G-006

集中治療期脳波の分類における ProtoPNet を用いた判断根拠の可視化
Visualizing reasoning behind critical care EEG classification using ProtoPNet

備後 拓真[1)] 矢野 肇[1)] 芦崎 太一朗[2)] 甲田 一馬[2)] 十河 正弥[2)]
Takuma Bingo Hajime Yano Taichiro Ashizaki Kazuma Koda Masaya Togo
松本 理器[3)2)] 滝口 哲也[1)]
Riki Matsumoto Tetsuya Takiguchi

1 はじめに

非けいれん性てんかん重積状態は，けいれんを伴わずにてんかん発作が持続し，多くの場合意識障害を呈する重篤な状態である [1]．この状態は意識障害を伴う集中治療期の患者に対して，しばしば疑われる病態であり，早期の診断と介入が治療の成否に大きく影響する．意識障害患者の治療においては脳波検査が極めて有効であり，特に非けいれん性てんかん重積状態では，脳波が特異的な異常を示すことが知られている [2]．しかし診断には専門医による持続的な脳波モニタリングが必要であり，少人数の医師が膨大な量の波形を判読することは困難である．そのため深層学習を用いた自動判読システムによる支援が期待されている．

近年，深層学習は脳波解析を含む様々な分野で用いられ，高い精度を示している．一方で，その推論過程はブラックボックスであり，その根拠を人間が理解，解釈することは難しい．そのため，特に医療や病理診断分野では説明可能な AI（XAI）に関する研究が注目を集めている [3]．

CNN に対する判断根拠の可視化手法としては，顕著性マップ [4] や CAM[5][6] などが代表的である．これらは主に勾配に基づいてモデルの出力に対する入力の重要度や寄与を可視化する手法である．しかし，これらの手法は学習済のモデルに対して事後的に解釈を与えるものであり，実際の推論過程を説明するものではない．また，解釈可能な構造を備えたモデルとしては，決定木や線形モデル，注意機構を用いた手法 [7] などが提案されている．しかしモデルを単純化すると解釈性は向上するものの，一般に予測精度は低下する傾向がある．さらに，注意機構を用いた手法では，モデルが入力のどの部分に着目したかを示すことはできるが，その部分がどのような典型的特徴と類似しているのかまでは明らかにできない．

Prototypical Part Network（ProtoPNet）[8] は，CNN をベースとした解釈可能な構造を持つモデルであり，各クラスにおける特徴的な局所パターンを複数のプロトタイプとして学習する．推論の際には入力データとプロトタイプの類似する領域の可視化に加え，分類に寄与したプロトタイプの可視化もできるため，従来の手法に比べ解釈性に優れる．実際に，ProtoPNet をてんかん発作予測に応用した研究 [9] では，EEG 信号のマルチスケール特徴を捉えるためにプロトタイプのサイズを多様化し，高い予測精度と解釈可能性を両立することが報告されている．

そこで，本論文では，ProtoPNet を用いて集中治療期脳波の周期性や空間的な広がりを持つパターンを検出する手法を提案する．この手法では，各クラスに属する複数のプロトタイプを個別に扱い，全ての時間，電極における脳波との類似度を計算し，高類似度のパターンを分類に利用する．これにより，集中治療期脳波に対する分類性能の向上を図り，判断根拠の可視化を可能とする．

2 ProtoPNet

2.1 原理

ProtoPNet は，入力画像の一部分と，各クラスの代表的な局所パターンを比較するという推論過程を構造的に組み込んだ，解釈性の高いモデルとして提案されている [8]．この手法では，各クラスの代表的な部位をプロトタイプとして学習し，入力画像の各部分をこれらのプロトタイプと比較することで分類を行う．

具体的には，CNN ベースの特徴抽出器で抽出された特徴マップに対し，プロトタイプ層において，各プロトタイプ p_j と特徴マップ Z の全てのパッチとの間で類似度マップ S_j を計算する．

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

ここで，$Z^{a,b}$ は特徴マップ Z の位置 (a,b) におけるパッチを表し，sim(・) は類似度関数である．得られた類似度マップに対して max-pooling を適用することで，各プロトタイプと特徴マップとの間の類似度スコア g_{p_j} を得る．

