Reasoning Visualization for Critical Care EEG Classification with Prototypical Part Networks

Takuma Bingo*, Hajime Yano* Taichiro Ashizaki†, Kazuma Koda†, Masaya Togo†
Riki Matsumoto‡† and Tetsuya Takiguchi*
* Graduate School of System Informatics, Kobe University, Japan
E-mail: bin5@stu.kobe-u.ac.jp
† Division of Neurology, Graduate School of Medicine, Kobe University, Japan
‡ Department of Neurology, Graduate School of Medicine, Kyoto University, Japan

***Abstract*—Automatic interpretation of electroencephalograms (EEGs) in critical care settings requires not only accurate classification but also clear explanations for model decisions. Prototypical part networks (ProtoPNets) offer inherent interpretability by learning typical local patterns (prototypes) for each class and visualizing the basis of their decisions. In this study, we adapt ProtoPNets for critical care EEG analysis by computing similarity between class-specific prototypes and EEG segments across all time windows and electrodes. Experimental results demonstrate that our method improves classification performance, particularly for periodic discharges (PD) and rhythmic delta activity (RDA), which exhibit temporally periodic and spatially widespread patterns. The proposed model achieved the highest recall for the RDA class (61.1%) and outperformed baselines in accuracy (+3.6%) and macro recall (+5.1%) over EEGNet + BiGRU. Furthermore, the model provides visually interpretable evidence that aligns with clinically typical waveforms, indicating its potential for decision support in critical care environments.**

I. Introduction

Nonconvulsive status epilepticus (NCSE) is a severe neurological condition in which epileptic seizures persist without observable convulsions. It is typically accompanied by impaired consciousness [1]. NCSE is frequently suspected in critically ill patients with altered mental status, where early diagnosis and intervention play a crucial role in determining treatment outcomes. Electroencephalography (EEG) is useful for evaluating such patients, as NCSE is known to exhibit typical abnormalities in EEG recordings [2]. However, diagnosis requires continuous EEG monitoring by clinicians, and it is difficult for a limited number of clinicians to interpret large volumes of EEG data. As a result, automatic EEG interpretation systems are in high demand.

In recent years, deep learning has been widely applied to EEG analysis, achieving high predictive accuracy. However, the reasoning process of these models is often a black box, making it difficult for humans to understand how decisions are made. This challenge has led to increasing interest in explainable artificial intelligence (XAI), particularly in medical and pathological diagnosis [3].

Several post-hoc visualization methods for Convolutional Neural Networks (CNNs) have been proposed, such as saliency maps [4] and class activation mapping [5]. Whereas these techniques can highlight important regions of the input, they do not reflect the actual inference process of the model. Alternatively, inherently interpretable architectures, such as decision trees, linear models, and attention-based methods, have also been explored [6]. Although these models improve interpretability, they often come at the cost of reduced predictive performance. Moreover, attention mechanisms can indicate which parts of the input were emphasized by the model, but they do not explain which prototypical features those parts resemble.

Prototypical part networks (ProtoPNets) [7] are CNN-based architectures designed with inherent interpretability. It visualizes both the input regions that match learned prototypes and the prototypes that contributed to classification. This dual visualization provides clearer reasoning than conventional methods. In a recent application to seizure prediction [8] demonstrated that varying prototype sizes enabled the model to capture multi-scale EEG features, achieving both high accuracy and interpretability. However, ProtoPNets assume that class-specific features are localized and rely on the most similar region for each prototype. This design is not always suited to critical care EEG, where clinically relevant patterns often repeat across time and multiple electrodes.

To address this issue, we modify the similarity scoring and classification process to leverage multiple highly similar regions. By doing so, the model can better capture the patterns often observed in critical care EEG. This approach allows for accurate EEG classification in critical care settings while providing visual explanations for the model' s decisions.

II. ProtoPNets

A. Methods

ProtoPNets are models that integrate interpretability into their architecture by comparing parts of the input with learned prototypes [7]. The model learns class-specific local patterns as prototypes and classifies inputs based on their similarity.

