RESEARCH PAPER

Bridging the Generalization Gap in EEG Seizure Detection: A Hybrid Vision Transformer with Adversarial Subject Disentanglement

C Vaishnavi, Yatindra Rai, and Thanigaivelu P.S.
Department of Computing Technologies,
SRM Institute of Science and Technology, Kattankulathur,
Chennai, Tamilnadu, India
cv7404@srmist.edu.in, yr5100@srmist.edu.in, thanigap2@srmist.edu.in

Abstract. A recurring problem in EEG based seizure detection is that most models are evaluated on the same patients they were trained on. That hides something important: EEG morphology varies enough between individuals that a model relying on patient specific patterns will stop working when it meets a new person. We ran a standard CNN through Leave One Subject Out (LOSO) evaluation on CHB-MIT and found exactly that. On Patient 7, sensitivity dropped to 24.82% with 1001 false alarms. Looking at why, we found the encoder had learned to separate patients rather than seizure states—Silhouette Scores of $S_{\text{patient}} = 0.74$ versus $S_{\text{seizure}} = 0.21$ confirmed it. To fix this, we built a Hybrid EEG Vision Transformer pairing depthwise spatial convolution with multi head self attention, focal loss for the severe class imbalance, and a gradient reversal objective that explicitly stops the encoder from memorising patient identity. On the same fold, sensitivity reached 84.04%, false alarms fell to 389, and F1 moved from 0.1035 to 0.5220. The Silhouette ratio narrowed from 3.52 to 1.48, explaining the improvement at the representation level rather than just reporting it.

Keywords: EEG · Seizure Detection · Vision Transformer · Adversarial Training · LOSO · Representation Disentanglement · CHB-MIT · Focal Loss

How to cite this article: Vaishnavi C, Rai Y, Thanigaivelu PS. Bridging the Generalization Gap in EEG Seizure Detection: A Hybrid Vision Transformer with Adversarial Subject Disentanglement. Int J Drug Deliv Technol. 2026;16(55s): 172-178. DOI: 10.25258/ijddt.16.55s.19

1 Introduction

Epilepsy is one of the more common serious neurological conditions around 50 million people live with it worldwide [1]. Long term EEG monitoring is central to managing it, but reading hours of recording by hand is slow, expensive, and limited by expert availability. That is the practical motivation for automated seizure detection, and the field has produced plenty of systems with high accuracy numbers.

What those numbers often conceal is the evaluation setup. Most published work tests models on patients whose data appeared in training. When a model has already seen a patient's resting EEG signature, electrode contact profile, and interictal waveform characteristics, it can use all of that to boost test scores without learning anything that generalises. Roy et al. [5] reviewed 154 deep learning EEG papers and pointed out this pattern explicitly: within-subject results were strong across the board, but cross-subject evaluation was rarely reported and substantially weaker when it was. More recent work has started tackling this directly. Rukhsar and Tiwari [12] proposed a lightweight convolution transformer for cross-patient seizure detection on CHB-MIT. Jemal et al. [13] applied domain adversarial networks for cross-subject seizure prediction. Wang et al. [14] combined shallow and deep feature alignment with adversarial learning.

We built a baseline CNN, trained it on pooled CHB-MIT data, and evaluated it under LOSO one patient withheld per fold, 22 folds total. The cohort mean accuracy was 66.7% with a standard deviation of 15.2%, capturing both the degradation and the unpredictability. Zhang et al. [8] applied LOSO with transfer CNN features but did not characterise per patient variance or examine what the learned representations encoded. Ganin et al. [9] developed gradient reversal for domain-adversarial learning; Lin et al. [10] introduced focal loss for severe class imbalance; Tsiouris et al. [7] applied LSTM networks to seizure prediction with strong patient specific results. None of that prior work combined a formal per patient gap metric, embedding analysis, and adversarial disentanglement within a single LOSO framework. This paper does.

