Spectral Gradients: Disentangled Time-Frequency Attributions to Explain Neural Networks for Time Series

Guido Gagliardi[a,*], Antonio Luca Alfeo[c,d], Mario G. C. A. Cimino[e], Wojciech Samek[f,g], Maarten De Vos[a,b]

[a]*Department of Electrical Engineering (ESAT), STADIUS, KU Leuven, Kasteelpark Arenberg 10, box 2446, Leuven, 3001, Belgium*
[b]*Department of Development and Regeneration, KU Leuven, Herestraat 49, box 805, Leuven, 3000, Belgium*
[c]*Department of Theoretical and Applied Sciences, eCampus University, Via Isimbardi 10, Novedrate (CO), 22060, Italy*
[d]*SMARTEST Research Center, eCampus University, Via Isimbardi 10, Novedrate (CO), 22060, Italy*
[e]*Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, Pisa, 56122, Italy*
[f]*Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, Berlin, 10587, Germany*
[g]*Department of Electrical Engineering and Computer Science, Technische Universität Berlin, Marchstraße 23, Berlin, 10587, Germany*

Abstract

Explainability is required to deploy deep learning models for time-series classification in high-stakes domains such as medicine. Most post-hoc attribution methods operate in the time domain and do not account for the spectral characteristics of the input, even when the model-relevant information lies in specific frequency bands. Existing time-frequency approaches based on the short-time Fourier transform (STFT) are constrained by the Gabor uncertainty principle, which imposes a fixed trade-off between temporal and spectral resolution. We propose Spectral Gradients (SG), a model-agnostic attribution method that produces joint time-frequency explanations by progressively ablating frequency bands by zeroing DFT coefficients and distributing band importance over time through path-integrated gradients. Because SG does not rely on windowed transforms, its frequency and time resolutions are set independently and are not coupled through an analysis window, hence the Gabor trade-off becomes a single band-width choice. We evaluate SG on 7 case studies spanning synthetic signals, ECG arrhythmia, speech commands and EEG sleep staging, against nine STFT-based configurations. Across 153

*Corresponding author
Email address: `guido.gagliardi@kuleuven.be` (Guido Gagliardi)

comparisons, no STFT configuration is significantly better than SG on both the time and the frequency variant of a metric simultaneously. SG reaches a single, window-free operating point that is competitive on both axes. It is the best method on 21 of 34 metric–domain pairs and never significantly worse than the assumption-free balanced window on 33 of 34.
Keywords: Explainable AI, Time-series classification, Time-frequency attribution, Spectral Gradients, Integrated Gradients, Gabor uncertainty principle

1. Introduction

Deep learning has become the dominant approach to time-series classification, achieving strong results across a wide range of tasks, from cardiac arrhythmia detection to automated sleep staging [1, 2, 3]. As these models are deployed in high-stakes domains such as medicine and safety-critical monitoring, explaining their predictions has become a practical requirement. Regulators, clinicians, and end-users increasingly demand explanations that are faithful to the model's internal reasoning while remaining interpretable in the terms of the application domain [4].

Many post-hoc attribution methods have been proposed for this purpose [5]. Gradient-based approaches such as Saliency maps and Input×Gradient [6, 7] exploit model differentiability to assign importance scores to individual input features. Layer-wise Relevance Propagation [8] obtains them through a modified backward pass that redistributes the output layer by layer [9]. Path-based methods such as Integrated Gradients [10] and its extensions [11] improve on simple gradients by accumulating them along a trajectory from a reference baseline to the input, satisfying axiomatic properties like completeness; their sensitivity to that baseline is well documented and has motivated weighted multi-baseline variants [12]. Model-agnostic perturbation-based methods, including SHAP [13] and LIME [14], estimate feature importance by observing how predictions change when subsets of the input are masked or perturbed. Despite their diversity, all these methods share a common limitation when applied to time-series data: they produce attributions in the time domain, assigning an importance score to each time point. This format is natural for image classification, where each pixel has a spatial meaning, but it is often inadequate for signals whose discriminative content can be spectral rather than temporal.

In many application domains, the features a classifier's decisions depend on are better characterised by their frequency content than by their temporal position [15]. In automated sleep staging, deep sleep (NREM stage N3) is defined by the presence of high-amplitude δ waves in the 0.5–4 Hz band, while sleep spindles in stage N2 occupy the σ band (12–16 Hz) [16, 17]. A time-domain attribution can reveal when the model attends to the signal,. yet it cannot reveal which frequency band the model considers relevant. This limits clinicians' ability to validate whether the model relies on the correct physiological features. Similarly, in speech processing phonemes are characterised by their spectral formants [18], and in cardiac electrophysiology arrhythmia subtypes are associated with distinct morphological patterns that manifest as frequency-domain signatures [19]. When a model's sensitivity is distributed across a frequency band rather than concentrated at specific time points, time-domain attributions become diffuse and difficult to interpret.

Explainability for time series has mostly been pursued through model design instead of post-hoc attribution. MTS2Graph [20] exposes the temporal evolution of the features a CNN learns from multivariate series. TimeVQVAE-AD [21] scores anomalies in a discrete time-frequency latent space so that they can be read per frequency band. JoCE [22] explains anomaly detectors by counterfactual repair of jointly selected temporal and feature regions. Each is tied to its model or task and a post-hoc attribution through which any trained classifier can be read on both axes remains open.

One approach to joint time-frequency attribution is to apply gradient methods through an explicit STFT–iSTFT pipeline, computing attributions in the STFT coefficient space [23]. The resolution of these explanations, however, is constrained by the Gabor uncertainty principle: improving frequency precision requires longer analysis windows, which degrade temporal resolution, and vice versa. The user must therefore choose a window size that favours one domain over the other. Rezaei et al. [24] have recently formalised this limitation, showing that under the uncertainty principle, time-domain and frequency-domain attributions can highlight features that are not direct counterparts of one another, establishing the need for jointly presenting both domains.

Frequency-only alternatives such as FreqATT [25], which applies perturbation-based occlusion in the spectral domain, confirm the growing interest in spectral explanations but produce one-dimensional frequency attributions instead of joint time-frequency maps.

Addresing the Gabor uncertainty principle we propose Spectral Gradients (SG). Instead of computing gradients through a windowed transform, SG constructs an integration path by progressively inserting frequency bands by zeroing DFT coefficients and distributing the resulting band importance over time through path-integrated gradients. Because the DFT operates globally on the entire signal, the frequency resolution (Δf) and time resolution ($1/f_s$) are set independently and are not coupled through an analysis window. The joint resolution limit is not violated, and it is reparameterised into a single band-width choice.

The main contributions of this work are:

1. We introduce SG, a post-hoc attribution method that produces joint time-frequency explanations without relying on windowed transforms, decoupling time and frequency resolution so that the frequency-analysis grid is set independently of the sampling grid; a single analysis window no longer constrains both.
2. We formulate a path-independent variant ($\overline{SG}$) that eliminates dependence on the spectral ablation order by averaging over random permutations, a construction inspired by Shapley values in cooperative game theory.
3. We conduct an extensive evaluation across seven case studies spanning four signal modalities (synthetic, ECG, speech, and EEG), comparing SG against nine STFT-based baselines on frequency-domain and time-domain metrics evaluated independently.
4. We provide empirical evidence that the Gabor trade-off manifests in the explainability setting: across 153 comparisons, no STFT configuration achieves significantly better performance than SG on both frequency and time metrics for the same evaluation criterion.

