This EEG Looks Like These EEGs: Interpretable Interictal Epileptiform Discharge Detection With ProtoEEG-kNN

Dennis Tang[1], Jon Donnelly[1], Alina Jade Barnett[1], Lesia Semenova[2], Jin Jing[3], Peter Hadar[4], Ioannis Karakis[5,6], Olga Selioutski[7], Kehan Zhao[3], M. Brandon Westover[3], and Cynthia Rudin[1]

[1] Department of Computer Science, Duke University, USA
[2] Microsoft Research, USA
[3] Beth Israel Deaconess Medical Center, Harvard Medical School, USA
[4] Massachusetts General Hospital, Harvard Medical School, USA
[5] Department of Neurology, Emory University School of Medicine, USA
[6] Department of Neurology, University of Crete School of Medicine, Greece
[7] Department of Neurology, Stony Brook University, USA
Dennis.tang@duke.edu

Abstract. The presence of interictal epileptiform discharges (IEDs) in electroencephalogram (EEG) recordings is a critical biomarker of epilepsy. Even trained neurologists find detecting IEDs difficult, leading many practitioners to turn towards machine learning for help. Although deep learning algorithms have shown state-of-the-art accuracy on this task, most models are uninterpretable and cannot justify their conclusions. Absent the ability to understand model reasoning, doctors cannot leverage their expertise to identify incorrect model predictions and intervene accordingly. To improve human-model interaction, we introduce ProtoEEG-kNN, an inherently interpretable IED-detection model that follows a simple case-based reasoning process. Specifically, ProtoEEG-kNN compares input EEGs to samples from the training set that contain similar IED morphology (shape) and spatial distribution (location). We show that ProtoEEG-kNN can achieve state-of-the-art accuracy while providing visual explanations that experts prefer over existing approaches.

Keywords: Interpretability · Epilepsy Diagnosis · Deep Learning

1 Introduction

Epilepsy, a chronic neurological disorder characterized by recurring seizures, affects approximately 50 million people worldwide [24]. Epilepsy significantly impairs quality of life, increases risk for injuries, and reduces life expectancy when inadequately managed. To diagnose epilepsy, clinicians look for electrophysiological events known as interictal epileptiform discharges (IEDs) in electroencephalogram (EEG) recordings. However, identifying IEDs among benign variations in brain activity is difficult, with disagreement being common even among trained neurologists [10]. To aid epilepsy diagnosis, clinicians and researchers have recently turned to deep learning models [3]. However, despite recent jumps in model accuracy, many of these models remain uninterpretable – providing no insight into how decisions are made. This paradigm is problematic because when a practitioner disagrees with a model, there is no way to check the model's reasoning for validity.

In contrast, interpretable models – models designed to explain the reasoning behind their decisions – allow practitioners to assess model predictions and incorporate machine learning insights into the diagnostic process. One such model is the Prototypical Part Network (ProtoPNet) [2], a family of interpretable neural networks that achieve accuracy on par with black box models. However, existing ProtoPNets are ill-equipped to handle the unique challenges of the EEG domain. Specifically, they are unable to handle uncertain labels, cannot capture the complex interplay between spatial relationships (location) and morphological patterns (shape) that characterize IEDs [16, 12], and struggle to learn semantically meaningful prototypes due to the extreme variability among IEDs.

To address these challenges, we introduce ProtoEEG-kNN, an interpretable IED-detection model that achieves state-of-the-art accuracy. Our model learns an effective EEG comparison space by training a ProtoPNet with a new similarity metric that incorporates selected interpretable statistical features (ISFs) and specialized spatial reasoning. Once this space is learned, we alter ProtoEEG-kNN to use k-Nearest Neighbors (kNN) reasoning over these learned embeddings, providing intuitive comparisons of the form "This IED-containing EEG looks like these IED-containing EEGs," (Fig. 1 (Top)) with coverage over the extreme diversity of IEDs. Specifically, our contributions are: **(1)** We adapt ProtoPNet into a kNN based probabilistic classification model and update the loss terms to reflect training under uncertain labels. **(2)** We define a new similarity metric that aligns our model's notion of EEG similarity with clinical practice by capturing both spike morphology and spatial distribution patterns. **(3)** We use channel masking to calculate channel-wise weights that allow the model to prioritize computations on medically relevant channels while revealing the spatial focus of the model's attention across the EEG. Our code is available at: https://github.com/DennisTang2000/ProtoEEG and data was released in [14].