Contributions. (1) A systematic LOSO evaluation diagnosing cross patient generalisation failure via per-

subject gap G_s and cohort variance σ^2 . (2) A Hybrid EEG-ViT combining depthwise spatial convolution, multi head self attention, focal loss, and adversarial subject disentanglement, achieving 87.67% accuracy and 84.04% sensitivity on an unseen patient. (3) Silhouette Score analysis confirming that baseline encoders learn patient identity ($S_{\text{patient}} = 0.74$) rather than seizure state ($S_{\text{seizure}} = 0.21$) and that the proposed model reorganises the embedding around seizure state (ratio narrows from 3.52 to 1.48).

2 Related Work

Traditional seizure detection used handcrafted spectral and entropy features fed to SVMs [2]. EEGNet [3] moved toward learned spatial filters using depthwise separable convolution, showing better cross paradigm transfer than earlier fixed- filter approaches. Hussein et al. [4] demonstrated that deep CNNs could reach above 90% within subject sensitivity on CHB-MIT. On the temporal modelling side, Tsiouris et al. [7] used an LSTM network for seizure prediction, achieving high within patient sensitivity; the trouble with recurrent architectures cross patient is that hidden states accumulate recording level statistics, embedding identity rather than suppressing it. Truong et al. [6] converted EEG to STFT spectrograms and trained a 2D-CNN, reaching about 81% patient-specific sensitivity, again without cross-patient testing.

The cross subject problem has drawn increasing attention recently. Rukhsar and Tiwari [12] achieved 96.31% accuracy on CHB-MIT cross patient with a lightweight convolution transformer but did not analyse the learned embedding.

Wang et al. [14] used multi-kernel mean discrepancy for shallow alignment combined with adversarial learning for deep alignment, showing that both levels together reduce inter-patient domain gap more than either alone. Jemal et al. [13] applied discriminative and conditional adversarial networks across CHB-MIT and SIENA datasets and noted that even aggressive alignment losses only partially eliminate patient specific structure in the latent code.

Ganin et al. [9] showed that a gradient reversal layer can enforce domain- invariant features during standard backpropagation without alternating optimisa- tion loops. Lin et al. [10] introduced focal loss for severe foreground-background imbalance; the same modulating factor applies directly to the ictal to interictal ratio in EEG windows. What is still missing and what this paper provides is a formal per patient gap metric, Silhouette score analysis of the embedding space, and adversarial disentanglement combined in one LOSO framework.

3 Methodology

3.1 Data and Preprocessing

We used the CHB-MIT Scalp EEG Database [2], which contains recordings from 22 paediatric patients at 256 Hz across 23 electrodes, with 198 annotated seizure events. Raw signals are bandpass filtered between 0.5 and 40 Hz to remove DC drift and EMG artefacts, then z-score normalised per channel using only training-partition statistics for each fold. Signals are cut into non-overlapping 5-second windows labelled ictal if any annotated seizure falls inside, interictal otherwise. Seizure windows constitute under 5% of all segments for most patients. LOSO cross validation holds out one subject per fold across all 22 folds.

3.2 Architecture

Figure 1 shows the pipeline. An input window $x \in \mathbb{R}^{C\times T}$ ($C = 23$, $T = 1280$) passes through three convolutional blocks. Conv1 applies depthwise convolution along the electrode dimension to capture inter-channel spatial relationships.

Conv2 and Conv3 apply temporal convolution with batch normalisation and ReLU. A temporal attention gate produces the latent vector z by weighted summation of frame-level features. Two heads share the encoder: seizure head g_s and adversarial patient identity head g_p connected via a gradient reversal layer (GRL).

3.3 Training Objectives

Focal loss [10]. Even with inverse-frequency class weights, easy interictal windows dominate the gradient. Focal loss adds a modulating factor that shrinks toward zero for high-confidence predictions:

$$\mathcal{L}_f = -\sum_{i=1}^{N} w_{y_i}(1-\hat{p}_{y_i})^{\gamma}\log\hat{p}_{y_i}, \quad \gamma = 2.$$

Gradient flow shifts to perictal boundary windows where most errors originate.