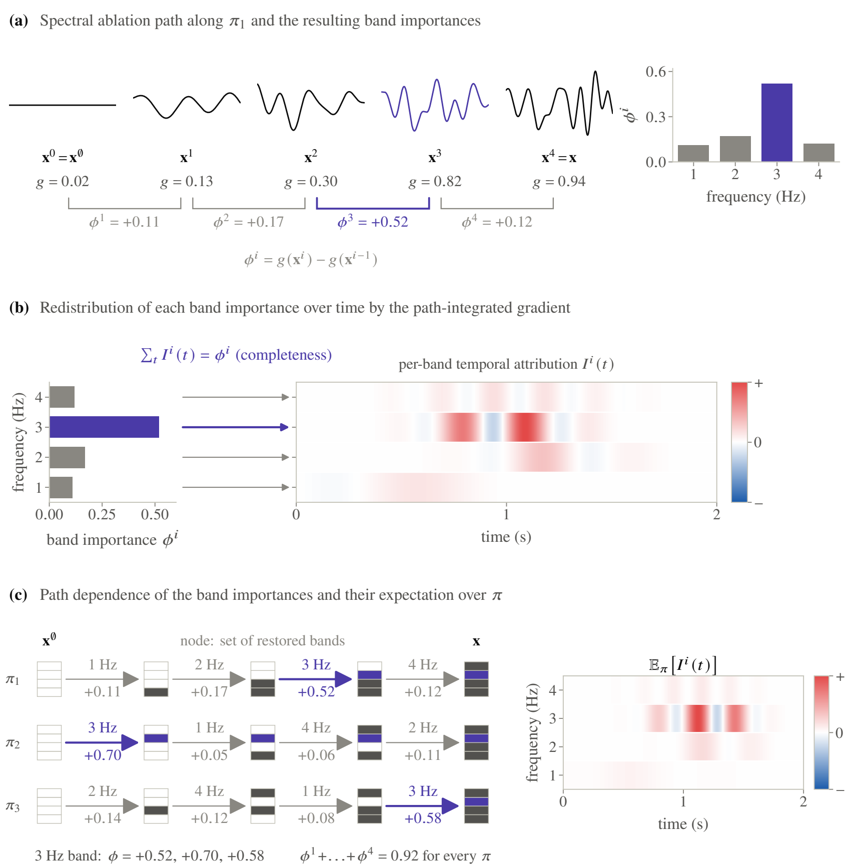

Figure 1: **Spectral Gradients computation schema**. **(a)**A spectral ablation path π_1 starts from the silent baseline $\mathbf{x}^\emptyset$ and restores one frequency band per step until the signal x is complete; at step i, the band's importance ϕ^i is the change in the model score g between consecutive path steps. **(b)** Each ϕ^i is redistributed over time by the path-integrated gradient forming the time-frequency attribution map. **(c)** Dependency by the specific ablation path π_1 is removed by computing the expected value of the time-frequency attribution map considering random permutations of the path steps.

2. Methods

2.1. Methodological Intuition

Path-based explanation methods attribute a model's output $g(\cdot)$ at input $\mathbf{x}$ by comparing it to a baseline $\mathbf{x}^{\emptyset}$, with total relevance given by $g(\mathbf{x}) - g(\mathbf{x}^{\emptyset})$. To distribute this relevance across input features $\mathbf{x} \in \mathbb{R}^n$, a path $r : [0, 1] \to \mathbb{R}^n$ from $\mathbf{x}^{\emptyset}$ to $\mathbf{x}$ is defined, and feature-wise attributions are computed via path-integrated gradients:

$$I(\mathbf{x}, g, i) = \int_0^1 \frac{\partial g(r(\alpha))}{\partial x_i} \cdot \frac{dr_i(\alpha)}{d\alpha} \, d\alpha \tag{1}$$

where $i \in \{0, 1, \ldots, n-1\}$. By the fundamental theorem of calculus, the attributions satisfy $\sum_{i=0}^{n-1} I(\mathbf{x}, g, i) = g(\mathbf{x}) - g(\mathbf{x}^{\emptyset})$, thus distributing the output difference across input features. Each $I(\mathbf{x}, g, i)$ quantifies the contribution of the difference $\mathbf{x}[i] - \mathbf{x}^{\emptyset}[i]$ to the model's output, and vanishes if $\mathbf{x}[i] = \mathbf{x}^{\emptyset}[i]$.

We extend this framework by constructing the path in the *spectral domain*. Instead of interpolating between $\mathbf{x}^{\emptyset}$ and $\mathbf{x}$ in the space of time samples, we build a *spectral ablation path* $\mathbf{x}^0 = \mathbf{x}^{\emptyset} \to \mathbf{x}^1 \to \cdots \to \mathbf{x}^{b-1} \to \mathbf{x}^b = \mathbf{x}$ of b steps. Each intermediate signal $\mathbf{x}^i$ is obtained by progressively restoring one frequency band in the DFT representation. At each step, the change in model output reflects the relevance of the restored band, which we further localise in time by computing path-integrated gradients between consecutive intermediates. Figure 1(a,b) illustrates this construction on a schematic case.

2.2. Spectral Gradients

Spectral Gradients (SG) is a model-agnostic attribution method that produces joint time-frequency importance maps by decomposing model sensitivity along a spectral ablation path.

Consider a neural network for time-series classification taking as input a signal $x(t) \in \mathbb{R}$ of duration s seconds, sampled at rate f_s, yielding a discrete signal $\mathbf{x} = [x(0), x(1), \ldots, x(n-1)] \in \mathbb{R}^n$ with $n = s \cdot f_s$. The network defines a function $g : \mathbb{R}^n \to \mathbb{R}$, i.e. the class-$c$ confidence. Let $\mathbf{X} = \mathrm{DFT}(\mathbf{x}) \in \mathbb{C}^{n/2+1}$ denote the one-sided discrete Fourier transform of $\mathbf{x}$. We partition the frequency axis into b contiguous bands $B = \{B_1, B_2, \ldots, B_b\}$, each spanning Δf Hz, such that $\bigcup_i B_i$ covers $[0, f_s/2]$.

To suppress a frequency band B_i, we zero out the corresponding DFT coefficients and reconstruct via inverse DFT:

$$\mathbf{x}' = \mathrm{iDFT}(\mathbf{X} \odot \mathbf{m}), \quad m_k = \begin{cases} 0 & \text{if } k \in B_i \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

Reconstruction itself is lossless: the DFT and its inverse form an exact pair, so $\mathrm{iDFT}(\mathrm{DFT}(\mathbf{x})) = \mathbf{x}$. The band suppression, by contrast, is exact only on the DFT grid. This procedure nulls the chosen bins exactly, but, since a bin is not a sinusoid, then a component whose frequency is not an integer multiple of the bin spacing f_s/n is not confined to a single bin, and its energy leaks across the spectrum with a Dirichlet-kernel envelope. Masking a band therefore removes such off-grid components only approximately: it partly erases tones centred inside the band and clips the leakage tails of tones centred outside it. The sharp spectral cut also introduces Gibbs-type ringing in time. The discretisation of the frequency axis at spacing f_s/n (and the chosen band width $\Delta f \geq f_s/n$) thus sets the granularity at which frequency content can actually be isolated. The baseline $\mathbf{x}^{\emptyset} = \mathrm{iDFT}(\mathbf{0})$ is the silent (zero) signal.