2 Related Works

There has been a dramatic increase in interest in IED detection using machine learning models [3], resulting in a wide variety of uninterpretable predictive approaches. Generally, IED detection operates at either the channel-level [4, 7, 20] or by analyzing entire EEGs at once [11, 13, 21].

In computer vision, a large body of work has emerged around interpretable neural networks, based on the Prototypical Part Network (ProtoPNet) [2]. ProtoPNet provides an interpretable alternative to traditional neural networks by forming predictions using a series of comparisons to learned prototypical parts. A ProtoPNet can explain its predictions by saying "this image is of class A be-

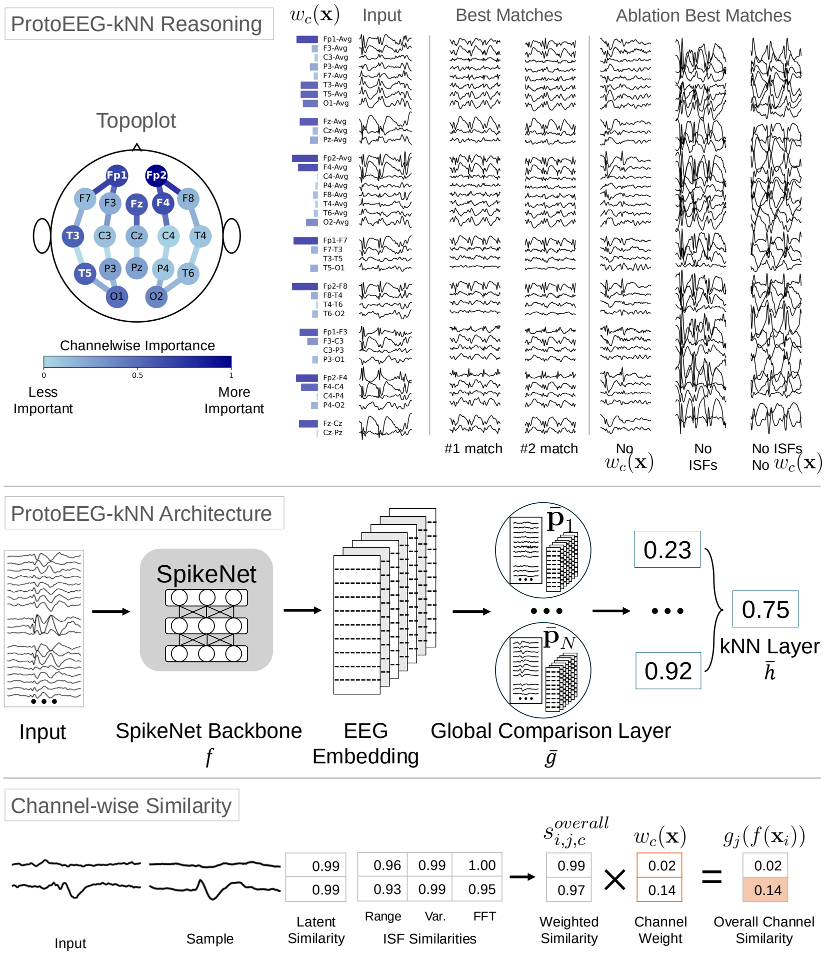

Fig. 1: **Top:** ProtoEEG-kNN reasoning. The topographic map ("topoplot") highlights important channels as calculated by the channel-wise weights ($w_c(\mathbf{x})$), which are also shown in bars to the left of the input channels. From left to right, we show the input sample, the best two matches selected by our model, and the best matches chosen by each of three ablated models. **Middle:** ProtoEEG-kNN architecture. An input is passed through the backbone f to produce a embedding. The Global Comparison Layer $\bar{g}$ computes the similarity between the embedding and each sample in the training set. The final prediction produced by $\bar{h}$ is the average label from the top-k most similar neighbors. **Bottom:** Channel-Wise Similarity Calculation. The similarity along each channel combines the latent and ISF similarities and then multiplies by $w_c(\mathbf{x})$.

cause it looks like this prototype from class A". Of particular interest to this work, Ukai et al. [22] introduce ProtoKNN, which performs kNN-style classification over the vector of prototype similarities. This is different from our kNN approach which computes a specialized similarity metric between an input and *every* training sample. Several papers have applied ProtoPNet style reasoning to IED detection [5, 6, 19]. Gao et al. [6], introduces Multi-Scale Prototypical Part Network for patient-specific seizure prediction, while Gao et al. [5] extends this work to cross-patient prediction. However, both are limited to single channel comparisons, thus failing to consider the spatial distribution of spikes, an important factor in how experts identify IEDs [16, 12]. In Tang et al. [19] prototypes represent full EEGs, but convolve every channel together which keeps their model from providing channel-level interpretability. Other papers, such as Lopez et al. [15] and Ozcan et al. [17] apply post-hoc methods to explain blackbox IED detection models, but these explanations are not necessarily faithful to how a model actually makes decisions, and may be misleading [1, 18].

