OPEN ACCESS
International Journal of Neural Systems, (2026) 2650033 (18 pages)

DOI: 10.1142/S0129065726500334

World Scientific
www.worldscientific.com

An Interpretability Framework for Convolutional Neural Network-Based Electroencephalography Analysis Discovers New Spatial and Spectral Epileptic Biomarkers

Vadim V. Grubov*
Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Stremyannyy Ln., 36, Moscow 236041, Russia
vvgrubov@gmail.com

Oleg E. Karpov
Pirogov National Medical and Surgical Center, 70, Nizhnyaya Pervomayskaya Str., Moscow 127051, Russia

Sergei I. Nazarikov and Semen A. Kurkin
Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Stremyannyy Ln., 36, Moscow 236041, Russia

Nikita P. Utyashev
Pirogov National Medical and Surgical Center, 70, Nizhnyaya Pervomayskaya Str., Moscow 127051, Russia

Denis A. Andrikov
Bauman Moscow State Technical University, 5, 2nd Baumanskaya Str., Moscow 105005, Russia

Alexander E. Hramov
Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Stremyannyy Ln., 36, Moscow 236041, Russia
Pirogov National Medical and Surgical Center, 70, Nizhnyaya Pervomayskaya Str., Moscow 127051, Russia

Received 14 August 2025
Accepted 17 March 2026
Published Online 18 April 2026

Deep learning (DL) models, particularly convolutional neural networks (CNNs), have shown promise in automated epileptic seizure detection from electroencephalogram (EEG). However, their "black-box" nature limits clinical adoption, as interpretability is critical for trust and validation in medical applications. A novel interpretability method for CNN-based seizure detection models, designed to uncover meaningful spatial and spectral EEG biomarkers, is proposed. The approach combines frequency- and spatial-domain interpretation to provide both global model behavior analysis and local, sample-specific explanations. It also accounts for the task-specific design, neurophysiological grounding and cross-framework validation — concepts often neglected by many state-of-the-art methods. Results are represented as heatmap matrices of feature importance

*Corresponding author.

(5 frequency bands * 5 brain regions) with important features determined through statistical testing. Interpretation is based on neurophysiological alignment of these features. The method is validated on three CNN architectures, demonstrating how each leverages distinct frequency bands and brain regions for seizure identification. Global interpretation reveals that the highest-performing model utilizes complementary biomarkers across multiple frequency bands, while local interpretation captures dynamic intra-seizure spectral shifts. The results align with known neurophysiological mechanisms, such as thalamocortical interactions (theta-band) and default mode network suppression (alpha/beta-bands), while also suggesting new biomarkers for seizure detection. The method bridges the gap between DL and clinical EEG analysis, offering a tool for model validation and discovery of electrophysiological signatures in epilepsy.

Keywords: Convolutional neural network; epileptic seizure detection; EEG; continuous wavelet transform; machine learning interpretability; global and local interpretation; frequency- and spatial-domain interpretation.

1. Introduction

Epilepsy, a chronic neurological disorder marked by recurrent seizures, remains a diagnostic challenge despite available treatments like antiepileptic drugs, surgery, and stimulation.[1] Electroencephalography (EEG) is the gold standard for seizure detection, but manual analysis is labor-intensive due to the rarity and heterogeneity of seizures.[2] Artificial intelligence (AI) is known for its ever-growing role in the automated diagnosis of various neurological disorders.[3] In this context, a promising approach is deep learning (DL), particularly convolutional neural networks (CNNs)[4,5] and their variations: Tiny-CNN,[6] temporal convolutional networks (TCNs),[7] variants with transfer learning,[8] or hybrid architectures.[9,10] However, clinical adoption of CNNs is hindered by their "black-box" nature — clinicians require interpretable predictions to trust and act upon them.[11] Thus, interpretability methods are an important step toward explainable AI (XAI) in medicine.[12]

Current interpretability methods for EEG-based CNNs often lack some important features, as the presented brief review suggests (see Sec. 2). To address these gaps, a novel interpretability method tailored to CNN-based seizure detection is proposed, with three key contributions:

- Task-specific design that combines frequency-domain (Grad-CAM) and spatial-domain (occlusion-based) analysis to reveal physiologically meaningful EEG patterns;
- Validation across frameworks, consisting of testing on three distinct CNN-based models to ensure robustness;
- Biophysical alignment, which involves interpreting the results in the context of known and potentially novel domain knowledge about epilepsy.

As shown in Fig. 1, the pipeline includes EEG data preprocessing (filtering, artifact removal, continuous wavelet transform (CWT)) (see Secs. 3.1 and 3.2), training three CNN-based models[13] for seizure classification (see Sec. 3.3), and interpreting predictions using the interpretability method, which evaluates these models through (see Sec. 4):

- Frequency-domain importance — Grad-CAM-derived saliency maps highlighting critical spectral bands;
- Spatial-domain importance — occlusion-based assessment of important EEG electrode regions;
- Integrated frequency–spatial-domain analysis — combined frequency–spatial heatmaps for global and local interpretation.

The results demonstrate how each model leverages distinct neurophysiological signatures (e.g. θ-band thalamocortical oscillations, α/β-band default mode network suppression), offering insights into both model behavior and potential biomarkers in terms of global and local interpretation. This work bridges the gap between DL and clinical EEG analysis, advancing interpretable machine learning (ML) for epilepsy diagnosis (see Sec. 6).

2. Related Works

Recent years have witnessed growing interest in interpretability methods for CNN-based EEG analysis. This section reviews key approaches across various EEG applications, highlighting methodological gaps specific to seizure detection.

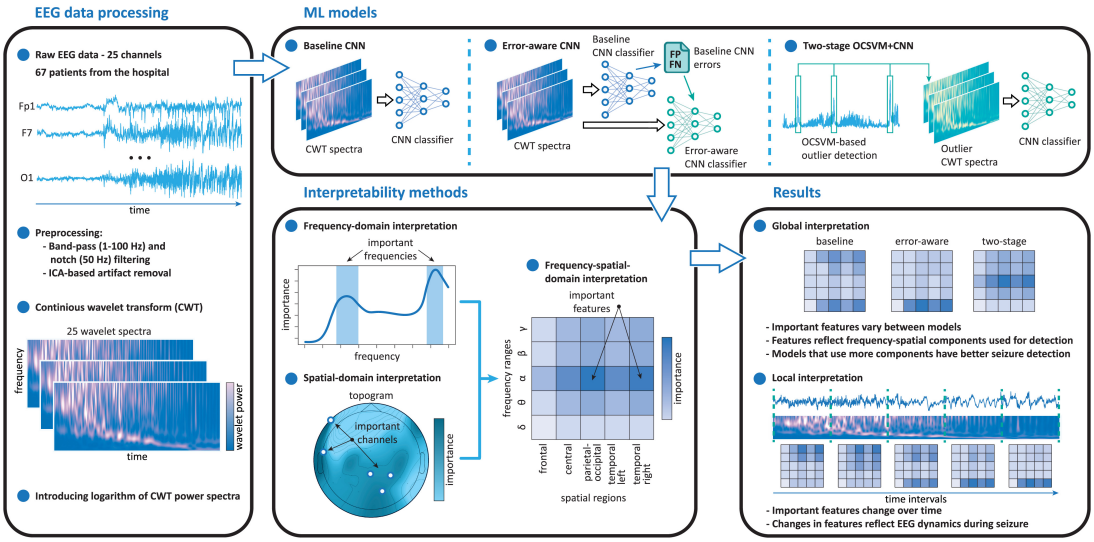

Fig. 1. Pipeline of EEG data processing and model interpretation that includes: (i) Data processing (bandpass/notch filtering, ICA artifact removal) and time–frequency representation via CWT and log-normalization; (ii) ML models: three CNN variants (Baseline, Error-aware, OCSVM+CNN) trained for binary seizure classification; (iii) Interpretability method: joint frequency–spatial analysis via Grad-CAM and occlusion to identify critical EEG features; (iv) Results: important frequency–spatial features of three CNN models assessed via global and local interpretation.

2.1. *Interpretability in EEG analysis*

Vilamala *et al.*[14] pioneered interpretable EEG sleep staging using sensitivity analysis. Their approach calculated input gradients to construct sensitivity maps from multitaper spectral features. Two VGG (Visual Geometry Group) variants were evaluated: VGG-FE (frozen convolutional layers) which had all convolutional layers fixed during training, and VGG-FT (fully trainable) which got updates for all weights during training. The analysis revealed dominant α (8–12 Hz), θ (4–7 Hz), σ (12–15 Hz), and low-frequency ($\leq$1.5 Hz) bands. However, the limitations included:

- outdated architecture lacking modern components (e.g. residual connections, batch normalization),
- potential bias from evaluating only the "best performing subject",
- unsubstantiated biophysical claims without domain-specific references.

