LLM-Powered Cross-Modal Alignment for Explainable Seizure Detection from EEG

Maryam Riazi[1], Deeksha M. Shama[1,2*], and Archana Venkataraman[1,2]

[1] Electrical and Computer Engineering, Boston University, Boston, USA
{mary79, archanav}@bu.edu
[2] Electrical and Computer Engineering, Johns Hopkins University, Baltimore, USA
dshama1@jhu.edu

Abstract. While artificial intelligence (AI) has revolutionized the field of epileptic seizure detection from electroencephalography (EEG), its clinical adoption remains limited, largely due to the lack of transparency in AI models and their inability to explain the underlying seizure etiology. This paper introduces SzXAI, a novel framework to enhance the reasoning abilities of AI models for EEG-based seizure detection. SzXAI employs a contrastive training mechanism, which uses cross-modality similarity layers to align the EEG encodings with textual concept embeddings derived from clinical notes using LLMs. Along with the alignment, SzXAI leverages an attention-weighted pooling mechanism to detect underlying seizure and baseline etiologies. We validate SzXAI via 10-fold cross validation on the publicly available Temple University Hospital dataset. Our results demonstrate that the alignment-powered training mechanism of SzXAI vastly outperforms direct etiology prediction, thus improving the reliability of the predicted seizure etiologies. Furthermore, structured sentence generation using the model output provided insights in a human-readable format. Thus, SzXAI provides an effective platform to boost clinical trust and AI usability in epilepsy management

Keywords: Epilepsy · Contrastive Learning · LLM · Explainable AI

1 Introduction

Artificial intelligence (AI) has emerged as a powerful tool to support clinical decision-making by automatically extracting information from complex and noisy data sources [1]. However, despite their strong performance in diagnostic tasks, AI models are not fully trusted, in large part due to their lack of transparency. As a result, clinicians still bear a significant burden in identifying the underlying disease etiology for treatment planning. Ultimately, this burden is both subject to human biases [6] and prolongs the decision-making process [11].

The challenges of AI integration are underscored in epilepsy, in which scalp electroencephalography (EEG) is the primary modality. Diagnosis from EEG is

* Corresponding author.

done by *manually scanning* hours of data for seizure patterns (e.g., spike-and-wave complexes or high-amplitude discharges [3,22]), a process that is labor-intensive and prone to human errors [2]. Over the past decade, AI-based methods have demonstrated a clear ability to detect the presence or absence of seizure activity in EEG data. Notable models include EEGNet [15,23], CNN-BLSTM [8], TGCN [7], and DeepSOZ [16], all of which implicitly capture spatio-temporal correlations in EEG data associated with seizure activity. However, the model outputs are limited to binary predictions of seizure versus baseline activity, and in some cases a categorical prediction of the seizure onset location [5,9,16]. These categorical predictions do not offer insight into the specific characteristics of the EEG data that indicate a seizure, also known as its *etiology*, which is crucial to bridging the gap between categorical prediction and diagnostic reasoning.

Large language models (LLMs) offer a compelling strategy to bridge this gap. For example, the EEGtoText model [4] is trained to generate reports for EEG recordings. However, this model lacks information about the timing of events, which limits its broader clinical utility. Concept-bottleneck models have become popular in medical image analysis [14,18,24,25] and offer a promising framework for explaining the underlying seizure etiologies. Recent studies have shown that concept-based explanations are preferred over other forms, such as heatmaps or example-based interpretations [20]. However, their adoption is hindered due to required architectural changes, limited information on temporal evolution, and a lack of structured concept datasets. The seminal work of [12] takes a first step in this direction by proposing EEG-GPT. This model prompts an LLM with structured outputs from a seizure detector to generate textual summaries of precomputed seizure predictions (e.g., onset time, duration) in EEG. While this strategy can be integrated with seizure detection systems, it merely summarizes the data statistics and cannot reason about the underlying seizure etiology.

This paper presents SzXAI to enhance the reasoning abilities of EEG-based seizure detection AI models. SzXAI uses cross-modality similarity layers to align the EEG representations with textual concept embeddings derived from clinical notes and uses an attention-weighted pooling mechanism to detect underlying seizure and baseline etiologies present in the EEG data. We use a supervised contrastive loss during training to enforce consistency between the *dynamic EEG data* and the *static clinical notes*. This procedure allows SzXAI to align the EEG representations with seizure and non-seizure etiologies at appropriate time points. We evaluate SzXAI with two seizure detection models on the publicly available Temple University Hospital (TUH) dataset. Our results demonstrate that alignment-powered training improves performance over direct etiology prediction. Reconstructed sentences from the etiologies predicted by SzXAI reveal key patient-specific insights, thus underscoring its clinical potential.

