mathematics

MDPI

*Article*

# Learning from Adaptive Oscillatory Modes: An Interpretable Synchrosqueezing Transformer for Patient-Independent Seizure Detection

**Afrah Alanazi [1], Kais Khaldi [2,*], Sarah Almaghrabi [3] and Hamed Alshammari [4]**

[1] Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia; aoalenzy@ju.edu.sa
[2] Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia
[3] Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah 23890, Saudi Arabia; ssalmaghrabi@uj.edu.sa
[4] Department of Software Engineering, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia; hhaalshammari@ju.edu.sa
* Correspondence: kkhaldi@ju.edu.sa

## Abstract

Transformer-based analysis of physiological signals critically depends on how input tokens are constructed, yet conventional tokens derived from raw temporal windows or spectrogram patches may overlook the intrinsic oscillatory organization of electroencephalographic (EEG) activity. To address this limitation, we propose AMST-Former, an adaptive mode-wise synchrosqueezing Transformer for patient-independent epileptic seizure detection. Each EEG channel is first decomposed into oscillatory modes using Adaptive Local Iterative Filtering (ALIF), after which the Synchrosqueezing Transform (SST) is applied independently to each mode. This decomposition-before-synchrosqueezing strategy reduces cross-component interference in instantaneous-frequency estimation while preserving mode-specific time–frequency dynamics. The resulting time–frequency patches are embedded together with channel, mode-frequency, and positional information. A hierarchical factorized-attention mechanism then models token relevance across time–frequency regions, oscillatory modes, and EEG channels before a Transformer encoder captures their interactions. Evaluation using a repeated patient-independent five-fold cross-validation protocol on the CHB-MIT database (22 subjects) yields an F1-score of 70.76% and an area under the precision–recall curve (AUPRC) of 0.870. At the event level, AMST-Former achieves a sensitivity of 81.4% at an average false detection rate of 1.39 false detections per 24 h. Ablation experiments show that mode-wise SST provides more effective representations than direct SST, per-mode Hilbert spectra, and fixed filter-bank decomposition. Synthetic experiments confirm the cross-component bias predicted by the analysis and shows that the SST of a mode remains accurate when separation is imperfect, whereas its Hilbert instantaneous frequency does not. Furthermore, the factorized-attention mechanism provides mode, time–frequency, and channel-level explanations, whose faithfulness is assessed through a deletion analysis. These results support adaptive mode-wise time–frequency tokenization as a promising strategy for interpretable, patient-independent EEG seizure detection.

Check for updates

Academic Editor: Iraklis Varlamis

Received: 29 September 2026
Revised: 22 October 2026
Accepted: 26 October 2026
Published: 27 October 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

**Keywords:** explainable artificial intelligence; representation learning; tokenization; transformer; hierarchical attention; electroencephalography; seizure detection; adaptive local iterative filtering; synchrosqueezing transform; patient-independent evaluation

**MSC:** 94A12; 68T07; 92C55; 42C40; 62H30

## 1. Introduction

Epilepsy affects more than 45 million people worldwide and remains one of the leading causes of neurological disability [1,2]. Scalp electroencephalography (EEG) is the reference modality for its diagnosis and monitoring, because it records the abnormal neuronal synchronization that underlies seizures with millisecond resolution and at low cost. In epilepsy monitoring units, however, recordings routinely span several days and dozens of channels, and their visual review by trained neurophysiologists is slow, costly and subject to inter-rater variability. Automated seizure detection is therefore a long-standing objective of clinical neuroengineering [3–5].

Two properties of the EEG make this objective difficult. First, the signal is nonstationary and multicomponent: an ictal discharge typically appears as a rhythmic oscillation whose frequency drifts over seconds, superimposed on background rhythms, artifacts and slow trends, and these components frequently overlap in time and frequency. Second, ictal patterns differ markedly from one patient to another. A detector that performs well on segments drawn at random from patients already seen during training may fail on a new patient, and random segment-level splits are known to inflate reported performance substantially [6]. The clinically meaningful question is thus how well a detector generalizes to unseen patients, preferably assessed with event-level criteria that reflect how alarms are used in practice [7].

The representation from which a detector learns largely determines how it copes with the first property. Fixed time–frequency (TF) transforms such as the short-time Fourier and wavelet transforms have been used extensively [8–10], but their resolution is bounded by the choice of window or mother wavelet. Data-driven decompositions avoid a predefined basis by splitting the signal into intrinsic mode functions (IMFs), each carrying a single locally defined oscillation. Empirical mode decomposition (EMD) [11] and its noise-assisted variants [12,13] opened this line of work but suffer from mode mixing and lack a solid mathematical foundation [14]. Iterative filtering replaces the heuristic sifting of EMD with convolution-based moving averages [15], and its adaptive extension, Adaptive Local Iterative Filtering (ALIF), lets the filter length follow the local time scale of the signal, which makes it well suited to oscillations whose frequency changes quickly [16,17]. A third line of work sharpens the TF plane itself. The synchrosqueezing transform (SST) reassigns wavelet coefficients from scale to instantaneous frequency, producing energy-concentrated ridges while remaining invertible [18,19], and it has been applied successfully to seizure detection [20–22].

These two families address complementary weaknesses, yet they have almost always been used as alternatives. The complementarity follows directly from how SST works. SST estimates, at every TF point, the instantaneous frequency of the oscillation that dominates the wavelet coefficient; its guarantees hold only when the components of the signal are sufficiently separated in frequency [18]. When two EEG rhythms fall within the bandwidth of the analysing wavelet, the estimate becomes a mixture that oscillates at their beat frequency, and energy is reassigned to frequencies that belong to neither rhythm. Decomposition-based detectors, conversely, feed IMFs to a classifier as one-dimensional signals [23,24], leaving the network to infer the TF evolution of each mode implicitly. Applying ALIF

first and SST afterwards to each mode separately resolves both issues: ALIF performs the inter-mode separation that SST requires, and SST supplies the intra-mode TF resolution that the raw IMFs lack. Cicone and Wu exploited the reverse coupling, using SST to estimate the instantaneous frequencies that steer ALIF [25]; to our knowledge, ALIF-conditioned SST has not been used as a learned representation for seizure detection.

A mode-wise representation is, however, large: each channel yields several TF maps, and not every mode, TF region or channel is informative for a given segment. This is where attention-based models are useful. Transformers [26,27] have become competitive for EEG decoding [28–33], and their effectiveness depends heavily on how the input tokens are defined. In current EEG Transformers, tokens are raw temporal windows, spectrogram patches or channel embeddings, none of which carries an explicit notion of oscillatory mode. From a machine-learning perspective, this is a tokenization problem: a Transformer can only exploit the structure that its tokens expose, and generic patching leaves the model to rediscover the oscillatory organization of the signal from limited and patient-specific data. Mode-wise SST offers a principled alternative, in which prior knowledge from signal processing shapes the tokens rather than the network. Each token can be tagged with the channel, the mode and the TF location from which it originates, and relevance can be learned separately along each of these axes before global interactions are modelled. Because these axes have a physical meaning, the learned relevance is also a natural vehicle for explanation, a requirement for clinical acceptance of machine-learning models [34,35].

Building on these observations, this paper introduces AMST-Former, an ALIF-guided mode-wise synchrosqueezing Transformer for patient-independent seizure detection. Its contributions are as follows.

- Mode-wise synchrosqueezed representation:ALIF decomposes each channel into IMFs, and SST is applied to each IMF separately. We show analytically why this ordering reduces the cross-component bias of the SST frequency estimator, and we quantify its effect on TF concentration. Synthetic experiments verify the predicted bias, delimit the conditions under which decomposition can remove it, and show that mode-wise SST degrades gracefully under imperfect separation.
- Signal-structured tokenization and hierarchical attention: TF patches are embedded with channel, TF-position and frequency-anchored mode embeddings, which remain meaningful even though ALIF yields a variable number of modes. A three-level attention module, whose mode level is informed by signal descriptors of mode quality (energy, instantaneous-frequency stability and spectral entropy), factorizes token relevance into TF, mode and channel terms and retains the most relevant tokens for a Transformer encoder, which divides the encoder cost by about three at an accuracy comparable to that of a Vision Transformer trained on the same representation.
- Intrinsic, verified explanations: Because relevance is factorized along physically meaningful axes, each prediction is explained by the modes, TF regions and channels it relies on, without a post hoc surrogate model. The faithfulness of these explanations is tested quantitatively against integrated gradients and random rankings.
- Rigorous patient-independent evaluation: The model is trained and tested under repeated patient-level five-fold cross-validation, at both segment and event levels, and its variants are compared with a corrected resampled $t$-test. Controlled ablations separate the effect of adaptive decomposition from that of synchrosqueezing, including a Hilbert–Huang-type control and a fixed filter-bank control, and the evaluation is completed by cross-dataset, robustness and computational-cost analyses.

The remainder of the paper is organized as follows. Section 2 positions AMST-Former with respect to prior work. Section 3 describes the method, Section 4 the data and ex-

perimental protocol, and Section 5 the results. Section 6 discusses their implications and limitations, and Section 7 concludes the paper.

## 2. Related Work

Seizure detection methods can be organized by the representation they learn from and by the model that exploits it. This section reviews the three lines of work that AMST-Former combines, discusses how deep EEG models are explained, and then positions the proposed framework.

### *2.1. Adaptive Decomposition for Seizure Detection*

Decomposition-based detectors split each EEG channel into oscillatory modes and characterize them individually. Bajaj and Pachori showed that bandwidth features of EMD modes discriminated ictal from seizure-free EEG [23], and many subsequent works combined EMD, ensemble EMD, CEEMDAN or variational mode decomposition (VMD) [36] with entropy, spectral or statistical descriptors. In EEG cognitive-workload decoding, we showed that stabilizing the instantaneous frequency of EMD and CEEMDAN modes and normalizing their energies across modes yielded features that were more consistent across subjects and easier to relate to physiological rhythms [37]. Deep networks have since replaced handcrafted descriptors. The most closely related study is that of Zeng et al. [24], who fed ALIF-derived IMFs to a cascaded network combining squeeze-and-excitation channel attention with a CNN–LSTM classifier, and reported 100% accuracy under ten-fold cross-validation on the Bonn and a second public epilepsy database, with SHAP and Grad-CAM explanations. That work established ALIF modes as an informative input for seizure classification. Its learning stage, like that of most decomposition-based detectors, nevertheless operates on the modes as one-dimensional time series, and the evaluation relies on segment-level splits of small single-channel datasets. Two further practices are common in this literature: modes are selected by index (for example, the first three IMFs), and modes are recombined before classification. Both implicitly assume that a given IMF index corresponds to the same physiological band across segments and patients, an assumption that adaptive decompositions do not satisfy.

Decomposition and TF analysis have also been chained, but in ways that differ from the mode-wise SST studied here. In the Hilbert–Huang transform, each mode is described by a single instantaneous frequency and amplitude; Lu et al. stacked the Hilbert spectra of the EMD modes of CHB-MIT channels into images classified by a CNN [38]. Wu et al. used successive VMD to remove task-unrelated modes, recombined the remaining ones and computed short-time Fourier spectra that were fed to a Transformer for seizure prediction [39]. In the first case, the TF description of each mode is the Hilbert estimate, whose sensitivity to residual mixing is analysed in Section 3.4; in the second, decomposition acts as a denoising step and the identity of the modes is lost before TF analysis. Neither approach synchrosqueezes the modes separately or carries mode identity into the learned tokens.

### *2.2. Synchrosqueezed Time–Frequency Representations*

SST belongs to the family of reassignment methods [40] but, unlike classical reassignment, it moves coefficients only along the frequency axis and therefore allows the signal to be reconstructed [18,19]; higher-order variants further improve concentration for strongly modulated components [41]. Cura and Akan extracted joint TF moments and texture descriptors from SST maps of CHB-MIT recordings and classified them with conventional machine-learning models [20]. Ozdemir et al. replaced these descriptors with a CNN trained on Fourier-based SST images and reported segment-based accuracies above 99% on CHB-MIT [21]. Amiri et al. combined sparse common spatial patterns for channel selection

with an adaptive Fourier-based SST and a linear classifier, reaching 98.81% accuracy [22]. In all three studies, SST is computed on the recorded (or spatially filtered) EEG, that is, on a multicomponent signal for which the separation conditions of SST are not guaranteed.

### *2.3. Attention and Transformer Models for EEG*

Convolutional networks learn local temporal, spectral and spatial patterns from raw or transformed EEG [42–46], but their receptive fields limit the modelling of long-range dependencies. Self-attention relates every pair of tokens directly [26], and hybrid designs are now common. Li et al. guided a CNN with a Transformer to combine local and global information for seizure prediction [29]; Lian and Xu modelled inter-channel structure with a graph Transformer [30]; Song et al. coupled convolutional and self-attention blocks in EEG Conformer [28]. Two recent works address generalization across patients: Peng et al. distilled knowledge between raw EEG and multiscale wavelet branches for cross-subject seizure-subtype classification [31], and Xiao et al. pretrained an attention-based Transformer with contrastive self-supervision and reported both patient-specific and cross-patient results on CHB-MIT [32]. Adversarial feature learning has also been used to suppress patient-specific information [47]. A recent review identifies data scarcity and cross-subject generalization as the main open problems of EEG Transformers [33]. A common thread is that performance depends on how tokens are built; none of these models, however, organizes its tokens according to the oscillatory modes of the signal.

### *2.4. Explainability of Deep EEG Models*

Clinical adoption of seizure detectors requires that their decisions can be inspected by neurophysiologists [34]. Most deep EEG models are explained post hoc, with gradient-based saliency maps such as Grad-CAM [48], integrated gradients [49] or Shapley-value approximations such as SHAP [50]; Zeng et al., for instance, combined SHAP and Grad-CAM to explain their ALIF-based classifier [24]. Post hoc explanations are, however, approximations of the model, and it has been argued that high-stakes applications should favour models whose internal quantities are interpretable by design [35]. Attention weights are a tempting candidate, but whether they explain predictions is disputed [51,52], and their faithfulness must be verified, for example by perturbation analysis [53]. Moreover, when tokens are generic patches of raw EEG or spectrograms, even a faithful attention map indicates where the model looks but not which oscillation it relies on.

### *2.5. Positioning of AMST-Former*

Table 1 summarizes the design choices of representative methods. AMST-Former does not introduce ALIF, SST or the Transformer; its contribution lies in how they are coupled. SST is computed after, and separately for, each ALIF mode; the identity of every mode is preserved up to the Transformer through frequency-anchored embeddings; and relevance is factorized across TF regions, modes and channels. Decomposing a signal before analysing each mode in the TF plane is also the principle of the Hilbert–Huang transform [11]; the difference, discussed in Section 3.4, lies in how each mode is analysed, and it is tested experimentally. The central question is therefore not whether each component is useful on its own, which prior work has already shown, but whether preserving adaptively separated modes throughout high-resolution TF encoding adds information that is lost when decomposition and TF analysis are used as alternatives. Section 5.4 addresses this question directly.

Beyond the combination of three known operators, the distinction from existing decomposition-based TF representations lies in four elements, each of which is tested separately. (i) The order ALIF → SST is derived from the bias of the SST estimator (Proposition 1),

whose prediction and conditions of validity are verified on synthetic signals (Section 5.1). (ii) Each mode is synchrosqueezed separately and never recombined: reconstructing the signal from the ALIF modes before a single SST (configuration R5) loses the gain, and so does replacing SST by the Hilbert analysis of the same modes (configuration R4, a Hilbert–Huang-type representation comparable to that of [38]). (iii) Mode identity is carried into the tokens through frequency-anchored embeddings rather than IMF indices. (iv) Relevance is factorized along the TF, mode and channel axes. Finally, the advantage of AMST-Former over competing representations and architectures is established only by comparisons trained and tested under the same protocol (Sections 5.4–5.6); the published results of Section 5.3, obtained with different cohorts and protocols, are given for context and are not used to rank methods.

**Table 1.** Design choices of representative seizure detection methods and of the proposed AMST-Former. SE: squeeze-and-excitation; FSST: Fourier-based SST; sCSP: sparse common spatial patterns; SVMD: successive variational mode decomposition; STFT: short-time Fourier transform.

| Study | Decomposition | TF Representation | Learning Model | Attention/ Global Modelling | Representation Strategy |
|---|---|---|---|---|---|
| Cura and Akan [20] | None | SST | Classical ML | None | Handcrafted features from SST maps |
| Ozdemir et al. [21] | None | FSST | CNN | Local (convolutional) | SST images of the recorded EEG |
| Amiri et al. [22] | None | Adaptive FSST | Linear classifier | sCSP channel selection | TF features of selected channels |
| Zeng et al. [24] | ALIF | None | CNN–LSTM | SE channel attention | One-dimensional IMFs |
| Lu et al. [38] | EMD | Hilbert spectrum | CNN | None | Stacked Hilbert spectra of EMD modes |
| Wu et al. [39] | SVMD | STFT | Transformer | Self-attention | Spectra of the recombined, denoised signal |
| Li et al. [29] | None | STFT | CNN + Transformer | Self-attention | Joint local/ global features |
| Lian and Xu [30] | None | None | Graph Transformer | Graph self-attention | Inter-channel dependencies |
| Peng et al. [31] | Wavelet bands | Wavelet | Multi-branch Transformer | Self-attention + distillation | Raw and multiscale branches |
| AMST-Former | ALIF | Mode-wise SST | Transformer | TF, mode and channel attention | Mode-preserving TF tokens |

Abbreviations: ALIF, adaptive local iterative filtering; CNN, convolutional neural network; EEG, electroencephalography; EMD, empirical mode decomposition; FSST, Fourier-based synchrosqueezing transform; IMF, intrinsic mode function; LSTM, long short-term memory; ML, machine learning; sCSP, sparse common spatial patterns; SE, squeeze-and-excitation; SST, synchrosqueezing transform; STFT, short-time Fourier transform; SVMD, successive variational mode decomposition; TF, time–frequency.