Cui *et al.*[15] worked on the task of detecting driver drowsiness. They trained a simple CNN on raw EEG signals, and used a method based on the CNN fixation idea with an alternative way of tracing important positions in feature maps. Their technique traced important features through network layers, applying Gaussian blurring to obtain the importance maps. Key findings included: α/θ bands and specific electrode groups drove true positives (TP), supported by neurophysiological evidence, and β/δ combinations characterized true negatives (TN). Error analysis identified three confounding factors:

- the sensor noise contained in the EEG signals,
- the presence of electromyogram (EMG) activity,
- the transient nature (between wakefulness and drowsiness) of misclassified examples.

Notably, the model utilized typically discarded artifacts as predictive features. The study's limitations include an oversimplified CNN architecture with two convolutional layers and one fully-connected layer and potential selection bias from few "representative" samples.

Zhao *et al.*[16] introduced interpretable CNN models for emotion recognition using EEG data. To solve the emotion classification task, they normalized the raw data using a baseline and employed a simple, 11-layer, VGG-like network with max-pooling layers that reduced only the temporal dimension of the data. They used the Grad-CAM approach for interpretation, which provided spatiotemporal heatmaps that revealed the physiological patterns learned by

the model. The analysis showed that the trained CNN model primarily relies on data from channels located on the right side of the frontal lobe, the left side of the parietal lobe, and some intermediate regions. This result is consistent with biological discoveries of biomarkers associated with emotion recognition. However, this method has two limitations: the lack of frequency domain interpretations and the usage of a single model with a trivial architecture.

2.2. *Seizure-specific interpretability*

Zhao *et al.*[17] introduced an unconventional approach: instead of the commonly considered raw EEG data or its time–frequency representation they used plot images of the raw signal obtained using the *Matplotlib* library. In their study the authors trained several CNN architectures: LeNet, VGG, ResNet,[18] and Visual Transformer (ViT). To interpret CNN-based models they used Grad-CAM[19] heatmaps, while for the attention-based model they visualized the attention layer. While results aligned with EEG waveform features, the interpretations lacked connections to established epilepsy biomarkers, representing a missed opportunity for clinical validation.

Gabeff *et al.*[20] considered global and local interpretation of DL models for epileptic seizure detection from EEG signals. The authors analyzed three CNN models that differed only in the kernel size of the first convolutional layer. The interpretability analysis included two approaches: (i) the first-layer kernels Activation Maximization via gradient ascent (80 epochs, step size 0.5); (ii) DeepLIFT to visualize the learned features back onto the EEG signals. Using the generated maximized inputs, the authors analyzed which frequencies they contained and theorized why this may be important for seizure detection from a biophysical perspective. However, there were several limitations for this work. The authors used different techniques for local (DeepLIFT) and global interpretation (Activation Maximization) which makes it difficult to understand the correlation between local and global representation. The DeepLIFT is a backpropagation-based approach based on the concept of difference from a reference activation obtained from a reference input. The choice of the reference input is highly dependent on domain knowledge. The authors used a 0-signal (with all channels set to zero) as the reference input. However, if the model was not exposed to such samples during training, the assumption that the training and test datasets have the same distribution would be violated, resulting in a phenomenon known as "domain shift".[21] Thus, the difference in activations in this case may be due to the "domain shift" rather than specific features being important.

Tuncer and Dogan[22] introduced a relation-based feature extraction function named Friend Pattern that builds a feature vector for further epilepsy classification and interpretation. The interpretation method itself is based on the Directed Lobish symbolic language, which generates a sentence by deploying the identities of the extracted features. A connectome diagram was built using the generated sentence, and frequency analysis was performed on the symbols of the generated sentence. The authors analyzed these results and drew conclusions about the type of epilepsy presented in the dataset and the spatial importance of specific regions. Despite this method's innovative approach, it has two limitations: the conclusions lacked connections to known biophysical manifestations of epilepsy, and interpretations were limited to spatial importance.

2.3. *Key methodological gaps*

The review reveals four critical requirements for EEG interpretability methods.

2.3.1. *Task-specific design*

Most approaches apply generic interpretation techniques without adapting to seizure detection's unique challenges.[17,20] Thus, task-specific design (TSD) is crucial, and this oversight motivates the domain-tailored solution.

2.3.2. *Neurophysiological grounding*

Effective interpretability requires mapping model decisions to known neurophysiological principles, i.e. neurophysiological grounding (NG).[23,24]

Interpretability is commonly described as *the ability to provide explanations in understandable terms to a human.*[23] In such a paradigm, *explanations* are seen as logical decision rules and *understandable terms* should be derived from the domain knowledge related to the task.[24] In the seizure detection task domain knowledge typically refers to

frequency ranges and EEG channels of interest, which is true for both manual and automated EEG analysis.[25] Thus, it is important to combine frequency interpretation (FI) and spatial interpretation (SI). Considering the above, it is crucial to build bridges between the explanations found and the epilepsy features known from the clinical practice, because understanding the logic of the successful model can provide valuable insights into the biomarkers of the disease.[26]

2.3.3. *Multi-scale interpretation*

Following Ref. 27, (i) global interpretation (GI) — analysis for classwise patterns,[28] and (ii) local interpretation (LI) capturing seizure evolution[29] are integrated.

Global analysis is able to provide information about which features/markers are important for a particular class as a whole, while local methods allow to understand why the model made a certain decision on a particular sample. Epileptic seizures possess features that help distinguish between ictal and interictal EEG in general, treating these two types of activity as belonging to two different classes.[28] At the same time, epileptic EEG activity can demonstrate evolution throughout the seizure resulting in different parts of a seizure having different spatial, temporal, and frequency features.[29] In this way, parts of the seizure (i.e. different samples) can be treated as belonging to different classes. In this context, both global and local interpretability can aid in the discovery of new biomarkers of epilepsy, and methods that incorporate them both are of paramount importance.

2.3.4. *Cross-framework validation*

Single-framework evaluations can lead to misleading conclusions.[30] "Framework" here refers to the overall design and specifics of the model's training. Testing interpretation methods on models within the same framework lacks verifiable common-sense expectations about model's behavior. Sturmfels *et al.*[30] demonstrated this by altering the training data to directly encode the target class, providing prior knowledge to verify interpretations. The authors tested two saliency methods — edge detection with subsequent smoothing and expected gradients. The first method more accurately reflected what humans think is the relationship between the image and the label. However, this result was misleading in this case. In contrast, only the second method revealed what the network had truly learned. Similarly, Adebayo *et al.*[31] used model changes to assess interpretation consistency with expectations. Only Grad-CAM exhibited expected changes, questioning the validity of other methods. Testing interpretations across different frameworks provides a more reasonable basis for evaluating their behavior. These findings motivate the cross-framework validation (CFV) across three CNN-based models to ensure robust interpretation analysis.

Table 1. Comparison of key characteristics of interpretability methods. Abbreviations: TSD — task-specific design, NG — neurophysiological grounding, LI — local interpretation, GI — global interpretation, SI — spatial interpretation, FI — frequency interpretation, CFV — cross-framework validation

Method	TSD	NG	LI	GI	SI	FI	CFV
Vilamala[14]	—	—	✓	✓	—	✓	—
Cui[15]	✓	✓	✓	—	✓	✓	—
Zhao[16]	✓	—	—	✓	✓	—	—
Zhao[17]	—	—	✓	—	—	✓	—
Gabeff[20]	✓	✓	✓	—	✓	✓	—
Tuncer[22]	—	✓	✓	✓	✓	—	—
Proposed	✓	✓	✓	✓	✓	✓	✓

Table 1 shows the presence of the aforementioned key features in the existing methods and the proposed method.

3. Materials and Methods

3.1. *Data acquisition*

The models of this study were trained on the dataset collected at the Pirogov National Medical and Surgical Center (Moscow, Russia). The data acquisition protocol follows the same methodology described in the previous works.[13,32] Key characteristics of the dataset include:

- 67 drug-resistant nonphotosensitive focal epilepsy patients after quality control;
- Total 132 spontaneous seizures recorded with 1–5 seizures per patient;
- EEG recordings: 25 channels, 128 Hz sampling rate, 10–20 EEG International System, "Micromed" EEG recorder (Micromed S.p.A., Italy);

- Total recording duration: 955.18 h;
- Total epileptic activity duration: 4.10 h;
- Seizure duration: 47–250 s (mean: 109.9 s).

While the size of the dataset may not be suitable for wide generalization, it is large enough to serve as an illustrative example and demonstrate the viability of the proposed approach. The data are intentionally left without any manual preprocessing such as visual analysis for "bad" epochs with noise or artifacts. This makes the dataset more similar to real data from a hospital, which facilitates in demonstrating the robustness of any tested method.