Let Z be the feature map extracted from the input by the convolutional backbone. For each prototype p_j, the model computes a similarity map S_j by calculating the similarity between the prototype and every spatial patch $Z^{a,b}$ in the feature map:

$$S_j^{a,b} = \mathrm{sim}(Z^{a,b}, p_j). \tag{1}$$

979-8-3315-7206-8/25/$31.00

Here, $\mathrm{sim}(\cdot)$ denotes a similarity function (e.g., negative squared l_2 distance) and (a, b) indexes the spatial locations in the feature map. The maximum value in the similarity map is taken as the similarity score between the prototype and the feature map:

$$g_{p_j}(Z) = \max_{a,b} S_j^{a,b}. \tag{2}$$

Finally, the model computes logits for each class $k \in \{1, \ldots, K\}$ by linearly combining the similarity scores of the prototypes $p_j \in P_k$ assigned to class k:

$$\hat{y}_k = \sum_{j\,:\,p_j \in P_k} w_h(k, j) \cdot g_{p_j}(Z), \tag{3}$$

where $w_h(k, j)$ are the weights of the final fully connected layer.

B. Receptive Field-Based Visualization

To accurately align prototypes with local image regions, PIXPNET has been proposed as an extension of ProtoPNet-based methods [9]. Previous ProtoPNet-based approaches typically employed simple linear upsampling to project activation maps onto the input image, often resulting in blurred or spatially imprecise visualizations. In contrast, this approach estimates the corresponding regions in pixel space using receptive fields derived from the backbone architecture, enabling sharper and more semantically grounded visual explanations.

A pixel-space heatmap $M \in \mathbb{R}^{H \times W}$ is initialized as a zero matrix. Given a similarity map S, where each score $s \in S$ represents the similarity between a prototype and a feature patch at a specific spatial location in the feature map. The heatmap is updated based on the receptive field $M^s \subseteq M$ corresponding to each score:

$$\forall s \in S, \quad M^s \leftarrow \max(M^s,\, s). \tag{4}$$

This max-based aggregation ensures that each receptive field region retains the highest activation value it receives, producing a localized and interpretable heatmap that highlights the most relevant image regions for the prediction.

III. Proposed Method

In conventional ProtoPNets and their variants, the similarity score between a prototype and a feature map is computed by applying max-pooling to the similarity map S_j, as shown in (2). The similarity scores for all prototypes are linearly combined to compute the logits for each class. This design is based on the assumption that each object part appears somewhere in an image without repetition. As a result, the models perform classification using only the most similar local patch for each prototype.

However, in EEG recordings from critical care patients, typical waveform patterns often exhibit temporal repetition and spatial spread across multiple electrodes. Max-pooling, which selects only the single most activated patch, may fail to capture the full extent of these clinically relevant patterns, resulting in underutilization of informative features.

To address this issue, we propose two major modifications to the scoring method in ProtoPNets. First, instead of using max-pooling, we compute the prototype score $g_{p_j}(Z)$ by averaging the top-$\alpha\%$ values in the similarity map S_j. Specifically, let S_j^α denote the set of top-$\alpha\%$ values in S_j; the score is then defined as:

$$g_{p_j}(Z) = \frac{1}{|S_j^\alpha|} \sum_{s \in S_j^\alpha} s. \tag{5}$$

Second, we modify the way logits are computed from prototype scores: instead of applying a linear combination, we use the maximum score among the prototypes assigned to each class:

$$\hat{y}_k = \max_{p_j \in P_k} g_{p_j}(Z). \tag{6}$$

This reflects our assumption that a single prototype can match relevant features at multiple positions.

A. Architecture

Fig. 1 illustrates the overall architecture of the ProtoPNet model used in this study. The model consists of three components:

1. **Backbone:** The input signal $x \in \mathbb{R}^{C \times T}$ is transformed into a feature map $Z \in \mathbb{R}^{D \times C \times T}$. Here, C is the number of electrodes, T is the time length, and D is the feature dimension.
2. **Prototype layer:** Each class k has m prototypes $\{p_j \in \mathbb{R}^{D \times 1 \times 1}\}_{j=1}^{m}$. Each prototype is compared with all patches of Z, and similarity scores $g_{p_j}(Z) \in \mathbb{R}^{K \times m}$ are computed using cosine similarity.
3. **Classification layer:** The similarity scores are aggregated and converted into logits.