3 Methods

Notation and Setup. We denote our training dataset $\mathcal{D} := \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^{C\times T}$, C is the number of channels in the EEG and T is the length (1 second sampling 128 Hz), and $y_i \in \{0/v, 1/v, \ldots, 1\}$ (in our case, $v = 8$). We treat this as a probabilistic classification problem because expert annotators often disagree on labels for this task (in 80.68% of samples in our dataset).

Our model architecture is inspired by that of ProtoPNet [2], and we train a specialized ProtoPNet to shape the latent space before replacing the learned prototype layer with a kNN module. During training, the architecture of our model consists of a feature extraction backbone $f : \mathbb{R}^{C\times T} \to \mathbb{R}^{L\times C}$, followed by prototype layer $g : \mathbb{R}^{L\times C} \to \mathbb{R}^{M}$, and a final class-connection layer $h : \mathbb{R}^{\mathrm{M}} \to [0, 1]$. Here, L and M are the latent dimension and number of prototypes respectively. For our backbone f, we use Spikenet, a pre-trained IED classification model. We remove SpikeNet's classifier head and alter the convolution layers to not convolve across EEG-channels, producing embeddings with C separated channels. At the end of training, we replace g and h with kNN style-components $\bar{g}$ and $\bar{h}$, which involves creating a prototype for every training sample and setting $M = N$. This results in the architecture shown in Fig. 1 (Middle).

Next, we introduce the novel features of our model: a new similarity metric that leverages ISFs, channel-wise weights, and a kNN layer.

ISFs and Prototype Similarity. Traditional ProtoPNets compare inputs to prototypes by computing their cosine similarity in latent space. For our model, we define each prototype $\mathbf{p}_j \in \mathbb{R}^{L\times C}$ from our set of prototypes $\mathcal{P}_g := \{\mathbf{p}_j\}_{j=1}^{M}$ in layer g to represent a complete, 37-channel EEG, and we denote channel c in prototype j with $\mathbf{p}_{j,c} \in \mathbb{R}^{L}$. To produce more semantically meaningful comparisons, we augment ProtoPNet's cosine similarity with additional comparisons between three ISFs that were selected based on domain expert feedback: the range, variance, and fast fourier transform (FFT) of each channel. These comparisons are then aggregated across channels with a weighted sum into an overall channel similarity. We introduce three learnable parameter tensors associated with each prototype $\mathbf{p}_j$: $\mathbf{p}_j^{range} \in \mathbb{R}^C$, $\mathbf{p}_j^{var} \in \mathbb{R}^C$, and $\mathbf{p}_j^{fft} \in \mathbb{R}^{C\times T}$, where each entry along the C dimension corresponds to the relevant statistic computed over each channel. This yields four similarity terms: s^{latent}, s^{range}, s^{var}, and s^{fft}, where the superscript defines which set of features the similarity scores are computed along. We define the four similarities between input i and prototype j for a single channel c as:

$$s_{i,j,c}^{latent} = \frac{f_c(\mathbf{x}_i)\cdot \mathbf{p}_{j,c}}{\|f_c(\mathbf{x}_i)\|_2\|\mathbf{p}_{j,c}\|_2}, \qquad s_{i,j,c}^{fft} = \frac{c_{fft}}{\||\mathbf{p}_{j,c}^{fft}| - |FT(\mathbf{x}_{i,c})|\|_2 + \epsilon},$$

$$s_{i,j,c}^{var} = 1 - \left|\frac{Var(\mathbf{x}_{i,c}) - Var(p_{j,c}^{var})}{V_{max} - V_{min} + \epsilon}\right|, \quad s_{i,j,c}^{range} = 1 - \left|\frac{R(\mathbf{x}_{i,c}) - R(p_{j,c}^{range})}{R_{max} - R_{min} + \epsilon}\right|,$$