Given a permutation π of the b bands, we construct a spectral ablation path by progressively inserting bands in the order prescribed by π, starting from $\mathbf{x}^{\emptyset}$. The importance of the i-th band along this path is:

Definition 1 (Band-Importance Φ along $\mathbf{x}^{\emptyset} \xrightarrow{\pi} \mathbf{x}$)**.**

$$\begin{aligned} \Phi_\pi(g, \mathbf{x}) &= [\phi_\pi(g, \mathbf{x}^1), \ldots, \phi_\pi(g, \mathbf{x}^b)] \\ \phi_\pi(g, \mathbf{x}^i) &= g(\mathbf{x}^i) - g(\mathbf{x}^{i-1}) \\ \mathbf{x}^i &= iDFT\Big(\mathbf{X} \odot \mathbf{m}_\pi^{(i)}\Big) \\ \mathbf{x}^0 &= \mathbf{x}^{\emptyset}, \quad \mathbf{x}^b = \mathbf{x} \end{aligned} \tag{3}$$

where $\mathbf{m}_\pi^{(i)}$ is the binary mask that retains only the first i bands in the order π, and $\Phi_\pi : \mathbb{R}^n \to \mathbb{R}^b$. The frequency resolution of SG depends solely on the band width Δf and is independent of the time resolution, which remains at the original sampling rate $1/f_s$.

To distribute each band's importance across time, we compute path-integrated gradients between consecutive intermediates:

Definition 2 (Spectral Gradients along π)**.**

$$SG_\pi(g, \mathbf{x}, t)^i = W_\pi(\mathbf{x})_i \cdot I_\pi(g, \mathbf{x}, t)^i \tag{4}$$

With:

$$\begin{aligned} I_\pi(g, \mathbf{x}, t)^i &= \int_0^1 \frac{\partial g(r(\alpha))}{\partial x_t} \cdot \frac{dr_t(\alpha)}{d\alpha}\, d\alpha \\ W_\pi(\mathbf{x})_i &= \frac{\left|\phi_\pi(g, \mathbf{x}^i)\right|}{\sum_{j=1}^{b} \left|\phi_\pi(g, \mathbf{x}^j)\right|} \end{aligned} \tag{5}$$

where $r : [0, 1] \to \mathbb{R}^n$ is a straight-line path from $\mathbf{x}^{i-1}$ to $\mathbf{x}^i$, and $W_\pi(\mathbf{x})_i$ weights time-domain gradients by the band's importance. The weight is the *magnitude* of the band's effect on the target confidence, hence sign-agnostic: a band that lowers the confidence is as informative about the decision as one that raises it, and only bands with no effect are suppressed. The denominator normalises the weights to sum to 1, making W_π a convex combination.

By the fundamental theorem of calculus, the per-step path-integrated gradient I_π satisfies the Completeness Axiom, distributing each band's importance ϕ_π exactly across time:

Lemma 1 (Completeness)**.**

$$\sum_{t=0}^{n-1} I_\pi(g, \mathbf{x}, t)^i = g(\mathbf{x}^i) - g(\mathbf{x}^{i-1}) = \phi_\pi(g, \mathbf{x}^i) \tag{6}$$

Proof. By definition, $r : [0, 1] \to \mathbb{R}^n$ is the straight-line path from $\mathbf{x}^{i-1}$ to $\mathbf{x}^i$, so $r(\alpha) = \mathbf{x}^{i-1} + \alpha(\mathbf{x}^i - \mathbf{x}^{i-1})$ and $\frac{dr_t}{d\alpha} = x_t^i - x_t^{i-1}$. Summing the path-integrated gradients over all time points:

$$\sum_{t=0}^{n-1} I_\pi(g, \mathbf{x}, t)^i = \int_0^1 \sum_{t=0}^{n-1} \frac{\partial g(r(\alpha))}{\partial x_t}(x_t^i - x_t^{i-1})\, d\alpha = \int_0^1 \frac{d}{d\alpha} g(r(\alpha))\, d\alpha \tag{7}$$

where the last equality follows from the chain rule. The fundamental theorem of calculus then gives $\int_0^1 \frac{d}{d\alpha} g(r(\alpha))\, d\alpha = g(r(1)) - g(r(0)) = g(\mathbf{x}^i) - g(\mathbf{x}^{i-1}) = \phi_\pi(g, \mathbf{x}^i)$. $\square$

The full SG attribution is a tensor $SG_\pi(g, \mathbf{x}) \in \mathbb{R}^{b\times n}$, computed as $SG_\pi = W_\pi \cdot I_\pi$, the product being element-wise along the band dimension. The band reweighting W_π does not preserve completeness. The weights $W_i \propto |\phi_\pi^i|$ are non-negative and renormalised to sum to one, so the reweighted map sums to a convex combination of the band importances, $\sum_{i,t} SG_\pi = \sum_i W_i\, \phi_\pi^i = \left(\sum_i (\phi_\pi^i)^2\right) / \sum_j |\phi_\pi^j|$, and not to $g(\mathbf{x}) - g(\mathbf{x}^\emptyset) = \sum_i \phi_\pi^i$. Completeness thus holds for the per-band term I_π (Lemma 1) but not for the reweighted attribution, which we treat as a relevance map rather than a conservative decomposition. The result has full temporal resolution ($1/f_s$) and a configurable frequency resolution (Δf).

2.3. Path-Independent Spectral Gradients

The attribution SG_π depends on the permutation π, i.e., the order in which frequency bands are inserted along the ablation path. To eliminate this path dependence, we define the path-independent Spectral Gradients as the expected value over all possible permutations:

Definition 3 (Path-Independent Spectral Gradients)**.**

$$\overline{SG}(g, \mathbf{x}) = \mathbb{E}_{\pi\sim\mathrm{Unif}(\Pi_b)} \left[SG_\pi(g, \mathbf{x}) \right] \tag{8}$$

where Π_b is the set of all $b!$ permutations of the b frequency bands. Averaging the band importances ϕ_π over uniform permutations yields, for each band, its *Shapley value* [26] in the cooperative game whose players are the frequency bands and whose payoff is the class confidence: this marginal-contribution-over-coalitions average satisfies efficiency, $\sum_i \mathbb{E}_\pi[\phi_\pi^i] = g(\mathbf{x}) - g(\mathbf{x}^\emptyset)$, by the telescoping of Φ_π. The attribution $\overline{SG}$ uses the same permutation-averaging construction but does not satisfy the corresponding axioms: the magnitude weighting in W_π and its per-permutation renormalisation break efficiency, linearity, and the null-player property, so $\overline{SG}$ is Shapley-*inspired* rather than a Shapley value. Averaging does guarantee order-independence, which is exact in the limit of all

Algorithm 1 Path-independent Spectral Gradients ($\overline{SG}$)

Require: signal $\mathbf{x} \in \mathbb{R}^n$, model g, band width Δf, permutations K, integration steps s