2 Novel Model-Agnostic Etiology Prediction in EEG

This section describes the three novel elements of SzXAI: (i) integrating cross-modality similarity layers and attention-weighted pooling into an existing seizure

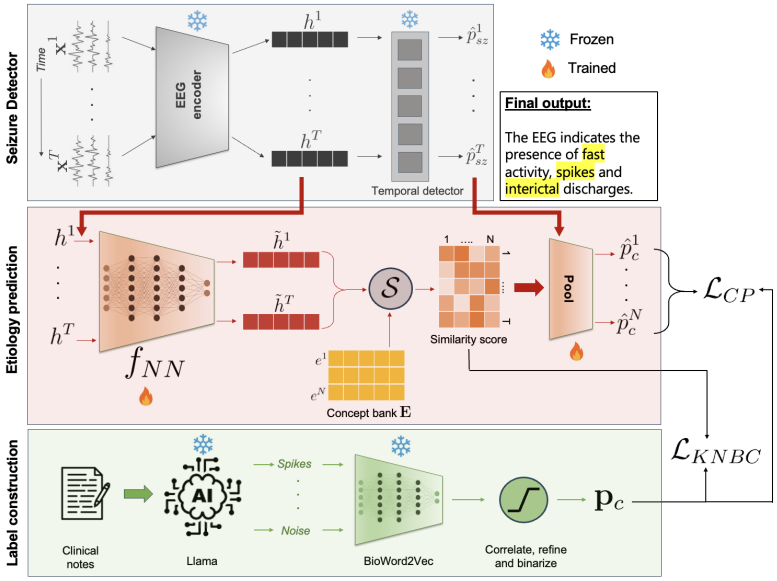

Fig. 1: Overall framework of SzXAI for seizure etiology prediction from EEG using any AI-based seizure detector, similarity computation, and pooling layers (**Top Two**) and the proposed LLM-driven label construction (**Bottom**)

detection system, (ii) a knowledge-based contrastive loss for temporal alignment of the EEG with static clinical notes, and (iii) ground-truth etiology extraction from clinical notes using pre-trained LLMs. Finally, an LLM is used to constuct summaries from predicted etiologies. Our overall framework is shown in Fig. 1.

Importantly, SzXAI can be *integrated into any AI-based seizure detection framework* by extracting an intermediate encoding of the EEG data. The concept banks can also be edited as per clinical needs. Moreover, given its small size, SzXAI can be trained efficiently without sacrificing the detection performance.

2.1 The SzXAI Framework

AI seizure detectors are designed to make a sequence of binary predictions of baseline vs. seizure activity for each short time windows of the EEG data [19]. As part of this process, most AI models will construct a latent representation of the EEG. Let T be the number of time windows in the recording, and let $\mathbf{h}^{1:T}$ denote the latent representations for each time window. These intermediate variables $\mathbf{h}^{1:T}$ are the key to our model-agnostic etiology prediction.

Concept Bank Generation: We construct a comprehensive *concept bank* of all expected seizure and baseline etiologies present in the EEG as follows:

1. Prompt the LLaMA model [10] to generate a list of EEG patterns across the time and frequency domains that cover both seizure-related and back-

ground activity. Prompt tuning and reference articles were used to guide the output [22]. Redundant phrases were eliminated to ensure distinct concepts.
2. Generate embeddings for each concept phrase using the BioWord2Vec model [27], which has been pre-trained on biomedical reference texts.
3. Extract 20 unique concepts by removing semantically similar terms based on their BioWord2Vec embedding space proximity.

We denote this concept bank as $\mathbf{E} := \mathbf{e}^{1:N} \in \mathbb{R}^{N_d}$, where $N = 20$ and N_d is encoding dimension. $\mathbf{E}$ is generated once and is then fixed across all experiments.