3.2. *Data processing*

The data processing pipeline is fully automated and follows the previously established methodology.[13,32] EEG signals were bandpass-filtered (1–60 Hz) with notch filtering at 50 Hz to remove low- and high-frequency noise as well as power grid interference. Next, artifact removal based on independent component analysis (ICA) was performed. ICA separates a multivariate EEG signal into linearly independent components. Many neurophysiological artifacts, such as eyeblinks, originate from a source independent of the EEG source (the brain) and are therefore localized in a separate component that can be removed to clear the EEG signal of such artifacts.[33] All EEG preprocessing was performed using the *FieldTrip* toolbox for MATLAB.[34]

EEG signals were then represented in the time–frequency domain using CWT with Morlet mother wavelet.[35] Besides some specifically designed wavelets, Morlet is often seen as the most appropriate mother wavelet to describe the time–frequency structure of epileptic EEG patterns.[36] Notably, time–frequency decomposition is known to improve DL-based epilepsy diagnosis.[37] Wavelet power (WP) is considered as

$$w_n(f,\tau) = |W_n(f,\tau)|^2, \tag{1}$$

where $W_n(f,\tau)$ are CWT coefficients, $n = 1, 2, \ldots N$ is the number of EEG channel ($N = 25$), f is the frequency and τ is the time shift of the mother wavelet.

The input for ML models was based on WP for each EEG channel in the frequency range of 1–40 Hz. The WP is segmented into nonoverlapping 10-s intervals following the previous works.[13,38] Additionally, a normalized logarithm of the WP is introduced as an input to the CNN because CNN models generally converge faster and more stably when the input data conforms to a normal distribution with a mean close to zero and a constrained variance[39]:

$$w_n^{\log}(f,\tau) = \ln(w_n(f,\tau)),$$
$$w_n^{\text{norm}}(f,\tau) = \frac{w_n^{\log}(f,\tau) - \mu(w_n^{\log})}{\sigma(w_n^{\log})}, \tag{2}$$

where $\mu(\cdot)$ is the mean value, and $\sigma(\cdot)$ is the standard deviation.

3.3. *Machine learning models*

Three models, developed in the previous studies, are considered: Baseline CNN,[40] Error-aware CNN [38] and Two-stage OCSVM+CNN.[13]

3.3.1. *Baseline CNN*

A CNN-based classifier with modified ResNet-18 architecture[40] for seizure detection, treating the wavelet spectrum as a pseudo-image, was employed. The architecture adaptations include: (i) modifying the first convolutional layer to accept 25 input channels corresponding to EEG electrodes, and (ii) reducing the final fully connected layer to a single output neuron for binary classification. While not state-of-the-art, ResNet-18 provides proven reliability for this image-like classification task.[18]

The ResNet-18 architecture has 18 layers and ∼11.3 million parameters. CNN models were trained for 10 epochs with a learning rate of 0.001, a batch size of 4, the Adam optimizer, and the Binary Cross Entropy (BCE) loss function. To ensure reasonable training time, 100 examples from each patient were chosen for each epoch. All parameter values were selected through experiments with the baseline CNN model and were fixed for all other models to ensure correct comparisons.

In the analyzed dataset the class of interest is rare (minority/majority ratio $\sim 1/233$), so it is important to address the class imbalance problem. First, the minority class (epileptic EEG) is oversampled and the majority class (normal EEG) is undersampled.[41] It is achieved by manipulating the likelihood of the data segments to be selected for the training. In a signal of total L segments, the probability of each segment to be selected for training is the same — P. However, when the aforementioned approach is used for an imbalanced dataset of L_n segments of normal activity and L_a segments of epileptic activity, the

probability of the normal segment to be selected becomes P_n, and the probability of the epileptic segment to be selected becomes P_a:

$$L = L_a + L_n, \quad P = \frac{1}{L}, \quad P_a = \frac{1}{2L_a}, \quad P_n = \frac{1}{2L_n}. \tag{3}$$

Second, to further increase the robustness of the model, augmentation — minor modifications to the dataset to artificially increase its size — is used. Here two augmentation techniques are utilized: (i) random mirroring of data segments in the temporal dimension; (ii) SpecAugment[42] that modifies WP spectrum by zeroing a random range in the temporal and/or frequency dimension. Each technique was applied to a training sample with a 50% probability, thus 75% of the training samples were augmented with one or both techniques.

3.3.2. *Error-aware CNN*

This model employs a cascade approach to enhance classification accuracy by reducing false positives.[38] First, a baseline CNN identifies challenging classification cases. Subsequently, a second CNN is trained with a modified sampling strategy: 50% standard samples and 50% Baseline CNN misclassifications. This balanced approach prevents overfitting while enabling the model to learn from complex cases, thereby improving overall performance. This is achieved by modifying probabilities of training sample selection. For error-aware CNN model, Eqs. (3) take the following form:

$$\begin{aligned} &L = L_a + L_n = (L_{\mathrm{TP}} + L_{\mathrm{FN}}) + (L_{\mathrm{TN}} + L_{\mathrm{FP}}), \\ &P_n = \frac{1}{4L_n}, \quad P_a = \frac{1}{4(L - L_n)}, \quad P_e = \frac{1}{2(L_{\mathrm{FP}} + L_{\mathrm{FN}})}, \end{aligned} \tag{4}$$

where P_e is the probability of segment with error to be selected, L_{FP} is the number of segments with false positive prediction by the initial CNN, L_{FN} is the number of segments with false negative prediction.

3.3.3. *Two-stage hybrid model*

This hybrid approach combines one-class support vector machine (OCSVM) with a CNN for refined seizure classification.[13] The methodology builds on the previous findings that epileptic seizures exhibit extreme event characteristics in specific frequency bands.[43–45] Although OCSVM-based outlier detection performed similarly to supervised methods,[40,46] further improvements required fundamental algorithmic changes beyond feature optimization. The proposed two-stage architecture addresses this by first identifying extreme events via OCSVM, then applying CNN-based classification for enhanced detection accuracy.[13]

To achieve two-stage classification, another variant of the modified probabilities of training sample selection is implemented:

$$\begin{aligned} &L = L_a + L_n = L_a + (L_{\mathrm{TN}} + L_{\mathrm{FP}}), \\ &P_a = \frac{1}{2L_a}, \quad P_{\mathrm{FP}} = \frac{1}{2L_{\mathrm{FP}}}, \quad P_{\mathrm{TN}} = 0, \end{aligned} \tag{5}$$

where P_{FP} is the probability of segment with OCSVM's false positive to be selected, P_{TN} is the probability of segment with OCSVM's true negative to be selected, L_{FP} is the number of segments with false positive prediction by OCSVM.

3.4. *Statistical analysis*

To statistically compare metrics of seizure detection quality between CNN models we used Repeated Measure Analysis of Variance (RM-ANOVA). First, we tested for the main effect, and if it was significant ($p < 0.05$), we performed pairwise comparisons (Baseline versus Error-aware, Baseline versus Two-stage, and Error-aware versus Two-stage) using the dependent samples t-test without and with the correction to multiple comparisons.

To identify the most discriminative features, we conducted a systematic statistical analysis of importance score across regions and frequency bands. We performed a one-sample, one-tailed t-test for each region-band combination to determine whether its average importance score significantly exceeded the 75th percentile of the global importance score distribution across all regions, bands, and models. This approach allowed us to identify features that consistently exhibited elevated importance levels relative to the overall distribution. All statistical tests were conducted at $\alpha = 0.05$.

4. Interpretability Method

An interpretability framework for CNN models processing multi-channel EEG data in the

time-frequency domain is proposed. Let $x = \{x_i\}_{i=1}^{N}$ represent raw signals from N electrodes, transformed into 2D spectra over frequency range $[f_{\min}, f_{\max}]$ (with $F = f_{\max} - f_{\min}$) using some transform (for example, CWT).

Formally, each preprocessed input sample is

$$G(x) = \{G(x_i)\}_{i=1}^{N} \in \mathbb{R}^{N\times F\times T}, \tag{6}$$

where G denotes the preprocessing and time–frequency transformation pipeline (CWT in this example), and T is the segment duration. CNN model $m : \mathbb{R}^{N\times F\times T} \to \mathbb{R}$ maps these inputs to seizure probability scores.

Following Ref. 24, the proposed interpretability method is *post hoc*, requiring no model m retraining or architectural changes. It explains the behavior of the model by attribution — in the process of construction, the method assigns importance scores to the input features. It provides both global and local interpretation through two complementary components: spatial region importance and frequency range importance.

4.1. *Spatial region importance*

The spatial region importance helps to understand which cortical areas were particularly important for the prediction. The importance of a region consists of the importance of the individual EEG channels forming this region. The channel importance calculation was inspired by the technique of occluding RGB images with patches proposed in Ref. 47. This approach is rather straightforward: if the model's prediction changes significantly after occluding part of an image, then the occluded part of the image was important. However, this technique cannot be applied to EEG data without certain adjustments. In EEG, occlusion means replacing the entire channel signal with zeros, which carries a high risk of "domain shift".[21]

The following approach to EEG data occlusion has been suggested. For an EEG signal x from a segment of epileptic activity, we introduce the baseline EEG signal $\bar{x}$, which contains nonepileptic activity for the same EEG channel. $\bar{x}$ is selected from the segments $\mathcal{X}^{\text{pre}}$ in the minute preceding the seizure, and the selection criterion is the minimum average spectrum power:

$$\bar{x} = \{\bar{x}_i\}_{i=1}^{N} = \text{argmin}_{x^{\text{pre}}\in\mathcal{X}^{\text{pre}}}\mu(G(x^{\text{pre}})). \tag{7}$$

Then the importance of the jth channel can be defined as

$$CI_j = |m(G(x)) - m(G(\hat{x}^j))|, \tag{8}$$

$$\hat{x}^j = \{x_1, \ldots, x_{j-1}, \bar{x}_j, x_{j+1}, \ldots, x_N\}, \tag{9}$$

where $\hat{x}^j$ is the original EEG signal with the jth channel being replaced with baseline signal $\bar{x}$. Thus, the importance of a channel is defined as the change in the model's prediction when the signal from this channel is replaced by its baseline signal.

