Modeling Clinical Decision Variability in Explainable Multimodal Seizure Detection

Asfandyar Azhar AAZHAR@ANDREW.CMU.EDU
Carnegie Mellon University

Amulya Mathur AMULYAM@ANDREW.CMU.EDU
Carnegie Mellon University

Sahil Jain SAHILJAI@ANDREW.CMU.EDU
Carnegie Mellon University

James Emilian JUVANJOS@ANDREW.CMU.EDU
Carnegie Mellon University

Shaurjya Mandal SMANDAL2@MGH.HARVARD.EDU
Harvard Medical School

Nidhish Shah NIDHISH@ASESA.LIFE
Asesa

Yongjie Jessica Zhang JESSICAZ@ANDREW.CMU.EDU
Carnegie Mellon University

Abstract

Electroencephalography (EEG) plays a critical role in the monitoring and diagnosis of neurological disorders, particularly in detecting seizures and other harmful brain activities. However, interpreting EEG signals is a complex task that often suffers from high variability and subjectivity among clinical experts. This study introduces the BiG-WaR architecture, a comprehensive multimodal framework designed to classify harmful brain activity using EEG signals. BiG-WaR combines several neural network models, including **Bi**LSTM, **G**NN, **Wa**veNet, and **R**esNet, to effectively leverage spatial and temporal dynamics inherent in EEG data. Our approach integrates curriculum learning and appropriate data preprocessing to address the challenges of EEG analysis, such as high variability and the need for robust feature extraction. Initial results demonstrate that BiG-WaR framework is a robust benchmark, enhancing reliability and interpretability—critical factors for clinical adoption—by integrating attention mechanisms and gradient-weighted class activation mappings to provide insights into model decisions.

Keywords: EEG, Brain Activity, Multimodal Learning, Explainable AI, Curriculum Learning

1. Introduction

Electroencephalography (EEG) is a pivotal tool in neurocritical care and epilepsy management, providing essential insights into the electrical activity of the brain. In neuroscience, EEG analysis has proven to be pivotal for studying brain activity, neural engineering, and psychomotor performance (Craik et al., 2019). In the realm of seizure care, EEG is essential for diagnosing and managing seizures, particularly in neonates (Abend and Wusthoff, 2012). Continuous EEG monitoring is recommended for high-risk patients to detect seizure activity promptly and facilitate timely treatment (Jan et al., 2017). Traditionally, the interpretation of EEG signals has depended on the manual expertise of specialized neurologists. However, several bottlenecks such as the potential for fatigue-based human error, inconsistencies between expert interpreters, and substantial demands on time and resources are quite prevalent in this domain; and these limitations highlight the pressing need for more scalable, efficient, and consistent analytical methods (Griffith et al., 2019).

In recent years, machine learning approaches have been increasingly successful in the medical field, especially in image-based diagnosis, disease prognosis, and risk assessment (de Bruijne, 2016; Litjens et al.,

2017). Machine learning models promise to predict future disease by finding new, early symptoms in image data from longitudinal studies (de Bruijne, 2016). In the EEG domain, current work has leveraged the potential of deep learning in interpreting EEG signals by extracting intricate features from raw data (Roy et al., 2019). This capacity has proven particularly valuable in tasks like neonatal seizure detection, where temporal EEG signals are core in training deep learning algorithms (O'Shea et al., 2020).

Interestingly, recent literature has presented improved diagnostic accuracy of neurological disorders, such as epilepsy in neonates, as a goal for the future (Okumura, 2020). Subsequently, deep learning algorithms have been effectively utilized to categorize EEG signals associated with various conditions, including alcoholism and driving fatigue (Farsi et al., 2020; Göker, 2023; Ma et al., 2019). These applications underscore the versatility and efficacy of deep learning in analyzing EEG data for diverse medical and research purposes. The fusion of machine learning and deep learning techniques has led to significant progress in decoding EEG signals and visualizing brain pathology, demonstrating the potential of these approaches in healthcare applications (Schirrmeister et al., 2017). Furthermore, promise has been shown in identifying biomarkers and significant patterns linked to neurological conditions like Alzheimer's disease (Fan et al., 2018) while (Tiwari et al., 2022) have done similar work for epilepsy combining the use of convolutional neural networks (CNNs) (LeCun et al., 2015) and signal processing techniques (Tiwari et al., 2022).

In the field of EEG signal classification, numerous efforts have been made to improve performance using unimodal approaches, such as LSTMs (Alhagry et al., 2017), GNNs (Demir et al., 2021), and WaveNet (Albaqami et al., 2023), among others. While each of these methods has its own advantages, to the best of our knowledge, no existing literature has explored fusing multiple models in a sequential manner to leverage the strengths of each. In this paper, we introduce a novel fusion approach, BiG-WAR (**Bi**LSTM-**G**NN-**Wa**veNet-**R**esNet), and discuss its merits in detail.