$\mathbf{X} \leftarrow \mathrm{DFT}(\mathbf{x})$; partition $[0, f_s/2]$ into $b = \lceil (f_s/2)/\Delta f \rceil$ bands
$A \leftarrow \mathbf{0} \in \mathbb{R}^{b \times n}$
for $k = 1$ to K **do**
  sample permutation π of the b bands uniformly; $\mathbf{x}^0 \leftarrow \mathrm{iDFT}(\mathbf{0})$
  for $i = 1$ to b **do**
    $\mathbf{x}^i \leftarrow \mathrm{iDFT}(\mathbf{X} \odot \mathbf{m}_\pi^{(i)})$ {restore the first i bands of π}
    $\phi^i \leftarrow g(\mathbf{x}^i) - g(\mathbf{x}^{i-1})$ {band importance}
    $I^i \leftarrow \frac{1}{s} \sum_{j=1}^{s} \nabla_{\mathbf{x}} g(\mathbf{x}^{i-1} + \frac{j}{s}(\mathbf{x}^i - \mathbf{x}^{i-1})) \odot (\mathbf{x}^i - \mathbf{x}^{i-1})$ {IG, s steps}
  end for
  $W_i \leftarrow |\phi^i| / \sum_j |\phi^j| \ \ \forall i$; reorder $\{W_i I^i\}_i$ to ascending frequency
  $A \leftarrow A + \frac{1}{K} [\, W_i \, I^i \,]_i$
end for
return time-frequency attribution $A \in \mathbb{R}^{b \times n}$

$b!$ permutations; Figure 1(c) shows three paths through the same band lattice: the same band is scored differently depending on when it enters, while the expectation over paths is not.

In practice, $\overline{SG}$ is approximated by sampling K permutations uniformly from Π_b and averaging the resulting attributions (Algorithm 1). The convergence of this approximation as a function of K is analysed empirically in Section 5. Each permutation performs b ablation steps, and each step approximates the path integral with s gradient evaluations, so $\overline{SG}$ costs $O(K\, b\, s)$ forward/backward passes; because the b intermediate signals of a step are processed as a single batch, wall-clock time scales as $O(K\, s)$.

2.4. STFT-Based Baselines

As baselines, we consider time-frequency attribution methods based on the short-time Fourier transform (STFT), following Vielhaben et al. [23]. Given a signal $\mathbf{x} \in \mathbb{R}^n$, these methods compute a real-valued STFT representation $\mathbf{X} = [\mathrm{Re}(\mathrm{STFT}(\mathbf{x}));\ \mathrm{Im}(\mathrm{STFT}(\mathbf{x}))] \in \mathbb{R}^{2F \times T}$, treat it as the differentiable input, reconstruct $\hat{\mathbf{x}} = \mathrm{iSTFT}(\mathbf{X})$, and compute gradient-based attributions of $g(\mathrm{iSTFT}(\mathbf{X}))$ with respect to $\mathbf{X}$. The real and imaginary attributions are summed per frequency bin to yield a time-frequency map of shape $F \times T$. With the hop set to $n_{\mathrm{fft}}/4$ (Table A.2), the analysis/synthesis windows satisfy the constant-overlap-add condition, so iSTFT $\circ$ STFT reconstructs the input up to boundary effects and the gradient is taken through a near-identity map. The STFT-coefficient attribution is therefore comparable, coefficient-for-coefficient, to the time-domain input.

We evaluate three gradient methods through this pipeline (Integrated Gradients, Saliency, and Input$\times$Gradient), each with three STFT window configurations (time-optimised, balanced, frequency-optimised) per domain, yielding $3 \times 3 = 9$ baseline configurations. The time-frequency resolution of these methods is constrained by the Gabor uncertainty principle $\Delta t \cdot \Delta f \geq 1/(4\pi)$ [27], whereas SG controls time and frequency resolution independently. The per-domain window sizes for all nine configurations are listed in Table A.2.

3. Evaluation Metrics

Since SG produces a 2D attribution map $\mathbf{A} \in \mathbb{R}^{b \times n}$ (frequency $\times$ time), we evaluate explanation quality independently along each axis using four metrics, drawn from the quantitative evaluation literature for post-hoc explanations [28, 29]. Because perturbation-based fidelity metrics are known to be statistically inconsistent across their design choices [30], we fix a single method-independent removal operator per axis and apply it identically to every explainer. For synthetic benchmarks where ground-truth masks are available, two additional localisation metrics are used (six total). All metrics are computed per sample; we report medians across test sets. Statistical significance is assessed via the two-sided Mann-Whitney U test [31] with Holm-Bonferroni correction [32] across the whole family of SG-vs-STFT comparisons ($\alpha = 0.05$). Because the full test sets are large enough to render negligible differences statistically significant, we additionally require a non-negligible effect size, Cliff's δ [33] with $|\delta| \geq 0.147$ (the negligible/small boundary of [34]), for a difference to count.

3.1. Complexity (↓)

A sparse explanation, concentrated on a few relevant features, is preferable to one spread uniformly across all of them. We measure this via the normalised Shannon entropy of the marginal attribution distribution [28].

Frequency Complexity. Let $\mathbf{a}_f \in \mathbb{R}^b$ be the frequency-marginal attribution (sum of $\mathbf{A}$ over time, one entry per band; only its magnitude enters the entropy). Define $p_i = |a_{f,i}| / \sum_j |a_{f,j}|$. The frequency complexity is:

$$C_f = \frac{-\sum_{i=1}^{b} p_i \ln p_i}{\ln b} \in [0, 1] \tag{9}$$

The normalisation by $\ln b$ ensures comparability across methods with different numbers of frequency bands.

Time Complexity. Defined analogously on the time-marginal attribution $\mathbf{a}_t \in \mathbb{R}^n$ (sum of $\mathbf{A}$ over frequency), normalised by $\ln n$.

3.2. Infidelity (↓)

Faithfulness is assessed by feature removal: a faithful ranking is one whose most-attributed features, once removed, degrade the model output fastest [35, 36, 37]. We ablate features in decreasing order of attributed importance and report the area under the normalised degradation curve. The removal operator is determined by the axis under test and applied identically to every method, so that all explanations are compared under the same, method-independent perturbation.

Frequency Infidelity. On the frequency axis, removing a feature means suppressing the spectral content of a band, which is exactly the quantity ranked by the frequency-marginal $\mathbf{a}_f$. We realise this by zeroing the band's DFT coefficients, which removes the selected components exactly on the DFT grid [38], with no filter order or transition band to tune. Ranking the b bands by $\mathbf{a}_f$, let $\mathbf{x}|_k$ be the signal with the top-k bands removed. The degradation curve $D(k) = g(\mathbf{x}|_k)/g(\mathbf{x})$ is integrated by the trapezoidal rule to a score in $[0, 1]$. A faithful ranking gives rapid degradation and an area near 0. The STFT baselines are scored under the same operator, with their time-frequency attributions marginalised to the same band grid.

Time Infidelity. On the time axis, removal means zeroing consecutive time-sample patches, ordered by the time-marginal $\mathbf{a}_t$; the score is defined as above.