Using the importance of the channel CI_j (Eq. (8)), the importance of the spatial region of interest R can be defined as the average change in the model's prediction when the signal of each individual channel j_r from R is replaced with the baseline signal $\bar{x}$:

$$\text{RI}^R = \frac{1}{|R|}\sum_{j_r\in R} CI_{j_r}, \tag{10}$$

where R is the subset of channels that make up the spatial region of interest.

The channel importance proposed in Eq. (8) can also be used to pick the most informative/representative EEG channel:

$$j^* = \text{argmax}_{j\in\{1,\ldots,N\}} CI_j. \tag{11}$$

4.2. *Frequency range importance*

The importance of frequency ranges is based on the Grad-CAM approach.[19] It is a simple and well-studied approach in the field of DL, and unlike other saliency map methods of interpretation, it passes important sanity checks described in Ref. 31.

In this approach, the CNN architecture decomposes into three mappings:

$$m = m^{\text{MLP}} \circ m^{\text{GAP}} \circ m^{\text{FE}}, \tag{12}$$

where these mapping are:

- high-level feature extractor (FE)
$$m^{\text{FE}} : \mathbb{R}^{N\times F\times T} \to \mathbb{R}^{\bar{N}\times\bar{F}\times\bar{T}};$$
- global average pooling (GAP) layer
$$m^{\text{GAP}} : \mathbb{R}^{\bar{N}\times\bar{F}\times\bar{T}} \to \mathbb{R}^{\bar{N}};$$
- classifier based on a multi-layer perception (MLP)
$$m^{\text{MLP}} : \mathbb{R}^{\bar{N}} \to \mathbb{R}.$$

In the seizure detection task, Grad-CAM provides a time–frequency importance map, which is essentially

a linear combination of high-level features, weighted using information about the gradient:

$$\mathrm{FTI} = \mathrm{ReLU}\left(\sum_{1}^{\bar{N}} \left\{ \frac{1}{\bar{F}\bar{T}} \sum_{ij} \frac{\partial S}{\partial \mathcal{F}_k^{ij}} \right\} \mathcal{F}_k \right), \quad (13)$$

where $\mathcal{F} = m^{\mathrm{FE}}(G(x))$ are the high-level features extracted by the trained model, $S = m^{\mathrm{MLP}}(m^{\mathrm{GAP}}(\mathcal{F}))$ is the prediction score that represents the model's level of confidence in the presence of epileptic activity within segment x.

Using the Grad-CAM saliency map, the importance of the frequency range $[f_0, f_1]$ is defined as

$$\mathrm{FI}_{[f_0,f_1]} = \frac{1}{(f_1 - f_0)T} \int_{f_0}^{f_1} \int_{0}^{T} \mathrm{FTI}^{\mathrm{US}}(\hat{f}, \tau)\, d\tau df, \quad (14)$$

where $\mathrm{FTI}^{\mathrm{US}}$ is the saliency map from Eq. (13) upscaled to match input sample size.

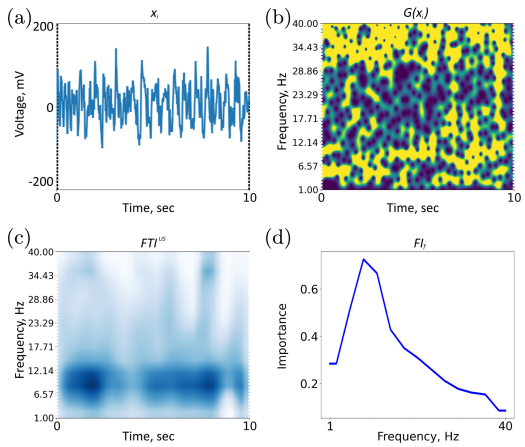

Fig. 2. Necessary steps to construct the importance of the frequency range $FI_{[f_0,f_1]}$: raw EEG signal, x_i (a); time–frequency representation $G(x_i)$ (b); time–frequency importance map *FTI* obtained with Grad-CAM approach and upscaled to match the size of $G(x_i)$ (c); frequency importance FI_f (d).

Figure 2 illustrates intermediate steps of building frequency range importance. Raw signal (Fig. 2(a)) is represented in the time–frequency domain (Fig. 2(b)) and then is fed to the model. Grad-CAM provides a time–frequency importance map FTI which is upscaled (Fig. 2(c)) and averaged over the time dimension to get frequency importance FI_f (Fig. 2(d)). Finally, by integrating the obtained curve in $[f_0, f_1]$, the proposed frequency range importance $\mathrm{FI}_{[f_0,f_1]}$ can be calculated.

4.3. *Frequency–spatial interpretation*

Frequency–spatial importance score is defined as the product of region importance and frequency range importance, as given by Eqs. (10) and (14):

$$\mathrm{FRI}^{R}_{[f_0,f_1]} = \mathrm{FI}_{[f_0,f_1]} \cdot \mathrm{RI}^{R}. \quad (15)$$

The proposed interpretation method is motivated by the principles of human pattern detection: clinicians often identify increased activity in specific frequency ranges and/or EEG channels.

Considering these factors, five common frequency bands are selected. They are frequently used by epileptologists for seizure detection and denoted by $\mathrm{Bands} = \{\delta, \theta, \alpha, \beta, \gamma\}$:

- $\delta = [1, 4]$ Hz,
- $\theta = [4, 8]$ Hz,
- $\alpha = [8, 14]$ Hz,
- $\beta = [14, 30]$ Hz,
- $\gamma = [30, 40]$ Hz.

Additionally, EEG electrodes are divided into five groups, $\mathrm{Regions} = \{F, \mathrm{TL}, \mathrm{TR}, C, \mathrm{PO}\}$, corresponding to distinct brain regions (see Fig. 3(a)):

- frontal: $F = \{Fp_1, Fp_2, F_9, F_7, F_3, F_z, F_4, F_8, F_{10}\}$,
- left temporal: $\mathrm{TL} = \{T_9, T_7, P_9, P_7\}$,
- right temporal: $\mathrm{TR} = \{T_{10}, T_8, P_{10}, P_8\}$,

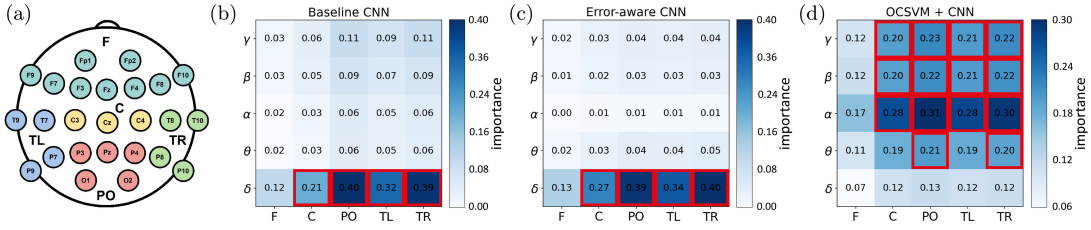

Fig. 3. (Color online) Results of global interpretation. EEG electrode locations and division of electrodes into regions, each indicated by a different color (a). Global interpretation matrices for Baseline CNN (b), Error-aware CNN (c), and OCSVM +CNN (d) models. Significant frequency–spatial features are shown with red frames.

- central: $C = \{C_3, C_z, C_4\}$,
- parietal-occipital: $\mathrm{PO} = \{P_3, P_z, P_4, O_1, O_2\}$.

Using these frequency bands and regions, $\mathrm{FRI}^{R}_{[f_0,f_1]}$ for each pair of elements is calculated from sets of Bands and Regions. The results are visualized as 5×5 heatmap matrices, enabling local interpretation of specific samples.

A global interpretation of the model's behavior can be obtained by averaging $\mathrm{FRI}^{R}_{[f_0,f_1]}$ over all segments identified by the model as containing epileptic activity. This average yields a heatmap representing the overall model behavior:

$$\mathrm{FRI}^{R,\mathrm{Global}}_{[f_0,f_1]} = \frac{1}{|\mathcal{X}^+|} \sum_{x \in \mathcal{X}^+} \mathrm{FRI}^{R}_{[f_0,f_1]}(x), \qquad (16)$$

where $\mathcal{X}^+$ represents the set of segments containing epileptic activity as predicted by the model.

5. Results

The proposed interpretation method is applied to the three models under consideration: Baseline CNN, Error-aware CNN, and OCSVM+CNN. Table 2 presents the key seizure detection performance metrics for each model: mean values across the tested patients for precision, recall, $F1$-score, and the total number of false negatives (FN), false positives (FP), and true positives (TP).