By integrating such an integrated multi-modular approach into EEG analysis, we can significantly reduce the workload on neurologists, enabling early detection of seizures and potentially life-threatening brain activity. This would not only accelerate the treatment process but also minimize human error, leading to more consistent and reliable diagnoses. However, developing an automated system carries significant implications - relying on such algorithms presents multiple challenges, especially in preserving the trust of clinicians and patients. This distrust is due to the often opaque nature of deep learning models - stakeholders could struggle to trust or comprehend how the algorithm makes its decisions, thus hindering their adoption. Addressing these issues, our research aims to answer the question: *How can a **viable** clinical decision support system (CDSS) be used to help diagnose harmful brain activity?* We define viability not only in terms of the system's **performance** but also its **explainability**. This dual focus ensures that the deep learning system we propose is not only effective but also transparent and understandable to clinicians which leads to the following sub-questions:

1. *What kind of technical components, in regards to data transformations and modeling, should be focused on to show context-based performance in a medical setting?*

2. *How can we show transparency within the model and garner trust from clients/stakeholders?*

2. Background

2.1. Primer on Measuring Brain Activity

Clinicians can acquire insights from EEG by detecting voltage variations caused by ionic currents in neurons (Niedermeyer, 2011). EEG data show diverse patterns indicating various brain functions. Periodic discharges and rhythmic delta activity are crucial for detecting neurological conditions like epilepsy. Recognizable by their frequency in hertz (Hz) or duration between peaks, periodic discharges exhibit consistent waveforms and shapes. In contrast, rhythmic delta activity has a continuous, rhythmic pattern without discrete intervals, sometimes observed across many brain regions. Patterns can be lateralized (predominantly in one hemisphere) or generalized (more uniform across the brain). For instance, lateralized rhythmic delta activity at 1 to 2 Hz indicates localized brain damage. These patterns are assessed for regularity and cycle duration to determine brain area functioning. Clinicians can better diagnose and manage neurological illnesses using objective metrics including frequency, amplitude, and cycle duration, with EEG data supporting efficient treatment options

(da Silva, 2013). However, there is a discrepancy among clinicians in classifying EEG patterns because these patterns can be ambiguous or present features that overlap between different categories, leading to varying interpretations based on personal expertise and experience. Additionally, practical challenges in classification include high inter-rater variability, fatigue, and the presence of artifacts or mixed patterns that complicate the identification of specific brain activities.

2.2. Dataset Description

The dataset was obtained from Harvard Medical School and consists of EEG recordings and related metadata of 1950 patients in the training set, meticulously annotated by 50 experts from the Critical Care EEG Monitoring Research Consortium to foster the development of models capable of classifying harmful brain activity (Jin Jing, 2024). The detailed components of this data is as follows:

1. **Metadata**: This file includes 15 columns relating to patient unique identifiers like: `patient_id`, `eeg_id`, `eeg_sub_id`, `spectrogram_id`, `spectrogram_sub_id`, `spectrogram_label_offset_seconds`, and `label_id` Additionally, there are expert annotations in the form of votes for the following 6 categories of brain activity: seizure (SZ), generalized periodic discharges (GPD), lateralized periodic discharges (LPD), lateralized rhythmic delta activity (LRDA), generalized rhythmic delta activity (GRDA), and "other. These annotations vary in consensus—from "idealized" patterns, where there is a high level of agreement among experts on the labeling, to "proto patterns" and "edge cases," where expert opinions diverge significantly, with some labeling as "other" or splitting between two of the five primary categories. These annotations serve as labels for the training and testing of the data.

2. **EEGs**: Linked to eeg id, this file contains 106800 50-second segments of EEG recordings sampled at 200 Hz, providing temporal brain activity data. Each patient has 8.76 EEGs on average and each EEG has 6.25 subsamples on average

3. **Spectrograms**: Associated with spectrogram id, this file includes frequency domain representations aggregated from 10-minute long EEG recordings. Each patient has 5.71 spectrograms on average and each spectrogram has 9.59 subsamples on average.

The test data consists of a single `eeg_id` and spectrogram id, which is 50 seconds and 10 minutes long respectively. This richly annotated dataset not only facilitates the accurate modeling of EEG patterns but also poses a unique challenge due to the varying degrees of expert agreement (label co-occurrence), requiring sophisticated machine learning approaches to interpret and classify the diverse and complex patterns effectively. Furthermore, to minimize bias, it is ensured that data from the same subject were not present in both the training and test sets. Temporal stratification was also used to ensure that sequences of EEG signals from the same time frame were grouped together, avoiding leakage of sequential information.

2.3. Classifiers & Kullback-Leibler Divergence

As various classes of deep learning networks have proven beneficial in EEG classification, it is important to note their unique utilities, specifically given the multimodal nature of the data. CNNs like ResNet, which use skip connections to preserve input features across layers, are effective for processing spatial hierarchies in data (He et al., 2016). In contrast, Recurrent Neural Networks (RNNs), particularly bidirectional LSTMs (BiLSTMs), excel in handling time-series data such as EEG signals by processing information both forward and backward to capture temporal dependencies from past and future states (Schuster and Paliwal, 1997). Graph Neural Networks (GNNs) are suited for EEG analysis where data is structured as graphs, as they iteratively update node states using node features and edge characteristics (Zhou et al., 2020). WaveNet, originally developed for audio waveform generation, has been adapted to EEG analysis, leveraging dilated convolutions to capture long-range temporal dependencies, making it effective for modeling complex and extended sequences (van den Oord et al., 2016).

In the context of this study, evaluating these models involves using the Kullback-Leibler divergence (KLD) metric. KLD measures the difference between the predicted probability distribution of the EEG classifications and the true distribution (expert labels), providing a mathematically grounded method to gauge model performance. The score is computed

as $D_{KL}(P \parallel Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$ where P is the true distribution and Q is the predicted distribution (Kullback and Leibler, 1951).