where $f_c(\mathbf{x}_i) \in \mathbb{R}^L$ denotes the the c-th channel of the latent representation of $\mathbf{x}_i$, $Var(\cdot)$ is the variance, $R(\cdot)$ is the range, $FT(\cdot)$ is the fourier transform, ϵ and c_{fft} are constants for numerical stability, and V_{min}, V_{max}, R_{min}, and R_{max} denote the minimum variance, maximum variance, minimum range, and maximum range across all channels in the training set respectively. An overall similarity score between two channels is calculated as: $s_{i,j,c}^{overall} = \lambda_1 s_{i,j,c}^{latent} + \lambda_2 s_{i,j,c}^{range} + \lambda_3 s_{i,j,c}^{var} + \lambda_4 s_{i,j,c}^{fft}$, where $\lambda_i := sm(\lambda_1', \lambda_2', \lambda_3', \lambda_4')$ for learned parameters $\lambda_1', \lambda_2', \lambda_3', \lambda_4'$, and sm denotes the softmax function.

Channel-wise Weights. To focus the model's similarity comparisons along relevant channels and to provide channel-level interpretability, we calculate a channel-wise weight for every channel in the input. We use a leave-one-channel-in masking approach and define the weight function $w_c : \mathbb{R}^{C\times T} \to \mathbb{R}$ such that $w_c(\mathbf{x}_i) = \frac{\tilde{w}_c(\mathbf{x}_i)}{\sum_{c\in C}\tilde{w}_c(\mathbf{x}_i)}, \tilde{w}_c(\mathbf{x}_i) = h_{spikenet}(f([\mathbf{0}^{c-1\times T}; \mathbf{x}_{i,c}; \mathbf{0}^{C-c\times T}]))$, where f is the backbone, $h_{spikenet}$ is the classifier head of SpikeNet, $\mathbf{0}^{A\times B}$ denotes an $A \times B$ dimensional matrix of zeroes, and ; indicates concatenation. Since each weight $w_c(\mathbf{x}_i)$ assigns a relative importance to the similarity score along channel c, an overall similarity score between input i and prototype j is calculated: $g_j(f(\mathbf{x}_i)) = \sum_{c=1}^{C} w_c(\mathbf{x}_i) s_{i,j,c}^{overall}$ (Fig. 1(Bottom)). Given our similarity function, we focus next on model training.

Weighted Loss Terms. We train our model to produce well calibrated predictions with binary-cross entropy loss $\mathcal{L}_{bce} = -y_i \log(\hat{y}_i) - (1 - y_i)\log(1 - \hat{y}_i)$ where y_i is the vote proportion. This way, we can retain the primary function of IED-classifcation with the added benefit of calibrating our model to also match the vote proportions.

Moreover, we adapt the loss terms (Cluster, Separation, Orthogonality) from ProtoPNet to handle uncertain labels. Let $cos(\mathbf{p}_j, \mathbf{p}_{j'}) := \frac{vec(\mathbf{p}_j)\cdot vec(\mathbf{p}_{j'})}{\|vec(\mathbf{p}_j)\|_2\|vec(\mathbf{p}_{j'})\|_2}$ denote the cosine similarity between two prototypes, where $vec(\mathbf{p}_j)$ denotes the vectorization of $\mathbf{p}_j$. We define the loss across a batch as:

$$\begin{aligned}
\mathcal{L}_{ortho} &= \sqrt{\sum_{j=1}^{M}\sum_{j'=1}^{M}\mathbf{1}_{[j\neq j']}cos^2(\mathbf{p}_j,\mathbf{p}_{j'})} + \sqrt{\sum_{j=1}^{M}\sum_{j'=1}^{M}\mathbf{1}_{[j\neq j']}cos^2(\mathbf{p}_j^{fft},\mathbf{p}_{j'}^{fft})}\ ,\\
\mathcal{L}_{clst} &= -\frac{1}{B}\sum_{i=1}^{B}\max_{j^\dagger:\text{class}(j^\dagger)=y_i} y_i g_{j^\dagger}(f(\mathbf{x}_i)),\\
\mathcal{L}_{sep} &= \frac{1}{B}\sum_{i=1}^{B} g_{j^*}(f(\mathbf{x}_i))\cdot|\text{class}\,(j^*)-y_i|, \text{where } j^* = \operatorname*{arg\,max}_{j:\text{class}(j)\neq y_i} g_j(f(\mathbf{x}_i)),
\end{aligned}$$

where $\mathbf{1}_{[\cdot]}$ denotes the indicator function, class $(j) \in [0,1]$ is the class associated with prototype j, and B is the batch size. Cluster and separation losses are adapted to scale based on the sample labels: cluster is higher for samples with higher vote proportions and separation is higher when vote proportions are wider apart. Finally, we add a regularization loss $\mathcal{L}_{CoefReg} = \lambda_1 - min(\lambda_2, \lambda_3, \lambda_4)$ to train balanced coefficients for ISFs.