3.3. *Localisation (↑, synthetic only)*

For synthetic benchmarks with ground-truth time-frequency masks, we measure the weighted ratio of positive attributions inside the mask to total positive attributions [39]:

$$\mu_w = \frac{R_{\text{in}}}{R_{\text{tot}}} \cdot \frac{S_{\text{tot}}}{S_{\text{in}}} \tag{10}$$

where R_{in} and R_{tot} are the positive attribution sums inside and total, and S_{in}, S_{tot} the number of features inside the mask and total. The second factor weights small masks more heavily. This metric is computed independently along the frequency and time axes.

4. Experimental Setup

We evaluate SG across seven case studies spanning four signal modalities (synthetic time series, ECG, speech, and EEG sleep staging). Sampling rates range from 100 Hz to 16 kHz, signal lengths from 200 to 16,000 samples, and classes per task from 5 to 35 (Table 1).

4.1. *Datasets and Models*

Table 1 summarises the datasets, models, and classification performance for each case study. The synthetic benchmarks use three controlled setups where class-discriminative information is localised in both time and frequency. Each signal (10 s, 100 Hz) contains a background noise component shared across classes and a class-discriminative component confined to a specific frequency band and randomly localised in time. Ground-truth time-frequency masks are available for supervised evaluation. The four real-world domains cover ECG arrhythmia classification on the MIT-BIH database [40], spoken word recognition on Google Speech Commands v0.02 [41], and EEG sleep staging on the MASS [42] and SleepEDF [43] datasets.

For the synthetic, ECG, and speech domains, the model to be explained is a one-dimensional convolutional classifier (Conv1D) whose architecture and training recipe are given in Table A.1. Depth, width, and the use of batch normalisation are adapted to

Table 1: Experimental setup. For each case study we report the dataset, signal properties, model, test-set performance, and the number of explained samples n_{expl} (correctly classified test samples with predicted-class confidence above 60%). The synthetic, ECG and speech classifiers were trained for this study (Table A.1); the sleep models, Chambon2018 [44] and TsinalisCNN [45], are pretrained checkpoints from PhysioEx [3]. Datasets: MIT-BIH [40], Speech Commands [41], MASS [42], SleepEDF [43]. For ECG, beats labelled unclassifiable (U) are excluded from the explained set.

	Synt-0/1/2	**ECG**	**Speech**	**Sleep-MASS**	**Sleep-EDF**
Dataset	Custom	MIT-BIH	SpeechCmd.	MASS	SleepEDF
f_s (Hz)	100	360	16,000	128	100
Length (samples)	1,000	200	16,000	3,840	3,000
Classes	5	5	35	5	5
Model	Conv1D	Conv1D	Conv1D	Chambon2018	TsinalisCNN
Parameters	1.5K	85K	795K	4.3K	6.4M
Test accuracy	98.3–99.7%	90.9%	92.9%	82.6%	77.3%
Cohen's κ	—	—	—	0.736	0.683
n_{expl}	5,000	37,346	11,005	42,835	135,257

the signal length and sampling rate, giving between 1.5K parameters on the synthetic signals and 795K on speech. These are standard convolutional baselines used only as the model under explanation. They are not a contribution of this work and follow no specific published design. We do not compare architectures across domains. For sleep staging, we use two established pretrained architectures from the PhysioEx model hub [3]. *Chambon2018* [44] is a compact temporal-spatial CNN with 4.3K parameters. *TsinalisCNN* [45] is a deeper model (6.4M parameters) whose first layer uses 200-sample filters acting as a learned spectral filterbank. Their training follows the original publications. The analysis configurations needed to reproduce every benchmark are given in Table A.2.

5. Results

Table 2: Comparison of SG against the nine STFT-based configurations (per-sample medians). * SG significantly better, † SG significantly worse (Holm-corrected two-sided Mann-Whitney U with |Cliff's δ| ≥ 0.147, Section 3). **Bold**: best median per metric and domain. Localisation on the synthetic setups is reported in Table 3.

Domain	**Expl.**	**Win.**	**Freq.Inf ↓**	**Freq.Cmplx ↓**	**Time.Inf ↓**	**Time.Cmplx ↓**
Synt-0	SG	—	0.1308	**0.6513**	**0.0901**	0.7860
	IG	time	**0.1274**	0.7663*	0.4536*	**0.7728**

Table 2 (continued)

Domain	**Expl.**	**Win.**	**Freq.Inf ↓**	**Freq.Cmplx ↓**	**Time.Inf ↓**	**Time.Cmplx ↓**
	IG	bal	0.2107*	0.7211*	0.3111*	0.8176*
	IG	freq	0.8818*	0.7165*	0.2837*	0.8985*
	Sal	time	0.7302*	0.9103*	0.8608*	0.8728*
	Sal	bal	0.8030*	0.8695*	0.8966*	0.8959*
	Sal	freq	0.8248*	0.8609*	0.8638*	0.9453*
	IxG	time	0.1610*	0.8733*	0.2856*	0.8466*
	IxG	bal	0.8026*	0.8405*	0.8819*	0.8672*
	IxG	freq	0.8912*	0.8239*	0.5417*	0.9318*
Synt-1	SG	—	0.1252	**0.6611**	**0.1961**	0.7823
	IG	time	**0.1251**	0.7677*	0.9177*	**0.7524**†
	IG	bal	0.4481*	0.7312*	0.3521*	0.8022
	IG	freq	0.9068*	0.7295*	0.2886	0.8889*
	Sal	time	0.7925*	0.8800*	0.8925*	0.8505*
	Sal	bal	0.8464*	0.8488*	0.8912*	0.8841*
	Sal	freq	0.8714*	0.8351*	0.6827*	0.9392*
	IxG	time	0.2550*	0.8805*	0.9171*	0.8232*
	IxG	bal	0.9023*	0.8437*	0.9290*	0.8612*
	IxG	freq	0.9078*	0.8225*	0.7263*	0.9356*
Synt-2	SG	—	0.1156	**0.6423**	**0.0321**	0.7878
	IG	time	**0.1034**	0.7726*	0.3360*	**0.7659**†
	IG	bal	0.1305*	0.7267*	0.2567*	0.8114*
	IG	freq	0.5764*	0.7240*	0.2642*	0.8936*
	Sal	time	0.6741*	0.8619*	0.7576*	0.8779*
	Sal	bal	0.8361*	0.8158*	0.7289*	0.8919*
	Sal	freq	0.8322*	0.8229*	0.4066*	0.9469*
	IxG	time	0.1306*	0.8649*	0.3172*	0.8603*
	IxG	bal	0.4599*	0.8219*	0.5576*	0.8924*
	IxG	freq	0.6746*	0.8061*	0.3572*	0.9463*
ECG	SG	—	0.0416	**0.2818**	0.0605	**0.8769**
	IG	time	0.2054*	0.6410*	**0.0417**	0.8952*
	IG	bal	0.0755	0.4292*	0.1168	0.9170*
	IG	freq	0.0405	0.3333*	0.1542*	0.9274*
	Sal	time	0.3626*	0.9048*	0.0615	0.9362*
	Sal	bal	0.2707*	0.8640*	0.0853	0.9514*
	Sal	freq	0.2674*	0.8634*	0.1078	0.9604*
	IxG	time	0.1399*	0.7325*	0.0444	0.9269*
	IxG	bal	0.0449	0.6012*	0.0781	0.9454*
	IxG	freq	**0.0333**	0.5635*	0.1042	0.9578*
Speech	SG	—	**0.0498**	**0.6798**	**0.0711**	**0.8666**
	IG	time	0.3198*	0.8001*	0.2654*	0.8748*
	IG	bal	0.2994*	0.7515*	0.2727*	0.8798*
	IG	freq	0.1407*	0.7118*	0.3884*	0.9055*