Given the latent representations $\mathbf{h}^{1:T}$ from the seizure detector and concept bank $\mathbf{E}$, SzXAI employs a neural network with three linear layers, batch normalization, LeakyReLU activation to obtain the encodings $\tilde{\mathbf{h}}^t = f_{NN}(\mathbf{h}^t) \in \mathbb{R}^{N_d}$, thus bringing the EEG representations into the same space as the elements of $\mathbf{E}$.

Etiology Prediction from EEG: We use the encodings $\tilde{\mathbf{h}}^t$ to predict which of the N concepts are present in the EEG data of a given patient via a two step procedure. First, we compute the cosine similarity between the encoding for each time window $\tilde{\mathbf{h}}^t$ and the N concept vectors $\mathbf{e}^n$ to get $S_{nt} = \langle \tilde{\mathbf{h}}^t, \mathbf{e}^n \rangle = \frac{\tilde{\mathbf{h}}^{t^T} \cdot \mathbf{e}^n}{\|\tilde{\mathbf{h}}^t\| \|\mathbf{e}^n\|}$.

Second, we construct an attention-weighted pooling of the similarity scores across time, where the attention weights is derived from the prediction of seizure activity for each time window by the original detector: $\hat{p}_{sz}^{1:T}$ such that $\hat{p}_{sz}^t \in [0, 1]$:

$$\hat{s}^n = \sum_{t=1}^{T} \hat{p}_{sz}^t \cdot S_{nt} \qquad \text{or} \qquad \hat{s}^n = \sum_{t=1}^{T} (1 - \hat{p}_{sz}^t) \cdot S_{nt} \tag{1}$$

where the left side of Eq. (1) is used for seizure concepts and the right side for non-seizure concepts. This strategy allows the encodings for seizure time windows to be used to predict seizure etiologies and likewise for the non-seizure time windows and etiologies. Finally, we use a linear layer to predict the probability of each etiology being present in the EEG data, denoted by $\hat{p_c}^{1:N}$.

2.2 Knowledge-driven Training for Etiology Prediction

SzXAI is trained using a combination of three loss terms as follows:

$$\mathcal{L} = \mathcal{L}_{CP} + \lambda_1 \mathcal{L}_{KNBC} + \lambda_2 \mathcal{L}_{reg}. \tag{2}$$

The final term $\mathcal{L}_{reg}$ is an L_2 penalty on the weights of all linear layers in SzXAI, and λ_1 and λ_2 are hyper-parameters. The first two terms are described below.

Prediction Loss ($\mathcal{L}_{CP}$)**:** We define the task of etiology prediction as a multi-label classification problem using a weighted binary cross entropy loss:

$$\mathcal{L}_{CP} = -\frac{1}{N} \sum_{n=1}^{N} [w^n \cdot p_c^n \log(\hat{p}_c^n) + (1 - w^n) \cdot (1 - p_c^n) \log(1 - \hat{p}_c^n)] \tag{3}$$

where w^n is the weight of the n^{th} concept and is inversely proportional to the frequency of occurrence in the training dataset. $p_c^n \in \{0, 1\}$ is the ground truth label. This loss guides SzXAI to detect the presence of etiologies in EEG.

Contrastive Alignment Loss ($\mathcal{L}_{KNBC}$)**:** The clinical notes and derived concepts are *static* at the patient level, whereas the seizure activity captured in EEG is *dynamic* over time. Thus, SzXAI must learn **an unknown temporal alignment** between the sequence of latent EEG representations $\tilde{\mathbf{h}}^{1:T}$ and the concept bank $\mathbf{E}$. We propose the Knowledge-Based Contrastive loss in Eq. (4) to map temporal information to patient etiologies. This loss encourages the model to learn discriminative features by contrasting positive pairs (e.g., embeddings from seizure time windows with seizure-related etiologies) against negative pairs (e.g., embeddings from seizure time windows with non-seizure-related etiologies).

$$\mathcal{L}_{KNBC} = \frac{-1}{|T_{sz}| \times |C_{sz}|} \sum_{t \in T_{sz}} \sum_{n \in C_{sz}} \log \left(\frac{\exp(\langle \tilde{\mathbf{h}}^t, \mathbf{e}^n \rangle)}{\sum_{k \in \mathbf{E}_{nsz}} \exp(\langle \tilde{\mathbf{h}}^t, \mathbf{e}^k \rangle)} \right) + \frac{-1}{|T_{nsz}| \times |C_{nsz}|} \sum_{t \in T_{nsz}} \sum_{n \in C_{nsz}} \log \left(\frac{\exp(\langle \tilde{\mathbf{h}}^t, \mathbf{e}^n \rangle)}{\sum_{k \in \mathbf{E}_{sz}} \exp(\langle \tilde{\mathbf{h}}^t, \mathbf{e}^k \rangle)} \right) \quad (4)$$