We minimize the overall loss function $\mathcal{L}_{overall} := \kappa_1\mathcal{L}_{bce} + \kappa_2\mathcal{L}_{ortho} + \kappa_3\mathcal{L}_{clst} + \kappa_4\mathcal{L}_{sep} + \kappa_5\mathcal{L}_{CoefReg}$, where each κ is a scalar hyperparameter, using Adam optimization. We denote the model $h \circ g \circ f$ as "EEG ProtoPNet" and train according to the regime described in [2] to produce a well-structured latent space when combined with our ISFs. Training lasted 200 epochs and stopped early if validation accuracy did not improve for two consecutive project epochs.

kNN Replacement Step. After EEG ProtoPNet training converged, we replaced the learned prototype layer g with a Global Comparison Layer $\bar{g} : \mathbb{R}^{L\times C} \to \mathbb{R}^N$ and the linear layer h with a kNN comparison layer $\bar{h} : \mathbb{R}^N \to [0,1]$. This is our final model, "ProtoEEG-kNN." The Global Comparison Layer $\bar{g}$ can be thought of as a prototype layer in which every training sample is a prototype. Formally, we set $\bar{\mathbf{p}}_i := f(\mathbf{x}_i)$, $\bar{p}_{i,c}^{range} := R(\mathbf{x}_{i,c})$, $\bar{p}_{i,c}^{var} := Var(\mathbf{x}_{i,c})$, and $\bar{\mathbf{p}}_{i,c}^{fft} := FT(\mathbf{x}_{i,c})$ for $i \in \{1, 2, \ldots, N\}$, and $\bar{g}$ operates as a prototype layer with prototypes $\mathcal{P}_{\bar{g}} := \{\bar{\mathbf{p}}_i\}_{i=1}^{N}$. This makes $\bar{g}_j(f(\mathbf{x}_i))$ the similarity between the j-th training sample and the input $\mathbf{x}_i$, using the weighted similarity metric defined previously. The kNN layer $\bar{h}$ averages the nearest neighbors' labels and is formalized as $\bar{h} \circ \bar{g} \circ f(\mathbf{x}_i) := \frac{1}{k}\sum_{j'\in\text{topk}(\bar{g}(f(\mathbf{x}_i)))} y_{j'}$, where topk returns the k largest indices in a vector and y_j denotes the label of the j-th training sample. ProtoEEG-kNN is therefore the composition $\bar{h} \circ \bar{g} \circ f$. We perform a grid search of k set to 5, 10, 15 and 20 and find $k = 10$ to have the highest accuracy on the validation set. Lastly, we note that the kNN replacement step increases the number of prototype comparisons during inference by a factor of N/M. Although this may seem computationally prohibitive, the single instruction, multiple data parallelism inherent in GPUs efficiently manages this overhead: inference over the test set only increases from ∼2 to ∼6 seconds. In Section 4, we demonstrate that the expanded number of comparisons increases accuracy and substantially improves interpretability.

4 Results

We train and evaluate ProtoEEG-kNN using a dataset of 16,499 1-second EEG segments labeled by 8 annotators. Participants were recruited from three settings: intensive care unit (n = 446), routine / outpatient EEG (n = 1,161), and epilepsy monitoring unit (n = 104). The data consists of 841 males (mean age = 36.56 years) and 921 females (mean age = 36.92 years). The data was split into 12,411 training, 2,151 validation, and 1,937 test samples, with no patient overlap between sets. This ensures that samples from the test set are compared only with EEGs from other patients. Samples are arranged in standard, 37-channel, "double-banana" format [9], were filtered (60-Hz notch, 0.5-Hz high-pass), and resampled to 128 Hz. Following the annotation procedure in Jing et al. [11], for each EEG sample, 8 subspecialist physicians independently annotated whether they observed an IED.

ProtoEEG-kNN was trained on a Nvidia P100 GPU for $\sim$ 5 clock hours. Class-balanced sampling was used during training and k was set to 10 in $\bar{h}$. We now describe our process to evaluate ProtoEEG-kNN's accuracy, assess its match-quality, and ablate its novel components.

ProtoEEG-kNN is Accurate. We evaluated model performance using binary classification accuracy, AUROC, and R^2. For binary classification and AUROC, we assigned a sample to the positive class if $y_i \geq 0.5$. To calculate R^2, we used y_i, the vote proportion. On the held-out test set, we evaluated ProtoEEG-kNN, SpikeNet, kNN over the FFT of the EEG samples, kNN over the ISFs of the EEG samples, Deep kNN [25], and EEG ProtoPNet.