The backbone architecture is based on the networks proposed by [8], with a key modification: the kernel size along the electrode axis was changed from 3 to 1. This change was made because the ordering of electrodes in EEG data does not reflect their true spatial adjacency. If the kernel size is greater than 1, features from multiple electrodes are combined. As a result, prototypes end up referencing signals across electrodes, which hinders interpretability. By setting the kernel size to 1, the signal at each electrode is processed independently. This ensures that prototypes refer only to waveforms from a single electrode and thereby allow for interpretation invariant to electrode ordering.

B. Training

The training process of the proposed ProtoPNets consists of the following two stages:

1. **Learning Backbone and Prototypes:** Both the backbone and the prototypes are jointly trained using the following loss function:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{clst} + \lambda_2 \mathcal{L}_{sep}, \tag{7}$$

$$\mathcal{L}_{clst} = -\frac{1}{N} \sum_{i=1}^{N} \max_{p_j \in \mathrm{P}_{y_i}} g_{p_j}(Z_i), \tag{8}$$

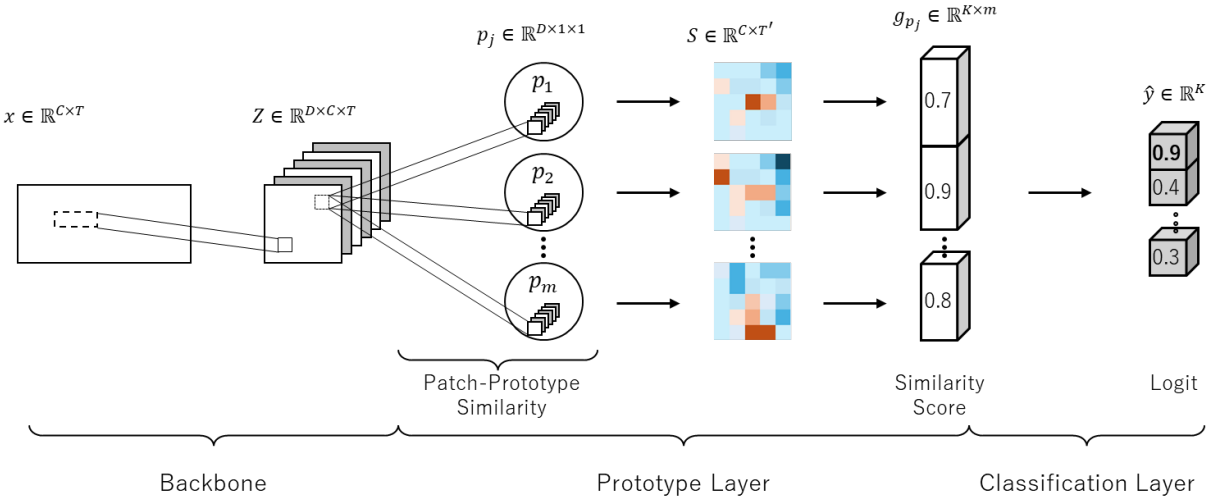


Fig. 1. The ProtoPNet Architecture.

$$\mathcal{L}_{\text{sep}} = \frac{1}{N}\sum_{i=1}^{N} \max\left(0, \max_{p_j \notin \mathrm{P}_{y_i}} g_{p_j}(Z_i) - \max_{p_j \in \mathrm{P}_{y_i}} g_{p_j}(Z_i) + \Delta\right). \tag{9}$$

Here, N denotes the batch size, and $\mathcal{L}_{CE}$ represents the cross-entropy loss. The clustering loss $\mathcal{L}_{clst}$ encourages the feature patches of the input to be close to prototypes of the same class, while the separation loss $\mathcal{L}_{sep}$ pushes them away from prototypes of other classes. Together, these losses promote the formation of class-specific distributions in the feature space.

2. **Prototype Replacement:** Each prototype is replaced with the most similar feature patch from the training data:

$$p_j \leftarrow \underset{z \in \text{patches}(Z)}{\arg\max}\ \text{sim}(z, p_j). \tag{10}$$

This ensures that each prototype corresponds to a visually interpretable segment of an EEG waveform from the training set.