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

最終的な分類スコアは，各クラス k に対応するすべてのプロトタイプの類似度を線形結合することで計算される．

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

ここで，$w_h(k,j)$ は全結合層のパラメータである．

2.2 受容野ベースの可視化

プロトタイプと画像の局所部分を正確に対応させる手法として，PIXPNET が提案されている [10]．従来までの ProtoPNet やその派生手法では，特徴パッチに対応する画像上のピクセル領域を単純な線形アップサンプリングによって求めていたため，ヒートマップによる可視化の結果が正確でない可能性があった．それに対し，PIXPNET では，特徴抽出器の受容野に基づいて対応領域を求めることで，正確な解釈を与えている．

1） 神戸大学大学院システム情報学研究科, Graduate School of System Informatics, Kobe University
2） 神戸大学大学院医学研究科 脳神経内科学, Division of Neurology, Graduate School of Medicine, Kobe University
3） 京都大学大学院医学研究科 臨床神経学, Department of Neurology, Graduate School of Medicine, Kyoto University

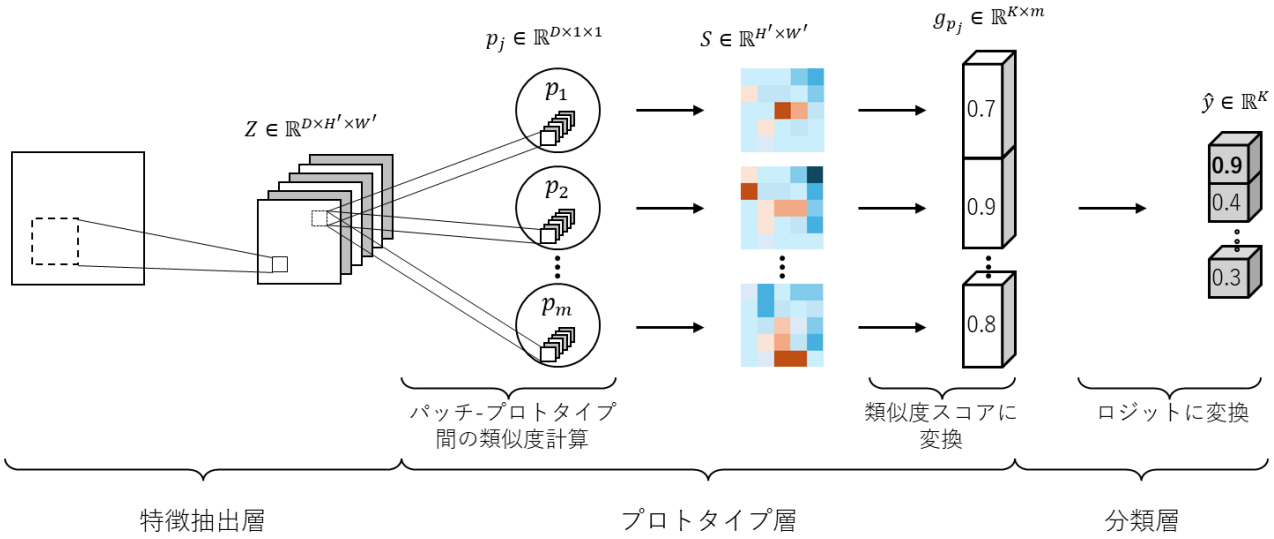


図 1: ProtoPNet のアーキテクチャ

PIXPNET では，ピクセル空間におけるヒートマップ $M \in \mathbb{R}^{H\times W}$ を零行列で初期化したのち，各位置におけるスコア $s \in S$, を基に以下の式で更新する．ここで，S はプロトタイプと入力画像の特徴マップの間で計算された類似度マップ，M^s は M の s に対応するピクセル領域である．

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

3 提案手法

3.1 アーキテクチャ

本論文における ProtoPNet のアーキテクチャを図 1 に示す．ProtoPNet は以下の 3 つの層で構成される．なお，K を分類クラス数とする．

(1) 特徴抽出層
先行研究 [9] に倣い，入力信号 $x \in \mathbb{R}^{C\times T}$ を電極数を保存したまま特徴マップ $Z \in \mathbb{R}^{D\times C\times T'}$ に変換する．なお，C は EEG 信号の電極数，T は時系列長，D は特徴マップとプロトタイプの次元数を表す．

(2) プロトタイプ層
各クラスごとに m 個のプロトタイプ $\{p_j\}_{j=1}^m$, $p_j \in \mathbb{R}^{D\times1\times1}$ を持つ．プロトタイプと特徴マップから，類似度マップ S_j，類似度スコア $g_{p_j}(Z)$ を計算する．本論文では，類似度関数にはコサイン類似度を用いた．