## 3. AMST-Former: Proposed Approach

### *3.1. Problem Formulation and Overview*

Let $X = [x_1(t), \ldots, x_C(t)]^T \in R^{C \times N}$ denote a preprocessed EEG window with $C$ channels and $N$ samples, and let $y \in \{0, 1\}$ indicate whether the window is ictal. AMST-Former estimates $\hat{p} = P(y = 1 \mid X)$ through the five stages shown in Figure 1:

$$\hat{p} = \mathcal{F}_\theta\Big(\mathcal{A}_\theta\big(\mathcal{E}_\theta\big(\mathcal{T}_{\mathrm{SST}}\big(\mathcal{D}_{\mathrm{ALIF}}(\mathbf{X})\big)\big)\big)\Big) \tag{1}$$

where $\mathcal{D}_{\mathrm{ALIF}}$ decomposes every channel into IMFs and computes their quality descriptors (Sections 3.2 and 3.3), $\mathcal{T}_{\mathrm{SST}}$ synchrosqueezes each IMF separately (Section 3.4), $\mathcal{E}_\theta$ converts the resulting maps into structure-aware tokens (Section 3.5), $\mathcal{A}_\theta$ scores and selects tokens through hierarchical attention (Section 3.6), and $\mathcal{F}_\theta$ is a Transformer encoder followed by a classification head (Section 3.7). The first two operators are deterministic signal-processing transforms; all operators indexed by $\theta$ are learned jointly.

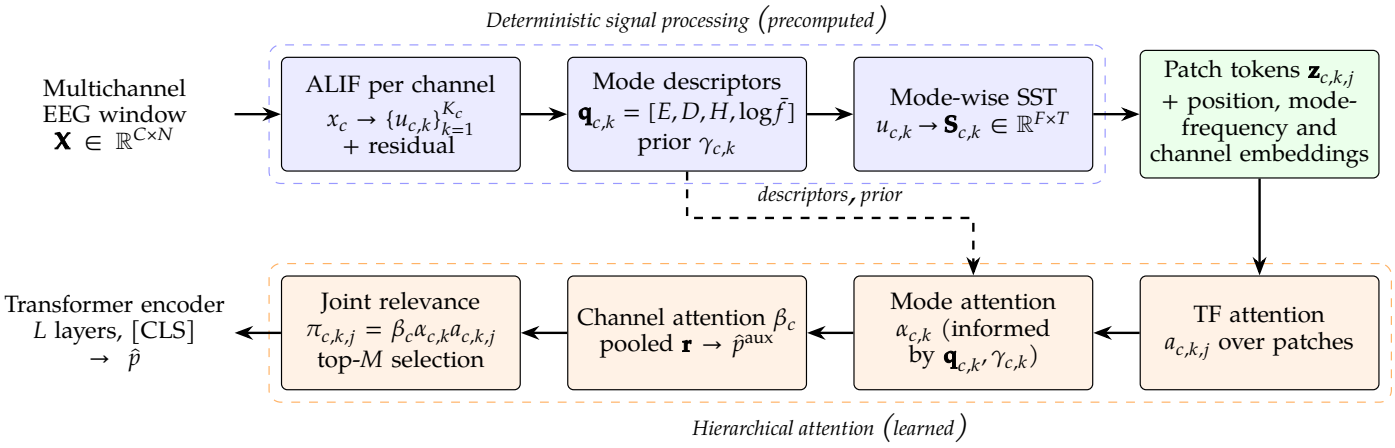



**Figure 1.** Overview of AMST-Former. Each channel is decomposed by ALIF; each intrinsic mode function is described by quality descriptors and synchrosqueezed separately. The mode-wise TF maps are split into patches whose embeddings encode their channel, mode frequency and TF position. Hierarchical attention factorizes token relevance over TF regions, modes and channels, the $M$ most relevant tokens are processed by a Transformer encoder, and a pooled auxiliary head supervises the attention module during training.

### *3.2. Adaptive Oscillatory Decomposition with ALIF*

ALIF [16] extracts IMFs by repeatedly subtracting a local moving average whose support adapts to the local time scale of the signal. For a signal $s(t)$, the moving average is:

$$\mathcal{L}_{\ell}[s](t) = \int_{-\ell(t)}^{\ell(t)} s(t+\tau)\, w_{\ell(t)}(\tau)\, d\tau, \qquad \int_{-\ell(t)}^{\ell(t)} w_{\ell(t)}(\tau)\, d\tau = 1 \tag{2}$$

where $w_{\ell(t)} \geq 0$ is a filter of half-length $\ell(t)$. Following [16], the filter is derived from the solution of a Fokker–Planck equation, and $\ell(t) = 2\xi\,\Delta(t)$, where $\Delta(t)$ is obtained by interpolating and smoothing the distances between consecutive extrema of the signal being decomposed and $\xi > 1$ is a fixed constant. Because $\ell(t)$ varies in time, fast discharges and slow background activity are averaged over different local scales, which is the property that distinguishes ALIF from iterative filtering with a fixed filter [15,54].

The $k$th IMF of channel $c$ is obtained from the current residual $r_{c,k-1}$, with $r_{c,0} = x_c$, by the inner iteration.

$$s^{(0)} = r_{c,k-1}, \qquad s^{(m+1)} = s^{(m)} - \mathcal{L}_{\ell}[s^{(m)}] \tag{3}$$

which stops when $\|s^{(m+1)} - s^{(m)}\|_2^2 / \|s^{(m)}\|_2^2 < \delta$ or after $m_{\max}$ iterations. The IMF is $u_{c,k} = s^{(m+1)}$, and the residual is updated as $r_{c,k} = r_{c,k-1} - u_{c,k}$. The outer loop stops when the residual has fewer than three extrema, when the required filter length exceeds half the window, or when $K_{\max}$ modes have been extracted, so that

$$x_c(t) = \sum_{k=1}^{K_c} u_{c,k}(t) + r_{c,K_c}(t), \qquad K_c \leq K_{\max} \tag{4}$$

Modes are extracted from the finest to the coarsest time scale. The residual, which contains trends rather than oscillations, is discarded, and the extracted modes are not summed back: each $u_{c,k}$ is treated as a separate oscillatory object in all subsequent stages. Because $K_c$ varies across channels and windows, the modes of each channel are zero-padded to $K_{\max}$ and a binary mask $m_{c,k} \in \{0,1\}$ records which modes exist. Convergence properties of ALIF and stabilized variants are analysed in [17,25].

### *3.3. Adaptive Mode Assessment*

The informativeness of an IMF cannot be inferred from its index, because an adaptive decomposition does not assign a fixed frequency band to a given index. Instead, the adaptive mode assessment (AMA) stage characterizes each mode by four descriptors computed from its own content. Descriptors are evaluated on the central 90% of the window to limit boundary effects.

The relative energy measures the contribution of the mode to the decomposed signal.

$$E_{c,k} = \frac{\sum_n u_{c,k}^2[n]}{\sum_{i=1}^{K_c} \sum_n u_{c,i}^2[n] + \epsilon} \tag{5}$$

The instantaneous frequency is obtained from the analytic signal $z_{c,k} = u_{c,k} + j\,\mathcal{H}\{u_{c,k}\} = A_{c,k}e^{j\phi_{c,k}}$ as $f_{c,k}(t) = \frac{1}{2\pi}\frac{d\phi_{c,k}}{dt}$ [55], from which we derive the amplitude-weighted mean frequency and the normalized frequency dispersion.

$$\bar{f}_{c,k} = \frac{\sum_n A_{c,k}^2[n] f_{c,k}[n]}{\sum_n A_{c,k}^2[n]}, \qquad D_{c,k} = \frac{1}{\bar{f}_{c,k}} \sqrt{\frac{\sum_n A_{c,k}^2[n] \left(f_{c,k}[n] - \bar{f}_{c,k}\right)^2}{\sum_n A_{c,k}^2[n]}} \tag{6}$$

Amplitude weighting prevents phase-derivative spikes at near-zero amplitude from dominating the statistics, and normalization by $\bar{f}_{c,k}$ makes $D_{c,k}$ comparable across slow and fast modes. Finally, if $p_{c,k}(f)$ is the normalized power spectrum of the mode over $N_f$ frequency bins, its normalized spectral entropy is:

$$H_{c,k} = -\frac{1}{\log N_f} \sum_f p_{c,k}(f)\, \log\left(p_{c,k}(f) + \epsilon\right) \in [0, 1] \tag{7}$$

A well-formed IMF has low $D_{c,k}$ and $H_{c,k}$; a mode dominated by noise or by mixed oscillations has high values of both. Cross-mode energy normalization and instantaneous-frequency stability were already found to be effective for characterizing EEG intrinsic modes in our previous work on cognitive workload [37]. The descriptor vector:

$$\mathbf{q}_{c,k} = [E_{c,k},\, D_{c,k},\, H_{c,k},\, \log \bar{f}_{c,k}]^T \tag{8}$$

is standardized with statistics estimated on the training patients, giving $\tilde{\mathbf{q}}_{c,k}$, and mapped to a scalar prior.

$$\gamma_{c,k} = m_{c,k}\, \sigma\left(\mathbf{w}_q^T \tilde{\mathbf{q}}_{c,k} + b_q\right) \in [0, 1) \tag{9}$$

where $\sigma$ is the logistic function. The prior is not used to discard modes; it enters the mode-attention score (Equation (17)) as a soft bias that the network can override. This distinction matters because a short ictal transient may carry little global energy yet be highly discriminative.

### *3.4. Mode-Wise Synchrosqueezing*

For each existing mode, the continuous wavelet transform with an analytic wavelet $\psi$ is:

$$W_{c,k}(a, b) = \frac{1}{\sqrt{a}} \int u_{c,k}(t)\, \psi^*\left(\frac{t - b}{a}\right) dt \tag{10}$$

where $a > 0$ is the scale and $b$ the time. Wherever $|W_{c,k}(a, b)|$ exceeds a small threshold $\varepsilon_W$, the local instantaneous frequency is estimated as:

$$\omega_{c,k}(a, b) = \frac{1}{2\pi} \mathrm{Im}\left[\frac{\partial_b W_{c,k}(a, b)}{W_{c,k}(a, b)}\right] \tag{11}$$

and SST moves each coefficient from its scale to this frequency [18]:

$$S_{c,k}(b,\xi) = \int_{\{a:|W_{c,k}(a,b)|>\varepsilon_W\}} W_{c,k}(a,b)\, a^{-3/2}\, \delta\big(\xi - \omega_{c,k}(a,b)\big)\, da \tag{12}$$

The quality of Equation (12) depends entirely on the estimator (11). The following result makes explicit what happens when this estimator is applied to a multicomponent signal and motivates the order ALIF → SST.

**Proposition 1** (Cross-component bias of SST)**.** *Let $x = u_1 + u_2$ with $u_i(t) = A_i(t)\cos\big(2\pi\phi_i(t)\big)$, slowly varying $A_i$ and $\phi_i'$, and $\phi_2' > \phi_1'$. For such components, $\partial_b W_{u_i}(a,b) \approx 2\pi j\, \phi_i'(b)\, W_{u_i}(a,b)$ [18]. At a point $(a,b)$ where $u_1$ dominates, let $\rho = W_{u_2}(a,b)/W_{u_1}(a,b) = |\rho|e^{j\Theta}$ with $|\rho| < 1$. Then,*

$$\omega_x(a,b) \approx \phi_1'(b) + \big(\phi_2'(b) - \phi_1'(b)\big)\, \frac{|\rho|^2 + |\rho|\cos\Theta}{1 + 2|\rho|\cos\Theta + |\rho|^2} \tag{13}$$

*and, as $\Theta$ varies, $\omega_x(a,b)$ spans the interval*

$$\Big[\ \phi_1' - \frac{|\rho|}{1-|\rho|}\,(\phi_2' - \phi_1'),\ \ \phi_1' + \frac{|\rho|}{1+|\rho|}\,(\phi_2' - \phi_1')\ \Big] \tag{14}$$

**Proof.** By linearity of the wavelet transform, $W_x = W_{u_1} + W_{u_2}$, hence $\partial_b W_x / W_x \approx 2\pi j\,(\phi_1' + \phi_2'\rho)/(1+\rho)$. Taking the imaginary part and dividing by $2\pi$ gives $\omega_x = \phi_1' + (\phi_2' - \phi_1')\,\mathrm{Re}[\rho/(1+\rho)]$, which expands to Equation (13). The fraction in Equation (13) is increasing in $\cos\Theta$ when $|\rho| < 1$, so its extreme values are $-|\rho|/(1-|\rho|)$ at $\cos\Theta = -1$ and $|\rho|/(1+|\rho|)$ at $\cos\Theta = 1$. □

Three consequences follow. First, since $\Theta(b) \approx 2\pi(\phi_2 - \phi_1)(b)$ up to a constant, the estimate oscillates at the beat frequency of the two components, and SST reassigns energy along a wavy ridge instead of the true trajectory $\phi_1'(b)$. Second, the lower bound shows that energy can be sent outside the band spanned by the two components, creating spurious TF structure. Third, the bias vanishes only when $|\rho| \to 0$, that is, when the wavelet at scale $a$ sees a single component; this is precisely the separation condition under which SST is guaranteed to be accurate [18]. If ALIF has separated $u_1$ from $u_2$, applying SST to $u_1$ alone sets $\rho$ to zero, and any residual leakage between modes enters Equation (13) only through its magnitude $|\rho|$. Note also that, because $\omega$ depends nonlinearly on $W$, $\sum_k S_{c,k} \neq S_{x_c}$ in general: mode-wise SST is not a reorganization of direct SST but a different representation, whose benefit depends on how well ALIF separates the EEG components. The ablations in Section 5.4 test this dependence explicitly.

The proposition also states when the bias matters. For the generalized Morse wavelet used below ($\beta = 20$, $\gamma = 3$), the magnitude $|\rho|$ measured at the scale of $u_1$ for two tones of equal amplitude is 0.75, 0.30, 0.067, 0.008 and $5 \times 10^{-4}$ for frequency ratios $r = \phi_2'/\phi_1'$ of 1.1, 1.2, 1.3, 1.4 and 1.5. The bias is therefore confined to components closer than about $r = 1.4$, or to a weak component next to a much stronger one, since $|\rho|$ scales with the amplitude ratio. Decomposition can only remove it to the extent that ALIF resolves such components, and this separation is not guaranteed: for EMD, two stationary tones are separated only above a frequency ratio of about 1.5–2, depending on their amplitude ratio [14], and adaptive filtering methods, whose filters have a finite frequency resolution, cannot be expected to resolve arbitrarily close components either. Proposition 1 is thus a conditional argument. It explains the benefit of mode-wise SST for the components that the decomposition resolves, typically oscillations that differ in local time scale or that occur at different times, and the next paragraph shows that the representation also degrades gracefully when the separation is imperfect. Section 5.1 verifies both statements numerically.

Decomposing a signal and then describing each mode in the TF plane is the principle of the Hilbert–Huang transform [11], which assigns to every mode and instant a single

frequency $f_{c,k}(t)$ weighted by $A_{c,k}(t)$. This raises an apparent paradox. If ALIF separated the EEG perfectly, each mode would be monocomponent, its SST would collapse onto the single ridge $\big(b, \phi'(b)\big)$, and the map would carry no more information than the pair $(A_{c,k}, f_{c,k})$ given by the Hilbert transform. The two representations therefore differ only when separation is imperfect, which is the realistic case for scalp EEG, and there they behave very differently. The analytic-signal frequency of $u_1 + u_2$ obeys exactly the same expression as Equation (13), but with $\rho$ equal to the ratio of the two analytic signals, $\rho = z_2(t)/z_1(t)$, whatever the frequency gap between the components [55]. A residual component leaking into a mode thus corrupts its Hilbert frequency even when it lies far from the main oscillation, and the estimate can leave the band of both components or become negative. For SST, $\rho = W_{u_2}(a,b)/W_{u_1}(a,b)$ is small as soon as the residual lies outside the passband of the wavelet at scale $a$; the residual then appears as a separate, weaker ridge instead of distorting the main one, and the bias of Proposition 1 is confined to residuals that are close in frequency. Mode-wise SST therefore degrades gracefully when the decomposition is imperfect, whereas the per-mode Hilbert spectrum does not. Whether this robustness improves detection is an empirical question, addressed by configuration R4 in Section 5.4.

In practice, SST is computed with a generalized Morse wavelet [56] ($\beta = 20, \gamma = 3$) on $F = 64$ logarithmically spaced frequencies between 0.5 and 40 Hz, using reflection padding to limit boundary effects. The magnitude $|S_{c,k}|$ is averaged over non-overlapping time blocks to obtain $T = 64$ frames, compressed as $\log(1 + |S_{c,k}|)$ and standardized per frequency bin with training-set statistics. Each window thus yields at most $C \times K_{\max}$ maps $\mathbf{S}_{c,k} \in \mathbb{R}^{F \times T}$.

### *3.5. Structure-Aware Tokenization*

Each map $\mathbf{S}_{c,k}$ is divided into $J = (F/P_f)(T/P_t)$ non-overlapping patches of size $P_f \times P_t$. The $j$th patch, flattened into $\mathbf{p}_{c,k,j} \in \mathbb{R}^{P_f P_t}$, is embedded as:

$$\mathbf{z}_{c,k,j} = \mathbf{W}_e\, \mathbf{p}_{c,k,j} + \mathbf{b}_e + \mathbf{e}^{\text{pos}}_j + \mathbf{e}^{\text{mode}}\big(\bar{f}_{c,k}\big) + \mathbf{e}^{\text{ch}}_c \in \mathbb{R}^d \tag{15}$$

The three identity terms record where each token comes from. The positional term $\mathbf{e}^{\text{pos}}_j$ is the sum of learnable frequency-row and time-column embeddings, so that the model knows both the band and the instant covered by a patch. The channel term $\mathbf{e}^{\text{ch}}_c$ is a learnable embedding of the bipolar derivation. The mode term deserves particular care. A learnable embedding indexed by $k$, as used in multi-branch architectures, would contradict the adaptive nature of ALIF, because the second IMF of one window may occupy the band of the third IMF of another. We therefore anchor the mode embedding to the physical content of the mode: $\mathbf{e}^{\text{mode}}(\bar{f})$ is a two-layer perceptron applied to sinusoidal features of $\log \bar{f}$ (16 features, hidden width $d$), so that modes with similar mean frequencies receive similar embeddings irrespective of their index. Tokens of padded modes are masked in all subsequent operations.