IV. Experiments

We conducted classification experiments using EEG recordings from 329 critically ill patients. The use of EEG data was approved by the Ethics Committee of Kobe University (Approval No. B220114).

A. Dataset

The dataset comprises approximately 108 hours of EEG recordings. Each recording was divided into 10-second segments and normalized using Z-score standardization, based on the mean and standard deviation of each patient's signals.

To prepare the data splits, we first divided the 329 patients into training-validation and test sets using an 80/20 ratio, ensuring no subject overlap. Then, for patients in the training-validation set, their EEG segments were split in temporal order into training and validation sets, again at an 80/20 ratio.

Each EEG segment was labeled with one of five annotation classes. Representative examples of each class are shown in Fig. 2.

- **Periodic discharges (PD):** Repetitive discharges with uniform morphology and duration, occurring at nearly regular intervals and typically lasting less than 0.5 seconds.
- **Rhythmic delta activity (RDA):** Sustained, uniform delta-range (0.5–4 Hz) rhythmic waveforms of a single morphology.
- **Seizure:** Epileptic discharges characterized by evolving high-amplitude spikes, sharp waves, or rhythmic activity, lasting several seconds or more with clear temporal and spatial progression.
- **Slowing:** Delta or theta slowing (0.5–8 Hz), typically associated with various forms of cerebral dysfunction.
- **Artifact:** High-amplitude or high-frequency noise caused by physiological or environmental sources such as patient movement.

B. Classification Performance Evaluation Metrics

To evaluate the performance of the classification model, we used accuracy, recall, and F1 score. In addition, we adopted macro recall and macro F1, which are the average recall and F1 score across all classes, as overall performance metrics. Recall was chosen over precision because minimizing the oversight of abnormal EEG patterns is critical in interpreting EEG signals in critical care settings.

C. Interpretability Evaluation Metrics

To evaluate the interpretability of ProtoPNets, we adopted three metrics.

1. **AvgDrop:** AvgDrop measures how much the logit for the predicted class drops when the activated regions are masked, thereby evaluating the contribution of those regions to the model's decision.
   We first computed similarity maps between the input and all prototypes. Then, we identified the prototypes that

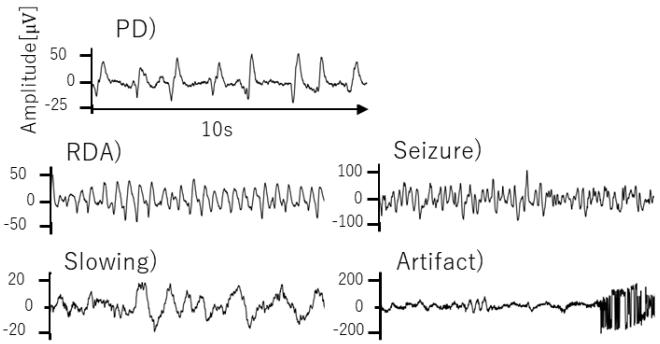


Fig. 2. Example EEG waveforms of segments from each class.

contributed to the computation of the predicted class's logit. We extracted the similarity maps corresponding to these prototypes and aggregated them into a single activation map. This activation map was then projected back to the input space using the receptive field, following (4). We identified the activated regions as the non-zero areas of the projected activation map, and masked them in the input to obtain x'.

Let $\hat{y}$, $\hat{y}'$, and $\hat{y}_{\min}$ denote the logits for the original input, the score after masking, and the score when the entire input is masked, respectively. AvgDrop is defined as:

$$\text{AvgDrop} = \max\left(0, \frac{\hat{y} - \hat{y}'}{\hat{y} - \hat{y}_{\min}}\right) \times 100. \quad (11)$$

A higher value indicates that the prediction relies more heavily on the masked activated regions.

2. **Redundancy:** Redundancy measures the degree of similarity among prototypes within the same class. For each class k, we compute the pairwise cosine similarity among its m prototypes and take the average:

$$\text{redundancy} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m(m-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{m}\left(\mathbf{p}_i^{(k)}\right)^{\top}\mathbf{p}_j^{(k)}. \quad (12)$$

Lower values indicate that the prototypes capture more diverse and less redundant representations.