(3) 分類層
類似度スコア $g_{p_j}(Z) \in \mathbb{R}^{K\times m}$ をロジット $\hat{y}$ に変換する．

特徴抽出器に関しては，文献 [9] で用いられている構成をベースとしつつ，電極方向の畳み込みカーネルサイズを 3 から 1 に変更した．電極方向のカーネルサイズが 1 より大きい時，複数電極の信号が畳み込まれるため，学習されるプロトタイプも複数電極にわたる信号パターンを参照するようになる．しかし，電極方向の隣り合う次元の関係性は，画像データと異なり必ずしも実際の電極の空間的な隣接関係を反映しないため，受容野に入力される電極の組み合わせには明確な意味付けが難しい．また，類似した波形パターンであっても，電極の組み合わせの違いで異なるプロトタイプとして学習され，解釈が困難になる可能性がある．カーネルサイズが 1 の時，各電極の信号は独立に処理され，プロトタイプは常に単一電極の波形パターンを参照するようになる．このため，電極順に依存しない解釈が可能となる．

3.2 最適化

本論文における ProtoPNet の最適化は以下の 2 ステージで構成される．

(1) 特徴抽出層およびプロトタイプの学習
ステージ 1 では，以下の損失関数に従って，特徴抽出層とプロトタイプの学習を行う．

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1\mathcal{L}_{clst} + \lambda_2\mathcal{L}_{sep} + \lambda_3\mathcal{L}_{div} \tag{5}$$

$$\mathcal{L}_{clst} = -\frac{1}{N}\sum_{i=1}^{N}\max_{p_j\in \mathrm{P}_{y_i}} g_{p_j}(Z_i) \tag{6}$$

$$\mathcal{L}_{sep} = \frac{1}{N}\sum_{i=1}^{N}\max\left(0, \max_{p_j\notin \mathrm{P}_{y_i}} g_{p_j}(Z_i) - \max_{p_j\in \mathrm{P}_{y_i}} g_{p_j}(Z_i) + \Delta\right) \tag{7}$$

$$\mathcal{L}_{div} = \left\|\mathbf{P}^\top\mathbf{P} - \mathbf{I}^{(Km)}\right\|_F^2 \tag{8}$$

ここで，N はバッチ内のデータ数，$\mathcal{L}_{CE}$ はクロスエントロピー損失である．クラスタリング損失 $\mathcal{L}_{clst}$ は入力データの特徴パッチが同じクラスのプロトタイプに近づくように，セパレーション損失 $\mathcal{L}_{sep}$ は異なるクラスのプロトタイプから遠ざかるように学習させることで，特徴空間上にクラスごとの分布を形成することを目的としている．また，$\mathcal{L}_{div}$ はプロトタイプ間の直交性を促進する損失である [11]．ここで，$\mathbf{P} \in \mathbb{R}^{D\times(Km)}$ はプロトタイプ行列，$\mathbf{I}^{(Km)}$ は $(Km)\times(Km)$ の単位行列，$\|\cdot\|_F$ はフロベニウスノルムである．

(2) プロトタイプの投影
ステージ 2 では，プロトタイプを最も近い訓練デー

タの特徴パッチに置き換える．

$$p_j \leftarrow \underset{z \in \text{patches}(Z)}{\arg\max}\ \text{sim}(z, p_j) \tag{9}$$

これにより，プロトタイプは視覚的に解釈可能な訓練データの脳波波形の一部と対応する．

3.3 類似度スコアの算出方法の変更

従来の ProtoPNet やその派生手法では，プロトタイプと特徴マップの間で計算された類似度マップ S_j に対し，式 2 のように max-pooling を適用することで，プロトタイプと特徴マップ間の類似度スコアを計算している．画像においては，物体の各パーツが画像内の異なる位置に現れるため，最も類似したパッチのみを用いて判断する設計となっている．

一方，集中治療期脳波では，同一の波形パターンが複数の電極に繰り返し現れることが多い．このようなデータに対して max-pooling を適用すると，局所的な一部分だけがスコアに反映され，全体の情報を十分に活用できない可能性がある．この問題に対処するために，本論文では ProtoPNet におけるスコア算出の方法を大きく 2 点にわたり改良した．