Table 2 (continued)

Domain	Expl.	Win.	Freq.Inf ↓	Freq.Cmplx ↓	Time.Inf ↓	Time.Cmplx ↓
	Sal	time	0.5704*	0.9184*	0.4125*	0.9459*
	Sal	bal	0.5199*	0.9203*	0.4135*	0.9525*
	Sal	freq	0.3924*	0.9236*	0.4668*	0.9651*
	IxG	time	0.2206*	0.8628*	0.1930*	0.8943*
	IxG	bal	0.1909*	0.8341*	0.1945*	0.8999*
	IxG	freq	0.1150*	0.8098*	0.3211*	0.9268*
Sleep-MASS	SG	—	0.0271	**0.5985**	**0.0259**	0.9373
	IG	time	0.0374*	0.8013*	0.0464*	0.9325†
	IG	bal	0.0321	0.7994*	0.1074*	0.9445*
	IG	freq	**0.0249**	0.7776*	0.2083*	0.9546*
	Sal	time	0.2471*	0.8864*	0.3985*	0.9336
	Sal	bal	0.7091*	0.8887*	0.5312*	0.9348
	Sal	freq	0.7750*	0.8920*	0.5836*	0.9415*
	IxG	time	0.0697*	0.8016*	0.0870*	**0.9133†**
	IxG	bal	0.0372*	0.7796*	0.1928*	0.9245†
	IxG	freq	0.0315*	0.7678*	0.2820*	0.9391
Sleep-EDF	SG	—	**0.0229**	**0.6123**	**0.0329**	0.9243
	IG	time	0.0678*	0.8620*	0.0855*	0.9260
	IG	bal	0.0290*	0.8262*	0.3154*	0.9410*
	IG	freq	0.0242	0.7942*	0.4618*	0.9507*
	Sal	time	0.0897*	0.8543*	0.6668*	0.9486*
	Sal	bal	0.7026*	0.8678*	0.7825*	0.9550*
	Sal	freq	0.6040*	0.8570*	0.7079*	0.9631*
	IxG	time	0.1178*	0.8610*	0.1663*	**0.9219**
	IxG	bal	0.0345*	0.8194*	0.4185*	0.9384*
	IxG	freq	0.0278*	0.7861*	0.5105*	0.9495*

5.1. Protocol

We evaluate $\overline{SG}$ against nine STFT-based explainers (three gradient methods × three window configurations) on seven case studies. For each domain we explain *every* correctly classified test sample whose predicted-class confidence exceeds 60% (seed 42); the resulting sample counts are listed in Table 1, and within each domain all explainers are compared on the identical sample set. All metrics are computed per sample; we report medians, with significance assessed as in Section 3 (Holm-corrected two-sided Mann-Whitney U with a non-negligible effect size, |Cliff's δ| ≥ 0.147). Table 2 presents the full results, and Table A.3 reports the corresponding effect sizes (signed Cliff's δ of SG against the best STFT rival per metric), so that every significance decision can be

Table 3: Localisation on the three synthetic setups, where ground-truth time-frequency masks are available (per-sample medians, higher is better). Markers and bold as in Table 2.

Domain	**Expl.**	**Win.**	**Freq.Loc ↑**	**Time.Loc ↑**
Synt-0	SG	—	**4.8379**	3.0551
	IG	time	3.0735*	**3.0965**
	IG	bal	3.9305*	3.0258
	IG	freq	3.9165*	2.3051*
	Sal	time	1.0816*	1.8927*
	Sal	bal	1.0954*	1.8127*
	Sal	freq	1.0724*	1.4912*
	IxG	time	1.9710*	2.4468*
	IxG	bal	3.0492*	2.7014
	IxG	freq	3.0960*	2.0644*
Synt-1	SG	—	**4.8443**	3.0425
	IG	time	3.6771*	**3.1979**
	IG	bal	3.9965*	2.9897
	IG	freq	3.9429*	2.2819*
	Sal	time	1.2499*	1.7670*
	Sal	bal	1.2233*	1.6804*
	Sal	freq	1.2063*	1.3732*
	IxG	time	2.9700*	2.6036
	IxG	bal	3.2841*	2.4416*
	IxG	freq	3.2790*	1.9172*
Synt-2	SG	—	**4.7695**	**2.9440**
	IG	time	3.0783*	2.9100
	IG	bal	3.5440*	2.8486
	IG	freq	3.5043*	2.1934*
	Sal	time	1.5128*	1.7102*
	Sal	bal	1.5213*	1.6545*
	Sal	freq	1.4876*	1.3806*
	IxG	time	2.4836*	2.3559*
	IxG	bal	2.9850*	2.3587*
	IxG	freq	2.9738*	1.8551*

checked directly.

5.2. Frequency axis: sparser and at least as faithful

On frequency complexity SG holds the best (lowest) median in all seven domains and significantly outperforms every STFT configuration. The strongest STFT competitor in each domain is always a frequency-optimised window (IG-freq or IxG-freq), and the gap is widest on the two sleep-EEG case studies, where the best window remains 0.17 above SG, and narrowest on Speech (0.03; Table 2). SG's sparser attributions are not less faithful. On frequency infidelity, which asks whether the highest-attributed bands are genuinely the most model-relevant, SG holds the best median in two domains (Speech and Sleep-EDF) and, under the corrected test, is never significantly worse than any STFT configuration in any domain. On the synthetic benchmarks, where ground-truth masks are available, it also attains the highest frequency localisation in all three setups (Table 3), so the concentrated attribution falls on the discriminative band.

5.3. Time axis and the window trade-off

No STFT configuration matches SG on both axes at once. For the IG and IxG families, in every domain, the window with the best frequency complexity is always the one with the worst time complexity, and conversely. On the infidelities the pattern is looser: a short window is sometimes best on both (IxG-time on the synthetic setups). On Speech, for instance, IG-freq attains the IG family's best frequency infidelity but its worst time infidelity, while IG-time reverses the ordering. IG-time achieves the best time complexity in three domains yet is never best on frequency complexity, where it is always significantly worse than SG. This reflects the Gabor trade-off and not a property of the estimator: a time-optimised window (n_{fft} = 16–256) spans 16 ms (Speech) to 0.3 s (synthetic, sleep) and resolves frequency no finer than 3–62 Hz. A frequency-optimised window (n_{fft} = 128–4096) does the opposite, so no single window can be precise on both axes at once. Counting over all 153 comparisons (17 metric groups $\times$ 9 configurations), no configuration is significantly better than SG on both the time and the frequency variant of a metric. SG itself holds the best median on 21 of the 34 metric–domain pairs, ties on 10, and is significantly worse on 3, all of them time complexity; against the assumption-free balanced window it is never significantly worse on 33 of 34 pairs.

5.4. Dependence on the signal regime

How SG's advantage is distributed depends on the signal regime. It is broadest on Speech, the longest signal (16,000 samples) and the only domain where SG wins all four metrics significantly, although with moderate effect sizes (δ = 0.16–0.39, Table A.3).