Table 3 shows the results of the statistical comparison of detection metrics between models. It contains p-values for the main effect and for the pairwise comparisons between conditions. The results for the latter are presented both without (the first value in a cell) and with (the second value in a cell) correction to multiple comparisons. Significant effects are shown with asterisk. For a more comprehensive analysis and detailed definitions of all metrics, please refer to Refs. 13, 38, and 40.

The resulting importance values, representing the contribution of each frequency–spatial characteristic to the model's decision-making process, are visualized in Figs. 3 and 4. The higher importance of a spatial–frequency feature within a single interpretation matrix for a certain model suggests that the neurophysiological processes occurring in this brain region and frequency range are more utilized by the model during seizure detection. These processes can be interpreted as being more effective at distinguishing epileptic seizures from normal EEG activity and are thus more closely connected to epileptic activity in the brain.

Important features were determined through the statistical testing (see Sec. 3.4) and are shown on Fig. 3 with red frames.

Both global and local interpretations have two primary applications. First, during model development for diagnostic purposes, since interpretations can be used to verify that the model is capturing meaningful physiological features. In this scenario, concordance between the model's features and known biomarkers enhances the explainability of ML methods. Second, during model analysis for exploratory purposes, since interpretations can be employed to uncover features characterizing particularly effective models. Such identified features may reveal previously unknown biomarkers.

All results were obtained using the following hardware: Win11; AMD Ryzen 9 8945HX; NVIDIA RTX 5070Ti Laptop; 32Gb RAM. The times achieved in seizure detection part were ∼12 h for training and 24+ h (all 67 patients) for inference. The time required for a frequency–spatial interpretation was 1.38 ± 0.04 s for a single 10-s EEG fragment and 35.43 ± 0.14 s for an entire seizure (a representative example consisting of 24 segments).

Table 2. Seizure detection results for the considered models

Model	prec	rec	$F1$	FN	FP	TP
Baseline	0.13	0.96	0.23	2	336	49
Error-aware	0.54	0.86	0.66	7	38	44
Two-stage	0.57	0.84	0.68	8	32	43

Table 3. Results of statistical analysis for the detection metrics. Significant effects are shown with asterisk. Abbreviations: B — Baseline, E — Error-aware, T — Two-stage.

Effect	prec	rec	$F1$
Main	***0.0011**	0.1408	***0.0015**
B versus E	***2.1595e−04**	—	***3.6532e−04**
	***6.4786e−04**		***0.0011**
B versus T	***0.0172**	—	***0.0218**
	0.0515		0.0654
E versus T	0.3354	—	0.3107
	1.0000		0.9321

Note: $^*p < 0.05$.

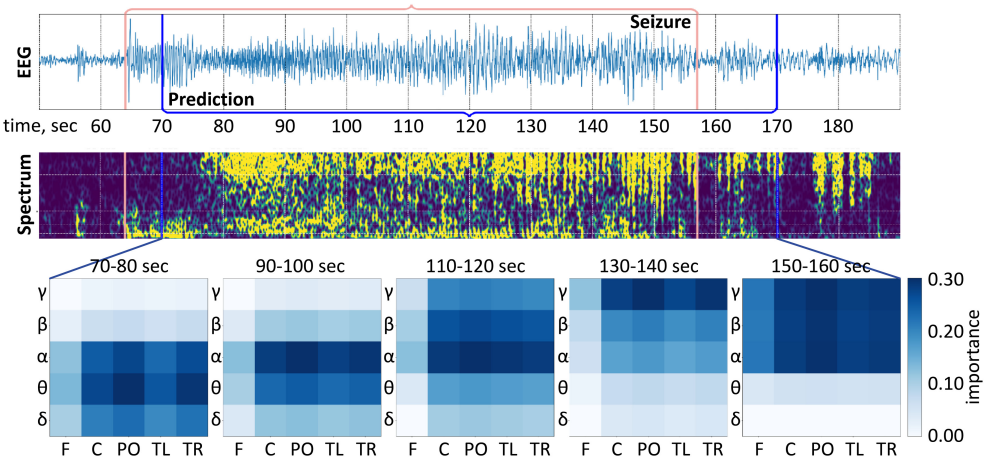

Fig. 4. (Color online) Example of local interpretation of a single seizure predicted by the OCSVM+CNN approach. Blue frame denotes prediction borders, while beige frame denotes the borders of real seizure. For compactness, the interpretation matrices are only visualized for a few 10-s segments.

5.1. *Global interpretation*

5.1.1. *Baseline CNN*

Figure 3(b) shows the Baseline CNN's predominant focus on the δ-band, with highest importance values across all spatial regions, particularly in occipital-parietal and temporal areas. While other bands (e.g. γ) show nonzero importance values, their contribution remains nonsignificant compared to the δ-band dominance.

The aforementioned results, in conjunction with the data segment selection procedure used for training, as described in Eq. (3), provide insights into the Baseline CNN's modest performance (see Tables 2 and 3). As discussed in the previous studies,[13,46] a major limitation of many ML-based seizure detection methods is their low precision due to a high number of FPs. Indeed, as shown in Table 2, there are significantly more FPs than TPs — 336 versus 49. However, collectively, FPs and TPs constitute only a small fraction of the overall dataset (approximately 2%). During Baseline CNN training, data segments from different classes are randomly selected in equal proportions without any specific strategy. This results in a relatively low probability of selecting a data segment containing potential FPs. Consequently, the model frequently commits errors when encountering such data during testing. This observation supports the hypothesis that modifications to the framework are necessary to achieve effective solutions, as implemented in the other models.

In this context, it becomes apparent why the Baseline CNN primarily relies on the δ-band while also attempting to extract relevant information from other bands. The primary frequency component of epileptic seizures lies within the δ-band, which serves as a reasonably reliable discriminator between seizures and normal EEG activity. However, other frequency bands can contribute to more refined seizure detection, as discussed upon in Sec. 6. The data segments used for training the Baseline CNN may lack the necessary diversity to encompass the full spectrum of epileptic activity manifestations. This limitation may impede the model's generalization capability due to the randomized nature of the sampling strategy. Consequently, the model fails to establish robust and reliable patterns across all available frequency bands.

5.1.2. *Error-aware CNN*

Figure 3(c) demonstrates that the Error-aware CNN shows near-zero importance for non-δ bands. This indicates that this model focuses predominantly on the δ-band, even more so than the Baseline CNN. This leads to improved performance, as evidenced in Tables 2 and 3. The hypothesis is that incorporating error awareness into the learning process has enabled the model to selectively prioritize the most salient frequency component, potentially enhancing robustness in noisy or ambiguous data scenarios.

Initially, it may seem paradoxical that the Error-aware CNN, despite its increased complexity compared to the Baseline CNN, utilizes fewer frequency bands. As shown in Tables 2 and 3, the Error-aware CNN surpasses the Baseline CNN in all key performance metrics except recall. Cascade approaches commonly exhibit a trade-off between increased precision and reduced recall. Conversely, the higher recall of the Baseline CNN may be attributed to its utilization (or attempted utilization) of multiple frequency bands. Therefore, frequency bands other than the δ-band may indeed contribute important information regarding epileptic seizures. However, the capacity of the CNN model under consideration may not be adequately leveraged to effectively integrate all frequency bands. By strategically selecting complex training examples, the Error-aware CNN was able to maximize the extraction of relevant information within a fixed number of iterations, and dedicate the model's full capacity to discerning subtle differences within the δ-band exclusively. The obtained results further reinforce the conclusion that the δ-band harbors critical biomarkers of epileptic seizures.

5.1.3. *OCSVM+CNN*

Figure 3(d) illustrates that the two-stage model exhibits minimal reliance on the δ-band. Instead, all other frequency bands demonstrate statistically high feature importance values, with the α-band exhibiting the highest. Simultaneously, the lowest feature importance values across all frequency bands are concentrated in the frontal brain region.

The observed importance of frequency bands can be explained by the design of the two-stage model (see Sec. 3.3.3). The OCSVM employed in the first stage is designed to identify trivial examples, readily distinguished by the dominant component of epileptic seizures in the δ-band (refer to Sec. 6). While OCSVM alone exhibits low precision but high recall,[32] it effectively filters the majority of the unbalanced dataset, isolating potential seizure segments. Consequently, the CNN implemented in the second stage processes data segments already characterized by pronounced activity in the δ-band. It is therefore logical that the CNN would de-emphasize the δ-band and focus on other frequency bands, as confirmed by the results. Given that the model is relieved of the burden of processing the δ-band, it can dedicate its full capacity to discerning subtle differences in other frequency bands.

According to Table 3, the two-stage model surpasses the Baseline CNN in precision and $F1$-score (comparison without the correction to multiple comparisons). At the same time, there are no significant differences between the two-stage model and the Error-aware CNN. On first glance, this suggests that these two models are identical in their seizure detection performance. However, Table 2 can provide additional insights. Although the two-stage model's recall is slightly lower than the error-aware CNN's, examining the number of TPs and FNs reveals that this difference stems from a single missed seizure. Consequently, these variations in recall can be attributed to random fluctuations. Notably, the two-stage model exhibits approximately 15% fewer FPs, and as we stated in our earlier work,[13] reducing FPs is the main focus in improving seizure detection methods. Considering this, as well as the two-stage model's tendency to have higher precision and $F1$-score, it can be seen as the most effective model.