第一に，プロトタイプと特徴マップ間の類似度スコアの計算方法において，従来の max-pooling の代わりに Top-α %の平均を用いた．具体的には，類似度マップ S_j に対して，スコア $g_{p_j}(Z)$ を以下のように定義した．ここで，S_j^{α} は，類似度マップ S_j における上位 α %の大きさを持つ類似度の集合である．

$$g_{p_j}(Z) = \frac{1}{|S_j^{\alpha}|} \sum_{s \in S_j^{\alpha}} s \tag{10}$$

第 2 に，類似度スコアからクラスごとの分類スコアを算出する際の処理を，線形結合ではなくクラスごとの最大値を取る方法に変更した．

$$\hat{y}_k = \max_{p_j \in P_k} g_{p_j}(Z) \tag{11}$$

従来手法では複数のプロトタイプがそれぞれ異なる特徴を捉え，それらを線形結合により統合して分類を行う．それに対し，本手法では単一のプロトタイプが複数位置の特徴を捉えるため，最も入力波形に合致する部分のみを分類に利用する．

4 実験

集中治療期における被験者の 329 人から収録した脳波を対象に，深層学習による分類を行った．脳波データの利用については神戸大学倫理委員会より承認を得た（課題番号 B220114）．

4.1 データセット

データは全体でおよそ 108 時間にわたり，以下の前処理を行い，データセットを作成した．まず，10 秒のセグメントに分割され，被験者ごとに脳波信号の平均と標準偏差を用いて Z スコア正規化を行い，各セグメントのスケールを揃えた．次に，被験者が重複しないように 8:2 の割合で分割し，学習・検証データとテストデータに割り当てた．次に，学習・検証データに割り当てられた被験者については，それぞれの脳波データを時系列順に 8:2 に分割して，学習データと検証データとした．各

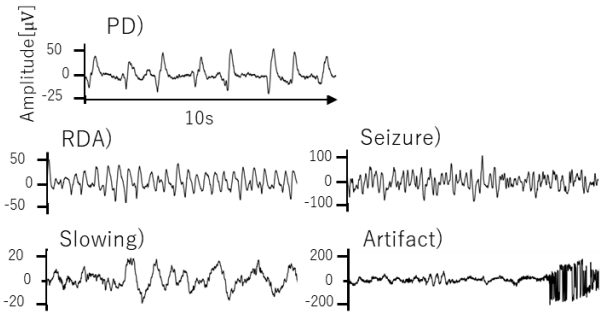


図 2: 各クラスのセグメントの脳波の波形例

セグメントには以下の 5 種類のアノテーションのいずれかがつけられた．各クラスの波形例を図 2 に示す．

- Periodic Discharges（PD）:
  持続時間が 0.5 秒以下の突発性発射が一定間隔で継続する．
- Rhythmic Delta Activity（RDA）:
  0.5–4 Hz の単一系の δ 律動波が出現する．
- Seizure:
  てんかん発作．数秒以上持続し，時間・空間的に発展する高振幅の鋭波・棘波・律動波．
- Slowing:
  0.5–8 Hz の徐波．種々の脳機能障害の際に出現する．
- Artifact:
  体動などによる高振幅または高周波のノイズ．

4.2 分類性能の評価指標

分類モデルの性能評価指標として，Accuracy，Recall，F1 score を用いた．また全体的な性能を反映する指標として，各クラスにおける Recall と F1 score を平均した値である Macro Recall と Macro F1 を採用した．なお，Precision ではなく Recall を評価指標として選択した理由は，集中治療期脳波の判読において異常脳波の見落としを最小限に抑えることが重要であるためである．

4.3 解釈性の評価指標

ProtoPNet の解釈性を評価するために，文献 [12] を参考に，3 つの指標を導入した．

4.3.1 平均スコア低下率（AvgDrop）

平均スコア低下率は，モデルの注目領域をマスクした際の，モデルが予測したクラスに対する分類スコアの減少率を計算することで，注目領域の予測への寄与を評価する指標である．

本論文では，次のようにモデルの注目領域を算出した．まず，モデルは入力に対して各プロトタイプとの類似度マップを計算し，予測クラスに対応するプロトタイプの類似度マップから分類スコアを算出する．この際，実際に分類スコア計算に寄与した特徴マップ上の領域を特定した．次に，特徴空間から入力空間への対応関係を求めるために，式 4 に基づき，特徴抽出器の受容野を考慮して，分類に寄与した特徴領域を入力空間上の注目領域として取得した．

本指標の算出手順として，上記で特定された注目領域をマスクした入力 x' を作成した．入力 x に対する予測クラスのスコアを $\hat{y}$ とし，マスク付き入力については $\hat{y}'$，また入力全体をマスクした場合の予測スコアを $\hat{y}_{min}$