3. Methods

3.1. Data-centric Transformations

Using the `eeg_id` key from the metadata, a hashmap was created, where each key maps to three values: the respective processed EEG signal dataframe, EEG spectrogram, and target labels.

EEG Signal Processing: The Discrete Wavelet Transform (DWT) was employed for effective time-frequency representation (TFR) of EEG data, crucial for identifying features like spikes and wave variations (Upadhyay et al., 2016). DWT decomposes the EEG signals into approximate and detailed wavelets, corresponding to low-frequency and high-frequency components, respectively. This decomposition is achieved through the formula: $\Psi(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-u}{s}\right)$ where $\Psi(t)$ is the mother wavelet, s represents the scale, and u denotes the translation parameter (Pradhan et al., 2022). After denoising using the Daubechies wavelet (db8), known for its precision in capturing frequency and location information (Daubechies, 1992), the EEG signals are reconstructed from selected wavelet coefficients. The reconstruction is formulated as: $X(t) = \Sigma_k c_k \phi(t-k) + \Sigma_{j,k} d_{j,k} \psi_{j,k}(t)$ where c_k and $d_{j,k}$ are the approximate and detailed coefficients, ϕ is the scaling function, and $\psi_{j,k}$ represents the wavelet function. Furthermore, statistical features are extracted from each wavelet sub-band, including mean average power, mean absolute value, variance, standard deviation, and the inter-channel ratio. Finally, Shannon entropy is calculated to quantify signal irregularity, a key indicator of seizure activity: $H = -\Sigma p(x) = \log(p(x))$ (Upadhyay et al., 2016).

Spectrogram Processing: EKG spectrograms were created using the time-series EKG data for use in analysis, however for EEG, the provided spectrograms in the dataset are visual representations of raw EEG waveforms, which initially contain missing values and lack denoising. To address these issues, spectrograms were recreated using the original EEG data. The recreation process leverages the bipolar double banana montage (BDBM) (Figure 1), which divides the 19 electrodes into four chains: left temporal (LT), left parasagittal (LP), right parasagittal (RP), and right temporal (RT). Spectrograms are computed for each electrode pairing within these chains and averaged to produce a composite spectrogram per chain. The mathematical expression for one such computation is: $\text{LL} = \frac{1}{4}[\text{S}(Fp1 - F7) + \text{S}(F7 - T3) + \text{S}(T3 - T5) + \text{S}(T5 - O1)]$ where S denotes the spectrogram function applied to the differential signal between paired electrodes. This approach enhances the localization of EEG analysis, aiding in the detection of asymmetries or focal activities. Missing data within the spectrograms was initially filled using mean value imputation, and later refined to k-nearest neighbors (KNN) imputation which interpolates based on the nearest data points in the feature space and enhances data continuity and integrity: $\text{Value}_{\text{missing}} = \frac{1}{k}\Sigma_{i=1}^{k}\text{Value}_{\text{ith-nearest}}$ where k is the number of nearest neighbors considered. Then, to improve signal clarity, denoising was performed using db8. The db8 wavelet is chosen for its high vanishing moments, effectively reducing noise while preserving significant features in the EEG data: $X_{\text{denoised}}(t) = \text{IDWT}(\text{DWT}(X(t) \cdot \text{db8}))$ and where IDWT is the inverse discrete wavelet transform, DWT is the discrete wavelet transform, and $X(t)$ is the original signal. These processed spectrograms, stored as arrays for efficient loading, provide a clearer and more accurate analysis of the frequency components over time.

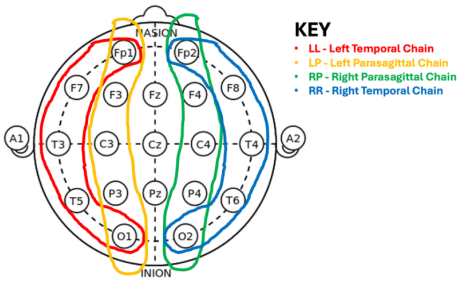

Figure 1: BDBM separating electrodes into temporal and parasagittal chains.

3.2. Architecture & Training Dynamics

The culmination of our architectural exploration led to the BiG-WaR (**Bi**LSTM-**G**NN-**Wa**veNet-**R**esNet) model displayed in Figure 2, integrating multiple modules.

Module 1: Utilizes a 2D ResNet architecture for extracting features from EEG signals, incorporating residual blocks to support deeper network training and gradient flow. Global Average Pooling (GAP) reduces feature map complexity, while a dual-layer, dual-head attention mechanism refines focus on rele-