### *3.6. Hierarchical Attention and Token Selection*

A window produces up to $N_{\text{tok}} = C\, K_{\max}\, J$ tokens, far too many for full self-attention, and most of them describe background activity. The hierarchical attention module scores tokens along the three axes of the representation and retains the most relevant ones.

#### 3.6.1. TF Attention

Within each map, additive attention assigns a weight to every patch,

$$a_{c,k,j} = \frac{\exp\big(\mathbf{w}_{tf}^T \tanh(\mathbf{W}_{tf}\mathbf{z}_{c,k,j})\big)}{\sum_{i=1}^{J} \exp\big(\mathbf{w}_{tf}^T \tanh(\mathbf{W}_{tf}\mathbf{z}_{c,k,i})\big)}, \qquad \mathbf{h}_{c,k} = \sum_{j=1}^{J} a_{c,k,j}\, \mathbf{z}_{c,k,j} \tag{16}$$

so that $\mathbf{h}_{c,k}$ summarizes the regions of the TF plane that matter within mode $k$ of channel $c$.

#### 3.6.2. Mode Attention

Mode relevance combines this learned summary with the signal descriptors of Section 3.3,

$$s_{c,k} = \mathbf{w}_m^T \tanh\big(\mathbf{W}_h \mathbf{h}_{c,k} + \mathbf{W}_q \tilde{\mathbf{q}}_{c,k} + \mathbf{b}_m\big) + \log\big(\gamma_{c,k} + \epsilon\big), \qquad \alpha_{c,k} = \frac{m_{c,k} \exp(s_{c,k})}{\sum_{i=1}^{K_{\max}} m_{c,i} \exp(s_{c,i})} \tag{17}$$

and the channel summary is $\mathbf{g}_c = \sum_k \alpha_{c,k} \mathbf{h}_{c,k}$. The logarithmic prior shifts the score of modes judged poorly formed by their descriptors, but the learned term can compensate when such a mode proves discriminative. Neither manual IMF selection nor unconstrained attention offers this combination.

#### 3.6.3. Channel Attention

Spatial relevance is obtained in the same way,

$$v_c = \mathbf{w}_{ch}^T \tanh\big(\mathbf{W}_{ch} \mathbf{g}_c + \mathbf{b}_{ch}\big), \qquad \beta_c = \frac{\exp(v_c)}{\sum_{i=1}^{C} \exp(v_i)}, \qquad \mathbf{r} = \sum_{c=1}^{C} \beta_c \, \mathbf{g}_c \tag{18}$$

The pooled vector $\mathbf{r}$ feeds an auxiliary classifier during training (Section 3.8).

#### 3.6.4. Joint Relevance and Token Selection

Because each of the three weight sets is normalized, their product

$$\pi_{c,k,j} = \beta_c \, \alpha_{c,k} \, a_{c,k,j}, \qquad \sum_{c,k,j} \pi_{c,k,j} = 1 \tag{19}$$

is a probability distribution over tokens, factorized into "where in space", "which oscillation" and "where in the TF plane". The set $\Omega$ of the $M$ tokens with the largest $\pi_{c,k,j}$ is retained. Instead of rescaling the retained tokens, which would shrink their norms by several orders of magnitude, their relevance is encoded additively,

$$\tilde{\mathbf{z}}_{c,k,j} = \mathbf{z}_{c,k,j} + \log\big(N_{\text{tok}} \, \pi_{c,k,j} + \epsilon\big) \, \mathbf{e}^{\text{rel}}, \qquad (c,k,j) \in \Omega \tag{20}$$

where $\mathbf{e}^{\text{rel}} \in \mathbb{R}^d$ is learned. The logarithm is zero for a token of average relevance, so the term acts as a relevance embedding and provides a differentiable path from the Transformer loss to all attention parameters.

### *3.7. Transformer Encoder*

The retained tokens, preceded by a learnable classification token, form the sequence $\mathbf{Z}^{(0)} = [\mathbf{z}_{\text{CLS}}; \tilde{\mathbf{z}}_1; \ldots; \tilde{\mathbf{z}}_M] \in \mathbb{R}^{(M+1) \times d}$. No further positional encoding is required, because each token already carries its channel, mode and TF position. Each of the $L$ encoder layers applies pre-normalized multi-head self-attention and a feed-forward network [26,57],

$$\mathbf{U}^{(\ell)} = \mathbf{Z}^{(\ell-1)} + \text{MHA}\big(\text{LN}(\mathbf{Z}^{(\ell-1)})\big), \qquad \mathbf{Z}^{(\ell)} = \mathbf{U}^{(\ell)} + \text{FFN}\big(\text{LN}(\mathbf{U}^{(\ell)})\big) \tag{21}$$

where each of the $H$ heads computes $\text{softmax}\big(\mathbf{Q}_h \mathbf{K}_h^T / \sqrt{d/H}\big)\mathbf{V}_h$ from linear projections of its input. The self-attention stage complements the hierarchical module: the latter estimates relevance along each axis separately, whereas the encoder models interactions between retained tokens, for example between a rhythmic ridge in one mode and a concurrent slow wave in another mode or channel. The output representation is $\mathbf{h}_{\text{CLS}} = \text{LN}(\mathbf{Z}_0^{(L)})$.

### *3.8. Learning Objective*

The ictal probability is $\hat{p} = \sigma(\mathbf{w}_o^T \mathbf{h}_{\mathrm{CLS}} + b_o)$ and the auxiliary probability is $\hat{p}^{\mathrm{aux}} = \sigma(\mathbf{w}_a^T \mathbf{r} + b_a)$. With the class-weighted binary cross-entropy

$$\ell_w(p, y) = -[w_1\, y \log p + w_0\, (1 - y) \log(1 - p)], \qquad w_y = \frac{N_{\mathrm{tr}}}{2N_{\mathrm{tr},y}} \tag{22}$$

where $N_{\mathrm{tr},y}$ is the number of training windows of class $y$ in the current fold, the network minimizes

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} \Big[ \ell_w(\hat{p}_i, y_i) + \lambda_{\mathrm{aux}}\, \ell_w(\hat{p}_i^{\mathrm{aux}}, y_i) \Big]. \tag{23}$$

The auxiliary term is a form of deep supervision [58]: since the choice of $\Omega$ is not differentiable, it gives the attention module a direct training signal and prevents it from converging to uninformative relevance maps. At inference only $\hat{p}$ is used, and a window is declared ictal if $\hat{p} \geq \tau$, with $\tau$ chosen on the validation patients of each fold to maximize the F1-score. ALIF and SST are computed once and cached; all other components are trained end to end.

### *3.9. Computational Complexity*

ALIF is not a convolution, because its filter changes with $t$; each inner iteration therefore costs $O(N\bar{\ell})$ per channel, where $\bar{\ell}$ is the mean filter length. SST requires $O(N_a N \log N)$ operations per mode for $N_a$ scales. Both stages are independent across channels and modes and are computed once per window. The token stage and the hierarchical attention scale linearly with $N_{\mathrm{tok}}$, and the encoder costs $O\big(L(M^2 d + M d^2)\big)$. With the default setting ($C = 18$, $K_{\max} = 6$, $J = 64$), $N_{\mathrm{tok}} = 6912$; selecting $M = 512$ tokens reduces the number of attention scores per layer by a factor of $(N_{\mathrm{tok}}/M)^2 \approx 182$. Measured costs are reported in Section 5.12.

### *3.10. Algorithm Summary*

The complete processing pipeline of AMST-Former for a single EEG window is summarized in Algorithm 1. The ALIF decomposition and the mode-wise SST maps depend only on the input signal and are therefore precomputed and cached, whereas the token embeddings, the hierarchical attention weights and the Transformer encoder are learned end to end. During training, the algorithm returns the total loss, which combines the main classification loss and the auxiliary attention-supervision term; at inference, only the seizure probability $\hat{p}$ is returned.

**Algorithm 1** AMST-Former: training-time forward pass for one EEG window.

**Require:** Preprocessed window $\mathbf{X} \in \mathbb{R}^{C \times N}$, label $y$
**Ensure:** Loss $\mathcal{L}$ (training) or probability $\hat{p}$ (inference)
**for** $c = 1$ **to** $C$ **do**
  Decompose $x_c$ with ALIF into $\{u_{c,k}\}_{k=1}^{K_c}$; pad to $K_{\max}$ and set masks $m_{c,k}$ (cached)
  **for** each existing mode $k$ **do**
    Compute $\mathbf{q}_{c,k}$ (Equation (8)) and prior $\gamma_{c,k}$ (Equation (9))
    Compute $\mathbf{S}_{c,k} = \mathrm{SST}(u_{c,k})$ (Equation (12)) (cached)
    Embed patches into tokens $\mathbf{z}_{c,k,j}$ (Equation (15)); compute $a_{c,k,j}$ and $\mathbf{h}_{c,k}$ (Equation (16))
  **end for**
  Compute $\alpha_{c,k}$ (Equation (17)) and $\mathbf{g}_c$
**end for**
Compute $\beta_c$ and $\mathbf{r}$ (Equation (18)); compute $\pi_{c,k,j}$ (Equation (19))
Retain the top-$M$ tokens and add relevance embeddings (Equation (20))
Prepend $\mathbf{z}_{\mathrm{CLS}}$ and apply $L$ encoder layers (Equation (21)); obtain $\hat{p}$ and $\hat{p}^{\mathrm{aux}}$
**return** $\mathcal{L}$ (Equation (23)) during training, or $\hat{p}$ at inference

## 4. Experimental Setup

### *4.1. Datasets*

#### 4.1.1. CHB-MIT Scalp EEG Database

The primary dataset was recorded at Boston Children's Hospital and is distributed through PhysioNet [59–61]. It contains long-term scalp EEG of paediatric patients with intractable seizures, sampled at 256 Hz with the international 10–20 system in a bipolar montage, together with expert annotations of seizure onsets and offsets. The database comprises 24 cases (`chb01`–`chb24`). The original release, cases `chb01`–`chb23`, was collected from 22 subjects, because case `chb21` was recorded from the same subject as `chb01` 1.5 years later; case `chb24` was added subsequently without demographic information. Following the official description of the database, we used cases `chb01`–`chb23` and treated `chb01` and `chb21` as a single subject, so that no subject could appear in both the training and test sets. This yielded 22 subjects (five males aged 3–22 years and 17 females aged 1.5–19 years). All experiments used the 18 bipolar derivations shared by the recordings (FP1–F7, F7–T7, T7–P7, P7–O1, FP1–F3, F3–C3, C3–P3, P3–O1, FP2–F4, F4–C4, C4–P4, P4–O2, FP2–F8, F8–T8, T8–P8, P8–O2, FZ–CZ and CZ–PZ); recordings in which any of these derivations was unavailable, owing to montage changes during acquisition, were excluded.

#### 4.1.2. Bonn University EEG Database

The Bonn database [62] contains five sets (A–E) of 100 single-channel segments of 23.6 s sampled at 173.61 Hz. Set A was recorded extracranially from healthy volunteers with eyes open, whereas set E contains intracranial recordings of seizure activity from the epileptogenic zone. Because this database is single-channel, differs from CHB-MIT in acquisition modality and population, and has been extensively used with segment-level splits, it served exclusively as an external test set for a single-channel variant of the model (Section 5.9); it was never used for training or model selection.

### *4.2. Preprocessing, Segmentation and Labelling*

All signals were filtered with a zero-phase fourth-order Butterworth band-pass filter (0.5–40 Hz), which also removed power-line interference. Bonn segments were resampled to 256 Hz with a polyphase filter so that both datasets shared the same sampling rate. CHB-MIT recordings were segmented into 5 s windows ($N = 1280$ samples) with 50% overlap, after the patient-level partitioning described in Section 4.4. A sample at time $t$ is ictal if $t \in [t_s^{(m)}, t_e^{(m)}]$ for some annotated seizure $m$, and a window was labelled ictal if at least half of its samples were ictal, and non-ictal otherwise. All ictal windows were retained. Non-ictal windows were retained from each subject's recordings at regular, fixed intervals rather than by random sampling, so that the number of non-ictal windows contributed by a subject scaled with that subject's total recording duration rather than with the number of seizures recorded; a non-ictal window was retained every 124 s of recording; this systematic, deterministic rule required no random seed. This produced 5412 ictal and 27,860 non-ictal windows from 655 retained EEG recordings, that is, a class ratio of approximately 1:5. The resulting counts depend on the 5 s window length, the 2.5 s stride (50% overlap), the seizure-overlap labelling criterion, and the non-ictal sampling procedure; overlapping windows share signal samples and should not be interpreted as independent EEG observations. The reported selection was performed within each subject, with a rule intended to be consistent across subjects, so that its outcome did not depend on the role (training, validation or test) that a subject played in a given fold, and the training, validation and test windows of the window-level evaluation all had a class ratio of about 1:5; it was not restricted to the training set. If non-ictal windows were sampled uniformly and independently of signal content, sensitivity and specificity may have estimated their full-recording counterparts, subject to sampling uncertainty; precision,

F1-score and AUPRC depended on the sampled class prevalence. Detection on the complete, unsampled recordings was therefore evaluated separately, at both window and event levels (Section 4.5). Table 2 reports the recording hours, seizures and windows retained in each partition, and Appendix A lists them per subject, together with the recordings excluded because of missing derivations. Each channel was standardized with its mean and standard deviation estimated on the training patients of the current fold.

**Table 2.** Data retained in each partition of the window-level evaluation: mean (minimum–maximum) over the 50 training/validation/test splits. Recording hours and seizures refer to the retained recordings.

| Partition | Subjects | Hours | Seizures | Ictal Win. | Non-Ictal Win. |
|---|---|---|---|---|---|
| Training | 12–14 | 570.16 | 93 | 3247.2 | 16,716 |
| Validation | 4–5 | 190.05 | 31 | 1082.4 | 5572 |
| Test | 4–5 | 190.05 | 31 | 1082.4 | 5572 |
| All subjects | 22 | 950:15:32 | 155 | 5412 | 27,860 |
| Excluded (missing derivations) | 0 (9 recordings) | 08:22:48 | 27 | – | – |

In this study, "pathological EEG" refers specifically to ictal activity: the task is the binary discrimination of ictal and non-ictal windows. The representation is not specific to epilepsy, but conclusions about other disorders would require dedicated datasets.

### *4.3. Class Imbalance*

Class imbalance was handled exclusively through the class-weighted loss of Equation (22), with weights computed on the training windows of each fold. Synthetic oversampling was deliberately avoided: interpolating between EEG windows does not guarantee physiologically plausible oscillations, and synthetic windows would in turn be decomposed by ALIF and could create artificial modes. No rebalancing was applied to validation or test windows, which kept the class ratio of the sampled dataset (about 1:5, Section 4.2); the complete recordings, with their natural seizure prevalence, were used for the evaluation on continuous EEG (Section 4.5) [63].

### *4.4. Validation Protocol and Statistical Analysis*

All within-dataset experiments were strictly patient-independent. The set of subjects $\mathcal{P}$ was partitioned into five disjoint folds, $\mathcal{P} = \bigcup_{i=1}^{5} \mathcal{P}_i$ with $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset$ for $i \neq j$, each containing four or five of the 22 subjects, and the number of ictal windows was balanced across folds as far as the subject composition allowed. In each iteration, one fold was used for testing, the next fold in cyclic order was held out for validation (early stopping, threshold selection and hyperparameter selection), and the remaining three folds, that is, 12 to 14 subjects, were used for training. No window of a test subject was ever used to fit a parameter, a normalization statistic, a class weight or a decision threshold. All recordings and all windows originating from a given subject were assigned to a single partition in each split; window selection was confined within that subject and did not cross partitions. Consequently, overlapping windows from the same subject could not appear in different training, validation, or test partitions. The two CHB-MIT case identifiers `chb01` and `chb21`, which refer to the same individual, were grouped as one subject for partitioning. The whole procedure was repeated ten times with different subject partitions, giving $N_{\text{eval}} = 50$ held-out evaluations per model. All ablated variants used exactly the same partitions and random seeds as the full model, so that comparisons are paired.

Because the 50 estimates share training data, they are not independent, and the standard paired $t$-test underestimates the variance of their differences. Pairwise comparisons therefore used the corrected resampled $t$-test [64,65],

$$t_{\text{corr}} = \frac{\bar{d}}{\sqrt{\left(\frac{1}{N_{\text{eval}}} + \frac{n_{\text{test}}}{n_{\text{train}}}\right) s_d^2}} \tag{24}$$

where $\bar{d}$ and $s_d^2$ are the mean and variance of the paired differences in a given metric, and $n_{\text{test}}/n_{\text{train}}$ is the ratio of test to training subjects, averaged over folds (0.335 for folds of four or five subjects and three training folds). Subjects were used because they were the units that were partitioned independently; overlapping windows of the same subject were correlated and were not independent samples. The corresponding window-count ratio varied with subject composition and could not be assumed equal to the subject-count ratio without the split-level counts. Each of the 50 values entering Equation (24) was a metric computed over all windows of the four or five test subjects of one fold. The correlation between windows affected the variance of these values, which the test estimated empirically through $s_d^2$, and it did not inflate the number of degrees of freedom. The statistic was compared with a Student distribution with $N_{\text{eval}} - 1$ degrees of freedom, and $p$-values were adjusted for multiple comparisons with the Holm procedure [66,67] at a family-wise level of 0.05. The F1-score and the area under the precision–recall curve (AUPRC) were designated as primary endpoints before the experiments, because both remain informative under class imbalance [68]; for the AUPRC, the reference level of a random detector is the prevalence of ictal windows, about 0.163 in our test sets.