Formally, let T_{sz} and T_{nsz} denote the set of seizure and non-seizure time windows respectively. Similarly, let C_{sz} and C_{nsz} denote the seizure and non-seizure etiologies *that are present for a given patient*. In contrast, $\mathbf{E}_{sz}$ and $\mathbf{E}_{nsz}$ represent the collection of seizure and non-seizure etiologies, respectively, in the concept bank $\mathbf{E}$. The first term of Eq. (4) states that the encoding $\tilde{\mathbf{h}}^t$ for a seizure time window should be more correlated with the present seizure concepts for that patient than to any of the non-seizure concepts. The second term of Eq. (4) encourages the inverse process for non-seizure time windows in EEG. At a high level, this loss encourages SzXAI to learn nuanced temporal representations and effectively distinguish between seizure and non-seizure etiologies.

2.3 LLM-powered Label Construction for Training and Evaluation

Inspired by [18,26], we construct labels for training and evaluation through the automated process illustrated in the bottom pane of Fig. 1. Specifically, we first prompt the open-source pre-trained LLaMA-8B with no further finetuning using the query: "List present etiologies based on this clinical note describing EEG" (and its semantic variations for stochasticity) to extract the relevant etiologies. The extracted text is then refined to obtain key terms. We utilize BioWord2Vec to generate embedding vectors minimally correlated with each other (Fig. 2(a)). Next, we iteratively compute the cosine similarity between each predicted term and every vector in our predefined concept bank, $\mathbf{E}$. If the similarity exceeds a threshold of 0.8, we classify that etiology as present for that patient in our ground-truth labels. This results in a binary vector ${p_c}^{1:N}$, where multiple etiologies may be simultaneously present. Notably, this process is computationally efficient, requiring only a few seconds when executed on one A100 GPU. We also confirmed the accuracy of the algorithm by comparing generated labels with clinical notes. Fig. 2(b) shows perfect recall and high specificity, with minor false positives within related etiologies.

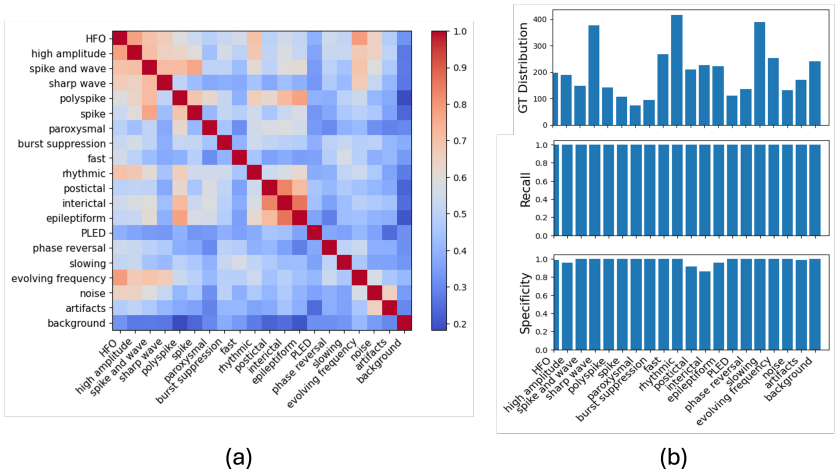

Fig. 2: **(a)** Self-Correlation of embeddings in **E** given by BioWord2Vec. **(b)** Class distribution of generated Ground Truth (GT) labels (Top). Recall (Middle) and Specificity (Bottom) of the labels compared with clinical notes. *HFO: High Frequency Oscillations. PLED: Partial Lateralized Epileptiform Discharge.*

2.4 Implementation Details

We compare SzXAI with two baseline concept prediction frameworks from the literature: [18] and [25]. We adopt their EEG-text similarity approach for etiology prediction while maintaining SzXAI's label generation and detection architectures. We also conduct several ablations of SzXAI: average pooling (SzXAI-avg) and max pooling (SzXAI-max) rather than attention-weighted, excluding the alignment loss (SzXAI-NAl), and standard BCE with (SzXAI-bce) and without alignment (SzXAI-bce-NAl). For thoroughness, we validate all models on two state-of-the-art seizure detectors: DeepSOZ [16] and CNN-BLSTM [8]. These networks remain frozen while training both SzXAI and the baselines.