3. **Coverage:** Coverage evaluates how well each input is represented by at least one prototype from its ground-truth class. For each input x_i, we check whether any prototype p_j belonging to the ground-truth class y_i yields a similarity map with values exceeding a threshold τ:

$$\text{coverage} = \frac{1}{N}\sum_{i=1}^{N}\max_{p_j\in P_{y_i}}\mathbf{I}\left(\max_{a,b} S_{ij}^{a,b} > \tau\right). \quad (13)$$

Here, S_{ij} denotes the similarity map between x_i and p_j, and $\mathbf{I}(\cdot)$ is the indicator function. In our experiments, we set $\tau = 0.6$.

These three metrics provide a multifaceted quantitative evaluation of the model's interpretability.

D. Baseline methods

To evaluate the effectiveness of our proposed method, we compared it with the following models:

1. **EEGNet + BiGRU:** This model first captures short-term temporal and spatial features using EEGNet [10], and then models long-term temporal dependencies with a bidirectional GRU [11]. It serves as a baseline that explicitly incorporates temporal dynamics.
2. **Backbone + Linear Classifier (BB + Linear):** This model uses the same backbone as our proposed method but the prototype layer is omitted. Instead, it appends a linear classifier directly on top of the extracted features. We include this model as a baseline to evaluate how well the backbone alone can perform classification without relying on the prototype mechanism.
3. **ProtoPNet (Base):** This model retains the overall architecture of our method but adopts the original ProtoPNets scoring scheme, applying max-pooling to similarity maps and computing logits via linear combination.

All methods, including our proposed one, were compared in terms of classification performance. For interpretability evaluation, we only assessed the ProtoPNet-based models.

E. Experimental Setup

All models were optimized using the Adam optimizer. The learning rate was scheduled using cosine annealing. Training was conducted for 50 epochs with a batch size of 64. To address class imbalance, we applied inverse class frequency weighting to all loss components: $\mathcal{L}_{CE}$, $\mathcal{L}_{clst}$, $\mathcal{L}_{sep}$. For the loss weights of the ProtoPNet (Base) model, we followed the original implementation [7] and fixed the number of prototypes per class to $m = 10$.

V. Results

A. Classification Performance

Table I summarizes the five-class classification results for each method. The proposed model with $m = 2$ prototypes per class and a similarity aggregation ratio of $\alpha = 5$, achieved the best overall performance.

The prototype-based model achieved performance comparable to the performance of the baseline models. The proposed model outperformed EEGNet+BiGRU by 3.6% in accuracy, 5.1% in macro recall, and 3.6% in macro F1-score. It also showed clear improvements over ProtoPNet (Base) despite using fewer prototypes. These results indicate that the proposed similarity-based aggregation enhances classification accuracy while preserving model interpretability. In addition, the class-wise recall values indicate that our method achieved the highest recall for the RDA class.

B. Interpretability

1) Qualitative Evaluation: Fig. 3 presents the activation map visualization corresponding to a waveform that has been annotated with PD. The left side shows the prototypes that contributed to the classification, while the right side overlays the highlighted regions on the input waveform. In ProtoPNet (Base), although all prototypes contributed to classification via the final fully connected layer, only those with positive weights are shown for visualization purposes. In contrast, ProtoPNet

TABLE I. Comparison of method performance and recall per class [%].

Model	Overall Performance			Recall per Class				
	Accuracy	Macro Recall	Macro F1	PD	RDA	Seizure	Slowing	Artifact
EEGNet + BiGRU	52.0	47.5	46.4	47.9	46.8	37.8	65.5	39.6
BB + Linear	52.2	46.2	46.5	52.4	34.2	40.7	64.6	39.0
ProtoPNet (Base)	47.6	45.4	42.3	57.6	54.6	20.4	46.5	48.1
ProtoPNet (Proposed)	55.6	52.6	50.0	54.2	61.1	35.5	64.7	47.6

(Proposed) simply selected the prototype with the highest similarity score. These highlighted regions are consistent with the PD's typical patterns captured by the prototypes, confirming that the ProtoPNet-based model focused on interpretable and relevant portions of the EEG signals during decision-making.