The optimal weighting coefficients for kNN over the ISFs were determined on the validation set by evaluating every combination of coefficients that sum to 1 in increments of 0.1. For Deep kNNs, we train the latent space of SpikeNet and copy Deep kNN's exact hyper-parameter and optimization configuration. As shown in Table 1 (Top), ProtoEEG-kNN substantially outperforms existing models for this task in terms of binary classification, AUROC and R^2.

ProtoEEG-kNN produces good matches. To demonstrate that ProtoEEG-kNN produces quality matches that align with medical intuition, we conducted a user study with four board-certified neurologists (with 2-16 years of clinical experience) and a clinical neurophysiology/EEG fellow. Experts were shown 100 'reference' EEG samples from the test set and ranked the similarity of four 'candidate' matches. Three candidates were the top matches identified by ProtoEEG-kNN, Deep kNN, and EEG-ProtoPNet, while the fourth was a randomly selected sample that shared the reference's classification label. For each ranking, the order of candidates was randomized and the selection method was hidden. We restricted reference samples to have label ≥ 0.75 to ensure clear IED patterns for matching.

To quantify each model's match quality, we used best-match frequency and Plackett-Luce model weights. Best-match frequency indicates how often each

Method	Binary Accuracy	AUROC	R^2
SpikeNet	77.12	0.844	0.429
kNN over FFT	70.72	0.720	0.209
kNN over ISFs	74.39	0.733	0.210
Deep kNN [25]	77.16 ± 0.01	0.805 ± 0.007	0.341 ± 0.019
EEG-ProtoPNet	80.24 ± 0.36	0.866 ± 0.006	0.207 ± 0.019
ProtoEEG-kNN (ours)	**81.15 ± 0.29**	**0.876 ± 0.000**	**0.529 ± 0.007**
Ablations			
Remove w_c	80.74 ± 0.08	0.878 ± 0.002	0.536 ± 0.003
Remove ISFs	80.91 ± 0.00	0.878 ± 0.001	**0.538 ± 0.005**
Remove w_c & ISFs	81.09 ± 0.61	**0.885 ± 0.004**	0.531 ± 0.027
ProtoEEG-kNN (complete)	**81.15 ± 0.29**	0.876 ± 0.000	0.529 ± 0.007

Table 1: Performance of ProtoEEG-kNN compared to baselines (**Top**) and ablated models (**Bottom**). For models that required additional training, we train with 3 different random seeds and report mean and standard deviation. For deterministic models and the model provided by another research group, the results of that single model is reported.

model was ranked first, while Plackett-Luce weights consider the full ranking distribution and represents the probability each model provides the best match [8]. Across both metrics, ProtoEEG-kNN produces matches that align the closest with expert opinion (Fig. 2 (Top)).

We also qualitatively evaluate the comparison space of our model by using the dimension reduction tool PaCMAP [23] to visualize the distribution of the test set under ProtoEEG-kNN's similarity metric. Relative to the comparison space based on the kNN over ISFs' similarity metric, ProtoEEG-kNN learns more distinct and well-separated classes (Fig. 2 (Bottom)).

Ablations. Finally, we evaluate ProtoEEG-kNN's performance without channel-wise weights and ISFs (Table 1 (Bottom)). The inclusion of channel-wise weights and ISFs marginally effects binary classification (↑ 0.06%), AUROC (↓ 0.0084), R^2 (↓ 0.0022), while resulting in much closer matches (Fig. 1 (Top)).

5 Conclusion

We introduced ProtoEEG-kNN, an interpretable model for IED detection that achieves state-of-the-art performance while providing interpretable reasoning for its decisions in the form of "This EEG looks like these EEGs". In addition to being interpretable, our model's kNN layer, similarity metric, and channel-wise weights scores constrain it to reason in a way that aligns with clinical intuition about spike morphology and spatial distribution, as shown through our

Method	Plackett-Luce Weight	Best-Match Frequency
Random	0.128 (0.111, 0.144)	0.104
EEG-ProtoPNet	0.078 (0.065, 0.088)	0.052
Deep kNN	0.333 (0.298, 0.368)	0.370
ProtoEEG-kNN	**0.462 (0.427, 0.501)**	**0.474**