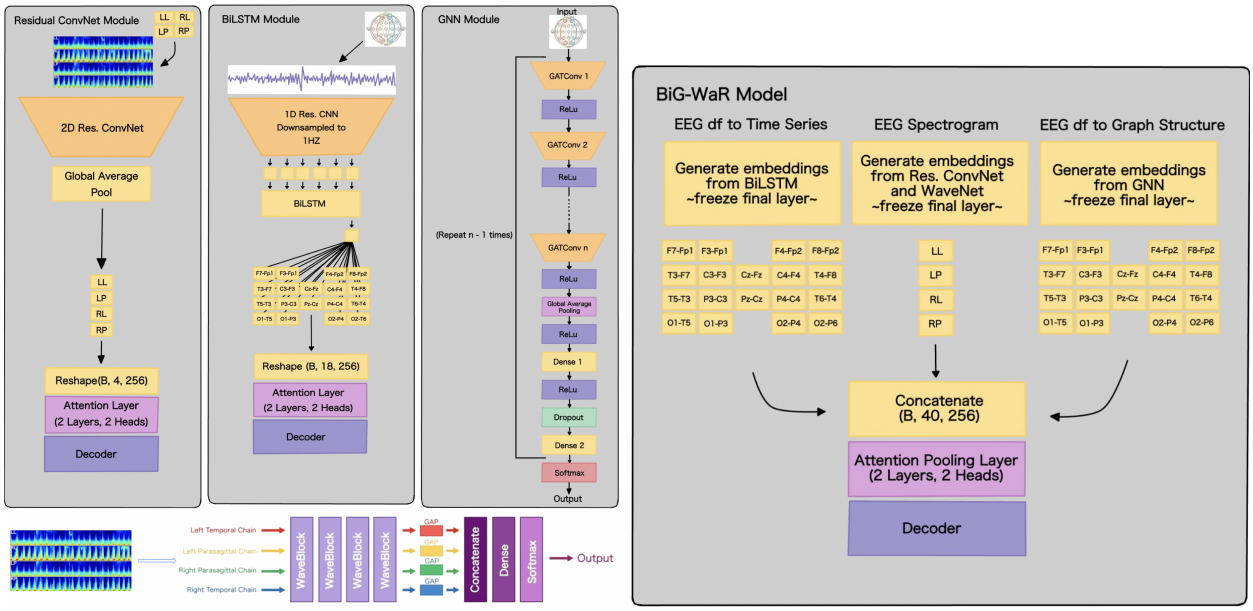

Figure 2: **Left**: All four modules represented separately. **Right**: BiG-WaR architecture – an amalgamation of the four modules resulting in a multimodal model.

vant visual-temporal features. Outputs are reshaped to $(\texttt{batch\_size}, 4, 256)$ and processed by a decoder to transform features into a probability distribution for KLD scoring.

Module 2: Integrates a BiLSTM network to process EEG data through a 1D ResNet, then downsampled to 1Hz to focus on critical signal frequencies. It assesses long-term dependencies in data sequences and maps outputs to specific EEG channels for targeted analysis. Data reshaped to $(\texttt{batch\_size}, 18, 256)$ undergoes processing through a two-layer, two-head attention system, enhancing the model's accuracy and feature integration before final probability distribution translation via a decoder.

Module 3: Employs a GNN to model EEG data as a non-Euclidean structure, using graph attentional operator (Veličković et al., 2018) (GATConv) layers followed by ReLU activations to enhance signal correlation analysis and feature extraction. Repeated GATConv and ReLU application enriches feature detection. Post-extraction, GAP condenses features, and dense layers further refine them with dropout for regularization. A softmax layer finalizes the process by converting features into a classification probability distribution.

Module 4: Adopts the WaveNet architecture, processing EEG signals through parallel chains for different regional analysis (LT, RT, LP, RP). It uses sequential WaveNet blocks with dilated convolutions to capture extensive temporal patterns. Features from all chains are condensed via GAP and concatenated for a comprehensive signal representation. Dense layers integrate these features, and a softmax layer outputs the final probability distribution.

Decoder: The decoder processes inputs shaped (b, c), where b is the batch size and c the number of channels. It follows two paths of linear transformations: one expands dimensions to 6, and another reduces them to 1. Both paths compute logits, or unnormalized log probabilities. The model then applies temperature scaling to adjust the confidence of predictions based on sample-specific parameters. After scaling, logits are normalized into a probability distribution over classes.

Training Phase: The final BiG-WaR model integrates these modules which intelligently handles the complexities of spatial layouts and modalities in data without making rigid assumptions. This flexibility allows the architecture to learn and adapt to new modalities and spatial configurations autonomously. The foundation of this approach involves decomposing the data into distinct nodes using, such as F8-Fp2 or RT. Each node is then embedded, and relationships between nodes are modeled using an attention-

style decoder that incorporates learned positional encodings.To streamline the learning process, we implemented a curriculum learning strategy (Bengio et al., 2009). Initially, we trained the model on 'soft' or 'easy' samples, which had expert votes or agreements ranging from 2 to 7 clinicians. During this phase, the four modules of our architecture were pretrained independently. After this, their backbones were frozen, and the model was fine-tuned on "hard" or "difficult" samples, characterized by having 10 or more clinician votes. This stage leveraged the pretrained embeddings from the initial modules. Subsequently, we unfroze the backbones and subjected the entire model to retraining on all available data. This retraining utilized a using a cosine annealing schedule (Loshchilov and Hutter, 2017) and early stopping to optimize performance and prevent overfitting. Data instances with only a single clinician vote were excluded from training due to their high noise levels and low relevance for predicting a distribution. Additionally, for each node in the banana montage, we embed it within the BiLSTM and GNN modules to utilize weight sharing during training. There was no robust hyperparameter tuning as standard values were used.