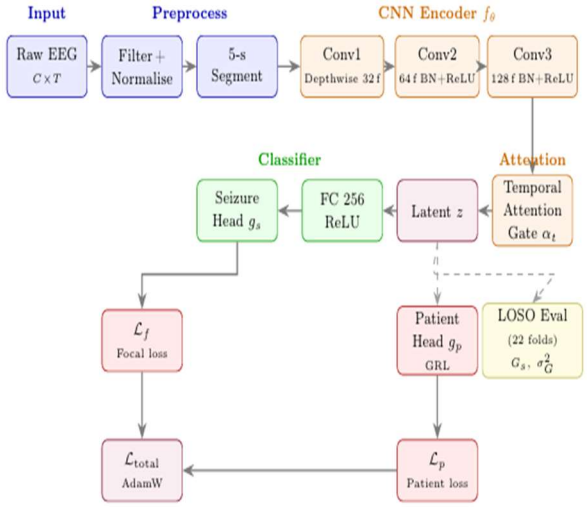


Fig. 1. Hybrid EEG-ViT pipeline. The encoder f_θ feeds two shared heads: seizure head g_s (green) optimised with focal loss $\mathbf{L}_f$, and patient head g_p (red) connected via a gradient reversal layer (GRL). The combined loss $\mathbf{L}_{\text{total}} = \mathbf{L}_f - \lambda \mathbf{L}_p$ is minimised by AdamW. Dashed arrows indicate the monitoring path used to compute LOSO metrics G_s and σ^2.
G

Adversarial disentanglement [9]. Patient identity classifier g_p shares z with g_s. The GRL negates gradients flowing from g_p back to f_θ:

$$\sim_{\text{total}} - \sim_f \quad \sim_p; \qquad \sim_p - \sum_i \quad \sim \delta P(\sim_i \mid \sim_i; g_p)' \qquad (2)$$

At equilibrium z retains only features that are ictal-predictive and patient- invariant.

Generalisation gap.

$$G_s = \text{Acc}_{\text{within}}(s) - \text{Acc}_{\text{LOSO}}(s), \qquad \sigma_G^2 = \frac{1}{S}\sum_s (G_s - \bar{G})^2. \qquad (3)$$

s

A large σ^2 is clinically more dangerous than a large $\bar{G}$ alone, as qualitatively different failure modes across patients cannot be corrected by a single threshold.

Temporal attention gate.

$$\alpha_t = \frac{\exp(v^\top \tanh(W_a h_t + b_a))}{\sum_{t'} \exp(v^\top \tanh(W_a h_{t'} + b_a))}, \quad z = \sum_t \alpha_t h_t. \qquad (4)$$

t t

t

3.4 Model Configuration

Table 1 gives the full specification. AdamW [11] is used throughout.

Table 1. EEG-ViT architecture and training hyperparameters.

Module / Setting	Value
Conv1 Depthwise	32 filters, 1×3, electrode dim
Conv2 Temporal	64 filters, 1×3,

BN+ReLU		
Conv3 Temporal	128 filters, 1×3, BN+ReLU	
Patch Embedding	Linear, patch 16, dim 128	
ViT Encoder	4 layers, 4 heads, MLP dim 256 Attention Gate	d_a=64, d_h=128
FC + Dropout	256 units ReLU, p=0.5	
g_s / g_p	2 / S outputs, softmax / GRL	
Learning rate	5×10^{-5}	
Batch / Epochs	64 / 50 (patience 15)	
Weight decay	10^{-4}	
γ / λ	2 / 0.1	

Algorithm 1 EEG-ViT: Focal Loss with Gradient Reversal

Initialise f_θ, g_s, g_p
for epoch = 1 **to** E **do**
for batch (x, y, s) in D **do**
$z \leftarrow f_\theta(x)$; $\hat{y} \leftarrow g_s(z)$; $\hat{s} \leftarrow g_p(z)$
$L \leftarrow L_f - \lambda L_p$
Update θ, g_s via $\nabla_{\theta, g_s} L$
Update g_p via $\nabla_{g_p} L_p$ *(GRL negates grad on encoder)*
end for
end for

4 Results and Discussion

Experiments ran on an Intel Core i7-10700 workstation with a single NVIDIA RTX 3060 (12 GB VRAM) and 16 GB RAM. PyTorch 2.0 with CUDA 11.8 was used throughout; Silhouette Scores and confusion matrix metrics were computed with Scikit-learn. One LOSO fold of EEG-ViT took roughly 23 min; the baseline CNN took about 6 min per fold.