There the STFT trade-off is most severe: resolving a few hertz requires $n_{\text{fft}} = 4096$ and a time resolution of a quarter of a second, while a 256-sample window gives a frequency resolution coarser than 60 Hz. SG requires no such window choice. On sleep EEG the outcome depends on the model more than on the signal length. The deeper TsinalisCNN, whose first layer of 200-sample filters behaves as a learned spectral filterbank, produces frequency-structured gradients that SG exploits (best on three of four metrics). The compact Chambon2018 yields less structured gradients, and SG is best on two of four. The shortest signal, ECG (200 samples), is, with Speech, the second case where SG is significantly the sparsest on both axes. Here the STFT windows cannot be sparse. The 128-sample frequency-optimised window covers two thirds of the beat in seven frames, the balanced window gives 13 frames, and even the 16-sample window resolves time in 4-sample hops with nine frequency bins. Every STFT marginal, once interpolated to the 200 samples, is therefore nearly flat. The windows still trade faithfulness, as elsewhere: the frequency-optimised IxG window is more faithful in frequency and the time-optimised IG window in time, in neither case significantly (Table A.3). On this short signal the infidelity metrics also saturate, with about half of the samples of SG and of its best rivals at the minimum attainable value, so SG's faithfulness advantage, present in every other case study, is absent here on both axes.

5.5. Where SG loses

Time complexity is SG's one recurrent weakness. It is the only metric on which SG loses more than once: significantly in three of the seven domains (Synt-1, Synt-2, Sleep-MASS), with a time-optimised STFT window as the best rival (on Sleep-MASS also to the balanced IxG window). SG wins on Speech and ECG and ties on Synt-0 and Sleep-EDF. We examine why SG loses sparsity in time in these three cases. The construction offers one parameter to act on, the band width: a DFT band component is spread over roughly $1/\Delta f$ in time, so a narrow band spreads each per-band attribution term beyond the event it explains. The three losses are the cases where $1/\Delta f$ is comparable to the discriminative event (0.5 Hz bands, i.e. about 2 s, against sleep graphoelements of about 1 s; 2 Hz bands, 0.5 s, against the short synthetic bursts). SG is sparsest where $1/\Delta f$ is negligible (Speech, 50 Hz) or the event spans the whole signal (ECG). We

Table 4: Sensitivity of SG to the band width Δf, on identical samples (paired). "Fine" is the benchmark setting of Table A.2, "coarse" a wider band (0.5 → 2 Hz on Sleep-MASS, 2 → 5 Hz on Synt-1/2); δ is Cliff's delta of coarse vs. fine (negative: lower with the coarser band); the last column is the best time-optimised STFT rival on the same samples. Sleep-MASS uses a strided 1/8 subset of the benchmark sample set.

Case	**Metric**	**SG fine**	**SG coarse**	δ	**Rival**	**Rival value**
Sleep-MASS	Freq.Inf	0.0274	0.0342*	+0.21	IxG-time	0.0668
	Freq.Cmplx	0.5981	0.6488*	+0.40		0.8009
	Time.Inf	0.0260	0.0194	-0.12		0.0843
	Time.Cmplx	0.9372	0.9290*	-0.38		0.9131
Synt-1	Freq.Inf	0.1252	0.1360	+0.11	IG-time	0.1251
	Freq.Cmplx	0.6611	0.7329*	+0.52		0.7677
	Time.Inf	0.1961	0.1848	-0.01		0.9177
	Time.Cmplx	0.7823	0.7668	-0.11		0.7524
Synt-2	Freq.Inf	0.1156	0.1606*	+0.34	IG-time	0.1034
	Freq.Cmplx	0.6423	0.7355*	+0.68		0.7726
	Time.Inf	0.0321	0.0291	-0.01		0.3360
	Time.Cmplx	0.7878	0.7771	-0.09		0.7659

tested this parameter directly by re-running SG with a coarser band on identical samples (Table 4). Widening the band lowers time complexity in all three cases, strongly on Sleep-MASS (0.937 → 0.929, $\delta = -0.38$) and only slightly on the synthetic setups ($\delta \approx -0.1$). It degrades frequency sparsity and frequency faithfulness substantially, and it does not reach the time-optimised window (0.913 on Sleep-MASS). The band width thus accounts for part of the gap, consistent with the trade-off being moved to Δf and not removed, while the remainder is a genuine advantage of a short analysis window on the time axis taken alone. The winning windows' extra sparsity comes with reduced spectral precision (on Sleep-EDF, IG-time's frequency complexity is far worse than SG's, Table 2) and, as the effect sizes in Table A.3 make explicit, reduced faithfulness. On the single balanced-window cell where SG is less sparse in time (Sleep-MASS), SG is about seven times more faithful on the same axis, so the loss reflects the sparsity–faithfulness trade-off.

5.6. A qualitative example

Figure 2 shows what the metrics summarise, on one real epoch. The SG map of an N2 epoch places about two thirds of its attribution in the σ band, which contains

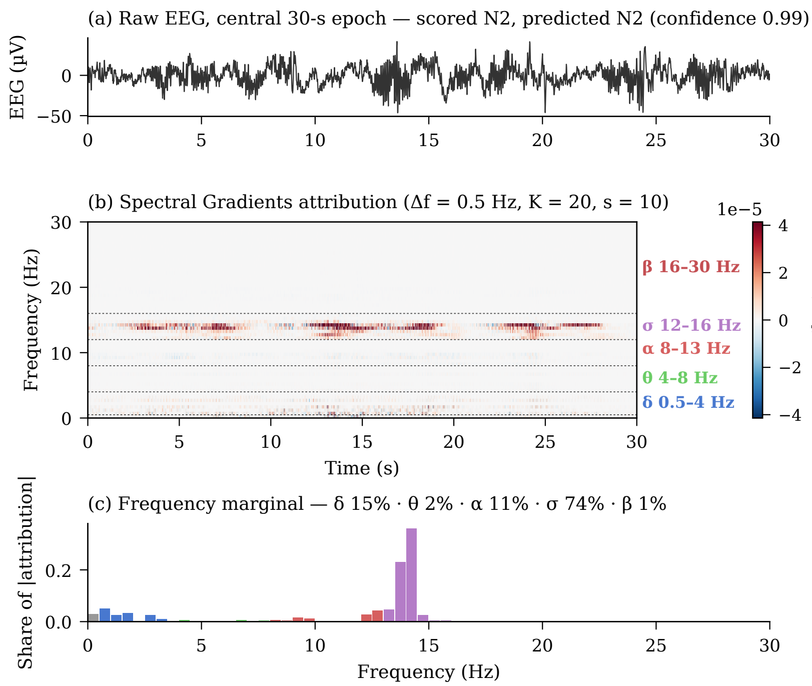

Figure 2: SG attribution for a single 30-s Sleep-EDF epoch scored N2 and classified as such by TsinalisCNN. Top: raw EEG. Middle: SG time-frequency map with the Sleep-EDF benchmark settings ($\Delta f = 0.5$ Hz, $K = 20$), with the clinical EEG bands (δ, θ, α, σ, β) marked on the frequency axis. Bottom: frequency marginal of the map, coloured by band, with the share of total attribution per band. The map is read directly in the terms of the scoring manual: the frequency axis shows which band the model relies on, and the time axis, at the original sampling resolution, shows when within the epoch it does so.

sleep spindles, the defining feature of N2 in the AASM rules [16]. It localises that attribution to a few one- to two-second bursts where the spindles occur. A secondary share falls in the δ band, consistent with the slow-wave activity that N2 epochs also contain, a small one in α, and the θ and β bands receive almost none (panel c). Because SG operates on global DFT coefficients, the band-level reading (frequency marginal) and the sample-level reading (time axis) come from the same map, with no window to choose between them. A clinician can therefore check both whether the model attends to the correct band and whether it does so at the correct moments, the joint question that motivates this work.