とする．このとき平均スコア低下率は以下のように定義される．

$$\text{AvgDrop} = \max\left(0, \frac{\hat{y} - \hat{y}'}{\hat{y} - \hat{y}_{min}}\right) \times 100 \tag{12}$$

この値が大きいほど注目領域が予測に重要な役割を果たしていることを示す．

4.3.2 プロトタイプ冗長性（Redundancy）

プロトタイプ冗長性は，同一クラスのプロトタイプ間の類似度を評価することで，学習されたプロトタイプにどの程度冗長性があるかを定量化する指標である．各クラス k について，そのクラスに属するプロトタイプ間のコサイン類似度を計算し，その平均値を求める．

$$\text{Redundancy} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m(m-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{m}\left(\mathbf{p}_i^{(k)}\right)^{\top}\mathbf{p}_j^{(k)} \tag{13}$$

ここで，$p^(k)$ はクラス k に属するプロトタイプである．この指標が小さいほどプロトタイプ間の冗長性が低く，より効率的で解釈しやすい特徴表現を学習していることを示す．

4.3.3 カバレッジ（Coverage）

カバレッジは，各入力データが対応するクラスのプロトタイプによって適切に表現されているかを評価する指標である．各入力データについて，その真のクラスに属するプロトタイプとの類似度が閾値を超えるかを判定する．

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N}\max_{j\in P_{y_i}}\mathbf{I}\left(\max_{a,b} S_{ij}^{a,b} > \tau\right) \tag{14}$$

ここで，S_{ij} は入力 x_i とプロトタイプ p_j の間で作成された類似度マップであり，y_i は入力 x_i の真のクラスを表す．また，$\mathbb{I}(\cdot)$ は指示関数である．この値が大きいほど，学習されたプロトタイプが入力データを広くカバーしていることを示す．本論文においては，$\tau = 0.6$ とした．

4.4 比較手法

本論文では，提案手法の有効性を検証するために，以下の比較手法を採用した．

(1) EEGNet + BiGRU モデル
EEGNet は EEG 信号の解析に広く用いられている CNN をベースとしたモデルである [13]．また，GRU（Gated Recurrent Unit）は，RNN をベースとした効率的なモデルである [14]．EEGNet + BiGRU は，EEGNet を特徴抽出器として用い，双方向 GRU によって時系列を処理する構成のモデルである．本論文では，提案手法の性能評価にあたり，時系列情報を考慮した従来手法として比較対象に含めた．

(2) FE + LC
本論文における ProtoPNet の特徴抽出器の部分を用い，分類にプロトタイプを用いず，代わりに線形分類器を付加したモデルである．この構成ではプロトタイプに基づく解釈性は失われるが，特徴抽出器の分類性能を評価するためのベースラインとして採用した．

(3) ProtoPNet (Base)
ネットワークのアーキテクチャは提案手法と同じとし，本来の ProtoPNet のように，類似度スコアを max-pooling によって，分類スコアを線形結合によって算出するモデルである．

分類性能の比較には，提案手法を含むすべての手法の比較を行った．一方，解釈性の評価に関しては，ProtoPNet を対象とした．

4.5 実験方法

本実験では，モデルは Adam によって最適化した．学習率はコサインアニーリングにより調整し，最大学習率を 10^{-3} に設定し，バッチサイズ 64 で 50 エポック間学習を行った．また，クラス不均衡の対応として，$\mathcal{L}_{CE}$，$\mathcal{L}_{clst}$，$\mathcal{L}_{sep}$ に，サンプルが属するクラスのデータ数の逆数を重みづけした．なお，損失重み $\lambda_1, \lambda_2, \lambda_3$ はそれぞれ，1.5，20，0.01 とし，$\mathcal{L}_{sep}$ のマージン $\delta = 0.2$ とした．また，ProtoPNet (Base) においては，原著 [8] に従い，各クラスのプロトタイプ数 $m = 10$ で固定した．全ての実験において 5 分割交差検証を用いて性能を評価した．

5 結果

5.1 分類性能

各手法の 5 クラス分類の結果を表 1 に示す．ここで，ProtoPNet (Proposed) は，各クラスのプロトタイプ数 $m = 2$，分類に利用する類似度の割合 $\alpha = 5$ で学習されたモデルである．