Why Spectrograms are Important: The inclusion of spectrograms is particularly valuable compared to using the EEG signal alone. Spectrograms enable the visual modules within BiG-WaR to process image data, adding an additional layer of information that enhances the model's predictive capabilities. In essence, this provides the model with a visual component from which it can learn and extract features. This approach aligns with the concept of multimodal machine learning, where different data inputs (e.g., temporal EEG signals and visual spectrograms) are combined to offer a richer, more comprehensive understanding. Multimodal methods have proven effective in other fields, such as large language models (LLMs) and vision-language models (VLMs), by leveraging diverse data types to boost overall performance. Including spectrograms follows this principle, allowing BiG-WaR to learn from both temporal and visual representations of EEG data.

3.3. The Black-to-White Box Transition

Despite the effectiveness of deep learning classifiers in EEG signal analysis, their opaque decision-making process poses a significant barrier to clinical acceptance (Ellis et al., 2021). To address this, we implement gradient-weighted class activation mapping (GradCAM) and attention mechanisms (Vaswani et al., 2023) to interpret which regions of a spectrogram influence the model's predictions. GradCAM utilizes the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions for predicting the brain activity pattern (Selvaraju et al., 2017). The class activation map is computed as follows: $L^c_{\text{GradCAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right)$ where A^k represents the feature maps of the last convolutional layer, α^c_k are the weights calculated by global average pooling (GAP) of the gradients flowing back, given by: $\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$ Here, y^c is the score for class c, and (i, j) are pixel indices. The attention mechanism in our model calculates a context vector that focuses on specific parts of the input by computing the relevance of each time step of the input signal. The mechanism is defined by: Context $= \sum_t \alpha_t h_t$, $\alpha_t = \frac{\exp(e_t)}{\sum_{t'} \exp(e_{t'})}$, and $e_t = a(h_{t-1}, s_t)$ where h_t are the hidden states, α_t are the attention weights, s_t is the input at time t, and a is the alignment model that scores how well the inputs around position t and the output at position t match. Attention helps in understanding which features of the input sequence are considered important for the output. These techniques provide visual explanations for the decisions made by our model, thus enhancing transparency and trust in its clinical applicability.

4. Experiments & Results[1]

4.1. Data Ablations

To validate the model's performance, ablation studies were performed using EEG signal features (described in Section 3) and spectrogram datasets. The model was trained on the following datasets: Raw signals (Base), DWT-processed signals (DWT), DWT-processed signals with statistical features (DWT w/s), DWT-processed signals with entropy features (DWT w/e), DWT-processed signals with both statistical and entropy features (DWT w/a), and finally, the same dataset as the previous one but with selected features based on variance thresholding (DWT w/a+v). Variance thresholding, a simple dimensionality reduction method, was applied to eliminate low-predictive features and retain only those with high variance (0.8). After testing thresholds from 0 to

1. Experiments used a TeslaV100, covering all pipeline stages. Results are reported using 5-fold cross-validation.

1, 0.8 was identified as optimal for preserving high-frequency seizure-related features while training the model. As shown in Table 1, each model achieved the lowest KL divergence score when trained on the final dataset (DWT w/a+v) and evaluated on the test set.

Model	Raw	DWT	DWT w/s	DWT w/e	DWT w/a	DWT w/a+v
LSTM	0.96 ± 0.05	0.74 ± 0.04	0.71 ± 0.03	0.69 ± 0.03	0.61 ± 0.02	**0.59 ± 0.02**
GNN	1.32 ± 0.06	1.03 ± 0.05	0.99 ± 0.04	0.99 ± 0.04	0.79 ± 0.03	**0.67 ± 0.03**

Table 1: EEG Signal Ablations

Model	Raw	Mean	KNN	db8	Mean w/db8	KNN w/db8
EfficientNet	0.56 ± 0.03	0.50 ± 0.02	0.48 ± 0.02	0.47 ± 0.02	0.44 ± 0.01	**0.41 ± 0.01**
WaveNet	0.78 ± 0.04	0.59 ± 0.03	0.53 ± 0.02	0.52 ± 0.02	0.48 ± 0.02	**0.42 ± 0.01**

Table 2: EEG Spectrogram Ablations

Table 2 shows similar data ablations with the spectrograms, where the following transformed spectrograms were tested: mean-imputed (Mean), KNN-imputed (KNN), db8-denoised (db8), mean-imputed with db8 denoising (Mean w/db8), KNN-imputed with db8 denoising (KNN w/db8). In short, using db8 denoising combined with KNN-imputation enhanced the quality of spectrograms compared to mean-imputation. This improvement is attributed to the granularity provided by filling missing values based on neighboring signals, rather than using the average value of the entire spectrogram, which had a "smudging" effect.

4.2. Model Ablations

Alluding to Table 3, we began our exploration by establishing baseline performances using two distinct architectures: EfficientNet (Tan and Le, 2019) and Boosting Trees (Friedman, 2001). EfficientNet, known for its efficiency and scalability, served as a comparative baseline for deep learning models, achieving a KLD of 0.41 in 93 minutes. Boosting Trees, implemented to evaluate non-deep learning methods, provided a KLD of 0.72 in 77 minutes. Our ablations progressed through various architectures, each chosen to address specific challenges observed in previous models. We first combined ResNet with LSTM, leveraging ResNet's ability to extract spatial features and LSTM's capacity to capture temporal dependencies, which improved the KLD to 0.4 over 165 minutes. To enhance focus on significant features, we incorporated an attention mechanism with BiLSTM, further reducing the KLD to 0.3 in 178 minutes. Adding WaveNet to the mix allowed us to model complex temporal patterns effectively, maintaining the KLD while reducing runtime. Our final ablation, the BiG-WaR model without attention, achieved a KLD of 0.27 in 235 minutes, demonstrating the combined strength of BiLSTM, GNN, and WaveNet in handling EEG data.