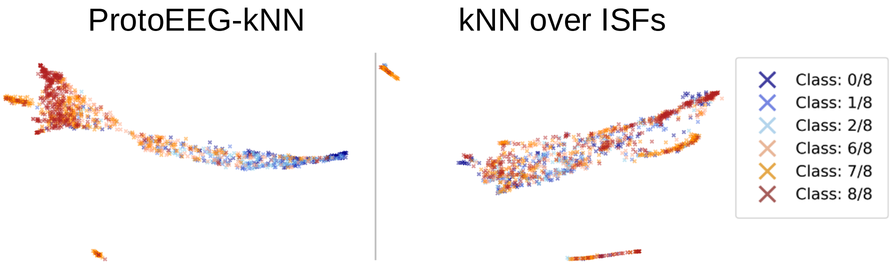

Fig. 2: **Top:** User Study Results. Bootstrapping with 1,000 iterations was used to calculate the mean and 95% confidence interval for Plackett-Luce weights. **Bottom:** PaCMAP visualization of the test set comparison spaces of ProtoPNet-kNN (left) and kNN over ISFs (right). Neighborhoods in high-dimensional space are preserved in two-dimensional PaCMAPs.

user study. While ProtoEEG-kNN demonstrated promising results, future work should externally validate ProtoEEG-kNN using different patient populations to confirm its generalizability. Nonetheless, ProtoEEG-kNN offers a promising path forward for the integration of machine learning into clinical practice.

Acknowledgments. This work was funded by the National Science Foundation under grant HRD-2222336, NSF Graduate Research Fellowship grant DGE 2139754, and the National Institutes of Health under NIDA R01 DA054994. We thank Luke Moffett for granting access to his at-the-time unpublished refactor of the ProtoPNet codebase, ProtoPNeXT and Eric Monson for his invaluable help with data visualization.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I.J., Hardt, M., Kim, B.: Sanity Checks for Saliency Maps. Advances in Neural Information Processing Systems **31** (2018)
2. Chen, C., Li, O., Tao, D., Barnett, A., Su, J.K., Rudin, C.: This Looks Like That: Deep Learning for Interpretable Image Recognition. Advances in Neural Information Processing Systems **32** (2019)

3. Diniz, J.B.C., Santana, L.S., Leite, M., Santana, J.L.S., Costa, S.I.M., Castro, L.H., Telles, J.P.M.: Advancing Epilepsy Diagnosis: A Meta-Analysis of Artificial Intelligence Approaches for Interictal Epileptiform Discharge Detection. Seizure: European Journal of Epilepsy **122**, 80–86 (2024)
4. Fürbass, F., Kural, M.A., Gritsch, G., Hartmann, M.M., Kluge, T., Beniczky, S.: An Artificial Intelligence-Based EEG Algorithm for Detection of Epileptiform EEG Discharges: Validation Against the Diagnostic Gold Standard. Clinical Neurophysiology **131**, 1174–1179 (2020)
5. Gao, Y., Liu, A., Cui, H., Qian, R., Chen, X.: An Interpretable and Generalizable Deep Learning Model for iEEG-based Seizure Prediction Using Prototype Learning and Contrastive Learning. Computers in Biology and Medicine **183**, 109257 (2024)
6. Gao, Y., Liu, A., Wang, L., Qian, R., Chen, X.: A Self-Interpretable Deep Learning Model for Seizure Prediction Using a Multi-Scale Prototypical Part Network. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 1847–1856 (2023)
7. Geng, D., Alkhachroum, A., Bicchi, M.M., Jagid, J.R., Cajigas, I., Chen, Z.S.: Deep Learning for Robust Detection of Interictal Epileptiform Discharges. Journal of Neural Engineering **18** (2021)
8. Hunter, D.R.: MM Algorithms for Generalized Bradley-Terry Models. Annals of Statistics **32**, 384–406 (2003)
9. Jadeja, N.: How to Read an EEG (2021), Chapter 3
10. Jing, J., Jing, J., Herlopian, A., Herlopian, A., Karakis, I., Ng, M., Halford, J.J., Lam, A.D., Maus, D., Chan, F., Dolatshahi, M., Muniz, C.F., Chu, C.J., Sacca, V., Pathmanathan, J.S., Pathmanathan, J.S., Ge, W., Sun, H., Dauwels, J., Cole, A.J., Hoch, D.B., Cash, S.S., Westover, M.B.: Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. JAMA Neurology **77**, 49–57 (2020)
11. Jing, J., Jing, J., Sun, H., Kim, J.A., Herlopian, A., Karakis, I., Ng, M.C., Halford, J.J., Maus, D., Chan, F., Dolatshahi, M., Muniz, C.F., Chu, C.J., Sacca, V., Pathmanathan, J.S., Ge, W., Dauwels, J., Lam, A.D., Cole, A.J., Cash, S.S., Westover, M.B.: Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation. JAMA Neurology **77**, 103–108 (2020)
12. Kural, M.A., Duez, L., Hansen, V.S., Larsson, P.G., Rampp, S., Schulz, R., Tankisi, H., Wennberg, R.A., Bibby, B.M., Scherg, M., Beniczky, S.: Criteria for Defining Interictal Epileptiform Discharges in EEG. Neurology **94**, e2139 – e2147 (2020)
13. Kural, M.A., Jing, J., Fürbass, F., Perko, H., Qerama, E., Johnsen, B., Fuchs, S., Westover, M.B., Beniczky, S.: Accurate Identification of EEG Recordings With Interictal Epileptiform Discharges Using a Hybrid Approach: Artificial Intelligence Supervised by Human Experts. Epilepsia **63**, 1064 – 1073 (2022)
14. Li, J., Goldenholz, D.M., Alkofer, M., Sun, C., Nascimento, F.A., Halford, J.J., Dean, B.C., Galanti, M., Struck, A.F., Greenblatt, A.S., Lam, A.D., Herlopian, A., Nwankwo, C., Weber, D., Maus, D., Haider, H.A., Karakis, I., Yoo, J.Y., Ng, M.C., Selioutski, O., Taraschenko, O., Osman, G., Katyal, R., Schmitt, S.E., Benbadis, S., Cash, S.S., Tatum, W.O., Sheikh, Z., Kong, W.Y., Bayas, G., Turley, N., Hong, S., Westover, M.B., Jing, J.: Expert-level detection of epilepsy markers in eeg on short and long timescales. NEJM AI **2**(7) (2025)
15. Lopes, M.K.S., Cassani, R., Falk, T.H.: Using CNN Saliency Maps and EEG Modulation Spectra for Improved and More Interpretable Machine Learning-Based Alzheimer's Disease Diagnosis. Computational Intelligence and Neuroscience **2023** (2023)