プロトタイプベースモデルは，従来の深層学習モデルと同程度の性能を達成した．EEGNet+BiGRU と比較すると，提案手法では Macro Recall で 1.5 %，MacroF1 で 0.6 %の向上が見られた．ProtoPNet (Base) と比較しても，類似度スコア算出方法の改良によって，少ないプロトタイプ数でほぼ同等の性能を達成した．

また，表 2 に各クラスにおける Recall 値を示す．提案手法は PD（55.2 %）で最高の Recall 値を達成した．また，RDA においては ProtoPNet (Base) を下回っているものの，FE + LC と比較して 28.7 %の大幅な改善を示した．

5.2 解釈性

5.2.1 定性的評価

図 3 に $m = 2$，$\alpha = 5$ で学習された提案手法におけるプロトタイプの例を示す．図 3a は各クラスのプロトタイプに対応する波形，図 3b はプロトタイプ間の類似度行列である．PD では特有のピーク，RDA では律動性，Artifact では高周波・高振幅な波形パターンが学習されており，これらは図 2 に示した各クラスの典型波形と類似している．また，図 3b の類似度行列において，プロトタイプ間の類似度はかなり抑制されている．

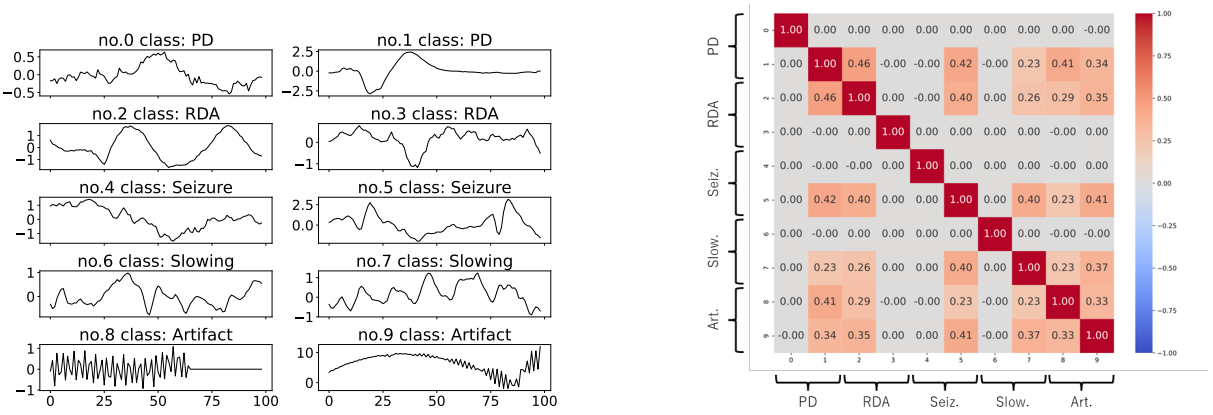
図 4 に PD のアノテーションが付与された波形に対する注目領域の可視化結果を示す．左側には分類に寄与したプロトタイプ，右側は注目領域を元の波形に重ねて表示している．従来手法である図 4a では，複数のプロト

表 1: 手法間の性能比較 [%]

Model	Acc.	Macro Rec.	Macro F1
EEGNet + BiGRU	54.7 ± 3.1	51.5 ± 3.1	49.0 ± 3.4
FE + LC	53.0 ± 4.2	44.3 ± 3.6	44.6 ± 4.4
ProtoPNet (Base)	52.0 ± 3.4	53.0 ± 5.3	48.4 ± 4.4
ProtoPNet (Proposed)	54.7 ± 3.4	52.8 ± 5.1	49.6 ± 4.5

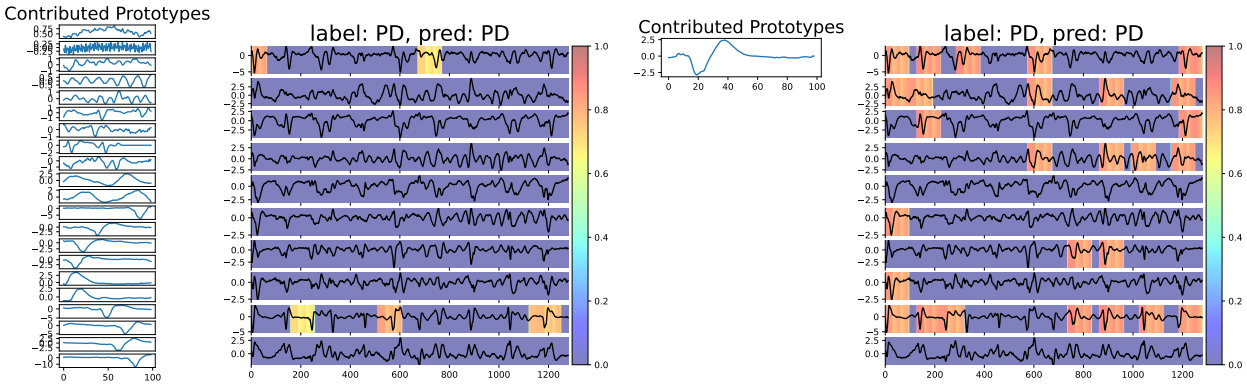
表 2: クラス間の Recall 値の比較 [%]