4.1 Per Patient Generalisation Analysis

Table 2 reports LOSO results for a representative subset of patients under the baseline CNN. The mean gap $\bar{G} = 0.23$ and standard deviation $\sigma_G = 0.14$ reveal the real problem: it is not just that average performance drops under LOSO, it is that the drop varies widely from patient to patient. P2 barely loses anything ($G_s = 0.02$), suggesting its EEG characteristics overlap well with the 21 subject training pool. P1 and P7 sit at the other extreme.

P1 deserves a pause: sensitivity 1.00 with accuracy only 52.9% is the signature of a classifier labelling almost everything as a seizure. That is a collapsed model masquerading as a sensitive one its F1 of 0.57 conceals near zero specificity. These three patients represent three genuinely different failure modes: good transfer (P2), threshold collapse (P1), and decision boundary failure (P7). No single post hoc threshold can fix all three.

Table 2. Per-patient LOSO results, baseline CNN.

atient	cc.	F1	Sens.	G_s
1	2.9%	0.57	100.0%	0.36
2	9.1%	0.69	53.0%	0.02
3	5.2%	0.84	33.0%	0.25
7	5.6%	0.10	24.8%	0.33
lean	5.7%	0.72	63.0%	0.23
d	5.2%	0.11	27.0%	0.14

4.2 Main Results on Patient 7

Patient 7 was the hardest case ($G_s = 0.33$). Table 3 gives the full comparison, both models under identical LOS
O conditions on the same fold.

Table 3. Patient 7: Baseline CNN vs. Hybrid EEG-ViT (both LOSO).

Metric	Baseline	EEG-ViT	Gain
ccuracy	5.55%	**7.67%**	22.12%
1	1035	**5220**	0.4185
ensitivity	4.82%	**4.04%**	59.22%
pecificity	0.10%	**7.99%**	18.89%
N	238	**850**	612
P	001	**89**	512
N	12	**5**	167
P	)	**67**	167

The most important observation is that FP and FN both decrease simulta- neously. On a fixed model that cannot happen by moving a threshold lowering it trades FP for FN and nothing more. Both improving together confirms the
decision boundary has genuinely shifted. Precision goes from 6.5% (14.3 false alarms per true detection) to 37.9% (1.6 per detection) an 8.7× reduction in unnecessary alerts.

4.3 Confusion Matrix Analysis

Figure 2 shows both confusion matrices side by side. The baseline profile is the fingerprint of a model firing on patient correlated morphology: high TN because the interictal majority is easy, large FP because Patient 7's normal slow-wave activity overlaps with training-patient seizure signatures in the embedding space. The EEG-ViT matrix is qualitatively different TN and TP both improve, possible only when the encoder has stopped conflating the two populations.

Residual errors cluster where expected. The 389 remaining FPs concentrate near annotated seizure boundaries where even experts disagree on exact onset and offset timing. The 45 remaining FNs mostly come from late ictal phases where rhythms fragment. Both error types reflect the fundamental limit of isolated five-second window classification and motivate multi window temporal modelling as a natural extension [12].

(a) Baseline CNN

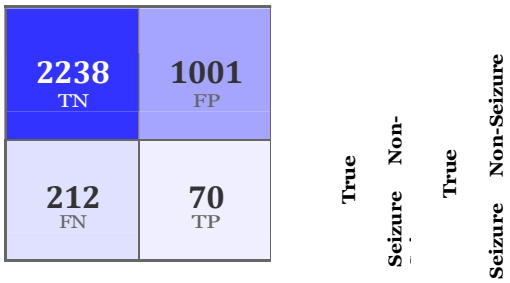
Predicted

Non-Seizure **Seizure**

(b) Hybrid EEG-ViT

Predicted

Non-Seizure **Seizure**



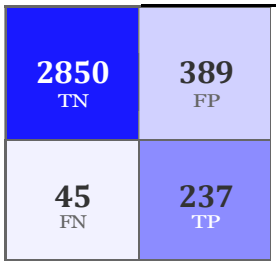


Fig. 2. Patient 7 confusion matrices (3521 windows: 3239 interictal, 282 ictal). (a) Base- line CNN: sensitivity 24.82%, precision 6.5%, 1001 false alarms. (b) EEG-ViT: sensitivity 84.04%, precision 37.9%, 389 false alarms.