We conduct experiments using a 10-fold cross-validation in Python v3.9.4 using Pytorch v2.2.1 [17] and Adam optimizer [13] with a batch size of one patient on two A100 GPUs. [3] All methods were trained in 200 epochs with early stopping ensuring no overfitting and thus fair comparison across methods. The hyperparameters were tuned within the cross validation in the ranges: $lr \in [10^{-4}, 10^{-5}]$, $\lambda_1 \in [0.5, 1]$ and $\lambda_2 \in [0.01, 0.001]$

3 Experimental Results

Dataset: We validate SzXAI on 642 EEG recordings from 120 adult epilepsy patients in the publicly available Temple University Hospital (TUH) corpus [21]

[3] All scripts available on the Github repository

Table 1: Performance of multi-label etiology prediction in 10-fold CV setup averaged across 20 etiologies. *Empirically-determined chance prediction is 0.05.*

Method	DeepSOZ[16]			CNN-BLSTM[8]		
	Recall	Precision	Specificity	Recall	Precision	Specificity
SzXAI-avg	0.42±0.09	0.29±0.05	0.57±0.01	**0.57±0.12**	0.32±0.10	0.50±0.08
SzXAI-max	0.50±0.07	0.29±0.05	0.45±0.01	0.35±0.12	0.28±0.13	0.52±0.08
SzXAI-bce-NAl	0.47±0.11	0.26±0.05	0.42±0.02	0.49±0.11	0.30±0.14	0.45±0.06
SzXAI-bce	0.44±0.08	0.28±0.04	0.54±0.02	0.45±0.10	0.31±0.12	0.45±0.06
SzXAI-NAl	0.42±0.06	0.28±0.04	0.48±0.01	0.43±0.07	0.3±0.09	0.54±0.09
Model of [18]	0.46±0.04	0.30±0.04	0.56±0.01	0.52 ±0.11	0.30±0.15	0.46±0.09
Model of [25]	0.42±0.10	0.27±0.05	0.49±0.04	0.43±0.08	0.29±0.11	0.5±0.04
SzXAI	**0.50±0.05**	**0.32±0.06**	**0.58±0.01**	0.52±0.14	**0.32±0.11**	**0.59±0.07**

within the age range of 19-91 years (average 55±16.6) who have clinically confirmed seizures of various types in their EEG with single-expert annotated and de-identified clinical notes describing the seizure characteristics. EEG recordings lasted an average of 79.8±135 min with 14.7±25.2 seizures per subject, each lasting 88.0±123.5 seconds. The extracted ground truth concept representation is imbalanced, ranging from 80 to 420 EEG recordings per concept (Fig. 2(Top)).

Raw EEG signals are resampled to 200 Hz, filtered between 1.6-30 Hz, and clipped at two standard deviations from the mean to remove muscle artifacts. To ensure uniformity, each recording is separately normalized to have a zero mean and unit variance. The recordings are cropped to 10 minutes around the seizure event, maintaining a uniform distribution of onset times. Finally, we segment the 10-minute recording into non-overlapping 1-second windows to be fed into the seizure detectors, which provide a binary prediction for each window of seizure versus non-seizure to aid SzXAI's attention-weighted etiology prediction.

Etiology Prediction Performance: Table 1 quantifies the etiology prediction performance (precision, recall, and specificity) across all experiments within our 10-fold cross-validation setup. We empirically determined the average chance prediction level to be 0.05 in our dataset. On average, SzXAI achieved the highest precision of 0.32 across all 20 etiologies, and the highest specificity of 0.58 in DeepSOZ and 0.59 in CNN-BLSTM while achieving comparable recall. The pooling ablations are severely affected by class imbalance, thus reducing specificity in SzXAI-max and recall in SzXAI-avg in DeepSOZ. A reverse trend is seen in CNN-BLSTM where SzXAI-avg achieves the highest recall at the cost of increased false positives. The remaining baselines show lower performance overall, highlighting the importance of our cross-modal alignment loss during training.