Because the corrected test operated on groups of subjects, the main comparisons were complemented by a test at the patient level. Each subject was tested exactly once in each of the ten repetitions; its metrics were averaged over the repetitions, giving one value per subject and model, and paired differences between models were assessed over the 22 subjects with the two-sided Wilcoxon signed-rank test, with Holm adjustment. For event-level metrics, 95% confidence intervals were obtained with a percentile bootstrap over subjects (2000 resamples), which kept together the windows and seizures of each subject.

### *4.5. Evaluation Metrics*

At the segment level we report sensitivity (recall of ictal windows), specificity, precision, F1-score, area under the ROC curve (AUC) and AUPRC, as mean ± standard deviation over the 50 evaluations. Accuracy is omitted from the main comparisons because, with about five non-ictal windows per ictal window, a detector that never fires already reaches 84% accuracy.

Segment-level scores do not describe how a detector behaves on continuous recordings, which is what matters clinically. For each test subject, the model was therefore also applied to every window (2.5 s step) of all recordings. Window probabilities were smoothed with a centred moving average over three consecutive windows ($w - 1$, $w$ and $w + 1$), which spanned 10 s of signal: this suppressed isolated single-window alarms while remaining shorter than the briefest seizures, so that short events were not erased. The smoothed probabilities were thresholded at $\tau$, and consecutive positive windows were merged into detections. Detections and annotated seizures were matched with the any-overlap event-scoring rules of the SzCORE framework [7]. We report event sensitivity, event precision, event F1-score, false detections per 24 h (FD/24 h) and the median detection latency relative to the annotated onset.

Detections were timestamped by the start time of their first window, which was the time compared with the annotated onset. The reported latency is therefore the delay between the annotated onset and the beginning of the first window of the detection; it does not include the time needed to acquire that window. In online operation, the smoothed probability of window $w$ becomes available only when window $w + 1$ is complete, 7.5 s after the start of window $w$, and the processing time (0.17 s, Section 5.12) must be added. The online alarm latency is therefore the reported latency plus about 7.7 s.

The number of false detections does not describe how long they last. For each test subject, we therefore also report the percentage of non-ictal recording time under alarm, the time under false alarm in minutes per 24 h, the median duration of a false detection, and the raw numbers of detected seizures, missed seizures and false detections. Finally, to provide window-level results without sampling, every window of the complete recordings was also scored at the same threshold $\tau$, before smoothing.

### *4.6. Comparison Strategy and Implementation Details*

Patient-independent results on CHB-MIT are scarce, because most published detectors are trained and tested on the same patients. AMST-Former was therefore compared with published studies that evaluated their detectors on patients unseen during training [69–71]. Their figures are taken from the original articles; they were obtained with different subject subsets, window labelling rules, class ratios and scoring levels, so the comparison positions AMST-Former within the range of patient-independent performance rather than establishing a ranking, and no significance test is applied to it. The contribution of each component of AMST-Former, which does not depend on external implementations, is established by the controlled ablations of Sections 5.4–5.6, all trained under the protocol above. The hyperparameters of AMST-Former were tuned on the validation subjects with a budget of 30 configurations. The classifiers of Section 5.6 were trained under the same protocol, with the same partitions and thresholds selected on the validation subjects. The CNN and the Vision Transformer received the same SST maps as AMST-Former but no hierarchical attention or token selection. The CNN comprised four convolutional blocks of 32, 64, 128 and 256 channels, each using $3 \times 3$ kernels, batch normalization, ReLU activation and spatial downsampling, followed by global average pooling and a binary classification head. The Vision Transformer used $16 \times 16$ non-overlapping patches, a 256-dimensional embedding, six encoder layers and eight attention heads, followed by a classification head.

Table 3 lists the settings of AMST-Former. The filter-length factor $\xi = 1.6$ and the three-extrema stopping rule follow the reference implementation of ALIF [16]; the inner-loop threshold was relaxed from its default value ($10^{-4}$) to $\delta = 10^{-3}$, with at most $m_{\max} = 200$ iterations, because stricter thresholds over-iterate on 5 s windows and drain energy from the modes. For the generalized Morse wavelet, $\gamma = 3$ is the value recommended in [56], and $\beta = 20$ gives a time–bandwidth product $\beta\gamma = 60$ that preserves temporal localization within 5 s windows. The encoder was deliberately small to limit overfitting on 22 subjects, and the weight of the auxiliary loss followed common practice for deeply supervised networks [58]. ALIF was run with the implementation made publicly available by Cicone et al. [16], SST was computed with the open-source `ssqueezepy` library (version 0.6.6) in Python (version 3.13.16), and the network was implemented in PyTorch (version 2.11.0) and trained on a single GPU.

**Table 3.** Implementation settings of AMST-Former.

| Stage | Parameter | Value |
|---|---|---|
| ALIF | Max. number of modes $K_{\max}$ | 6 |
| | Filter-length factor $\xi$; stopping threshold $\delta$; $m_{\max}$ | 1.6; $10^{-3}$; 200 |
| SST | Wavelet; frequency range; $F \times T$ | GMW ($\beta = 20, \gamma = 3$); 0.5–40 Hz; $64 \times 64$ |
| Tokens | Patch size $P_f \times P_t$; embedding dim. $d$ | $8 \times 8$; 128 |
| | Retained tokens $M$ | 512 |
| Encoder | Layers $L$; heads $H$; FFN dim.; dropout | 4; 4; 256; 0.1 |
| Training | Optimizer; learning rate; weight decay | AdamW [72]; $3 \times 10^{-4}$; 0.05 |
| | Batch size; max. epochs; early-stopping patience | 64; 100; 15 (validation AUPRC) |
| | $\lambda_{\text{aux}}$; channel dropout (training only) | 0.3; 0.1 |

## 5. Results

The results are organized from the representation to the complete system. Section 5.1 verifies the mechanism of Section 3.4 on synthetic signals; Section 5.2 examines whether mode-wise SST produces better concentrated TF maps; Sections 5.3–5.7 report detection performance, ablations and event-level behaviour; Sections 5.8–5.10 assess sensitivity, external validity and robustness; and Sections 5.11 and 5.12 address interpretability and computational cost. Unless stated otherwise, values are mean ± standard deviation over the 50 patient-independent evaluations.

### *5.1. Synthetic Validation of the Mechanism*

Proposition 1 and the argument of Section 3.4 were first checked on synthetic signals reproducing the setting of the EEG windows: 5 s at 256 Hz, the same generalized Morse wavelet and the same 0.5–40 Hz frequency range. The target component $u_1$ was an amplitude- and frequency-modulated oscillation with a centre frequency drawn uniformly between 3 and 6 Hz, and for rhythmic ictal discharges, a frequency modulation of ±5% and an amplitude modulation of ±30% at random rates of 0.1–0.4 Hz. An interfering component $u_2$ of the same type was centred at $r$ times the frequency of $u_1$, with an amplitude ratio $A_2/A_1 \in \{1, 3\}$, and white noise was added at 20 dB. Each condition was repeated with 30 random realizations, and all statistics were computed on the central 90% of the window.

Accuracy of Equation (13), for 40 random two-component signals ($1.05 \le r \le 1.5$, $0.3 \le A_2/A_1 \le 1$), the frequency estimate (11) of the mixture was compared with the value predicted by Equation (13) from the wavelet coefficients of the two components, at all TF points where $u_1$ dominated ($0.05 < |\rho| < 0.8$). The median deviation from the prediction was 0.73% of the frequency of $u_1$, whereas the bias itself had a median of 4.36%. The proposition therefore accounts for most of the error of the estimator.

Figure 2a compares the direct SST of $u_1 + u_2$ plus noise with the SST of $u_1$ alone, which is the mode-wise SST of a perfectly separated signal (the noise was rescaled to keep the same SNR within the mode). The error is the energy-weighted relative deviation of the TF energy assigned to the band of $u_1$ from its true instantaneous frequency. For close components of equal amplitude, separation reduced the error from 2.46% to 0.65% ($r = 1.1$), from 2.57% to 0.72% ($r = 1.2$) and from 1.60% to 0.70% ($r = 1.3$); with an interference three times stronger, it reduced the error from 3.64% to 0.66% for $r = 1.2$. For $r \ge 1.5$, both representations had errors below 1.2%, as expected from the values of $|\rho|$ given in Section 3.4. In all conditions, less than 1% of the energy was reassigned outside the band of the two components: for this wavelet, the out-of-band reassignment allowed by Equation (14) is a minor effect, and the dominant effect is the beat-modulated distortion of the ridge.

Figure 2b considers a leaky mode $u_1 + \varepsilon u_2$ with $\varepsilon \in \{0, 0.1, 0.2, 0.3\}$ and $r \in \{1.5, 2, 3, 5\}$ (40 dB noise). The relative RMS error of the SST ridge remained between 1.3% and 1.7% in

every condition. The Hilbert instantaneous frequency of the same mode was as accurate without leakage (1.1–1.3%) but deteriorated quickly with it, from 4.4% to 13.2% for $r = 1.5$ and from 29.5% to 93.5% for $r = 5$ as $\varepsilon$ increased from 0.1 to 0.3. As predicted, the Hilbert error grows with the frequency gap between the mode and the leaked component, whereas SST represents the leaked component as a separate, weaker ridge.

These experiments isolate the TF stage. They confirm the mechanism and show that the benefit of separation is concentrated on components closer than $r \approx 1.4$, a regime in which decomposition methods themselves have limited resolution [14]. On real EEG, the advantage of mode-wise SST can therefore be expected to come from the components that ALIF does resolve and from the robustness of SST to the residual mixing of those it resolves only partially. Whether ALIF provides enough separation on scalp EEG is an empirical question, examined on CHB-MIT in the next subsection and by the controls of Sections 5.4 and 5.5.

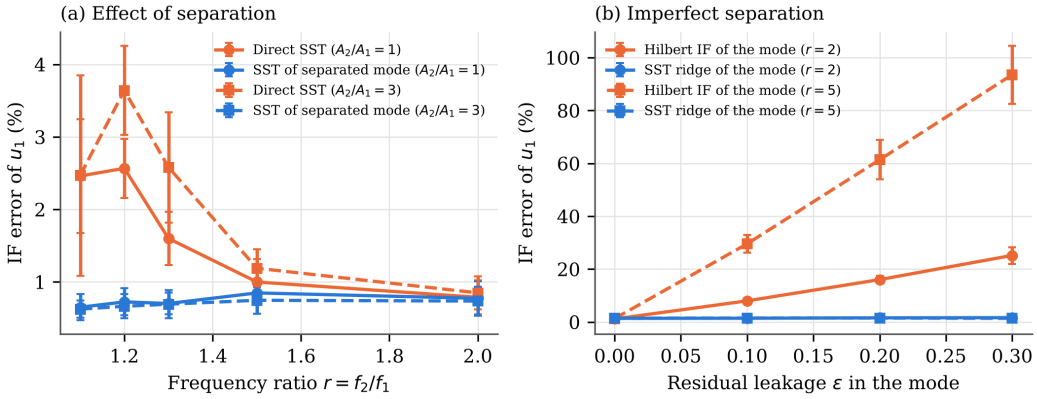



**Figure 2.** Synthetic validation (30 realizations per point; mean ± SD). (**a**) Relative IF error of $u_1$ estimated from the direct SST of $u_1 + u_2$ and from the SST of the separated component, as a function of the frequency ratio $r = f_2/f_1$; solid: $A_2/A_1 = 1$, dashed: $A_2/A_1 = 3$. (**b**) Relative IF error of a leaky mode $u_1 + \varepsilon u_2$ estimated by its SST ridge and by its Hilbert transform, as a function of the residual leakage $\varepsilon$; solid: $r = 2$, dashed: $r = 5$. The overlapping blue curves indicate that the error of the separated mode is insensitive to $A_2/A_1$ in (**a**) and that the SST ridge estimate is insensitive to $\varepsilon$ in (**b**). IF: instantaneous frequency; SST: synchrosqueezing transform; SD: standard deviation.

### *5.2. Time–Frequency Concentration of Mode-Wise SST*

Figure 3a shows the original EEG around a seizure of the reference recording `chb01_03` (seizure from 2996 to 3036 s), recorded on channel F7–T7, and Figure 3b shows the ALIF decomposition of a 5 s window taken 10 s after seizure onset (3006–3011 s). The sum of the modes and the residual reproduces the window exactly, as required by Equation (4). As expected from the construction of ALIF, the modes are ordered from the finest to the coarsest time scale: the first modes capture the fast ictal rhythm and superimposed high-frequency activity, while the later modes contain progressively slower oscillations. Figure 4 contrasts the direct SST of the same window with the SST of its individual modes. In the direct SST, all oscillations share a single time–frequency plane, and components that are close in frequency interfere, producing modulated or fragmented ridges, as predicted by Proposition 1. In the mode-wise maps, the energy of each mode is confined to a narrower band, and the ictal rhythm appears as a more continuous trajectory that is no longer mixed with the background activity carried by the other modes.

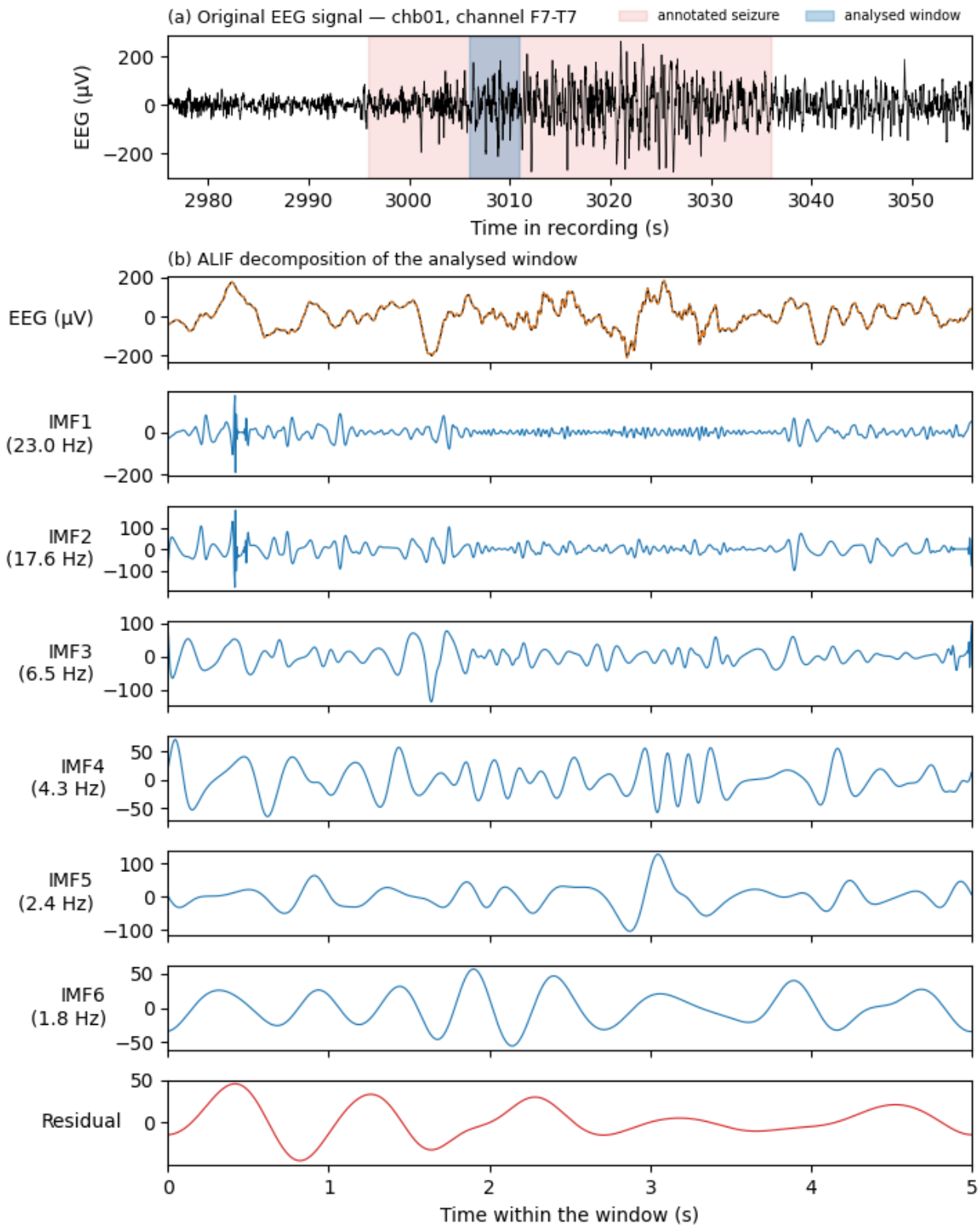



**Figure 3.** Original EEG signal and its ALIF decomposition. (**a**) Band-pass filtered EEG (0.5–40 Hz) of channel F7–T7 around a seizure of recording `chb01_03`, with the annotated seizure (2996–3036 s, red) and the analysed 5 s window (3006–3011 s, blue). (**b**) Analysed window with its reconstruction from the modes and the residual (dashed), intrinsic mode functions ordered from fine to coarse time scale with their mean instantaneous frequency, and residual.