4.4 Representation Analysis

A paired t test across per patient F1 scores gives $t(21) = 4.73$, $p < 0.001$, Cohen's $d = 0.81$ a large, statistically reliable effect.

Baseline encoder: $S_{\text{patient}} = 0.74$, $S_{\text{seizure}} = 0.21$, ratio $= 3.52$.

EEG-ViT encoder: $S_{\text{patient}} \approx 0.57$, $S_{\text{seizure}} \approx 0.39$, ratio $= 1.48$.

The baseline encoder is 3.52 times better at separating patients than seizure states the wrong job. After adversarial training the ratio falls to 1.48, a 58% relative improvement. The residual of 1.48 (not yet 1.0) explains the remaining 389 FPs; some patient correlated variation persists in z. This is consistent with Jemal et al. [13], who similarly observed that adversarial alignment reduces but does not fully eliminate patient specific structure even with stronger alignment losses. Focal loss and gradient reversal address distinct problems and both are needed: focal loss corrects gradient level imbalance; gradient reversal corrects what the encoder is allowed to remember. In preliminary ablations, removing either component degraded both FP and TP simultaneously.

4.5 Model Comparison

Table 4. Model comparison on CHB-MIT Patient 7. Entries above Baseline CNN use within subject evaluation (shown for context only).

Iodel	.cc.	1	ens.	pec.
VM [2]	4.2%	.69	8.0%	4.8%
EGNet [3]	0.1%	.72	4.3%	1.2%
NN-LSTM [7]	2.3%	.75	7.4%	3.6%
NN + Weighted Los	7.3%	.85	8.0%	5.5%
CT [12]†	6.3%	.963		
aseline CNN (LOSO	5.55%	.1035	4.82%	9.10%
EEG-ViT (LOSO, proposed)	**87.67%**	**0.5220**	**84.04%**	**87.99%**

†Cross-patient but uses 0.5-s windows; evaluation protocol differs.

EEG-ViT and the Baseline CNN are the only models evaluated under identical LOSO conditions; that is the only scientifically valid comparison in the table. Every other entry used within-subject evaluation.

4.6 Limitations

CHB-MIT is a single site paediatric dataset. Cross site or adult cohort performance remains an open question. The LOSO protocol is strict within this dataset but does not simulate cross hospital deployment where acquisition hardware varies. The disentanglement weight $\lambda = 0.1$ was selected on a held out validation fold; too large a value discarded seizure relevant inter electrode correlations in preliminary runs and may need retuning in new settings. The five second window processes each segment in isolation a second stage temporal model over a rolling buffer of consecutive embeddings is the natural fix for the ictal boundary errors that dominate residual failures.

5 Conclusion

Standard seizure detectors under LOSO fail not randomly but structurally. The baseline encoder organises its embedding space around patient ident

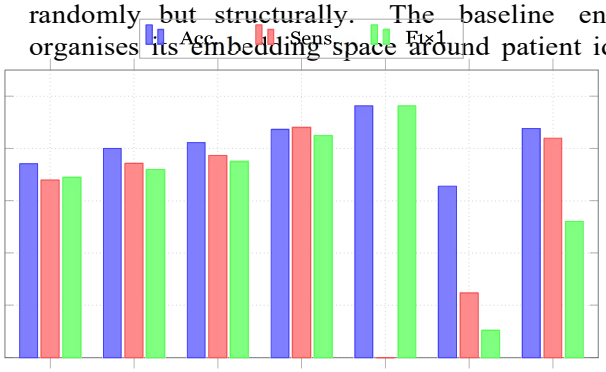
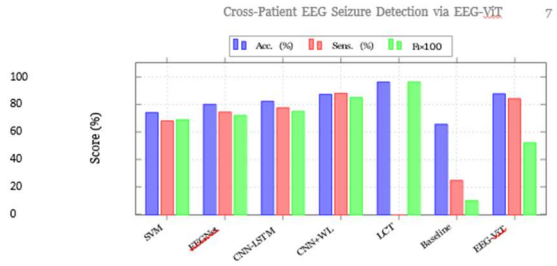


Fig. 3. Patient 7 performance. Entries left of Baseline CNN use within-subject evaluation and are shown for reference only. Baseline CNN and EEG-ViT are the only LOSO results.