Model	KLD Score	Runtime (mins)
EfficientNet (Baseline)	0.41 ± 0.03	93
Boosting Trees	0.72 ± 0.05	77
ResNet + LSTM	0.40 ± 0.03	165
ResNet + BiLSTM (w/ Attn)	0.30 ± 0.02	178
EfficientNets + ResGRU + WaveNet	0.30 ± 0.02	113
ResNet + BiLSTM + GNN	0.29 ± 0.02	204
BiG-WaR (w/o Attn)	0.27 ± 0.01	235
BiG-WaR (w/ Attn)	**0.25 ± 0.01**	260

Table 3: Model Ablations

5. Discussion

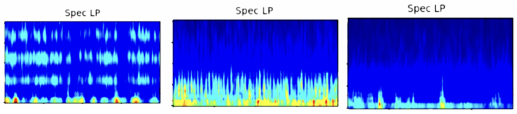

Figure 3: GradCAM output on instances of the seizure (left), periodic (middle), and rhythmic (right) delta activities in the left parasagittal chain respectively.

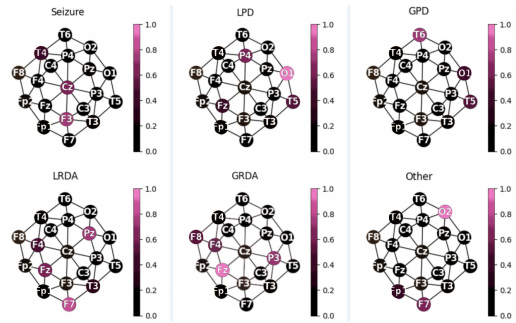

Figure 4: EEG attention weight maps for each class.

Viability Objectives	Yes	No
KLD performance	X	
Training runtime	X	
Visual interpretability	X	
Temporal interpretability		X
Spatial interpretability	X	
Signal interpretability	X	
Data optimization		X
Data explainability	X	

Table 4: Evaluation of Goals and Client Objectives (green = performance, pink = transparency)

GradCAM is used to highlight the relevant signals the model looks at to classify brain activity. As shown in Figure 3, when activity is classified as a seizure, signals in the high Hz range play a larger role in the analysis. When activity is classified as periodic, signals in the low-to-medium range are analyzed with greater importance instead. Lastly, when classifying activity as rhythmic, the relevant signals are primarily in the low Hz range. Figure 4 outlines the difference in overall attention weight mappings (extracted from the attention pooling layer of BiG-WaR) across all of the 10-20 electrode locations for each class. Each class places emphasis on a different set of electrodes - for example in GRDA, we see a fairly diffuse color and no specific highlights; and this is in line with the "generalized" pattern of activity that is expected in the GRDA class. This emphasizes differences between every class quite specifically and provides spatial and signal interpretability for BiG-WaR. When reflecting on the goals and client objectives originally set, not every original objective was reached (Table 4). Performance-wise, a significant KLD score was achieved, although the overall model and final dataset were not entirely suited for use in a practical setting. However, we achieved most of our goals concerning explaining what caused the model to decide upon a particular classification for a given EEG signal segment. Hence, clients and investors would greatly value the transparency and explainability of the model.

5.1. Deployment Challenges

EKG spectrograms were not incorporated into the BiG-WaR model, though they are thought to contain valuable information that could potentially enhance model performance. Initially, our model employed exclusively 2D architectures such as ResNet and WaveNet, which were incompatible with the EKG data. However, later developments introduced a 1D Residual CNN in module 2 of BiG-WaR, which downsamples the data. This advancement could potentially facilitate the integration of EKG data into future iterations of the model. Moreover, the lack of a domain expert significantly constrained our data-centric approach, limiting our ability to draw meaningful insights and validate the relevance of features used by the model. As highlighted in Table 3 - compared to other published solutions that employ simpler, more transparent methods facilitated by effective data-centric strategies, BiG-WaR highlights the classic trade-offs between model accuracy/sophistication and runtime. In a practical setting, this could imply that BiG-WaR would be more time-consuming to run especially if the deployment targets (clinics) aim to run the EEG diagnostic model on an edge device. However, with recent advancements in edge computing (for example, the NVIDIA Jetson Orin series with runtime and compute equivalent to that of powerful workstations) - the significance of this shortcoming is diminished.

5.2. Deciding on The Clinically Acceptable Level of Performance

The clinically acceptable level of predictive performance in our study is benchmarked using the ground-truth test data provided by Jin Jing (2024). This dataset reflects inter-rater variability among expert annotations, making it a suitable standard for evaluating clinical applicability. To ensure alignment with expert consensus, we employ the KLD score as an evaluation metric, which measures how well the model's output probability vectors match the distribution of expert votes across classes. By optimizing the model to approximate or replicate this consensus level of performance, we provide a realistic measure of the model's clinical relevance.