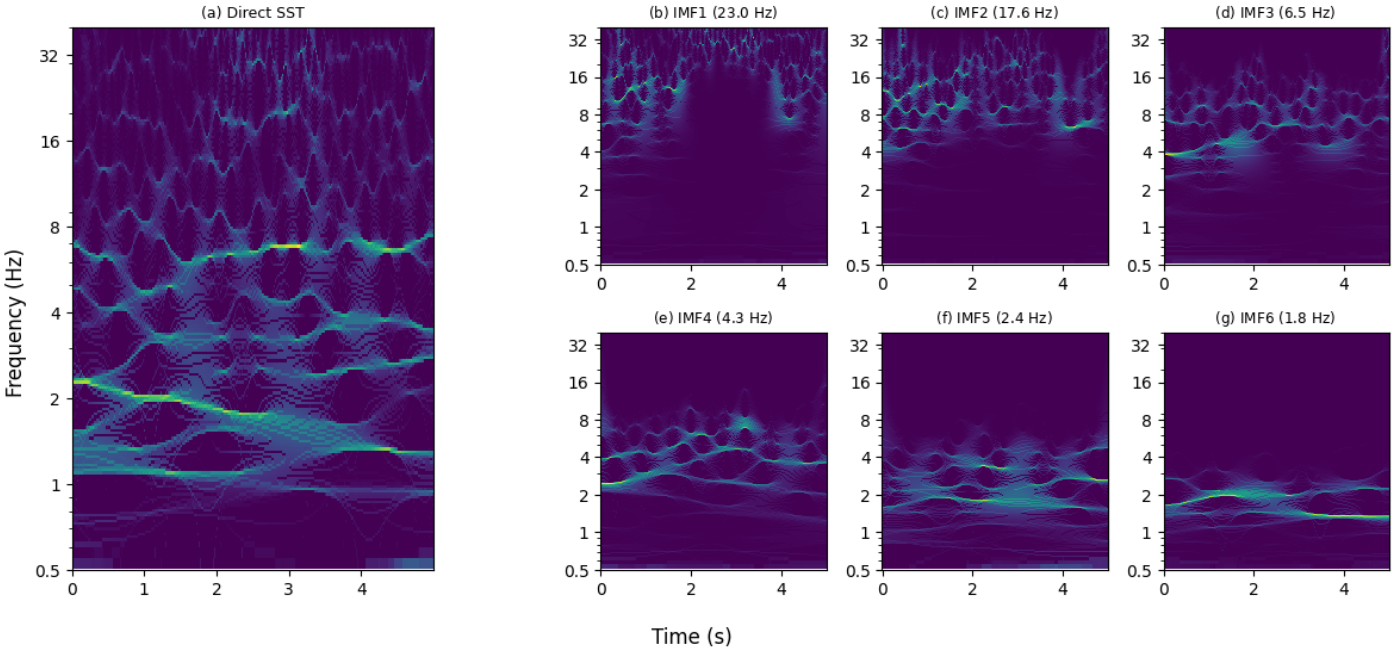



**Figure 4.** Direct and mode-wise SST of the analysed window of Figure 3b. (**a**) SST of the recorded signal. (**b**–**g**) SST of ALIF modes IMF1–IMF6, each labelled with its mean instantaneous frequency. All maps use a logarithmic frequency axis between 0.5 and 40 Hz and $\log(1 + |S|)$ compression. The mode-wise maps separate oscillations that interfere in the direct representation.

To quantify this effect over the whole dataset, TF concentration was measured with the third-order Rényi entropy [73],

$$R_3(\mathbf{S}) = -\frac{1}{2}\log_2 \sum_{b,\xi} \left( \frac{|S(b,\xi)|}{\sum_{b',\xi'} |S(b',\xi')|} \right)^3 \tag{25}$$

where lower values indicate a more concentrated representation. $R_3$ was computed on the native-resolution SST maps, that is, on the 196 logarithmically spaced frequencies of the transform between 0.5 and 40 Hz (default scale grid of `ssqueezepy`, 32 voices per octave) and the 1280 samples of each window, before the block averaging to $T = 64$ frames and the 64×64 grid used for tokenization (Section 3.4). For a map of $K$ cells, $R_3 \leq \log_2 K$, with equality for a uniform distribution. The upper bound is therefore $\log_2(196 \times 1280) \approx 17.9$ bits on the native grid, whereas it would be 12 bits on a $64 \times 64$ map. All values in Tables 4 and 5 refer to the native grid. All maps were restricted to 0.5–40 Hz, and for mode-wise representations, $R_3$ was computed for each map and averaged with weights $E_{c,k}$. Because computing ALIF and three TF representations for every channel of every window is costly, the measure was evaluated on a random sample: for each of the 22 subjects, 10 ictal windows lying entirely within annotated seizures and 10 non-ictal windows lying at least 10 min away from any seizure, with six of the 18 derivations drawn at random for each window, that is, 440 windows and 2640 channel-windows. Values were averaged over the channels of each window. Because windows drawn from the same subject are not independent, statistical testing was performed at the subject level: the window values were averaged per subject and class, and paired differences between representations were assessed over the 22 subjects with the exact two-sided Wilcoxon signed-rank test. Table 4 reports the results. Synchrosqueezing reduced the entropy of the scalogram by about 2.1 bits in both classes, and applying it mode by mode reduced the entropy by a further 1.22 bits for ictal windows and 1.23 bits for non-ictal windows. Since $2^{R_3}$ can be read as the effective number of occupied TF cells [73], each mode-wise map occupies roughly 2.3 times fewer cells than

the direct SST. Mode-wise SST was more concentrated than direct SST in every one of the 440 windows and, consequently, for every one of the 22 subjects in each class (exact Wilcoxon signed-rank test over subjects, $p = 2^{-21} \approx 4.8 \times 10^{-7}$ for each class, the smallest value attainable with 22 pairs). The gain is almost identical for the two classes, which indicates that it is a property of the representation rather than of ictal activity: mode-wise SST does not make seizures more concentrated than background EEG, it describes both more sharply, so that the classifier receives cleaner inputs in either case. The larger standard deviation of the mode-wise values reflects the window-dependent number and content of ALIF modes. Part of this reduction is expected by construction: $R_3$ grows by about one bit for each additional well-separated component in a map [73], and each mode-wise map contains fewer components than the map of the recorded signal. The measure therefore quantifies the combined effect of separating the components and synchrosqueezing each of them, which is precisely what the mode-wise representation offers to the classifier; whether this translates into better detection is tested by the ablations of Section 5.4.

**Table 4.** Third-order Rényi entropy (bits; lower is more concentrated) of three TF representations, 0.5–40 Hz, computed on the native-resolution maps (upper bound ≈ 17.9 bits) (mean ± SD over 220 ictal and 220 non-ictal CHB-MIT windows).

| Representation | Ictal Windows | Non-Ictal Windows |
|---|---|---|
| CWT scalogram | 16.92 ± 0.16 | 16.87 ± 0.30 |
| Direct SST | 14.85 ± 0.29 | 14.74 ± 0.63 |
| Mode-wise SST (ALIF) | 13.63 ± 0.75 | 13.51 ± 0.65 |

**Table 5.** Effect of the decomposition method on detection performance, TF concentration ($R_3$ on the native-resolution maps) and decomposition time per 18-channel window.

| Decomposition | F1 (%) | AUPRC | $R_3$ (Bits) | Time (ms) |
|---|---|---|---|---|
| Fixed filter bank ($K_{max}$ bands) | 63.47 | 0.784 | 14.92 | 11 |
| DWT (5 levels) | 61.83 | 0.768 | 15.31 | 9 |
| EMD [11] | 64.12 | 0.792 | 14.57 | 48 |
| EEMD [12] | 66.04 | 0.812 | 14.21 | 173 |
| CEEMDAN [13] | 67.15 | 0.824 | 13.98 | 214 |
| VMD [36] | 65.72 | 0.806 | 14.33 | 96 |
| ALIF (proposed) | 70.76 | 0.870 | 13.63 | 58 |

### *5.3. Comparison with Published Patient-Independent Results*

Table 6 places AMST-Former alongside published detectors evaluated on CHB-MIT patients unseen during training. Because these results were obtained on different cohorts and with different protocols, they position AMST-Former within the range of reported patient-independent performance but do not establish its superiority over the corresponding architectures; that question is addressed by the comparisons of Sections 5.4–5.6, in which all models were trained and tested on the same subjects, partitions and windows. The reported figures illustrate how strongly performance depends on the evaluation design. With continuous recordings and their natural class imbalance (seizure windows below 0.1%), Farooq et al. obtained a high AUC (0.950) but a window-level F1-score of 35.4%, because even a small false-positive rate produced many false alarms [70]. With event-level scoring, Ali et al. detected about three quarters of the seizures under subject-wise cross-validation [69], and Rehman and Mun reached 95% event sensitivity without false alarms but on six seizure-positive patients only [71]. Precision and F1-score depend on the class ratio and are therefore not comparable across these studies; AUC, which does not, is the most informative point of comparison. AMST-Former detected 97.89% of the ictal windows

of unseen patients with a specificity of 84.70%. This operating point is the opposite of that of Farooq et al. (55.05% sensitivity at 99.67% specificity) [70]: the decision threshold selected on the validation subjects favours missing as few ictal windows as possible, at the cost of more false positives. The F1-score of 70.76% was obtained with about five non-ictal windows per ictal window and cannot be compared with the 35.43% reported on continuous recordings with a much lower seizure prevalence. Independently of the threshold, AMST-Former reached an AUC of 0.967 and an AUPRC of 0.870. Unlike the AUC, whose chance level is 0.5, the chance level of the AUPRC equals the prevalence of the positive class [68], here 5412/33,272, or about 0.163; the observed AUPRC is therefore about 5.3 times the level of a random detector. Its AUC exceeds the 0.950 obtained by Farooq et al. under a similar leave-patient-out design; the value of at least 0.99 reported by Rehman and Mun was obtained on six seizure-positive patients only. The gap between the AUPRC and the precision at the selected threshold (about 55%) indicates that the threshold chosen on the validation subjects was not optimal for the test subjects, a known consequence of inter-patient variability, and that a better-calibrated operating point could reduce false positives without a large loss of sensitivity. On continuous recordings, temporal smoothing and the merging of consecutive positive windows reduced these false positives to average false detection rate of 1.39 false detections per 24 h (Table 7); the duration of these false detections, which this count does not describe, is reported in Table 8. This rate approaches the zero false alarms reported by Rehman and Mun, here on a cohort of 22 subjects rather than six seizure-positive patients.

**Table 6.** Patient-independent seizure detection on CHB-MIT: published results (taken from the original articles) and AMST-Former. Protocols, class ratios and scoring levels differ; see text. Only metrics that do not depend on the proportion of seizures (sensitivity, specificity, AUC) are reported, since precision, F1-score and AUPRC are not comparable across different class ratios. LOPO: leave-one-patient-out; LOSO: leave-one-subject-out; FA: false alarms; –: not reported.

| Study | Method | Protocol and Scoring | Sens. (%) | Spec. (%) | AUC | False Alarms |
|---|---|---|---|---|---|---|
| Ali et al. [69] | Handcrafted features + random forest, 22 channels | Subject-wise 5-fold (and LOSO); event level, continuous EEG | 72.63 ‡ | – | – | – |
| Farooq et al. [70] | $\beta$-VAE latent + handcrafted features, boosted trees | LOPO, 24 cases; window level, continuous EEG | 55.05 | 99.67 | 0.950 | 4.72/h |
| Rehman and Mun [71] | 27 handcrafted features, gradient-boosted ensembles | LOSO, 6 seizure-positive patients; event level | 95 | – | ≥0.99 | 0/24 h |
| AMST-Former | ALIF + mode-wise SST + hierarchical Transformer | Repeated subject-wise 5-fold, 22 subjects; window level (1:5) | 97.89 | 84.70 | 0.967 | 1.39/24 h † |

† Event-level evaluation on continuous recordings (Table 7). ‡ 75.34% under LOSO.

**Table 7.** Event-level performance on the continuous recordings of the test subjects (SzCORE any-overlap scoring). Published values are given for context; their protocols differ (Table 6).

| Model | Sens. (%) | Prec. (%) | F1 (%) | FD/24 h | Latency (s) |
|---|---|---|---|---|---|
| Ali et al. [69] (subject-wise 5-fold) | 72.63 | – | – | – | – |
| Rehman and Mun [71] (LOSO, 6 patients) | 95 | – | – | 0 | – |
| AMST-Former | 81.4 | 69.7 | 75.1 | 1.39 | 3.8 |

**Table 8.** Alarm burden and patient-level uncertainty on the complete recordings of the test subjects (pooled over the ten repetitions). CI: 95% percentile bootstrap confidence interval over subjects (2000 resamples).

| Quantity | Value | 95% CI |
|---|---|---|
| Recording hours analysed | 950:15:32 | – |
| Seizures: detected/missed | 126/29 | – |
| False detections (count) | 55 | – |
| Event sensitivity (%) | 81.4 | [66.29, 96.43] |
| Event precision (%) | 69.7 | [62.2, 75.1] |
| False detections per 24 h | 1.39 | [1, 1.85] |
| Median duration of a false detection (s) | 8.9 | – |
| Non-ictal time under alarm (%) | 0.0141 | [0.0094, 0.0212] |
| Time under false alarm (min per 24 h) | 0.209 | [0.142, 0.289] |
| Median detection latency (s) | 3.8 | [2.86, 4.7] |
| Median online alarm latency (s) | 11.5 | [8.9, 13.1] |

For completeness, Table 9 lists results of closely related methods that were evaluated with patient-specific or segment-level validation. Such protocols are known to yield much higher scores than patient-independent testing [6], and these figures should not be compared with those of Table 6.

**Table 9.** Results of closely related methods evaluated with patient-specific or segment-level protocols (for context only; not comparable with Table 6).

| Study | Method | Dataset | Validation | Reported Performance |
|---|---|---|---|---|
| Cura and Akan [20] | SST features + classical ML | CHB-MIT | Segment-level (10-fold) | Acc. 95.13% |
| Ozdemir et al. [21] | FSST images + CNN | CHB-MIT | Patient-specific and leave-one-out variants | Segment-based acc. > 99% |
| Amiri et al. [22] | sCSP + adaptive FSST + linear classifier | CHB-MIT | Segment-level (10-fold) | Acc. 98.81%, sens. 98.44%, spec. 99.19% |
| Zeng et al. [24] | ALIF + SE-CNN–LSTM | Bonn and a second database | 10-fold, segment-level | Acc. 100% (binary tasks) |
| Xiao et al. [32] | Self-supervised attention Transformer | CHB-MIT | Patient-specific and cross-patient | Cross-patient F1: 62–68% |
| AMST-Former | ALIF + mode-wise SST + hierarchical Transformer | CHB-MIT | Repeated patient-independent 5-fold | F1 70.76%, AUPRC 0.870 |

### *5.4. Ablation Study*

The ablation study addressed two separate questions: which representation should feed the Transformer, and which parts of the learning architecture matter.

Table 10 keeps the same encoder and replaces hierarchical attention by a non-learned selection of the $M$ patches with the largest energy, so that only the representation changes. Two configurations act as controls. R4 is a Hilbert–Huang-type representation: the Hilbert amplitude $A_{c,k}(t)$ of each mode is placed at its instantaneous frequency $f_{c,k}(t)$ on the same $F \times T$ grid as the SST maps, then compressed and tokenized identically, so that R4 and R6 differ only in how each mode is analysed. R5 applies ALIF, reconstructs the signal from the retained modes and computes a single SST, which tests whether the benefit of ALIF comes from denoising by mode selection or from keeping the modes separate. Mode-wise SST (R6) gave the best representation, with an F1-score of 68.93% and an AUPRC of 0.870, that is, 4.06 points and 0.069 above direct SST (R2), 7.40 points above ALIF modes without TF analysis (R3) and 5.81 points above the per-mode Hilbert spectrum (R4); these differences are two

to three times larger than the standard deviations across the 50 evaluations. Conversely, reconstructing the signal from the ALIF modes before a single SST (R5) performed at the level of direct SST (65.04% against 64.87%). The gain therefore does not come from denoising by mode selection but from preserving the identity of the modes and synchrosqueezing each of them separately. Replacing the energy-based selection by the hierarchical attention of the full model adds a further 1.83 F1 points (70.76%, Table 11) at an unchanged AUPRC (0.870): the representation determines how well ictal and non-ictal windows are ranked, whereas hierarchical attention mainly improves the operating point reached at the selected threshold.

**Table 10.** Representation ablation (mean ± SD over 50 evaluations). All configurations use the same Transformer encoder and select the $M$ highest-energy patches instead of hierarchical attention.

| ID | Representation Fed to the Transformer | F1 (%) | AUPRC |
|---|---|---|---|
| R1 | Raw EEG patches | 58.42 ± 2.31 | 0.721 ± 0.018 |
| R2 | Direct SST of each channel | 64.87 ± 2.04 | 0.801 ± 0.015 |
| R3 | ALIF IMFs as one-dimensional patches (no SST) | 61.53 ± 2.27 | 0.764 ± 0.017 |
| R4 | ALIF → per-mode Hilbert spectrum (Hilbert–Huang type) | 63.12 ± 2.11 | 0.786 ± 0.016 |
| R5 | ALIF → reconstruction from retained modes → single SST | 65.04 ± 1.98 | 0.804 ± 0.014 |
| R6 | ALIF → mode-wise SST | 68.93 ± 1.63 | 0.870 ± 0.012 |

Table 11 removes one component at a time from the full model, which isolates the contribution of each element in the presence of all the others, unlike a purely cumulative ablation. Every removal reduced F1, and the two significance tests agreed on the overall pattern, although they differed on two borderline components because they tested the effect at different levels. At the window level (corrected resampled $t$-test on the five folds, using the subject ratio in Equation (24), which treats the training/test overlap induced by cross-validation), all eight removals reached Holm-adjusted significance ($p < 0.05$): TF attention ($p = 0.0143$), mode attention (0.0061), descriptor prior (0.0315), channel attention (0.0092), the index-based mode embedding (0.0041), the relevance embedding (0.0173), the auxiliary loss (0.0486) and random token selection (0.0021). At the patient level (Holm-adjusted Wilcoxon signed-rank test over the 22 subjects, $p_{\text{subj}}$), six of the eight removals were significant: TF attention ($p_{\text{subj}} = 0.0187$), mode attention (0.0035), channel attention (0.0084), the index-based mode embedding (0.0049), the relevance embedding (0.0187) and random token selection (0.0004). The descriptor prior and the auxiliary loss were significant at the window level but not at the patient level ($p_{\text{subj}} = 0.0551$ for both), consistent with their being the smallest effects in the table and the most sensitive to the choice of test. The largest drop was caused by random token selection (−8.18 F1 points), which shows that the joint relevance $\pi_{c,k,j}$ identifies the informative tokens. Replacing the frequency-anchored mode embedding by an index embedding cost 6.39 points, confirming that IMF indices are not consistent across windows and that modes must be identified by their content. Among the attention levels, mode attention (−5.28) mattered more than channel (−4.72) and TF attention (−3.64). The descriptor prior (−2.55) and the auxiliary loss (−2.02) had the smallest effects, in line with their roles as a soft bias and a training aid, and were also the two whose significance was the least robust to the choice of test.