In all cases, the seizure detection performance is maintained at AUROC of 0.92±0.03 for DeepSOZ and 0.90±0.03 for CNN-BLSTM, which are frozen.

To better understand the results, we group the etiologies into four types: (i) Spiking, (ii) Rhythmic, (iii) Interictal discharges, and (iv) Background. The first two strongly indicate seizures, interictal discharges occur around seizures, and the background covers non-seizure activity. Fig. 3(a) shows a confusion ma-

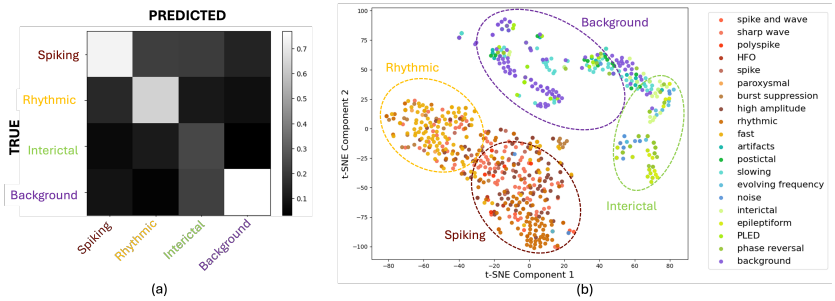

Fig. 3: **(a)** Joint distribution of predicted vs true etiologies in four groups. **(b)** t-SNE scatter plot of EEG encodings in the validation set. HFO: High Frequency Oscillations. PLED: Partial Lateralized Epileptiform Discharge.

trix when comparing these groups in the true and predicted labels for SzXAI applied to DeepSOZ. The large diagonal entries indicate a high recall ~ 0.7; however the interictal class has a has lower sensitivity due to class imbalance and similarity with background activity. Fig. 3(b) presents a t-SNE plot from the best-performing split in the cross validation, showing a clear separation between seizure and non-seizure etiologies and strong intra-group clustering with "background" being distributed possibly due to its varied features. These results demonstrate the strong discriminative ability of crossmodal alignment in SzXAI.

Fig. 4 presents the temporal alignment between the SzXAI encodings $\tilde{\mathbf{h}}^{1:T}$ and textual concept embeddings for an EEG recording with two seizures. Clinical notes identified seizure-related patterns ("spike-and-wave", "sharp wave", "rhythmic") and postictal features ("slowing", "evolving frequency"), both highly correlated with EEG embeddings at respective time intervals only. SzXAI also predicted related spiking and background etiologies. Finally, we can use LLaMA to generated a coherent summary from the predicted etiologies, as shown in green.

4 Conclusion

We have presented SzXAI , a novel explainable AI framework that can be added to any EEG-based seizure detector. SzXAI predicts seizure and baseline etiologies from EEG using cross-modal similarity and attention-weighted pooling across two deep networks. It successfully aligns EEG dynamics with clinical concepts through a novel knowledge-based loss function, which creates a distinctive embedding space. We propose a training approach that leverages pretrained LLMs for label construction and uses an LLM to generate human-readable summaries of the seizure detection reasoning. Compared to existing methods, SzXAI proved to be model-agnostic and explainable, taking a step toward clinical integration.

Acknowledgments. This work was supported by the National Institutes of Health awards 1R01HD108790 (PI Venkataraman) and 1R01EB029977 (PI Caffo) and the National Science Foundation CAREER award 1845430 (PI Venkataraman).