Based on these observations and the insights gained from feature importance analysis, it can be concluded that the θ-, α-, β-, and γ-bands all play crucial roles in distinguishing true seizures from other patterns exhibiting high δ-band activity.

The second observation regarding the least important brain region (frontal) holds true across all three models. The frontal region appears to be the least influential, despite its substantial size (see Fig. 3(a)). This is an intriguing point that will be further explored in Sec. 6.

5.2. *Local interpretation*

In this study, local interpretation is employed to analyze individual seizure predictions generated by the top-performing model: OCSVM+CNN. This model not only exhibits high seizure detection performance but also engages with a diverse set of frequency–spatial features, as demonstrated in the global interpretation results. By examining the temporal evolution of feature importance during a seizure, the insights into the internal structure and progression of seizure events can be gained.

A representative example is shown in Fig. 4, which illustrates the evolution of importance values

for frequency–spatial features throughout a single seizure episode. Figure 4 reveals that epileptic seizures exhibit a complex and dynamic structure, characterized by the dominance of different frequency components at varying stages. Specifically, for the particular seizure from Fig. 4:

(i) at the seizure onset, activity in the θ-band is most important;
(ii) as the seizure progresses, the α- and β-bands gain increased importance;
(iii) toward the seizure termination, the γ-band becomes most important.

This example of progression suggests a potential temporal scenario for seizure evolution. While this pattern probably varies between patients, epilepsy types, or even seizures within a patient, the key takeaway is that fine-grained temporal dynamics in EEG signals can reveal intra-seizure variation and contribute to seizure detection. In clinical or research settings, this level of detail may aid in characterizing diverse seizure types or identifying patient-specific biomarkers. Ultimately, local interpretation complements global analysis by providing a window into the dynamic unfolding of seizures over time, based on physiologically interpretable features.

6. Discussion

The results of the proposed interpretation method are discussed below in the context of both existing and emerging domain knowledge.

6.1. *Brain regions and epileptic foci*

The analysis of important brain regions revealed a low importance assigned to the frontal region. Given that spatial features are typically localized in areas of epileptic foci,[48,49] this result is consistent with the fact that the patients in the present dataset had epileptic foci primarily in the temporal and occipital regions (see Sec. 3.1). While frontal lobes may play a crucial role in seizure detection for frontal lobe epilepsy,[50] their relative unimportance in the present study supports the validity of the model. This exemplifies how interpretation can be used to verify model behavior; alignment with existing domain knowledge regarding expected feature locations suggests that the model is indeed capturing meaningful physiological features.

6.2. *Frequency bands and biomarkers*

The analysis of important frequency bands demonstrated that multiple bands (δ, θ, α, β, γ) can contribute to seizure detection, and that integrating information from all these bands leads to superior performance, as seen in the OCSVM+CNN model. This presents an opportunity to delve deeper into the features of this successful model to uncover both established and novel biomarkers of seizures.

6.3. *Neurophysiological processes and feature importances*

The observed feature importances across the models align with specific neurophysiological processes. Some of these processes are commonly linked to seizure initiation and propagation, while others may have a more complex relationship with epilepsy.

Activity in the δ-band (1–4 Hz) is frequently associated with a hypersynchronous state of the cerebral cortex,[51,52] becoming prominently visible on most EEG channels during a seizure. The presence of such characteristic patterns on EEG during epileptic seizures[51] may lead one to reduce seizure detection to identifying EEG patterns with a core rhythm of 1–4 Hz. Indeed, two of the three models support this notion, as their important features are found almost exclusively in the δ-band. However, seizure patterns are typically far more complex, with potential changes in their main frequency over time,[53] as evidenced by the results from local interpretation. Therefore, considering other frequency bands is crucial.

The θ-band (4–8 Hz) is second important frequency range. The θ-rhythm, particularly in the occipital region of the cortex, is often linked to activity of the thalamocortical neural network. Enhanced EEG power in the θ-band and significant θ-coherence between EEG and local field potentials in the thalamus have been observed in patients with neurogenic pain, movement disorders, and epilepsy.[54] This suggests that enhanced θ-rhythmicity occurs within tight functional thalamocortical loops, and that this network plays an important role in a number of disorders, including epilepsy.[55] Thus, important

features identified in the θ-band may reflect underlying mechanisms of seizure generation and propagation.

Another important combination to consider is the α and β frequency bands (8–30 Hz). Changes within this combined frequency range in the occipital and temporal regions can be associated with the activation/suppression of the default mode network (DMN). The DMN is known to be consistently active during the resting state and deactivated during task engagement.[56] Notably, this same network is selectively impaired during epileptic seizures associated with altered states or loss of consciousness.[57] Although the specific mechanisms of seizure onset and propagation vary considerably between seizure types, the resulting loss of consciousness is consistent due to active inhibition of subcortical arousal systems that normally maintain DMN activity in the awake state.[57] This DMN-related activity can represent a valuable biomarker for seizure detection primarily captured by the best-performing model — OCSVM+CNN.

Finally, the γ-band (30–40 Hz and above) should also be considered. This band is commonly associated with motor activity EEG patterns that emerge during the convulsive phases of a seizure. Such patterns are typically considered an obstacle to seizure detection rather than a biomarker. However, some studies have demonstrated that, with the appropriate approach and methodology, motor activity can be leveraged for seizure detection. Luca *et al.*[58] detected seizures in acceleration data collected by 3D acceleration sensors, achieving 80% sensitivity and 89% precision. The fact that two of the three models consider the γ-band further supports the validity of motor activity as a potential biomarker.

7. Limitations and Future Research

The proposed interpretability framework, while providing valuable insights into CNN-based seizure detection, has several limitations. The method is currently limited to CNN models processing time–frequency EEG representations, excluding raw EEG data or other modalities like magnetoencephalography (MEG) or functional magnetic resonance imaging (fMRI). The fMRI data and wavelet spectra of MEG are similar in presentation to the wavelet spectra of EEG used in the framework. Thus, transitioning to these modalities appears feasible after extensive, albeit not conceptual, changes to the preprocessing, seizure detection, and feature interpretation pipelines. However, transitioning to raw EEG data is much more challenging since its structure is quite different from wavelet spectra, requiring a complete reimagining of the frequency-domain interpretation pipeline.

ResNet-18, the architecture used in seizure detection is not the state-of-the-art for neural networks. Future work should explore extensions to more advanced architectures such as transformers. The proposed interpretability framework is compatible with these architectures and can account for their attention mechanisms. However, implementing the large state-of-the-art architecture would require advanced training techniques, such as SSL,[59] or MAE[60] which can be challenging to set up and debug. Furthermore, a large quantity of "images" (EEG spectra) would be required for training, which could pose an additional challenge in the seizure detection task, where data volume is limited due to the natural rarity of epileptic events.

Spatial importance analysis depends critically on the choice of baseline signal for occlusion, whether patient-specific, population-averaged, or synthetic. A systematic investigation of baseline selection strategies in the future could enhance robustness.

The computational overhead of $(N + 1)$ forward passes for N EEG channels presents another challenge, potentially addressed through parallelization or gradient-based approximations. In general, optimization techniques in future research may include the following: caching the power spectrum after CWT, approximating CWT with simpler alternatives, parallelizing the channel importance calculation, and reducing the number of EEG channels that represent the region of interest. Solving the problem of high computational costs would also make sensitivity analysis for leave-one-patient-out possible.

The framework currently lacks explicit mapping to neurophysiological concepts (e.g. thalamocortical θ-synchronization), unlike methods like TCAV.[61] Future adaptations could incorporate EEG-specific concepts as annotated spectral patterns.[62]

The method also does not provide metrics for interpretability quality; instead, it relies on the

neurophysiological alignment of the features. Properly calculating these metrics requires the existence of "ground truth" (GT) for the dataset, i.e. information about the "correct interpretation". Our dataset, like most others, lacks these for a reason. Obtaining GT for important frequency ranges and brain regions suggests detailed analysis of EEG for each seizure, which can only be performed by a highly trained professional in the field of epilepsy diagnostics. This task is beyond the scope of common epilepsy diagnostics and is rarely addressed by clinicians. Creating GT of frequency–spatial features for an epileptic dataset is challenging but could be attempted in the future. In fact, interpretability methods like ours could greatly assist human experts with this task.