However, a natural limitation arises from the model's reliance on current expert-generated labels, as it may be unable to surpass expert-level performance—a potential bottleneck. To address this, we propose incorporating an active learning pipeline with a human-in-the-loop (HITL) framework, as illustrated in Figure 5. Through iterative feedback from multiple experts, this pipeline has the potential to refine the model beyond the accuracy of any single expert. Studies have demonstrated that integrating AI

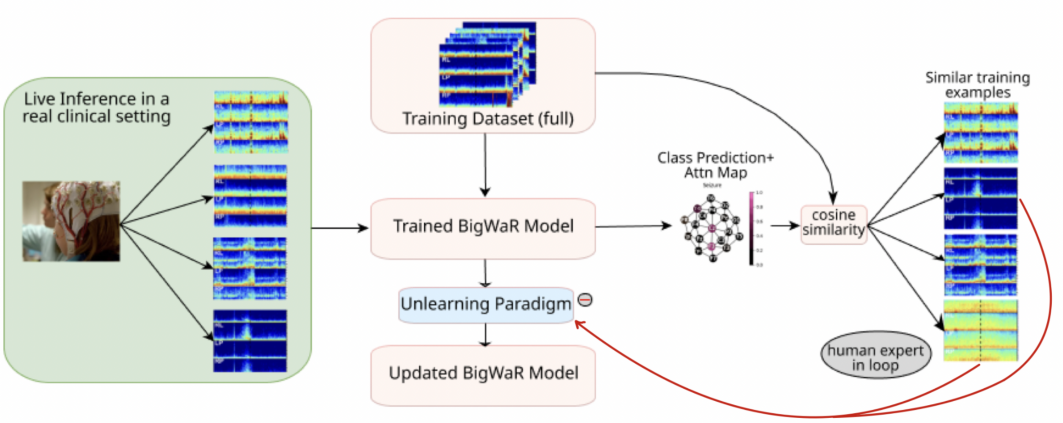

Figure 5: Proposed XAI Active Learning Pipeline.

predictions with expert knowledge often yields superior results (AI + Expert > Expert > AI) (Tschandl et al., 2020), a principle that forms the foundation of our approach. By iteratively enhancing the model with expert input, our goal is to develop an AI assistant capable of not only matching but exceeding expert decision-making. This approach would enable a robust CDSS, ensuring the model remains reliable, accurate, and valuable in clinical practice.

6. Conclusion & Future Work

The importance of data quality in the performance of deep learning models cannot be understated. Though we were able to extensively study the efficacy of various types of data preprocessing in such models, we plan to investigate more advanced data augmentation techniques such as time warping (Iwana and Uchida, 2021), channel shifting (Huang et al., 2019), or synthetic minority over-sampling (Forestier et al., 2017) to address class imbalance and improve generalization. A significant innovation discussed in our results section was the extraction of graphical attention maps from the attention pooling layer of BiGWaR, which illustrate where the model focuses its attention, node-wise, during inference. Building on this, we envision the development of an active learning pipeline (Monarch, 2021) by using cosine similarity to match the live inference data sample with instances from the training data with and present them for review to expert clinicians, who can then provide feedback on the accuracy of the inferences (see Figure 5). Data points identified as misinterpreted by the model can subsequently be fed into an unlearning process, wherein sparsity-based unlearning algorithms (Liu et al., 2024) can be employed to adjust the trained model weights to specifically "forget" these misleading examples. This approach not only enhances model accuracy and clinical fidelity, but also creates continuous improvement through an expert-in-the-loop system, increasing general clinician trust in the automated diagnostics provided by said system.

Given the performance of the final model, we envision that it can potentially be deployed in real-time clinical settings for live inference on EEG data for improved diagnosis speed. To this end, model pruning techniques (Jiang et al., 2022) to reduce the complexity and computational cost of the model without sacrificing performance, should be explored. Moreover, the deployment of such models in clinical environments raises significant data privacy concerns. To address these, future works should investigate federated learning frameworks (Sheller et al., 2020), which allow for model training on decentralized data, ensuring patient privacy and compliance with regulatory standards.

Overall, our main contribution is the development of a novel multimodal fusion framework, BiGWaR, which seamlessly integrates diverse models to comprehensively capture the spatial, temporal, and graph-based features of EEG data. Furthermore, our pipeline includes advanced preprocessing techniques, such as DWT and KNN imputation, along with curriculum learning, resulting in a model that is both more accurate and interpretable compared to existing single-modality approaches.

References

Nicholas S Abend and Courtney J Wusthoff. Neonatal seizures and status epilepticus. *Journal of clinical neurophysiology*, 29(5):441–448, 2012.

Hezam Albaqami, Ghulam Mubashar Hassan, and Amitava Datta. Automatic detection of abnormal eeg signals using wavenet and lstm. *Sensors*, 23 (13):5960, 2023.

Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. Emotion recognition based on eeg using lstm recurrent neural network. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

Fernando Lopes da Silva. Eeg and meg: relevance to neuroscience. *Neuron*, 80(5):1112–1128, 2013.

Ingrid Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

Marleen de Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, 33:94–97, jan 2016. ISSN 1361-8415. doi: 10.1016/j.media.2016.06.032. URL https://www.sciencedirect.com/science/article/pii/S1361841516301098. 20th anniversary of the Medical Image Analysis journal (MedIA).