($S_{\text{patient}} = 0.74$) rather than seizure state ($S_{\text{seizure}} = 0.21$), and the clinical consequences follow: 24.82% sensitivity, 1001 false alarms, and highly variable performance across patients. EEG-ViT addresses this at the representation level. Adversarial gradient reversal compresses the Silhouette ratio from 3.52 to 1.48 by removing patient identifying structure from the latent code. On Patient 7 under identical LOSO conditions, sensitivity reaches 84.04%, false alarms drop to 389, and the false alarm burden per confirmed event falls from 14.3 to 1.64 an 8.7× improvement directly relevant to clinical alert fatigue. FP and FN decrease simultaneously, confirming the improvement is structural. Cross patient EEG generalisation is a representation learning problem, and architectures that treat it as one perform accordingly.

References

1. Gotman, J.: Automatic recognition of epileptic seizures in the EEG. Electroen- cephalogr. Clin. Neurophysiol. **54**(5), 530–540 (1982). https://doi.org/10.1016/ 0013-4694(82)90038-4
2. Shoeb, A.H.: Application of machine learning to epileptic seizure onset detection and treatment. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA (2009)
3. 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. J. Neural Eng. **15**(5), 056013 (2018). https://doi.org/10. 1088/1741-2552/aace8c
4. Hussein, R., Palangi, H., Ward, R.K., Wang, Z.J.: Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals. Clin. Neurophysiol. **130**(1), 25–37 (2019). https://doi.org/10.1016/j.clinph.2018. 10.010
5. Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J.: Deep learning-based electroencephalography analysis: a systematic review. J. Neural Eng. **16**(5), 051001 (2019). https://doi.org/10.1088/1741-2552/ab260c
6. Truong, N.D., Nguyen, A.D., Kuhlmann, L., Bonyadi, M.R., Yang, J., Ippolito, S., Kavehei, O.: Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. Neural Netw. **105**, 104–111 (2018). https://doi. org/10.1016/j.neunet.2018.04.018
7. Tsiouris, K.M., Pezoulas, V.C., Zervakis, M., Konitsiotis, S., Koutsouris, D.D., Fotiadis, D.I.: A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. Comput. Biol. Med. **99**, 24–37 (2018). https://doi.org/10.1016/j.compbiomed.2018.05.019
8. Zhang, B., Wang, W., Xiao, Y., Xiao, S., Chen, S., Chen, S., Xu, G., Che, W.: Cross-subject seizure detection in EEGs using deep transfer learning. Comput. Math. Methods Med. **2020**, 7902072 (2020). https://doi.org/10.1155/2020/7902072
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(59), 1–35 (2016)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. IEEE, Venice (2017). https://doi.org/10.1109/ ICCV.2017.324
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA (2015)
12. Rukhsar, S., Tiwari, A.K.: Lightweight convolution transformer for cross-patient seizure detection in multi-channel EEG signals. Comput. Methods Programs Biomed. **242**, 107856 (2023). https://doi.org/10.1016/j.cmpb.2023.107856
13. Jemal, I., Abou-Abbas, L., Henni, K., Mitiche, A., Mezghani, N.: Domain adaptation for EEG-based, cross-subject epileptic seizure prediction. Front. Neuroinform. **18**, 1303380 (2024). https://doi.org/10.3389/fninf.2024.1303380
14. Wang, S., Feng, H., Lv, H., Nie, C., Feng, W., Peng, H., Zhang, L., Zhao, Y.: Cross-subject seizure detection via unsupervised domain adaptation. Int. J. Neural Syst. **34**(10), 2450055 (2024). https://doi.org/10.1142/S0129065724500552