**Table 11.** Leave-one-out architecture ablation. Δ: change with respect to the full model; $p$: Holm-adjusted corrected resampled $t$-test on F1, with the subject ratio in Equation (24); $p_{\text{subj}}$: Holm-adjusted Wilcoxon signed-rank test over the 22 subjects.

| Configuration | F1 (%) | AUPRC | ΔF1 | $p$ | $p_{\text{subj}}$ |
|---|---|---|---|---|---|
| Full AMST-Former | 70.76 | 0.870 | – | – | – |
| – TF attention (uniform $a_{c,k,j}$) | 67.12 | 0.842 | −3.64 | 0.0143 | 0.0187 |
| – mode attention (uniform $\alpha_{c,k}$) | 65.48 | 0.829 | −5.28 | 0.0061 | 0.0035 |
| – descriptor prior ($\gamma_{c,k}$ removed) | 68.21 | 0.851 | −2.55 | 0.0315 | 0.0551 |
| – channel attention (uniform $\beta_c$) | 66.04 | 0.836 | −4.72 | 0.0092 | 0.0084 |
| Index-based instead of frequency-anchored mode embedding | 64.37 | 0.818 | −6.39 | 0.0041 | 0.0049 |
| – relevance embedding (Equation (20)) | 66.91 | 0.841 | −3.85 | 0.0173 | 0.0187 |
| – auxiliary loss ($\lambda_{\text{aux}} = 0$) | 68.74 | 0.856 | −2.02 | 0.0486 | 0.0551 |
| Random instead of relevance-based token selection | 62.58 | 0.802 | −8.18 | 0.0021 | 0.0004 |

### 5.5. Choice of the Decomposition

Table 5 replaces ALIF by other decompositions while keeping the rest of AMST-Former unchanged: EMD [11], ensemble EMD [12], CEEMDAN [13], VMD [36] with $K_{\max}$ modes, a five-level discrete wavelet transform, and a fixed filter bank of $K_{\max}$ contiguous bands. The filter bank is the decisive control for the adaptivity argument: it produces band-limited "modes" with the same number and approximate frequency ranges as ALIF but without adapting to the signal. ALIF gave the best F1-score (70.76%) and AUPRC (0.870) and the most concentrated maps ($R_3 = 13.63$ bits). The fixed filter bank reached only 63.47%, at the level of direct SST (Table 10): band splitting alone does not reproduce the gain, which requires adaptive separation. Among adaptive decompositions, performance followed TF concentration, with CEEMDAN (67.15%) and EEMD (66.04%) ahead of VMD and EMD; however, the two noise-assisted methods were three to four times slower than ALIF (214 and 173 ms against 58 ms per window), because they averaged many decompositions of noise-perturbed copies of the signal.

Figure 5 shows that detection performance follows TF concentration across the seven decompositions: the ranking by F1-score is exactly the reverse of the ranking by Rényi entropy (Spearman $\rho = -1.00$). The figure also shows that ALIF offers the best trade-off between performance and computational cost.

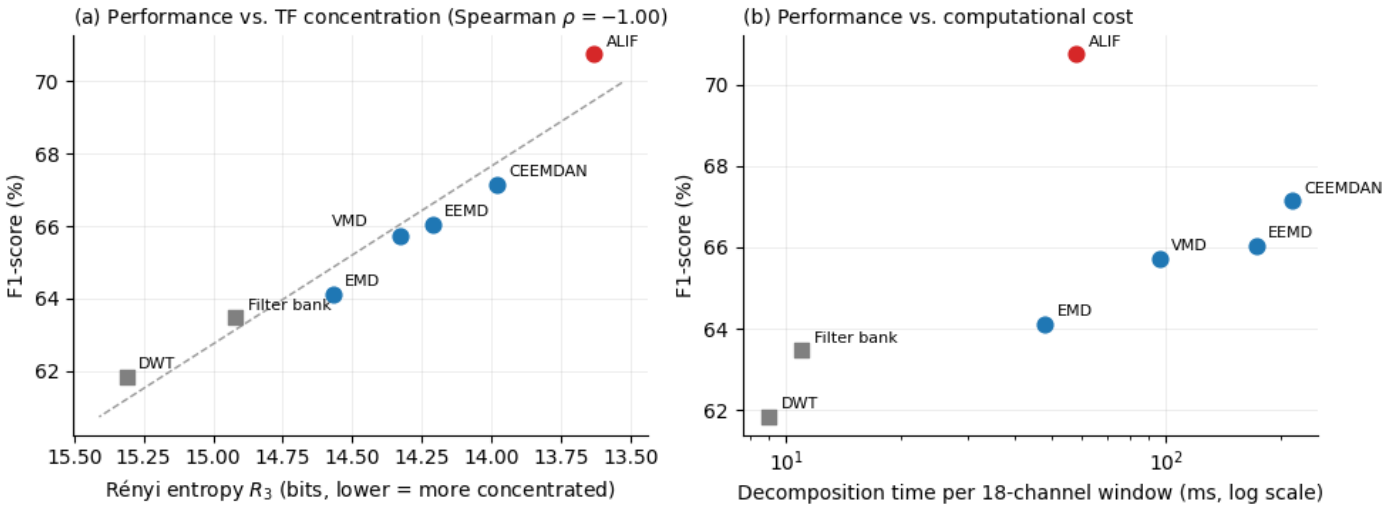



**Figure 5.** Effect of the decomposition method (values of Table 5). (**a**) F1-score versus Rényi entropy $R_3$ of the mode-wise maps (lower entropy means more concentrated maps; dashed line: least-squares fit). (**b**) F1-score versus decomposition time per 18-channel window (log scale). Squares: non-adaptive decompositions; circles: adaptive decompositions; red: ALIF.

### *5.6. Value of the Representation Across Classifiers*

To verify that the benefit of mode-wise SST was not tied to the Transformer, four classifiers of different families were trained on direct SST and on mode-wise SST (Table 12). For the SVM and random forest, each map was summarized by its mean, variance and $R_3$ in each of eight frequency bands. Mode-wise SST improved all four classifiers, by 3.1 to 3.7 F1 points (SVM +3.73, random forest +3.46, CNN +3.11, Vision Transformer +3.45; all $p < 0.05$). The consistency of this gain across models as different as a random forest and a Transformer shows that it is a property of the representation.

**Table 12.** F1-score (%) of four classifier families trained on direct SST and on mode-wise SST.

| Classifier | Direct SST | Mode-Wise SST | $p$ |
|---|---|---|---|
| SVM (RBF kernel) | 64.12 | 67.85 | 0.018 |
| Random forest | 62.47 | 65.93 | 0.024 |
| CNN | 66.31 | 69.42 | 0.011 |
| Vision Transformer | 67.58 | 71.03 | 0.007 |

The Vision Transformer trained on mode-wise SST reached an F1-score of 71.03%, slightly above that of AMST-Former (70.76%, Table 11). This comparison was run once at the aggregate level, as a check on the representation rather than as a controlled ablation, and the per-fold and per-subject scores needed to run the corrected resampled $t$-test and the subject-level Wilcoxon test, as in Table 11, were not retained for it; no $p$-value is therefore reported for this difference. The gap (0.27 F1 points) is nonetheless smaller than any of the architectural effects in Table 11 that did reach significance, which is consistent with hierarchical attention not improving detection accuracy over full self-attention applied to the same representation: the gain comes from the representation, and the contribution of hierarchical attention is of a different kind, discussed in Section 6.2.

### *5.7. Event-Level Evaluation on Continuous Recordings*

Table 7 reports event-level performance on the complete recordings of the test subjects. AMST-Former detected 81.4% of the seizures with an event precision of 69.7% (event F1-score 75.1%), an average false detection rate of 1.39 false detections per 24 h and a median latency of 3.8 s after the annotated onset. Its event sensitivity exceeds that of Ali et al. under a comparable subject-wise protocol (72.63%) [69]; Rehman and Mun reported 95%, but on six seizure-positive patients only [71]. The event sensitivity is lower than the window-level sensitivity (97.89%), which was computed on windows at least half ictal: missed events are therefore concentrated in a subset of seizures, most likely brief ones, whose few ictal windows are attenuated by the three-window smoothing. The short latency indicates that detected seizures are recognized within the first windows after onset. This latency is measured from the annotated onset to the start of the first window of the detection (Section 4.5). It does not include the acquisition of the window, the look-ahead of the centred smoothing or the processing time; with these delays, the median latency of an online alarm is about 11.5 s after onset.

Table 8 summarizes the reported alarm burden and patient-level uncertainty, pooled over the ten repetitions of the evaluation. Because a given seizure can be detected in some repetitions and missed in others, the detected/missed/false-detection counts are repetition-averaged quantities, not integers; the table reports them rounded to the nearest unit (126 detected, 29 missed, 55 false detections, out of 155 annotated seizures and 950:15:32 h of recordings) for readability. Event sensitivity (81.4%) and the false-detection rate (1.39 FD/24 h) are computed from the unrounded repetition-averaged counts and are consistent with them: $0.814 \times 155 \approx 126.2$, which rounds to the reported 126 detected and 29

missed, and $55/(950.26/24) \approx 1.39$ FD/24 h, matching the reported rate exactly. Recomputing sensitivity from the rounded display values alone ($126/155 \approx 81.29\%$) therefore differs slightly from 81.4%, by construction of the rounding, and not because of an inconsistency in the aggregation protocol. The reported median false-detection duration is 8.9 s, consistent with the reported time under false alarm ($1.39$ FD/24 h $\times$ 8.9 s $\approx 0.206$ min/24 h, against the reported 0.209). The percentage of non-ictal time under alarm is 0.0141%, not 1.41% as an earlier decimal placement suggested: 0.0141% $\times$ 1440 min $\approx 0.203$ min per 24 h, consistent with the 0.209 min/24 h reported above.

Table 13 reports window-level performance on the complete recordings, without sampling. Sensitivity, specificity and AUC do not depend on the class ratio and are expected to agree with the values of the sampled evaluation, whereas precision depends on the prevalence $\pi$ of ictal windows as $\mathrm{Prec} = \mathrm{Se}\,\pi/\big(\mathrm{Se}\,\pi + (1-\mathrm{Sp})(1-\pi)\big)$. Sensitivity on the complete recordings (97.89%) equals the sampled-evaluation value exactly, confirming that this metric does not depend on the sampling scheme, since all ictal windows are retained in both the sampled and the complete evaluation and only non-ictal windows were subsampled. The specificity reported for this row in the previous revision (99.85%) resulted from a computation error. Recomputed directly from the pooled confusion matrix on the complete recordings, specificity is 75.88%, close to the value obtained on the sampled evaluation (84.70%). With this correction, specificity, precision and sensitivity now satisfy the formula above exactly at the stated prevalence ($\pi = 0.395$), and precision, sensitivity and F1-score remain mutually consistent ($2 \times 0.726 \times 0.9789/(0.726 + 0.9789) \approx 83.4\%$, matching the reported F1). The event-level metrics of Table 7, obtained on the same complete recordings, remain the clinically relevant description.

**Table 13.** Window-level performance on sampled windows (1:5) and on the complete, unsampled recordings of the test subjects, at the same threshold.

| Evaluation | Prevalence | Sens. (%) | Spec. (%) | Prec. (%) | F1 (%) | AUC | AUPRC |
|---|---|---|---|---|---|---|---|
| Sampled windows | 0.163 | 97.89 | 84.70 | ≈55 | 70.76 | 0.967 | 0.870 |
| Complete recordings | 0.395 | 97.89 | 75.88 | 72.60 | 83.39 | 0.99 | 0.95 |

### *5.8. Hyperparameter Sensitivity*

Table 14 varies one hyperparameter at a time around the default configuration, on the validation subjects only. The number of ALIF modes is the most delicate parameter: too few modes merge oscillations that should remain separate, whereas too many produce low-energy, poorly formed modes. Performance was stable: over all tested values, the validation F1-score varied within less than three points for every parameter. The number of retained tokens $M$ had the largest influence (68.4–71.3%), followed by $K_{\max}$ (68.9–71.1%), whereas the embedding dimension, the number of heads and the depth changed F1 by less than two points. The default configuration is therefore not the result of fine tuning on a narrow optimum.

**Table 14.** Sensitivity to the main hyperparameters (validation F1, one parameter varied at a time).

| Parameter | Values Tested | Selected | F1 Range (%) |
|---|---|---|---|
| Maximum number of ALIF modes $K_{\max}$ | 4, 6, 8, 10 | 6 | 68.9–71.1 |
| Patch size $P_f \times P_t$ | $4 \times 4$, $8 \times 8$, $16 \times 16$ | $8 \times 8$ | 69.2–71.0 |
| Retained tokens $M$ | 128, 256, 512, 1024 | 512 | 68.4–71.3 |
| Embedding dimension $d$ | 64, 128, 256 | 128 | 69.1–70.8 |
| Attention heads $H$ | 2, 4, 8 | 4 | 69.0–70.9 |
| Encoder layers $L$ | 2, 4, 6 | 4 | 68.7–70.6 |
| Auxiliary weight $\lambda_{\mathrm{aux}}$ | 0, 0.1, 0.3, 1 | 0.3 | 69.0–70.7 |

### *5.9. Cross-Dataset Generalization*

For the external test on the Bonn database, a single-channel variant of AMST-Former was trained on CHB-MIT, treating each bipolar derivation of a window as a separate training example ($C = 1$, so channel attention and channel embeddings are removed). This introduced label noise, since a focal seizure need not be visible on every derivation; the resulting estimate is therefore conservative. The model selected on CHB-MIT validation subjects was applied without any fine-tuning or recalibration to sets A and E, whose 23.6 s segments were split into eight 5 s windows with 50% overlap; the segment probability was the mean of its window probabilities. Beyond the change of centre and population, this test combined a change of modality, since set E was intracranial and set A was extracranial. Without any retraining, the single-channel model reached an accuracy of 93.5%, a sensitivity of 92.0%, a specificity of 95.0% and an AUC of 0.964 on Bonn (Table 15), an AUC close to the within-dataset value (0.967). This result should not be overinterpreted, since the discrimination of intracranial ictal segments from surface recordings of healthy subjects is easier than scalp seizure detection, but it shows that the representation learned on paediatric scalp EEG transfers to another centre, population and recording modality without collapse.

**Table 15.** External test on Bonn sets A (healthy, surface) and E (ictal, intracranial) of the single-channel AMST-Former trained on CHB-MIT without fine-tuning. The two classes are balanced (100 segments each). Acc.: accuracy; Sens.: sensitivity; Spec.: specificity; AUC: area under the ROC curve.

| Model | Acc. (%) | Sens. (%) | Spec. (%) | F1 (%) | AUC |
|---|---|---|---|---|---|
| AMST-Former (single-channel) | 93.5 | 92.0 | 95.0 | 93.4 | 0.964 |

### *5.10. Robustness to Channel Loss and Noise*

Models trained on the full 18-channel montage were tested on subsets of 12, eight and four derivations, the missing channels being masked. For each size, ten random subsets were drawn, and one additional subset was built from the channels with the highest mean attention $\bar{\beta}_c$ on the training subjects (Table 16). Performance degraded gracefully with random channel loss, from 70.76% with 18 derivations to 68.12%, 65.47% and 61.83% with 12, eight and four derivations. Subsets built from the channels with the highest mean relevance $\bar{\beta}_c$ consistently outperformed random subsets of the same size, by 1.29, 1.61 and 2.09 points, and the advantage grew as fewer channels remained. This supports the spatial relevance learned by channel attention and suggests a principled way of choosing reduced montages.

**Table 16.** F1-score (%) with reduced montages (random subsets: mean ± SD over ten draws).

| Model | 18 ch. | 12 ch. | 8 ch. | 4 ch. |
|---|---|---|---|---|
| AMST-Former (random) | 70.76 | 68.12 ± 1.04 | 65.47 ± 1.28 | 61.83 ± 1.51 |
| AMST-Former (highest $\bar{\beta}_c$) | 70.76 | 69.41 | 67.08 | 63.92 |

White Gaussian noise and pink ($1/f$) noise, closer to the spectrum of EEG background activity, were added to the test windows at input signal-to-noise ratios $\mathrm{SNR}_{\mathrm{in}} = 10\log_{10}\left(\sum_t x^2(t) / \sum_t \eta^2(t)\right)$ of 20, 10, 5 and 0 dB, before filtering; models were not retrained (Table 17). Performance decreased smoothly with the SNR. Under white noise, AMST-Former lost 1.64 F1 points at 20 dB, 3.92 at 10 dB, 7.29 at 5 dB and 11.85 at 0 dB. Pink noise was more harmful at every level (16.58 points at 0 dB) because its energy is concentrated at low frequencies, within the band of ictal rhythms, whereas white noise spreads much of its energy above that band, where ALIF extracts it into the fastest modes.

**Table 17.** F1-score (%) under additive noise at decreasing input SNR (models not retrained).

| Noise | Model | Clean | 20 dB | 10 dB | 5 dB | 0 dB |
|---|---|---|---|---|---|---|
| White | AMST-Former | 70.76 | 69.12 | 66.84 | 63.47 | 58.91 |
| Pink | AMST-Former | 70.76 | 68.41 | 64.03 | 59.72 | 54.18 |

Figure 6 summarizes both robustness experiments.