Validation has been primarily limited to focal epilepsy with temporal/occipital foci, and generalization to other epilepsy types remains to be established. The CNN model detects seizures by learning frequency–spatial features from EEG, and the interpretability method reveals these features. Therefore, this approach should be effective in various clinical contexts, such as different seizure types and patient ages, as long as the EEG signal contains patterns specific to epileptic seizures. However, further testing is required. Such task is closely related to validating on widely known and used datasets, such as those from Temple University Hospital (TUH) or Children's Hospital Boston–Massachusetts Institute of Technology (CHB–MIT). However, there are two major obstacles to such a task. On one hand, each dataset has a distinct structure that includes the number and placement of EEG channels, the sampling rate, the approach to data segmentation and labeling, etc. For example, the CHB–MIT dataset presents its EEG channels in a bipolar montage, while the dataset in this study uses a monopolar montage. Details like this require major alterations to the data processing, seizure detection, and feature interpretation pipelines. On the other hand, analyzing any sizable dataset is time-consuming because it requires retraining and revalidating three CNN-based models, which are then analyzed for feature importance. The bottleneck of this procedure is the feature extraction process, which uses CWT — the limitation of the seizure detection approach. Thus, extending the developed approach to other datasets and seizure types is a large-scale task for future research.

Local interpretation has been demonstrated only as a proof of concept. Additional research is required to study any possible generalization of temporal pattern of feature importance evolution across patients or seizure types.

Finally, the approach treats frequency bands and spatial regions independently, neglecting potential cross-band coupling (e.g. theta-gamma phase-amplitude coupling) or inter-regional synchrony. Graph-based representations could capture these dynamic features in future implementations. For example, Pitsik *et al.*[63] demonstrated that a hypergraph representation of a multilayer brain network helps account for cross-frequency interactions in EEG signals. This, in turn, enhances the detection of autism spectrum disorder.

8. Conclusion

This study presented a novel interpretability framework for CNN-based seizure detection, combining frequency-domain and spatial-domain analyses to uncover physiologically meaningful EEG biomarkers. Validated across three distinct architectures the method demonstrated that optimal performance arises from complementary use of multiple frequency bands and brain regions rather than reliance on single biomarkers. Key contributions include:

- *Task-Specific Design.* The integrated frequency–spatial interpretation revealed biomarkers aligning with known neurophysiological mechanisms (e.g. thalamocortical θ-band interactions, DMN suppression in α/β-bands) while suggesting new candidates for seizure detection.
- *Cross-Model Insights.* Global interpretation showed that the highest-performing model (OCSVM+CNN) leveraged multi-band biomarkers, whereas less robust models (Baseline/Error-aware CNN) over-relied on δ-band features. Local interpretation further captured dynamic intra-seizure spectral shifts, highlighting the method's clinical utility for analyzing seizure evolution.
- *Clinical Alignment.* The framework bridged DL with clinical EEG analysis by providing:

(i) model validation through neurophysiologically plausible explanations;

(ii) discovery of potential biomarkers (e.g. γ-band motor activity);
(iii) tools for comparing model's behavior across architectures.

According to the results, the two-stage OCSVM +CNN model is the best candidate to be used in real clinical practice. It provides the best tradeoff between performance and latency, and also captures more important biomarkers associated with epileptic seizures than the other two models.

Interpretability findings suggest that the next-generation seizure detection algorithms could benefit from incorporating temporal state models, sequence-aware architectures, or stage-specific detectors that reflect seizure onset, propagation, and termination. Additionally, architectures should explicitly target different frequency bands, which can be implemented as band-specific subnetworks alongside cross-band interaction modules, for example.

Despite limitations in computational efficiency and input specificity, this work advances interpretable DL approach for epilepsy by demonstrating how model decisions can be mapped to domain knowledge. Future directions include extending the method to raw EEG models, incorporating dynamic feature interactions, and validating clinical utility in surgical planning. The proposed framework not only enhances trust in CNN-based seizure detection but also opens avenues for collaborative biomarker discovery between DL and neuroscience.

Code Availability

All code for the proposed method has been publicly available at *https://github.com/snazau/seizure-detection-cnn-interpretability*.

Acknowledgments

This research was supported by the Russian Science Foundation, Grant No. 23-71-30010.

ORCID

Vadim V. Grubov https://orcid.org/0000-0003-2491-2592
Oleg E. Karpov https://orcid.org/0000-0002-5227-0657
Sergei I. Nazarikov https://orcid.org/0000-0002-7056-3373
Semen A. Kurkin https://orcid.org/0000-0002-3438-5717
Nikita P. Utyashev https://orcid.org/0000-0002-0770-2983
Denis A. Andrikov https://orcid.org/0000-0003-0359-0897
Alexander E. Hramov https://orcid.org/0000-0003-2787-2530

References

1. R. D. Thijs, R. Surges, T. J. O'Brien and J. W. Sander, Epilepsy in adults, *Lancet* **393**(10172) (2019) 689–701.
2. R. Cooper, J. W. Osselton and J. C. Shaw, *EEG Technology* (Butterworth-Heinemann, 2014).
3. U. Raghavendra, U. R. Acharya and H. Adeli, Artificial intelligence techniques for automated diagnosis of neurological disorders, *Eur. Neurol.* **82**(1–3) (2020) 41–64.
4. U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan and H. Adeli, Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals, *Comput. Biol. Med.* **100** (2018) 270–278.
5. L.-C. Lin, C.-S. Ouyang, R.-C. Wu, R.-C. Yang and C.-T. Chiang, Alternative diagnosis of epilepsy in children without epileptiform discharges using deep convolutional neural networks, *Int. J. Neural Syst.* **30**(5) (2020) 1850060.
6. Y. Zhang, H. Feng, S. Wang, H. Lv, T. Xiao, Z. Wang and Y. Zhao, Tiny convolutional neural network with supervised contrastive learning for epileptic seizure prediction, *Int. J. Neural Syst.* **35**(7) (2025) 2550034.
7. X. Dong, Y. Wen, D. Ji, S. Yuan, Z. Liu, W. Shang and W. Zhou, Epileptic seizure detection with an end-to-end temporal convolutional network and bidirectional long short-term memory model, *Int. J. Neural Syst.* **34**(3) (2024) 2450012.
8. H. S. Nogay and H. Adeli, Detection of epileptic seizure using pretrained deep convolutional neural network and transfer learning, *Eur. Neurol.* **83**(6) (2021) 602–614.
9. Y. Wang, S. Yuan, J.-X. Liu, W. Hu, Q. Jia and F. Xu, Combining EEG features and convolutional autoencoder for neonatal seizure detection, *Int. J. Neural Syst.* **34**(8) (2024) 2450040.
10. J. Wang, H. Li, C. Li, W. Lu, H. Cui, X. Zhong, S. Ren, Z. Shang and W. Zhou, Efficient seizure detection by complementary integration of convolutional neural network and vision transformer, *Int. J. Neural Syst.* **35**(7) (2025) 2550023.
11. N. Hallowell, S. Badger, A. Sauerbrei, C. Nellåker and A. Kerasidou, "I don't think people are ready to trust

these algorithms at face value": Trust and the use of machine learning algorithms in the diagnosis of rare disease, *BMC Med. Ethics* **23**(1) (2022) 112.