Model	PD	RDA	Seizure	Slowing	Artifact
EEGNet + BiGRU	53.0 ± 11.7	53.4 ± 13.1	38.5 ± 13.3	63.0 ± 3.3	49.6 ± 3.1
FE + LC	52.8 ± 8.8	24.8 ± 8.8	30.7 ± 14.9	69.8 ± 6.2	43.5 ± 7.6
ProtoPNet (Base)	54.5 ± 9.2	62.1 ± 9.6	44.1 ± 23.1	47.6 ± 10.6	56.8 ± 7.9
ProtoPNet (Proposed)	55.2 ± 6.4	53.5 ± 13.5	42.2 ± 15.6	57.1 ± 7.4	56.1 ± 6.0



(a) プロトタイプに対応する波形　　(b) プロトタイプ間の類似度

図 3: 学習されたプロトタイプの例



(a) ProtoPNet (Base), m=10　　(b) ProtoPNet (Proposed), m=2, α=5

図 4: 注目領域の可視化

タイプが分類に寄与しているが，各プロトタイプによる注目領域は狭く分散的である．これに対し，提案手法の結果である図 4b では，単一のプロトタイプのみが分類に寄与し，Top-α %平均により広範囲にわたり反復的に注目領域を形成している．

表 3 に，テストデータに対し，各プロトタイプが分類に寄与した回数，すなわち式 (11) において最大値を取った回数を示す．RDA，Seizure，Slowing のプロトタイプに，分類に利用されていないものがあることが確認された．

5.2.2　定量的評価

表 4 に解釈性の定量評価結果を示す．提案手法は従来手法と比較して平均スコア低下率とプロトタイプ冗長性を改善した．特に $m = 2, \alpha = 5$ の設定では最も高い平均スコア低下率を達成し，注目領域の予測への寄与が向上していることが確認された．プロトタイプ冗長性についても大幅な削減が実現された．

6　考察

6.1　クラス別性能差

プロトタイプベースの手法において，PD と RDA で大幅な性能向上が得られたのは，プロトタイプによる表現がしやすい，特徴的な波形が現れるクラスであったためと考えられる．Artifact についても改善が見られたが，これは視覚的には多様な波形パターンを持つものの，特徴抽出層において高振幅や高周波なノイズ成分が共通の特徴量として捉えられたためと推測される．

一方，Seizure，Slowing については大きな改善が見られなかった．Seizure は波形・周波数の時間的変化があるクラスであり，Slowing は他クラスに該当しない波形全般を含むため，本手法に適さなかったと考えられる．

これらの結果から，特徴的な時間パターンを持つ脳波とそうでない脳波を事前に分類し，特徴的な時間パター

表 3: プロトタイプが分類に寄与した回数

クラス	PD		RDA		Seiz.		Slow.		Art.	
プロトタイプ番号	0	1	2	3	4	5	6	7	8	9
寄与回数	367	4,101	1,966	0	0	1,806	0	4,818	1,457	2,419

表 4: 解釈性の定量評価

Model	m	α	AvgDrop↑	Redundancy↓	Coverage↑
ProtoPNet (Base)	10		0.11 ± 0.01	0.28 ± 0.03	0.98 ± 0.01
ProtoPNet (Proposed)	1	1	0.17 ± 0.02	0.00 ± 0.00	0.88 ± 0.02
	1	5	0.26 ± 0.02	0.00 ± 0.00	0.91 ± 0.02
	2	1	0.19 ± 0.01	0.01 ± 0.02	0.86 ± 0.03
	2	5	0.29 ± 0.04	0.04 ± 0.03	0.89 ± 0.02