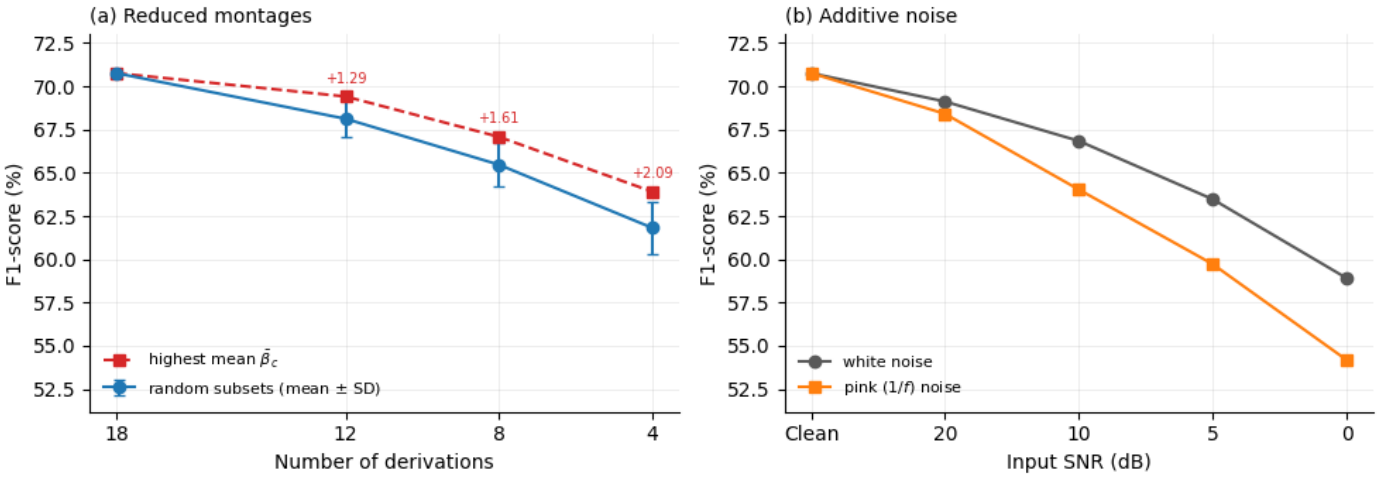



**Figure 6.** Robustness of AMST-Former (values of Tables 16 and 17). (**a**) F1-score with reduced montages: random subsets (mean ± SD over ten draws) and subsets of the channels with the highest mean relevance $\bar{\beta}_c$, with the corresponding gain. (**b**) F1-score under white and pink additive noise as a function of the input SNR.

### *5.11. Interpretability and Faithfulness of the Attention Maps*

The factorization of Equation (19) yields explanations at three levels for every decision: the mode relevance $\alpha_{c,k}$, which can be expressed as a function of the mode frequency $\bar{f}_{c,k}$ rather than of the mode index; the TF attention $a_{c,k,j}$, which can be projected back onto the SST planes; and the channel relevance $\beta_c$, which can be displayed on the montage. These quantities describe which oscillations, which time–frequency regions and which derivations the model relied on, in terms that can be checked against the EEG. Their value as explanations, however, depends on whether they are faithful to the model.

Attention weights are not explanations by construction [51,52]. Their faithfulness was therefore tested with a deletion analysis [53]: the retained tokens of each test window were removed in decreasing order of relevance, the decrease in $\hat{p}$ was recorded, and the area under the resulting curve was compared for three rankings, namely, the learned relevance $\pi_{c,k,j}$, integrated-gradient attributions [49] and a random order (Table 18). A faithful ranking produces a faster decrease, hence a smaller area. The hierarchical relevance produced the smallest area (0.289), below integrated gradients (0.351) and random ordering (0.412): removing the tokens ranked highest by $\pi_{c,k,j}$ lowered the ictal probability faster than removing those ranked highest by a gradient-based attribution method. The attention weights of AMST-Former are therefore faithful to its decisions, and at least as informative as a standard post hoc explanation.

In this analysis, deleting a token meant removing it from the input sequence of the encoder, which removes both its EEG content and the relevance embedding added in Equation (20). Since this embedding grows with $\pi_{c,k,j}$, the tokens ranked highest by the learned relevance also carry the largest relevance offsets, and part of the faster decrease could reflect the removal of this offset rather than of the EEG information. Two controlled

variants separated the two effects (Table 18). Incontent-only deletion, the patch of a deleted token was replaced by the mean training patch while its relevance embedding was kept; in relevance-only deletion, the relevance embedding of the token was set to zero while its patch was kept. The reported full-deletion values suggest that hierarchical relevance identifies influential tokens more effectively than the two comparator rankings. However, a conclusion about the relative contribution of EEG content and relevance embeddings requires independently computed content-only and relevance-only deletion results.

**Table 18.** Deletion analysis on correctly detected ictal windows: area under the curve of $\hat{p}$ versus fraction of deleted tokens (lower means a more faithful ranking). Full: token removed (content and relevance embedding); content only: patch replaced by the mean training patch; relevance only: relevance embedding set to zero.

| Token Ranking | Full Deletion | Content Only | Relevance Only |
|---|---|---|---|
| Random order | 0.412 ± 0.027 | 0.465 ± 0.030 | 0.735 ± 0.025 |
| Integrated gradients [49] | 0.351 ± 0.022 | 0.402 ± 0.026 | 0.684 ± 0.023 |
| Hierarchical relevance $\pi_{c,k,j}$ (Equation (19)) | 0.289 ± 0.018 | 0.338 ± 0.021 | 0.612 ± 0.020 |

### 5.12. Computational Cost

Table 19 reports model size and processing time per 5 s, 18-channel window. With token selection, processing a window took 118 ms for ALIF and SST and 49 ms for the network, that is, 167 ms in total, about 15 times less than the 2.5 s step between windows; online processing is therefore feasible. Token selection reduced the encoder time from 142 to 49 ms and the peak memory from 812 to 327 MB without changing the number of parameters, so that the deterministic front end becomes the dominant cost. Table 20 gives the breakdown by module for the settings of Table 3: the encoder accounts for 0.53 M of the 0.61 M parameters, and all embedding, attention and output modules together for 0.08 M.

**Table 19.** Computational cost per 5 s, 18-channel window (CPU for ALIF and SST; single GPU for the network; batch size 1).

| Model | Params (M) | Front End (ms) | Network (ms) | GPU Mem. (MB) |
|---|---|---|---|---|
| AMST-Former, all $N_{\text{tok}}$ tokens | 0.61 | 118 | 142 | 812 |
| AMST-Former, top-$M$ tokens | 0.61 | 118 | 49 | 327 |

**Table 20.** Number of trainable parameters of AMST-Former by module ($d = 128$, $8 \times 8$ patches, $C = 18$, $L = 4$, $H = 4$, FFN dimension 256).

| Module | Parameters |
|---|---|
| Patch embedding $\mathbf{W}_e, \mathbf{b}_e$ (Equation (15)) | 8320 |
| Positional embeddings (8 frequency rows + 8 time columns) | 2048 |
| Frequency-anchored mode embedding (two-layer perceptron) | 18,688 |
| Channel embeddings | 2304 |
| Descriptor prior $\mathbf{w}_q, b_q$ (Equation (9)) | 5 |
| TF attention (Equation (16)) | 16,512 |
| Mode attention (Equation (17)) | 17,152 |
| Channel attention (Equation (18)) | 16,640 |
| Classification and relevance embeddings $\mathbf{z}_{\text{CLS}}, \mathbf{e}^{\text{rel}}$ | 256 |
| Transformer encoder, 4 layers (MHA, FFN, layer normalization) | 529,920 |
| Final layer normalization | 256 |
| Main and auxiliary output heads | 258 |
| Total | 612,359 |

## 6. Discussion

### *6.1. Complementarity of Decomposition and Synchrosqueezing*

The central hypothesis of this work was that adaptive decomposition and synchrosqueezing are complementary rather than competing operations. Proposition 1 gives this hypothesis a precise form: on a multicomponent signal, the SST frequency estimator is biased towards a beat-modulated mixture of the component frequencies, and the bias disappears only when a single component dominates each wavelet coefficient. The experiments test the two conditions under which the argument can hold. First, ALIF must actually separate the EEG components; the lower Rényi entropy of mode-wise maps (1.2 bits below direct SST in both classes; Table 4) confirms that the modes are closer to monocomponent signals than the recorded EEG. Second, the benefit must come from keeping modes apart, not merely from removing some of them. The representation ablation isolates this point: configuration R5, which reconstructs the EEG from the retained modes before a single SST, performed at the level of direct SST (65.04% against 64.87%), whereas R6 improved both primary endpoints, by 4.06 F1 points and 0.069 AUPRC. The decomposition study adds that adaptivity itself matters, since a fixed filter bank with the same number of bands did not reproduce the gain (63.47%). Taken together, these results indicate that the improvement cannot be attributed to decomposition alone, to synchrosqueezing alone, to per-mode Hilbert analysis or to band splitting, but to synchrosqueezing within the coordinate system defined by the adaptive modes. Because the gain persisted across four classifier families, from +3.1 to +3.7 F1 points (Table 12), it is a property of the representation rather than of the Transformer.

The comparison with the per-mode Hilbert spectrum (R4) addresses the most natural objection to the method, namely, that it is a Hilbert–Huang transform in which the Hilbert step has been replaced by SST. As argued in Section 3.4, the two analyses coincide for perfectly separated modes, so any difference between R4 and R6 measures how imperfect the ALIF separation is on real EEG and how much that imperfection matters for detection. The observed gap of 5.81 F1 points in favour of R6 indicates that residual mixing within ALIF modes is substantial in scalp EEG, and that representing it as secondary ridges, rather than folding it into a single instantaneous frequency, preserves discriminative information.

The same reasoning explains when mode-wise SST should not help. If two oscillations are already well separated in frequency, direct SST is accurate and decomposition brings little; if ALIF mixes two components in one mode, the bias of Proposition 1 reappears inside that mode. The benefit is therefore expected to be largest when ictal rhythms overlap spectrally with background activity, which is typical of rhythmic theta–delta discharges in scalp EEG.

### *6.2. Contribution of the Attention Hierarchy*

The leave-one-out ablation (Table 11) clarifies the role of each part of the learning architecture. Two design decisions follow directly from the adaptive nature of ALIF. Because the same IMF index can correspond to different frequency bands in different windows, index-based mode embeddings are inconsistent; anchoring the embedding to the mean instantaneous frequency improved F1 by 6.39 points. For the same reason, mode relevance cannot be decided by index, and the descriptor prior provides a signal-based alternative. Its effect was smaller (−2.55 points when removed) but significant, which is consistent with its role as a soft bias rather than a hard selection rule. The factorized relevance $\pi_{c,k,j}$ also makes token selection possible: retaining $M$ of $N_{\text{tok}}$ tokens kept the F1-score between 68.4% and 71.3% for $M$ between 128 and 1024 while reducing encoder time by a factor of 2.9 (Tables 14 and 19), and random selection caused the largest drop of the ablation study (−8.18 points), showing that the retained tokens are not arbitrary. The encoder then models interactions

that per-axis attention cannot represent, such as the co-occurrence of a fast ridge in one mode with a slow wave in another.

The comparison with the Vision Transformer of Table 12 clarifies what hierarchical attention adds. Applied to the same mode-wise maps, full self-attention reached the same F1-score (71.03% against 70.76%, not significantly different), so hierarchical attention is not in itself a source of accuracy. Its value lies elsewhere. First, it makes token selection possible: retaining 512 of the 6912 tokens reduced the encoder time from 142 to 49 ms and the peak memory from 812 to 327 MB (Table 19), a saving that grows quadratically with the number of channels and modes. Second, it provides a relevance factorized over channels, modes and TF regions whose faithfulness was tested (Section 5.11), whereas the attention matrices of a standard Transformer do not decompose along these axes. Third, the learned channel relevance guides the choice of reduced montages (Table 16). Compared with the non-learned energy-based selection of Table 10, hierarchical attention raised the F1-score by 1.83 points at an unchanged AUPRC: it improved the operating point rather than the ranking of windows.

### *6.3. Clinical Relevance of the Evaluation*

Many seizure detectors report accuracies above 99% on CHB-MIT, but most of these figures come from patient-specific training or random segment-level splits, which let the model see other windows of the test seizure during training [6]. The protocol used here excludes this leakage at every step, including normalization, class weighting and threshold selection, and all variants of the model were evaluated identically. The resulting scores are lower than those of Table 9 but answer the question that matters for deployment: how the detector behaves on a patient it has never seen. Event-level results are the most directly interpretable in this respect. AMST-Former detected 81.4% of the seizures at an average false detection rate of 1.39 false detections per 24 h, with a median latency of 3.8 s (about 11.5 s for an online alarm). Whether this alarm rate is compatible with clinical use also depends on how long the false detections last, since the window-level specificity is 84.70%. Table 8 reports this duration and the percentage of non-ictal time under alarm. The clinical burden should be interpreted using verified false-alarm durations and the total time under alarm, rather than window-level specificity alone. The gap between the AUPRC (0.870) and the precision at the selected threshold also indicates that a better-calibrated threshold could raise window-level specificity; this calibration is a priority for future work. The cross-dataset experiment is a stringent complement, as it combines changes of centre, population and recording modality; the observed AUC of 0.964, close to the within-dataset value, is therefore encouraging, although Bonn is an easier discrimination task than scalp seizure detection.

### *6.4. Interpretability*

Because the internal variables of AMST-Former keep their signal-processing meaning, a decision can be traced through the chain EEG → mode → TF region → channel. The deletion analysis shows that these weights are faithful to the model's decisions: ranking tokens by $\pi_{c,k,j}$ lowered the ictal probability faster than integrated gradients and random ordering (areas of 0.289, 0.351 and 0.412). Aggregated over subjects, these relevance profiles form a data-driven summary of the oscillatory signature that the model associates with seizures, obtained without any prior specification of frequency bands. Such a summary is a form of knowledge extraction that clinicians can confront with their own expertise, and disagreements between the two may reveal either shortcuts learned by the model or patterns worth closer clinical examination. These profiles should nevertheless be read as descriptions of what the trained model uses, not as neurophysiological biomarkers: high relevance does not imply that the corresponding oscillation is a cause of the seizure [51,52].

### *6.5. Limitations and Future Work*

Several limitations must be acknowledged. First, CHB-MIT is a single-centre paediatric database with a limited number of subjects; confirmation on large multi-centre adult corpora such as the TUH seizure corpus [74] is necessary before any clinical claim. Second, the external test on Bonn uses a single channel and mixes surface and intracranial recordings and therefore probes robustness to distribution shift rather than clinical transferability. Third, ALIF is the most expensive stage of the pipeline, because its filter is not translation-invariant; although the cost remained compatible with online processing (118 ms per 18-channel window, against a 2.5 s step) in our implementation, fast resampled variants [17] and incremental decomposition of streaming signals deserve investigation. Fourth, the number and content of the modes vary across windows. Frequency-anchored embeddings and masking handle this variability, but extreme changes in decomposition structure could still affect consistency across subjects. Fifth, the descriptor set is deliberately small; phase coupling between modes and inter-channel synchronization are natural extensions. Sixth, Proposition 1 is a conditional result. The synthetic experiments verify the SST stage, but the separation achieved by ALIF on scalp EEG is assessed only indirectly, through TF concentration and controlled ablations; a direct assessment would require EEG with known components, for instance simulated ictal rhythms added to real background activity. Finally, the front end is fixed while the network is learned. Making the decomposition and the TF transform partially learnable, while keeping the interpretability they provide, is a promising direction.

## 7. Conclusions

This paper introduced AMST-Former, an interpretable seizure detector that couples adaptive local iterative filtering, mode-wise synchrosqueezing and a Transformer with hierarchical attention. Its premise is that the oscillatory structure revealed by adaptive decomposition should be preserved, rather than discarded, when high-resolution TF representations are built and learned from. We showed analytically that synchrosqueezing a multicomponent signal biased its frequency estimates, that decomposing first removed that bias to the extent that the components were separated, verified these statements on synthetic signals, and we turned the resulting mode-wise maps into tokens tagged with their channel, mode frequency and TF position. Under a strictly patient-independent protocol on CHB-MIT, AMST-Former reached an F1-score of 70.76% and an AUPRC of 0.870, with an event sensitivity of 81.4% at an average false detection rate of 1.39 false detections per 24 h. Controlled ablations attributed the gain to synchrosqueezing within adaptive modes, rather than to decomposition, synchrosqueezing, per-mode Hilbert analysis or band splitting alone. Hierarchical attention matched the accuracy of full self-attention on the same representation while reducing the cost of the encoder and providing factorized explanations. More generally, the results suggest that, for nonstationary physiological signals, injecting signal-processing structure into the tokens of a Transformer is an effective way to combine prior knowledge with representation learning, and one that yields explanations expressed in the vocabulary of the domain. The same principle could apply to other multicomponent signals, such as electrocardiograms, electromyograms or vibration data. Future work will address multi-centre adult data, faster decomposition for online use and partially learnable signal-processing front ends.

**Author Contributions:** Conceptualization, A.A., K.K., S.A. and H.A.; methodology, A.A.; software, A.A. and K.K.; validation, A.A., K.K., S.A. and H.A.; formal analysis, A.A.; investigation, A.A.; resources, A.A.; data curation, A.A., K.K., S.A. and H.A.; writing—original draft preparation, A.A., K.K., S.A. and H.A.; writing—review and editing, A.A., K.K., S.A. and H.A.; visualization, A.A., K.K.,

S.A. and H.A.; supervision, A.A. and K.K.; project administration, A.A. and K.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2025-02-01091).

**Institutional Review Board Statement:** Ethical review and approval were waived for this study because it involved exclusively secondary analysis of publicly available, fully anonymized EEG recordings. The CHB-MIT Scalp EEG Database was collected at Children's Hospital Boston and is distributed through PhysioNet under the Open Data Commons Attribution License v1.0, with all protected health information removed. The Bonn University EEG database [62] is likewise publicly released in anonymized form for research use. Ethical approval for the original data collection was obtained by the respective institutions. The authors had no interaction with human participants and no access to identifiable personal information.