In our proposed method, classification is driven by a single prototype, and the use of top-$\alpha\%$ averaging results in a wide range of highlighted regions. This allows the model to clearly capture the typical repetitive peaks of PD.

Fig. 4 shows the waveforms corresponding to the prototypes learned by the proposed method with $m = 2$ and $\alpha = 5$. In the PD class, typical peaks were captured; in the RDA class, rhythmic patterns were learned; and in the artifact class, high-frequency and high-amplitude waveforms were observed. These learned patterns resemble the typical waveforms of each class shown in Fig. 2.

2) Quantitative Evaluation: Table II presents the results of the interpretability evaluation. The proposed model achieved a higher AvgDrop and lower redundancy than the baseline. In particular, with $m = 2$ and $\alpha = 5$, it reached the highest AvgDrop and a redundancy of 0.09. However, coverage dropped to 0.00, meaning that for all inputs, no patch had similarity above the threshold τ = 0.6 with any prototype from the ground-truth class.

VI. Discussion

Significant performance improvements were observed for PD and RDA in the prototype-based models, which likely reflected their typical waveform patterns well captured by prototypes. Artifact also showed improvement, possibly because the backbone had extracted shared features such as high-amplitude or high-frequency noise, despite visual variability.

Although the prototype-based models made decisions based on only limited regions within each EEG segment, they achieved classification performance comparable to other models without a prototype layer. This suggests that class-relevant information is not necessarily spread across the entire segment, but can be effectively captured by focusing on specific regions where typical waveforms appear. Such targeted reliance may contribute to robustness against distributional variability in clinical data.

In contrast, no notable improvements were found for seizure and slowing. Both seizure and slowing lack consistent temporal structure; seizure exhibits temporal variability, and slowing includes a variety of patterns that are not assigned to any of the other defined classes. In this study, we assigned the same number of prototypes to each class for simplicity and comparability, but this uniform allocation may be suboptimal for seizure. Possible extensions include allocating prototypes adaptively depending on class complexity and introducing multi-scale prototypes [8] that can better capture temporal dynamics. An alternative direction is to adopt a staged system: first identifying whether EEG segments exhibit typical temporal structures and then applying the prototype-based method only to those segments.

The improvement in AvgDrop can be attributed to the top-$\alpha\%$ average, which allowed the model to capture broader and more distributed features compared to conventional max-pooling. In addition, since our method used a smaller number of prototypes for classification, redundancy was inherently reduced.

The low coverage is likely due to the increased α value. When α was large, prototypes were likely trained to maximize average similarity across multiple patches rather than achieving high similarity with any single patch. This may have resulted in lower maximum similarity values $\max(S_{ij}^{a,b})$, which could have reduced coverage despite improved overall performance.

The relatively modest AvgDrop of 53% is mainly attributed to the use of a small α, which was chosen to focus on local features. Given that the optimal α might vary depending on the class or individual input, making α a learnable parameter could further enhance performance.

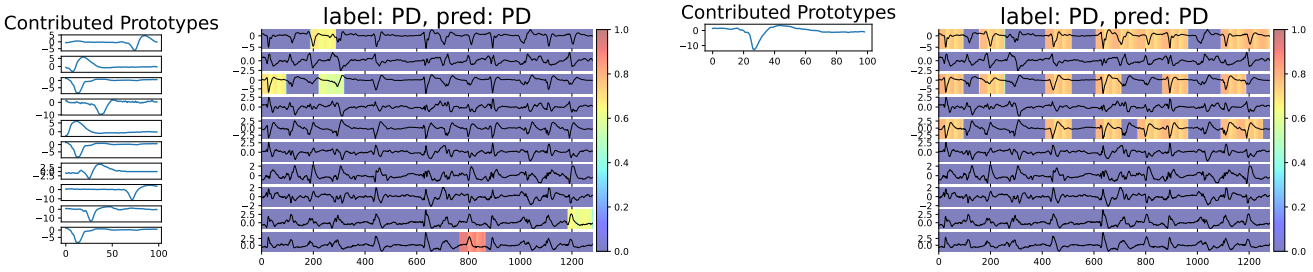
VII. Conclusion

In this study, we proposed an interpretable classification method for EEG signals in critical care settings based on ProtoPNets. By introducing a top-$\alpha\%$ average pooling for similarity scoring and adopting a single-prototype-based classification scheme, our method improved interpretability while preserving overall classification performance. Our evaluation, however, was limited to a small set of prototype numbers and aggregation ratios. A more systematic study of these parameters is necessary to clarify the balance between accuracy and interpretability.