ンを持つ脳波に対してのみプロトタイプベース手法を適用するといったような，段階的なシステムの構築が有効と考えられる．

6.2 注目領域の解釈性

図 3b で観察された低い類似度は，多様性損失 $\mathcal{L}_{\text{div}}$ による直交性促進の効果を示している．図 4 の注目領域可視化において，ProtoPNet (Base) では複数のプロトタイプによる分散的な注目領域が観察された．これは類似したプロトタイプが学習されることで注目領域が重複し，解釈の複雑化を招いている可能性がある．提案手法では，単一プロトタイプと類似した波形パターンのみに注目するため，反復的に表れる PD のピークを明確に捉えていると考えられる．ただし，従来手法においても PD，RDA の Recall は高くなっており，これは人間が重要と考える脳波の特徴と，実際の分類に用いられた特徴が必ずしも一致しないことを示唆している．

定量的評価における，提案手法の平均スコア低下率の改善は，Top-α %平均により従来手法の max-pooling と比較してより広範囲の特徴を捉えられたためと考えられる．さらに，提案手法では少数のプロトタイプで学習・分類を行うため，必然的にプロトタイプ冗長性も低くなる．

6.3 手法の改良方針

平均スコア低下率が 29 %程度にとどまったのは局所化を優先して小さな α を用いたためであるが，クラスやデータによって最適な局所化の程度は異なる．α を学習可能パラメータとして可変にすることで，さらなる性能向上が期待される．

図 3b で観察された異なるプロトタイプ間の低い類似度は，多様性損失 $\mathcal{L}_{\text{div}}$ による直交性促進の効果を示している．その一方で，表 3 の結果から分類に使用されないプロトタイプの存在も示された．直交制約によって同一クラスのプロトタイプ間の類似度さえも過度に小さくしたことで，分類に使用されないプロトタイプを獲得した可能性も考えられる．今後は使用頻度に基づくプロトタイプ選択や，データの分布に応じた適応的なプロトタイプ数決定手法の検討が必要である．

7 おわりに

本論文では，集中治療期脳波の分類における ProtoPNet を用いた判断根拠の可視化手法を提案した．Top-α %平均による類似度計算と単一プロトタイプによる分類により，分類性能を維持したうえで，周期的な特徴を持つクラスで解釈性の改善を実現した．

今後の課題として，波形が時間変化する Seizure や多様な特徴を持つ Slowing に対する性能改善が挙げられる．また，実臨床での運用を想定し，学習されたプロトタイプが実際の診断で医師が重視する特徴と一致するかどうかの検証が重要である．専門医による評価を通じて，プロトタイプの臨床的妥当性や注目領域の診断価値について検討を行う必要がある．

参考文献

[1] Krumholz et al., ”Complex partial status epilepticus accompanied by serious morbidity and mortality,” *Neurology*, vol.45, no.8, pp.1499–1504, 1995.

[2] Sutter et al., ”The neurophysiologic types of nonconvulsive status epilepticus: EEG patterns of different phenotypes,” *Epilepsia*, vol.54, no.6, pp.23–27, 2013.

[3] Singh et al., ”Explainable Deep Learning Models in Medical Image Analysis,” *Journal of Imaging*, vol.6, no.6, 52, 2020.

[4] Itti et al., ”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] Zhou et al., ”Learning Deep Features for Discriminative Localization,” *CVPR*, pp.2921–2929, Las Vegas, USA, 2016.

[6] Selvaraju et al., ”Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *ICCV*, pp.618–626, Venice, Italy, 2017.

[7] Xie et al., ”ViT-CX: causal explanation of vision transformers,” *IJCAI*, pp.1569–1577, Macao, S.A.R, 2023.

[8] Chaofan et al., ”This Looks Like That: Deep Learning for Interpretable Image Recognition,” *NeurIPS*, vol.32, pp.8930–8941, Vancouver, Canada, 2019.

[9] Gao et al., ”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.

[10] Carmichael et al., ”Pixel-grounded prototypical part networks,” *WACV*, pp.4768–4779, Waikoloa, Hawaii, 2024.

[11] Donnelly et al., ”Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes,” *CVPR*, pp.10255–10265, New Orleans, LA, USA, 2022.

[12] Poppi et al., ”Revisiting The Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis,” *CVPRW*, pp.2299–2304, Nashville, TN, USA, 2021.

[13] Lawhern et al., ”EEGNet: A Compact Convolutional Neural Network for EEG-based Brain-Computer Interfaces,” *J. Neural Eng.*, vol.15, no.5, 056013, 2018.

[14] Moschitti et al., ”Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” *EMNLP*, pp.1724–1734, Doha, Qatar, 2014.