16. Nascimento, F.A., Barfuss, J.D., Jaffe, A., Westover, M.B., Jing, J.: A Quantitative Approach to Evaluating Interictal Epileptiform Discharges Based on Interpretable Quantitative Criteria. Clinical Neurophysiology **146**, 10–17 (2022)
17. Ozcan, A.R., Erturk, S.: Seizure Prediction in Scalp EEG Using 3D Convolutional Neural Networks With an Image-Based Approach. IEEE Transactions on Neural Systems and Rehabilitation Engineering **27**, 2284–2293 (2019)
18. Rudin, C.: Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
19. Tang, D., Willard, F., Tegerdine, R., Triplett, L., Donnelly, J., Moffett, L., Semenova, L., Barnett, A.J., Jing, J., Rudin, C., Westover, B.: ProtoEEGNet: An Interpretable Approach for Detecting Interictal Epileptiform Discharges. Medical Imaging Meets NeurIPS Workshop (2023)
20. Tjepkema-Cloostermans, M.C., de Carvalho, R., van Putten, M.J.A.M.: Deep Learning for Detection of Focal Epileptiform Discharges From Scalp EEG Recordings. Clinical Neurophysiology **129**, 2191–2196 (2018)
21. Tveit, J., Aurlien, H., Plis, S., Calhoun, V.D., Tatum, W.O., Schomer, D.L., Arntsen, V., Cox, F.M., Fahoum, F., Gallentine, W.B., Gardella, E., Hahn, C.D., Husain, A.M., Kessler, S.K., Kural, M.A., Nascimento, F.A., Tankisi, H., Ulvin, L.B., Wennberg, R.A., Beniczky, S.: Automated Interpretation of Clinical Electroencephalograms Using Artificial Intelligence. JAMA Neurology **80**, 805 – 812 (2023)
22. Ukai, Y., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: This Looks Like it Rather Than That: ProtoKNN for Similarity-Based Classifiers. In: The Eleventh International Conference on Learning Representations (2022)
23. Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y.: Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering T-Sne, UMAP, TriMAP, and PaCMAP for Data Visualization. Journal of Machine Learning Research **22**, 201:1–201:73 (2020)
24. World Health Organization: Epilepsy (2 2024), accessed 20 December 2024
25. Zhuang, J.X., Cai, J., Wang, R., Zhang, J., Zheng, W.: Deep kNN for Medical Image Classification. International Conference on Medical Image Computing and Computer-Assisted Intervention **12261**, 127–136 (2020)