12. F. C. Morabito, C. Ieracitano and N. Mammone, An explainable artificial intelligence approach to study MCI to AD conversion via HD-EEG processing, *Clin. EEG Neurosci.* **54**(1) (2023) 51–60.
13. V. V. Grubov, S. I. Nazarikov, S. A. Kurkin, N. P. Utyashev, D. A. Andrikov, O. E. Karpov and A. E. Hramov, Two-stage approach with combination of outlier detection method and deep learning enhances automatic epileptic seizure detection, *IEEE Access* **12** (2024) 122168–122182.
14. A. Vilamala, K. H. Madsen and L. K. Hansen, Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring, in *2017 IEEE 27th Int. Workshop on Machine Learning for Signal Processing (MLSP)* (IEEE, 2017), pp. 1–6.
15. J. Cui, Z. Lan, O. Sourina and W. Müller-Wittig, EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network, *IEEE Trans. Neural Netw. Learn. Syst.* **34** (10) (2022) 7921–7933.
16. Y. Zhao, L. Cao, Y. Ji, B. Wang and W. Wu, Interpretable EEG emotion classification via CNN model and gradient-weighted class activation mapping, *Brain Sci.* **15**(8) (2025) 886.
17. X. Zhao, N. Yoshida, T. Ueda, H. Sugano and T. Tanaka, Epileptic seizure detection by using interpretable machine learning models, *J. Neural Eng.* **20**(1) (2023) 015002.
18. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
19. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *Proc. IEEE Int. Conf. Computer Vision* (IEEE, 2017), pp. 618–626.
20. V. Gabeff, T. Teijeiro, M. Zapater, L. Cammoun, S. Rheims, P. Ryvlin and D. Atienza, Interpreting deep learning models for epileptic seizure detection on EEG signals, *Artif. Intell. Med.* **117** (2021) 102084.
21. J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer and N. D. Lawrence, *Dataset Shift in Machine Learning* (MIT Press, 2022).
22. T. Tuncer and S. Dogan, An explainable EEG epilepsy detection model using friend pattern, *Sci. Rep.* **15**(1) (2025) 16951.
23. F. Doshi-Velez and B. Kim, Towards a rigorous science of interpretable machine learning, preprint (2017), arXiv:1702.08608.
24. Y. Zhang, P. Tino, A. Leonardis and K. Tang, A survey on neural network interpretability, *IEEE Trans. Emerg. Top. Comput. Intell.* **5** (2021) 726–742.
25. V. K. Harpale and V. K. Bairagi, Time and frequency domain analysis of EEG signals for seizure detection: A review, in *2016 Int. Conf. Microelectronics, Computing and Communications (MicroCom)* (IEEE, 2016), pp. 1–6.
26. F. Wang, Q. Su and C. Li, Identification of novel biomarkers in non-small cell lung cancer using machine learning, *Sci. Rep.* **12**(1) (2022) 16693.
27. C. Molnar, G. Casalicchio and B. Bischl, Interpretable machine learning — a brief history, state-of-the-art and challenges, in *ECML PKDD 2020 Workshops* (Springer International Publishing, Cham, 2020), pp. 417–431.
28. K. D. Tzimourta, A. T. Tzallas, N. Giannakeas, L. G. Astrakas, D. G. Tsalikakis, P. Angelidis and M. G. Tsipouras, A robust methodology for classification of epileptic seizures in EEG signals, *Health Technol.* **9** (2019) 135–142.
29. S. S. Spencer, P. Guimaraes, A. Katz, J. Kim and D. Spencer, Morphological patterns of seizures recorded intracranially, *Epilepsia* **33**(3) (1992) 537–545.
30. P. Sturmfels, S. Lundberg and S.-I. Lee, Visualizing the impact of feature attribution baselines, Distill (2020), https://distill.pub/2020/attribution-baselines.
31. J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt and B. Kim, Sanity checks for saliency maps, in *NIPS'18: Proc. 32nd Int. Conf. Neural Information Processing Systems* (ACM, 2018), pp. 9525–9536.
32. O. E. Karpov, V. V. Grubov, V. A. Maksimenko, S. A. Kurkin, N. M. Smirnov, N. P. Utyashev, D. A. Andrikov, N. N. Shusharina and A. E. Hramov, Extreme value theory inspires explainable machine learning approach for seizure detection, *Sci. Rep.* **12**(1) (2022) 11474.
33. D. Mantini, M. G. Perrucci, S. Cugini, A. Ferretti, G. L. Romani and C. Del Gratta, Complete artifact removal for EEG recorded during continuous fMRI using independent component analysis, *Neuroimage* **34**(2) (2007) 598–607.
34. R. Oostenveld, P. Fries, E. Maris and J.-M. Schoffelen, FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data, *Comput. Intell. Neurosci.* **2011** (2011) 1–9.
35. A. E. Hramov, A. A. Koronovskii, V. A. Makarov, V. A. Maksimenko, A. N. Pavlov and E. Sitnikova, *Wavelets in Neuroscience* (Springer Nature, 2021).
36. E. Sitnikova, A. E. Hramov, A. A. Koronovsky and G. Van Luijtelaar, Sleep spindles and spike–wave discharges in EEG: Their generic features, similarities and distinctions disclosed with Fourier transform and continuous wavelet analysis, *J. Neurosci. Methods* **180**(2) (2009) 304–316.
37. P. Thangavel *et al.*, Time–frequency decomposition of scalp electroencephalograms improves deep learning-based epilepsy diagnosis, *Int. J. Neural Syst.* **31**(8) (2021) 2150032.

38. V. Grubov, S. Nazarikov, N. Utyashev and O. E. Karpov, Error-aware CNN improves automatic epileptic seizure detection, *Eur. Phys. J. Spec. Top.* **234** (2025) 3871–3881.
39. L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu and L. Shao, Normalization techniques in training DNNs: Methodology, analysis and application, *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(8) (2023) 10173–10196.
40. S. I. Nazarikov, Mathematical model for epileptic seizures detection on an EEG recording, *Izv. VUZ. Appl. Nonlin. Dyn.* **31**(5) (2023) 628–642.
41. J. Hernandez, J. A. Carrasco-Ochoa and J. F. Martínez-Trinidad, An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congr., CIARP 2013*, Havana, Cuba, 20–23 November 2013, Proceedings, Part I 18 (Springer, 2013), pp. 262–269.
42. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk and Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, preprint (2019), arXiv:1904.08779.
43. A. Pisarchik, V. Grubov, V. Maksimenko, A. Lüttjohann, N. Frolov, C. Marqués-Pascual, D. Gonzalez-Nieto, M. Khramova and A. Hramov, Extreme events in epileptic EEG of rodents after ischemic stroke, *Eur. Phys. J. Spec. Top.* **227** (2018) 921–932.
44. N. S. Frolov, V. V. Grubov, V. A. Maksimenko, A. Lüttjohann, V. V. Makarov, A. N. Pavlov, E. Sitnikova, A. N. Pisarchik, J. Kurths and A. E. Hramov, Statistical properties and predictability of extreme epileptic events, *Sci. Rep.* **9**(1) (2019) 7243.
45. O. E. Karpov, V. V. Grubov, V. A. Maksimenko, N. Utaschev, V. E. Semerikov, D. A. Andrikov and A. E. Hramov, Noise amplification precedes extreme epileptic events on human EEG, *Phys. Rev. E* **103**(2) (2021) 022310.
46. O. E. Karpov, M. S. Khoymov, V. A. Maksimenko, V. V. Grubov, N. Utyashev, D. A. Andrikov, S. A. Kurkin and A. E. Hramov, Evaluation of unsupervised anomaly detection techniques in labelling epileptic seizures on human EEG, *Appl. Sci.* **13**(9) (2023) 5655.
47. M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, in *Computer Vision — ECCV 2014* (Springer International Publishing, 2014), pp. 818–833.
48. M. O. Baud, J. K. Kleen, G. K. Anumanchipalli, L. S. Hamilton, Y.-L. Tan, R. Knowlton and E. F. Chang, Unsupervised learning of spatiotemporal interictal discharges in focal epilepsy, *Neurosurgery* **83**(4) (2018) 683–691.
49. H. Daoud and M. Bayoumi, Deep learning approach for epileptic focus localization, *IEEE Trans. Biomed. Circuits Syst.* **14**(2) (2019) 209–220.
50. J. O'Muircheartaigh and M. P. Richardson, Epilepsy and the frontal lobes, *Cortex* **48**(2) (2012) 144–155.
51. U. Seneviratne, M. J. Cook and W. J. D'Souza, Electroencephalography in the diagnosis of genetic generalized epilepsy syndromes, *Front. Neurol.* **8** (2017) 499.
52. P. Jiruska, M. de Curtis, J. G. R. Jefferys, C. A. Schevon, S. J. Schiff and K. A. Schindler, Synchronization and desynchronization in epilepsy: Controversies and hypotheses, *J. Physiol.* **591** (2012) 787–797.
53. D. Bosnyakova, A. Gabova, A. Zharikova, V. Gnezditski, G. Kuznetsova and G. Van Luijtelaar, Some peculiarities of time–frequency dynamics of spike–wave discharges in humans and rats, *Clin. Neurophysiol.* **118**(8) (2007) 1736–1743.
54. J. Sarnthein, A. Morel, A. Von Stein and D. Jeanmonod, Thalamic theta field potentials and EEG: High thalamocortical coherence in patients with neurogenic pain, epilepsy and movement disorders, *Thalamus Relat. Syst.* **2**(3) (2003) 231–238.
55. B. E. Lindquist, C. Timbie, Y. Voskobiynyk and J. T. Paz, Thalamocortical circuits in generalized epilepsy: Pathophysiologic mechanisms and therapeutic targets, *Neurobiol. Dis.* **181** (2023) 106094.
56. M. F. Mason, M. I. Norton, J. D. Van Horn, D. M. Wegner, S. T. Grafton and C. N. Macrae, Wandering minds: The default network and stimulus-independent thought, *Science* **315**(5810) (2007) 393–395.
57. N. B. Danielson, J. N. Guo and H. Blumenfeld, The default mode network and altered consciousness in epilepsy, *Behav. Neurol.* **24**(1) (2011) 55–65.
58. S. Luca, P. Karsmakers, K. Cuppens, T. Croonenborghs, A. V. de Vel, B. Ceulemans, L. Lagae, S. V. Huffel and B. Vanrumste, Detecting rare events using extreme value statistics applied to epileptic convulsions in children, *Artif. Intell. Med.* **60**(2) (2014) 89–96.
59. R. Balestriero *et al.*, A cookbook of self-supervised learning, preprint (2003), arXiv:2304.12210.
60. K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, Masked autoencoders are scalable vision learners, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (IEEE, 2022), pp. 16000–16009.
61. B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas and R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in *Int. Conf. Machine Learning* 2017.
62. A. G. Madsen, W. T. Lehn-Schiøler, A. Jónsdóttir, B. Arnardóttir and L. K. Hansen, Concept-based explainability for an EEG transformer model, in *2023 IEEE 33rd Int. Workshop on Machine Learning for Signal Processing (MLSP)* (IEEE, 2023), pp. 1–6.
63. E. Pitsik, S. Kurkin, O. Martynova, G. Portnova and A. E. Hramov, Hypergraph representation of multilayer brain network enhances autism spectrum disorder detection, *Chaos* **35**(7) (2025) 071104.