Andac Demir, Toshiaki Koike-Akino, Ye Wang, Masaki Haruna, and Deniz Erdogmus. Eeg-gnn: Graph neural networks for classification of electroencephalogram (eeg) signals. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1061–1067. IEEE, 2021.

Charles A Ellis, Mohammad SE Sendi, Robyn L Miller, and Vince D Calhoun. A gradient-based spectral explainability method for eeg deep learning classifiers. *bioRxiv*, pages 2021–07, 2021.

Miaolin Fan, Albert C Yang, Jong-Ling Fuh, and Chun-An Chou. Topological pattern recognition of severe alzheimer's disease via regularized supervised learning of eeg complexity. *Frontiers in neuroscience*, 12:390782, 2018.

Leila Farsi, Siuly Siuly, Enamul Kabir, and Hua Wang. Classification of alcoholic eeg signals using a deep learning method. *IEEE Sensors Journal*, 21 (3):3552–3560, 2020.

Germain Forestier, François Petitjean, Hoang Anh Dau, Geoffrey I Webb, and Eamonn Keogh. Generating synthetic time series to augment sparse datasets. In *2017 IEEE international conference on data mining (ICDM)*, pages 865–870. IEEE, 2017.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Hanife Göker. Welch spectral analysis and deep learning approach for diagnosing alzheimer's disease from resting-state eeg recordings. *Traitement du Signal*, 40(1), 2023.

Brent Griffith, Nadja Kadom, and Christopher M. Straus. Radiology Education in the 21st Century: Threats and Opportunities. *Journal of the American College of Radiology*, 16(10):1482–1487, 2019. doi: 10.1016/j.jacr.2019.04.003.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Liang Huang, Weijian Pan, You Zhang, Liping Qian, Nan Gao, and Yuan Wu. Data augmentation for deep learning-based radio modulation classification. *IEEE access*, 8:1498–1506, 2019.

Brian Kenji Iwana and Seiichi Uchida. Time series data augmentation for neural networks by time warping with a discriminative teacher. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3558–3565. IEEE, 2021.

Saber Jan, Frances J Northington, Charlamaine M Parkinson, and Carl E Stafstrom. Eeg monitoring technique influences the management of hypoxic-ischemic seizures in neonates undergoing therapeutic hypothermia. *Developmental neuroscience*, 39 (1-4):82–88, 2017.

Yuang Jiang, Shiqiang Wang, Victor Valls, Bong Jun Ko, Wei-Han Lee, Kin K Leung, and Leandros Tassiulas. Model pruning enables efficient federated learning on edge devices. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Chaoqi Yang Ashley Chow Sohier Dane Jimeng Sun M. Brandon Westover. Jin Jing, Zhen Lin. Hms - harmful brain activity classification, 2024. URL https://kaggle.com/competitions/hms-harmful-brain-activity-classification.

Solomon Kullback and Richard A. Leibler. *On information and sufficiency*, volume 22. Institute of Mathematical Statistics, 1951.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, PRANAY SHARMA, Sijia Liu, et al. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

Yuliang Ma, Bin Chen, Rihui Li, Chushan Wang, Jun Wang, Qingshan She, Zhizeng Luo, and Yingchun Zhang. Driving fatigue detection from eeg using a modified pcanet method. *Computational intelligence and neuroscience*, 2019, 2019.

Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Ernst Niedermeyer. *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2011.

Akihisa Okumura. Electroencephalography in neonatal epilepsies. *Pediatrics International*, 62(9):1019–1028, 2020.

Alison O'Shea, Gordon Lightbody, Geraldine Boylan, and Andriy Temko. Neonatal seizure detection from raw multi-channel eeg using a fully convolutional architecture. *Neural Networks*, 123:12–25, 2020.

Bikash K. Pradhan, Maciej Jarzebski, Anna Gramza-Michałowska, and Kunal Pal. Automated detection of caffeinated coffee-induced short-term effects on ecg signals using emd, dwt, and wpd. *Nutrients*, 14(4), 2022. ISSN 2072-6643. doi: 10.3390/nu14040885. URL https://www.mdpi.com/2072-6643/14/4/885.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

Robin Tibor Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology. *arXiv e-prints*, pages arXiv–1708, 2017.

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598, 2020.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Shailendra Tiwari, Syamsundararao Thalakola, A. Selvarani, R. Rathi, N. Vini Antony Grace, J. Selvaraj, Khalid M. A. Almutairi, and V. Alonazi. An efficient signal processing algorithm for detecting abnormalities in eeg signal using cnn. *Contrast Media & Molecular Imaging*, 2022:1502934, 9

2022. ISSN 1555-4309. doi: 10.1155/2022/1502934. URL https://doi.org/10.1155/2022/1502934.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Josep Malvehy, Susana Puig, Cliff Rosendahl, H. Peter Soyer, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26:1229–1234, 2020. doi: 10.1038/s41591-020-0942-0. URL https://doi.org/10.1038/s41591-020-0942-0.

R. Upadhyay, P.K. Padhy, and P.K. Kankar. A comparative study of feature ranking techniques for epileptic seizure detection using wavelet transform. *Computers Electrical Engineering*, 53:163–176, 2016. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2016.05.016. URL https://www.sciencedirect.com/science/article/pii/S0045790616301495.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Li Wang. Graph neural networks: A review of methods and applications. *AI Open*, 1: 57–81, 2020.