Future work includes improving performance for classes such as seizure, which exhibit temporal variability, and slowing, which includes a wide variety of waveform patterns. Furthermore, to support clinical use, it is important to confirm that the learned prototypes and their highlighted regions correspond to clinically meaningful waveform patterns for each class. This requires validation by clinicians, who will assess the medical relevance of the prototypes and the diagnostic significance of the highlighted regions against manual annotations.

VIII. Acknowledgments

This work was supported in part by JSPS KAKENHI (Grant No. JP22K18423).



(a) ProtoPNet (Base) (b) ProtoPNet (Proposed)

Fig. 3. Visualization examples illustrating the model' s reasoning process.

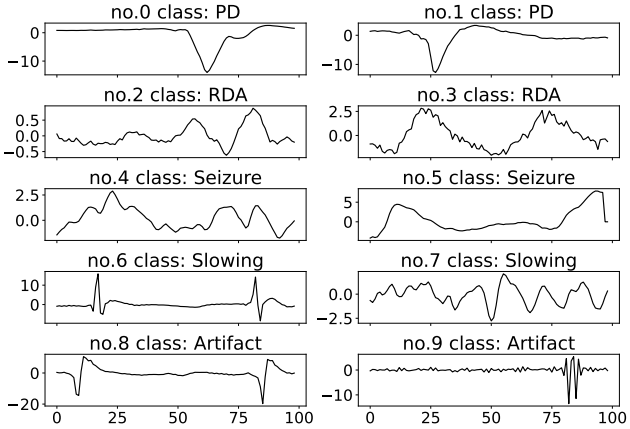


Fig. 4. Waveforms corresponding to learned prototypes

TABLE II. Evaluation of interpretability using quantitative metrics.

ProtoPNet	m	α	AvgDrop↑	Redund.↓	Cov.↑
Base	10	-	0.0	0.33	1.00
Proposed	1	1	0.26	0.00	0.92
	1	5	0.30	0.00	0.01
	2	1	0.36	0.16	0.34
	2	5	0.53	0.09	0.00

REFERENCES

[1] A. Krumholz, G. Y. Sung, R. S. Fisher, E. Barry, G. K. Bergey, and L. M. Grattan, "Complex partial status epilepticus accompanied by serious morbidity and mortality," *Neurology*, vol. 45, no. 8, pp. 1499–1504, 1995.

[2] R. Sutter and P. W. Kaplan, "The neurophysiologic types of nonconvulsive status epilepticus: Eeg patterns of different phenotypes," *Epilepsia*, vol. 54, no. 6, pp. 23–27, 2013.

[3] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imaging*, vol. 6, no. 6, p. 52, 2020.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, USA, 2016, pp. 2921–2929.

[6] W. Xie, X. Li, C. C. Cao, and N. L. Zhang, "Vit-cx: Causal explanation of vision transformers," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Macao, S.A.R, 2023, pp. 1569–1577.

[7] C. Chaofan, L. Oscar, D. Tarlow, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, Vancouver, Canada, 2019, pp. 8930–8941.

[8] Y. Gao, A. Liu, L. Wang, R. Qian, and X. Chen, "A self-interpretable deep learning model for seizure prediction using a multi-scale prototypical part network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1847–1856, 2023.

[9] Z. Carmichael, S. Lohit, A. Cherian, M. J. Jones, and W. J. Scheirer, "Pixel-grounded prototypical part networks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, Hawaii, 2024, pp. 4768–4779.

[10] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, p. 056 013, 2018.

[11] A. Moschitti, B. Pang, and W. Daelemans, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734.