iteratively refine the model. Both members also independently implemented and ran models, facilitating robust comparisons and the integration of insights, which enhanced the understanding and performance of the overall model. Their combined efforts were crucial for the project's success.

References

Alsubaie et al. Alzheimer's disease detection using deep learning on neuroimaging: A systematic review, 2024a. URL https://www.mdpi.com/2504-4990/6/1/24.

Chung et al. Deep learning-based automated detection and multiclass classification of focal interictal epileptiform discharges in scalp electroencephalograms. *Scientific Reports*, 13 (1):6755, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-33906-5. URL https://www.nature.com/articles/s41598-023-33906-5. Publisher: Nature Publishing Group.

Ding et al. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain | Radiology, 2018a. URL https://pubs.rsna.org/doi/abs/10.1148/radiol.2018180958?journalCode=radiology.

Han et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021. ISSN 2666-6510. doi: 10.1016/j.aiopen.2021.08.002. URL https://www.sciencedirect.com/science/article/pii/S2666651021000231.

Jing et al. HMS - Harmful Brain Activity Classification, 2024b. URL https://kaggle.com/competitions/hms-harmful-brain-activity-classification.

Nadhan et al. Enhancing healthcare security in the digital era: Safeguarding medical images with lightweight cryptographic techniques in IoT healthcare applications. *Biomedical Signal Processing and Control*, 88:105511, 2024c. ISSN 1746-8094. doi: 10.1016/j.bspc.2023.105511. URL https://www.sciencedirect.com/science/article/pii/S1746809423009448.

Shoeb et al. Application of machine learning to epileptic seizure onset detection and treatment, 2009. URL https://dspace.mit.edu/handle/1721.1/54669. Accepted: 2010-04-28T17:17:43Z.

Tsiouris et al. A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals. *Computers in Biology and Medicine*, 99:24–37, 2018b. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2018.05.019. URL https://www.sciencedirect.com/science/article/pii/S001048251830132X.

Thodoroff, Pineau, and Lim. Learning robust features using deep learning for automatic seizure detection, 2016. URL http://arxiv.org/abs/1608.00220. arXiv:1608.00220 [cs].