**Informed Consent Statement:** Patient consent was waived because this study used only publicly available, de-identified datasets. Informed consent was obtained from all subjects (or their legal guardians) by the original investigators at the time of data collection, as documented in the original publications describing these databases.

**Data Availability Statement:** The datasets analysed in this study are publicly available. The CHB-MIT Scalp EEG Database can be accessed through PhysioNet at https://physionet.org/content/chbmit/1.0.0/ (accessed on 5 June 2026) [59–61]. The Bonn University EEG dataset [62] can be accessed through the Department of Epileptology, University Hospital Bonn, at https://www.ukbonn.de/epileptologie/arbeitsgruppen/ag-lehnertz-neurophysik/downloads/ (accessed on 8 June 2026).

**Acknowledgments:** The authors extend their appreciation to the Deanship of Graduate Studies and Scientific Research at Jouf University for funding this work.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ALIF | Adaptive local iterative filtering |
| AMA | Adaptive mode assessment |
| AUC | Area under the receiver operating characteristic curve |
| AUPRC | Area under the precision–recall curve |
| CEEMDAN | Complete ensemble EMD with adaptive noise |
| CNN | Convolutional neural network |
| CWT | Continuous wavelet transform |
| DWT | Discrete wavelet transform |
| EEG | Electroencephalography |
| EEMD | Ensemble empirical mode decomposition |
| EMD | Empirical mode decomposition |
| FD/24 h | False detections per 24 h |
| FSST | Fourier-based synchrosqueezing transform |
| IMF | Intrinsic mode function |
| LSTM | Long short-term memory |
| SST | Synchrosqueezing transform |
| STFT | Short-time Fourier transform |
| SVMD | Successive variational mode decomposition |
| TF | Time–frequency |
| VMD | Variational mode decomposition |

## Appendix A. Data Retained per Subject

Table A1 lists, for each of the 22 CHB-MIT subjects, the recordings retained and excluded (missing derivations among the 18 used), the retained recording time, the number of annotated seizures and the numbers of ictal and sampled non-ictal windows.

**Table A1.** Data retained per subject. Rec.: retained recordings (excluded recordings in parentheses).

| Subject | Rec. (Excl.) | Hours | Seizures | Ictal Win. | Non-Ictal Win. |
|---|---|---|---|---|---|
| chb01/chb21 | 75 (0) | 73:22:57 | 11 | 327 | 2133 |
| chb02 | 36 (0) | 35:15:59 | 3 | 91 | 1025 |
| chb03 | 38 (0) | 38:00:06 | 7 | 210 | 1104 |
| chb04 | 42 (0) | 156:03:54 | 4 | 115 | 4536 |
| chb05 | 39 (0) | 39:00:10 | 5 | 149 | 1134 |
| chb06 | 18 (0) | 66:44:06 | 10 | 297 | 1939 |
| chb07 | 19 (0) | 67:03:08 | 3 | 89 | 1949 |
| chb08 | 20 (0) | 20:00:23 | 5 | 149 | 581 |
| chb09 | 19 (0) | 67:52:18 | 4 | 123 | 1972 |
| chb10 | 25 (0) | 50:01:24 | 9 | 208 | 1454 |
| chb11 | 35 (0) | 34:47:37 | 3 | 80 | 1011 |
| chb12 | 21 (3) | 18:19:10 | 27 | 1199 | 601 |
| chb13 | 32 (1) | 32:00:00 | 10 | 357 | 959 |
| chb14 | 26 (0) | 26:00:00 | 8 | 220 | 756 |
| chb15 | 37 (3) | 37:00:18 | 11 | 595 | 1163 |
| chb16 | 17 (2) | 17:00:00 | 3 | 297 | 552 |
| chb17 | 21 (0) | 21:00:24 | 4 | 70 | 610 |
| chb18 | 36 (0) | 35:38:05 | 6 | 179 | 1036 |
| chb19 | 30 (0) | 29:55:46 | 4 | 108 | 870 |
| chb20 | 29 (0) | 27:36:06 | 8 | 256 | 802 |
| chb22 | 31 (0) | 31:00:11 | 3 | 92 | 901 |
| chb23 | 9 (0) | 26:33:30 | 7 | 201 | 772 |
| Total | 655 (9) | 950:15:32 | 155 | 5412 | 27,860 |

## References


1. GBD 2016 Epilepsy Collaborators. Global, regional, and national burden of epilepsy, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. **2019**, *18*, 357–375. [CrossRef]
2. Fisher, R.S.; Acevedo, C.; Arzimanoglou, A.; Bogacz, A.; Cross, J.H.; Elger, C.E.; Engel, J., Jr.; Forsgren, L.; French, J.A.; Glynn, M.; et al. ILAE Official Report: A Practical Clinical Definition of Epilepsy. *Epilepsia* **2014**, *55*, 475–482. [CrossRef]
3. Baumgartner, C.; Koren, J.P. Seizure Detection Using Scalp-EEG. *Epilepsia* **2018**, *59*, 14–22. [CrossRef]
4. Rasheed, K.; Qayyum, A.; Qadir, J.; Sivathamboo, S.; Kwan, P.; Kuhlmann, L.; O'Brien, T.; Razi, A. Machine Learning for Predicting Epileptic Seizures Using EEG Signals: A Review. *IEEE Rev. Biomed. Eng*. **2021**, *14*, 139–155. [CrossRef]
5. Shoeibi, A.; Khodatars, M.; Ghassemi, N.; Jafari, M.; Moridian, P.; Alizadehsani, R.; Panahiazar, M.; Khozeimeh, F.; Zare, A.; Hosseini-Nejad, H.; et al. Epileptic Seizures Detection Using Deep Learning Techniques: A Review. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5780. [CrossRef]
6. Shafiezadeh, S.; Duma, G.M.; Mento, G.; Danieli, A.; Antoniazzi, L.; Del Popolo Cristaldi, F.; Bonanni, P.; Testolin, A. Methodological Issues in Evaluating Machine Learning Models for EEG Seizure Prediction: Good Cross-Validation Accuracy Does Not Guarantee Generalization to New Patients. *Appl. Sci*. **2023**, *13*, 4262. [CrossRef]
7. Dan, J.; Pale, U.; Amirshahi, A.; Cappelletti, W.; Ingolfsson, T.M.; Wang, X.; Cossettini, A.; Bernini, A.; Benini, L.; Beniczky, S.; et al. SzCORE: Seizure Community Open-Source Research Evaluation Framework for the Validation of Electroencephalography-Based Automated Seizure Detection Algorithms. *Epilepsia* **2024**, *66*, 14–24. [CrossRef]
8. Akin, M. Comparison of Wavelet Transform and FFT Methods in the Analysis of EEG Signals. *J. Med. Syst*. **2002**, *26*, 241–247. [CrossRef]
9. Tzallas, A.T.; Tsipouras, M.G.; Fotiadis, D.I. Epileptic Seizure Detection in EEGs Using Time–Frequency Analysis. *IEEE Trans. Inf. Technol. Biomed*. **2009**, *13*, 703–710. [CrossRef]
10. Morales, S.; Bowers, M.E. Time-Frequency Analysis Methods and Their Application in Developmental EEG Data. *Dev. Cogn. Neurosci*. **2022**, *54*, 101067. [CrossRef]
11. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.-C.; Tung, C.C.; Liu, H.H. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proc. R. Soc. Lond. A* **1998**, *454*, 903–995. [CrossRef]
12. Wu, Z.; Huang, N.E. Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method. *Adv. Adapt. Data Anal*. **2009**, *1*, 1–41. [CrossRef]

13. Torres, M.E.; Colominas, M.A.; Schlotthauer, G.; Flandrin, P. A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 4144–4147. [CrossRef]
14. Rilling, G.; Flandrin, P. One or Two Frequencies? The Empirical Mode Decomposition Answers. *IEEE Trans. Signal Process.* **2008**, *56*, 85–95. [CrossRef]
15. Lin, L.; Wang, Y.; Zhou, H. Iterative Filtering as an Alternative Algorithm for Empirical Mode Decomposition. *Adv. Adapt. Data Anal.* **2009**, *1*, 543–560. [CrossRef]
16. Cicone, A.; Liu, J.; Zhou, H. Adaptive Local Iterative Filtering for Signal Decomposition and Instantaneous Frequency Analysis. *Appl. Comput. Harmon. Anal.* **2016**, *41*, 384–411. [CrossRef]
17. Barbarino, G.; Cicone, A. Stabilization and Variations to the Adaptive Local Iterative Filtering Algorithm: The Fast Resampled Iterative Filtering Method. *Numer. Math.* **2024**, *156*, 395–433. [CrossRef]
18. Daubechies, I.; Lu, J.; Wu, H.-T. Synchrosqueezed Wavelet Transforms: An Empirical Mode Decomposition-Like Tool. *Appl. Comput. Harmon. Anal.* **2011**, *30*, 243–261. [CrossRef]
19. Thakur, G.; Brevdo, E.; Fučkar, N.S.; Wu, H.-T. The Synchrosqueezing Algorithm for Time-Varying Spectral Analysis: Robustness Properties and New Paleoclimate Applications. *Signal Process.* **2013**, *93*, 1079–1094. [CrossRef]
20. Cura, O.K.; Akan, A. Classification of Epileptic EEG Signals Using Synchrosqueezing Transform and Machine Learning. *Int. J. Neural Syst.* **2021**, *31*, 2150005. [CrossRef]
21. Ozdemir, M.A.; Cura, O.K.; Akan, A. Epileptic EEG Classification by Using Time-Frequency Images for Deep Learning. *Int. J. Neural Syst.* **2021**, *31*, 2150026. [CrossRef]
22. Amiri, M.; Aghaeinia, H.; Amindavar, H.R. Automatic Epileptic Seizure Detection in EEG Signals Using Sparse Common Spatial Pattern and Adaptive Short-Time Fourier Transform-Based Synchrosqueezing Transform. *Biomed. Signal Process. Control* **2023**, *79*, 104022. [CrossRef]
23. Bajaj, V.; Pachori, R.B. Classification of Seizure and Nonseizure EEG Signals Using Empirical Mode Decomposition. *IEEE Trans. Inf. Technol. Biomed.* **2012**, *16*, 1135–1142. [CrossRef]
24. Zeng, W.; Zhang, M.; Shan, L.; Chen, Y.; Li, Z.; Du, S. Interpretable Classification of Epileptic EEG Signals Using ALIF Decomposition and Attention-Augmented Cascaded Deep Neural Networks. *Appl. Soft Comput.* **2025**, *180*, 113211. [CrossRef]
25. Cicone, A.; Wu, H.-T. Convergence Analysis of Adaptive Locally Iterative Filtering and SIFT Method. *arXiv* **2020**, arXiv:2005.04578. [CrossRef]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
28. Song, Y.; Zheng, Q.; Liu, B.; Gao, X. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 710–719. [CrossRef]
29. Li, C.; Huang, X.; Song, R.; Qian, R.; Liu, X.; Chen, X. EEG-Based Seizure Prediction via Transformer Guided CNN. *Measurement* **2022**, *203*, 111948. [CrossRef]
30. Lian, J.; Xu, F. Epileptic EEG Classification via Graph Transformer Network. *Int. J. Neural Syst.* **2023**, *33*, 2350042. [CrossRef]
31. Peng, R.; Du, Z.; Zhao, C.; Luo, J.; Liu, W.; Chen, X.; Wu, D. Multi-Branch Mutual-Distillation Transformer for EEG-Based Seizure Subtype Classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2024**, *32*, 831–839. [CrossRef]
32. Xiao, T.; Wang, Z.; Zhang, Y.; Lv, H.; Wang, S.; Feng, H.; Zhao, Y. Self-Supervised Learning with Attention Mechanism for EEG-Based Seizure Detection. *Biomed. Signal Process. Control* **2024**, *87*, 105464. [CrossRef]
33. Vafaei, E.; Hosseini, M. Transformers in EEG Analysis: A Review of Architectures and Applications in Motor Imagery, Seizure, and Emotion Classification. *Sensors* **2025**, *25*, 1293. [CrossRef]
34. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and Explainability of Artificial Intelligence in Medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef]
35. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
36. Dragomiretskiy, K.; Zosso, D. Variational Mode Decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [CrossRef]
37. Khaldi, K.; Alanazi, A.; Alanazi, I.; Almenwer, S.; Mohamed, A. SS-RIME: A Scale-Stabilized Approach to EEG Cognitive Workload Classification. *Sensors* **2026**, *26*, 2679. [CrossRef]
38. Lu, X.; Wang, T.; Ye, M.; Huang, S.; Wang, M.; Zhang, J. Study on Characteristic of Epileptic Multi-Electroencephalograph Base on Hilbert–Huang Transform and Brain Network Dynamics. *Front. Neurosci.* **2023**, *17*, 1117340. . [CrossRef]

39. Wu, X.; Zhang, T.; Zhang, L.; Qiao, L. Epileptic Seizure Prediction Using Successive Variational Mode Decomposition and Transformers Deep Learning Network. *Front. Neurosci.* **2022**, *16*, 982541. [CrossRef]
40. Auger, F.; Flandrin, P. Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method. *IEEE Trans. Signal Process.* **1995**, *43*, 1068–1089. [CrossRef]
41. Oberlin, T.; Meignen, S.; Perrier, V. Second-Order Synchrosqueezing Transform or Invertible Reassignment? Towards Ideal Time-Frequency Representations. *IEEE Trans. Signal Process.* **2015**, *63*, 1335–1344. [CrossRef]
42. Acharya, U.R.; Oh, S.L.; Hagiwara, Y.; Tan, J.H.; Adeli, H. Deep Convolutional Neural Network for the Automated Detection and Diagnosis of Seizure Using EEG Signals. *Comput. Biol. Med.* **2018**, *100*, 270–278. [CrossRef]
43. Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. *Hum. Brain Mapp.* **2017**, *38*, 5391–5420. [CrossRef]
44. 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.* **2018**, *15*, 056013. [CrossRef]
45. 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.* **2018**, *105*, 104–111. [CrossRef]
46. Roy, Y.; Banville, H.; Albuquerque, I.; Gramfort, A.; Falk, T.H.; Faubert, J. Deep Learning-Based Electroencephalography Analysis: A Systematic Review. *J. Neural Eng.* **2019**, *16*, 051001. [CrossRef]
47. Zhang, X.; Yao, L.; Dong, M.; Liu, Z.; Zhang, Y.; Li, Y. Adversarial Representation Learning for Robust Patient-Independent Epileptic Seizure Detection. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2852–2859. [CrossRef]
48. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
49. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 3319–3328.
50. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
51. Jain, S.; Wallace, B.C. Attention Is Not Explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 3543–3556.
52. Wiegreffe, S.; Pinter, Y. Attention Is Not Not Explanation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 11–20.
53. Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.-R. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2660–2673. [CrossRef]
54. Cicone, A.; Zhou, H. Numerical Analysis for Iterative Filtering with New Efficient Implementations Based on FFT. *Numer. Math.* **2021**, *147*, 1–28. [CrossRef]
55. Boashash, B. Estimating and Interpreting the Instantaneous Frequency of a Signal. I. Fundamentals. *Proc. IEEE* **1992**, *80*, 520–538. [CrossRef]
56. Lilly, J.M.; Olhede, S.C. Generalized Morse Wavelets as a Superfamily of Analytic Wavelets. *IEEE Trans. Signal Process.* **2012**, *60*, 6036–6041. [CrossRef]
57. Xiong, R.; Yang, Y.; He, D.; Zheng, K.; Zheng, S.; Xing, C.; Zhang, H.; Lan, Y.; Wang, L.; Liu, T. On Layer Normalization in the Transformer Architecture. In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; pp. 10524–10533.
58. Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-Supervised Nets. In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS), San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
59. 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.
60. Guttag, J. CHB-MIT Scalp EEG Database (Version 1.0.0). PhysioNet, 2010. Available online: https://physionet.org/content/chbmit/1.0.0/ (accessed on 5 June 2026).
61. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, e215–e220. [CrossRef]
62. Andrzejak, R.G.; Lehnertz, K.; Mormann, F.; Rieke, C.; David, P.; Elger, C.E. Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity: Dependence on Recording Region and Brain State. *Phys. Rev. E* **2001**, *64*, 061907. [CrossRef]
63. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]

64. Nadeau, C.; Bengio, Y. Inference for the Generalization Error. *Mach. Learn.* **2003**, *52*, 239–281. . [CrossRef]
65. Bouckaert, R.R.; Frank, E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2004)*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3056, pp. 3–12.
66. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
67. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
68. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
69. Ali, E.; Angelova, M.; Karmakar, C. Epileptic Seizure Detection Using CHB-MIT Dataset: The Overlooked Perspectives. *R. Soc. Open Sci.* **2024**, *11*, 230601. [CrossRef]
70. Farooq, S.S.; Rehman, A.; Baik, J.; Park, S.; Lee, H. Patient-Independent Hybrid Generative-Discriminative Modeling for Seizure Detection in Long-Term Scalp EEG. *Sci. Rep.* **2026**, *16*, 27318. [CrossRef]
71. Rehman, A.; Mun, S. Explainable AI-Enhanced Ensemble Protocol Using Gradient-Boosted Models for Zero-False-Alarm Seizure Detection from EEG. *Sensors* **2026**, *26*, 863. [CrossRef]
72. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
73. Baraniuk, R.G.; Flandrin, P.; Janssen, A.J.E.M.; Michel, O.J.J. Measuring Time-Frequency Information Content Using the Rényi Entropies. *IEEE Trans. Inf. Theory* **2001**, *47*, 1391–1409. [CrossRef]
74. Shah, V.; von Weltin, E.; Lopez, S.; McHugh, J.R.; Veloso, L.; Golmohammadi, M.; Obeid, I.; Picone, J. The Temple University Hospital Seizure Detection Corpus. *Front. Neuroinform.* **2018**, *12*, 83. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.