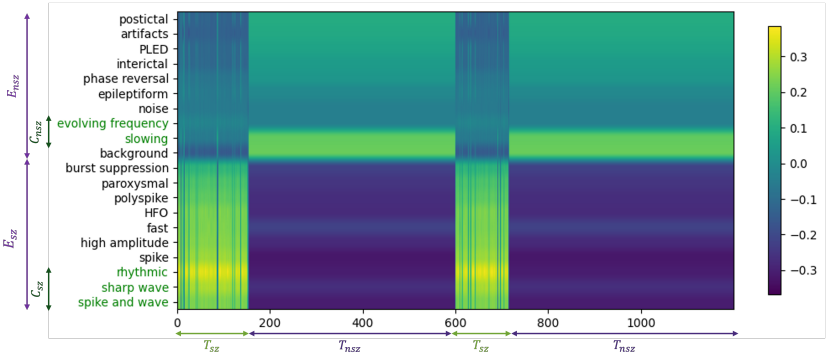

Fig. 4: Temporal heatmap of correlation of EEG encodings of a patient with all etiologies in **E** along with the sentence generated by prompting Llama with predicted etiologies in bold. True etiologies are presented in green in the legend.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adlung, L., Cohen, Y., Mor, U., Elinav, E.: Machine learning in clinical decision making. Med **2**(6), 642–665 (2021)
2. Amin, U., Benbadis, S.R.: The role of eeg in the erroneous diagnosis of epilepsy. Journal of clinical neurophysiology **36**(4), 294–297 (2019)
3. Binnie, C.D., Stefan, H.: Modern electroencephalography: its role in epilepsy management. Clinical Neurophysiology **110**(10), 1671–1697 (1999)
4. Biswal, S., Xiao, C., Westover, M.B., Sun, J.: Eegtotext: learning to write medical reports from eeg recordings. In: Machine Learning for Healthcare Conference. pp. 513–531. PMLR (2019)
5. de Borman, A., Vespa, S., El Tahry, R., Absil, P.A.: Estimation of seizure onset zone from ictal scalp eeg using independent component analysis in extratemporal lobe epilepsy. Journal of neural engineering **19**(2), 026005 (2022)
6. Corrao, S., Argano, C.: Rethinking clinical decision-making to improve clinical reasoning. Frontiers in Medicine **9**, 900543 (2022)
7. Covert, I.C., et al.: Temporal graph convolutional networks for automatic seizure detection. In: Machine Learning for Healthcare Conference. pp. 160–180. PMLR (2019)
8. Craley, J., et al.: Automated inter-patient seizure detection using multichannel convolutional and recurrent neural networks. Biomedical signal processing and control **64**, 102360 (2021)
9. Craley, J., et al.: Szloc: A multi-resolution architecture for automated epileptic seizure localization from scalp eeg. In: Medical Imaging with Deep Learning (2022)
10. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

11. Hah, H., Goldin, D.S.: How clinicians perceive artificial intelligence–assisted technologies in diagnostic decision making: Mixed methods approach. Journal of Medical Internet Research **23**(12), e33540 (2021)
12. Kim, J.W., Alaa, A., Bernardo, D.: Eeg-gpt: exploring capabilities of large language models for eeg classification and interpretation. arXiv preprint arXiv:2401.18006 (2024)
13. Kingma, D.P., et al.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
15. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)
16. M. Shama, D., Jing, J., Venkataraman, A.: Deepsoz: A robust deep model for joint temporal and spatial seizure onset localization from multichannel eeg data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 184–194. Springer (2023)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Patrício, C., Teixeira, L.F., Neves, J.C.: Towards concept-based interpretability of skin lesion diagnosis using vision-language models. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
19. Paul, Y.: Various epileptic seizure detection techniques using biomedical signals: a review. Brain informatics **5**, 1–19 (2018)
20. Ramaswamy, V.V., Kim, S.S., Fong, R., Russakovsky, O.: Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10932–10941 (2023)
21. Shah, V., et al.: The temple university hospital seizure detection corpus. Frontiers in neuroinformatics **12**, 83 (2018), https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml
22. Shinnar, S., Kang, H., Berg, A.T., Goldensohn, E.S., Hauser, W.A., Moshé, S.L.: Eeg abnormalities in children with a first unprovoked seizure. Epilepsia **35**(3), 471–476 (1994)
23. Shoji, T., Yoshida, N., Tanaka, T.: Automated detection of abnormalities from an eeg recording of epilepsy patients with a compact convolutional neural network. Biomedical Signal Processing and Control **70**, 103013 (2021)
24. Wu, Y., Liu, Y., Yang, Y., Yao, M.S., Yang, W., Shi, X., Yang, L., Li, D., Liu, Y., Gee, J.C., et al.: A concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data. arXiv preprint arXiv:2403.05606 (2024)
25. Yan, A., Wang, Y., Zhong, Y., He, Z., Karypis, P., Wang, Z., Dong, C., Gentili, A., Hsu, C.N., Shang, J., et al.: Robust and interpretable medical image classifiers via concept bottleneck models. arXiv preprint arXiv:2310.03182 (2023)
26. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19187–19197 (2023)

27. Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z.: Biowordvec, improving biomedical word embeddings with subword information and mesh. Scientific data **6**(1), 52 (2019)