5.7. *Convergence in the number of permutations*

Finally, because $\overline{SG}$ approximates the permutation average by sampling K orderings and K enters the cost linearly ($O(K \cdot b \cdot s)$ forward passes), we ask how small K can be without degrading quality.

Figure A.1 tracks median frequency infidelity as K grows on the three synthetic setups. Below $K = 20$ the estimate is unstable: a single permutation gives values two to three times higher than the converged one, and two permutations can be worse still, because the variance across orderings dominates the mean. The estimate settles by $K = 20$ and plateaus by $K = 50$, beyond which $K = 100$ or 200 shifts it by under 5%; the rate is case-dependent (Setup 2 plateaus at $K = 20$, Setups 0 and 1 at $K = 50$). We therefore use $K = 20$ as the default for the synthetic and sleep case studies, and the more conservative $K = 100$, fixed before this analysis, for ECG and Speech (Table A.2); both settings lie on the plateau. Because infidelity is computed from the same band ablations that SG already performs, its stabilisation across K can be monitored at negligible extra cost, which gives a practical rule for choosing the smallest sufficient K in a new domain.

6. Discussion

The pattern in Section 5 has a simple structure. Each STFT window must favour one axis, so none of the nine configurations is significantly better than SG on both the time and the frequency variant of any metric. In 117 of the 153 cells SG is significantly better on both; the remaining cells are ties, and there are no losses on both axes. SG does at a single window-free operating point what the windows cannot do jointly, and its only recurrent losses are in time complexity, with a time-optimised window as the best rival. The result is the existence of one parametrisation competitive on both axes at once, in a region of the joint quality space that no single fixed window attains; it is not an average metric win.

The comparison a practitioner can actually run reinforces this. The time- and frequency-optimised windows are oracle baselines: each wins its specialised axis only by committing to it in advance and losing the other axis. The single assumption-free choice is the balanced window, and against it SG is never significantly worse on 33 of

34 pairs. On the only one where a balanced window is sparser in time (Sleep-MASS, time complexity) SG is about seven times more faithful on the same axis, so neither dominates. When an explanation must be read on both axes at once, as in checking whether a model attends to the correct band at the correct time, no single window suffices. Among the compared methods, SG's joint map is the only one that needs no such choice.

A natural explanation is the Gabor uncertainty principle ($\Delta t \cdot \Delta f \geq 1/(4\pi)$): any windowed transform imposes a coupled resolution trade-off, and this coupling propagates to the attributions computed through it, whereas SG operates on global DFT coefficients where Δf and $1/f_s$ are independent. Super-resolution time-frequency decompositions such as hyperlets [46] are designed to relax the same limit for short-bursting neural signals; SG avoids it on the attribution side by not windowing at all. Our evidence establishes the *pattern* of non-domination; attributing it solely to the uncertainty principle, as opposed to the specific window grid or the integration path, would require a dedicated ablation that we leave to future work.

The argument is not specific to the STFT. Any method that computes attributions through a windowed or multi-scale transform, including wavelet-based approaches using the invertible CWT, should in principle remain subject to the Gabor bound at every analysis scale. The CWT redistributes the trade-off across scales, but the product $\Delta t \cdot \Delta f$ stays bounded from below at each one. If so, the advantage of SG would reflect a structural limitation of windowed transforms in general, and would not depend on the specific choice of STFT as a baseline. We did not run a wavelet-based attribution baseline, so this generalisation is a hypothesis rather than an empirical result; testing it against an invertible-CWT pipeline is an important direction for future work.

This view is consistent with the analysis of Rezaei et al. [24], who apply the uncertainty principle to XAI and show that time-domain and frequency-domain attributions can highlight different features that are not direct counterparts of one another, establishing the need for joint multi-domain explanations. SG addresses this need by producing a single time-frequency attribution map whose resolutions are not coupled; Figure 2 shows the practical effect: a band-level and a sample-level reading of the same decision from one map.

Our results also carry a methodological implication: current evaluation practices for time-series explainability may be incomplete. Most existing benchmarks we are aware of evaluate attribution quality in a single domain, either the time domain or, more rarely, the frequency domain. As we have shown, a method can achieve excellent frequency metrics while performing poorly on time metrics (and vice versa), and this trade-off remains invisible if only one domain is evaluated.

We therefore recommend that future benchmarks for time-frequency attribution report both the frequency- and time-domain variant of each metric, evaluate STFT baselines across multiple window configurations rather than a single balanced choice, and test explicitly whether any method dominates across both domains at once.

6.1. Limitations

The computational cost of $\overline{SG}$ is $O(K \cdot b \cdot s)$ forward passes (K permutations, b bands, s integration steps), against $O(s)$ for STFT-IG, i.e. a factor of $K \cdot b$ more evaluations. For the configurations used here (K of 20 or 100 and b between 26 and 161, Table A.2) this factor ranges from about 500 to 16,000. In wall-clock terms the gap is far smaller: the mega-batched GPU implementation evaluates all b bands of a step in parallel, so runtime scales as $O(K \cdot s)$, between twenty and one hundred times the STFT-IG cost.

SG's frequency resolution Δf must be chosen per domain. While this is a single scalar parameter rather than the coupled $(n_{\text{fft}}, \text{hop}, \text{window})$ triple required by STFT methods, it still requires domain knowledge or tuning. Similarly, the number of permutations K must be sufficient for convergence, which can be verified by monitoring infidelity stabilisation but adds to the computational budget.

SG's only recurrent weakness is time complexity, where it is significantly outperformed by a time-optimised STFT on three domains (Synt-1, Synt-2, Sleep-MASS); against the balanced parameterisation the losses reduce to a single cell (Sleep-MASS, time complexity). In each of these cells SG is markedly more faithful on the same axis, so no STFT configuration is simultaneously sparser and more faithful than SG in time: SG is not Pareto-dominated once complexity and infidelity are read jointly. The corrected ECG row shows the opposite pattern: there the time-optimised IG window is somewhat more faithful in time while SG is sparser, again with no domination.

7. Conclusion

We introduced Spectral Gradients (SG), a post-hoc attribution method for time-series classifiers whose explanations are read on the time and on the frequency axis at once. Its integration path is built by restoring DFT frequency bands from a silent baseline, and the importance of each band is redistributed over time by path-integrated gradients; the path-independent variant $\overline{SG}$ averages over band orderings, so that the band importances are Shapley values. Because no windowed transform is involved, the frequency grid and the sampling grid are chosen separately, and the Gabor limit reappears as one interpretable parameter, the band width Δf, rather than as a coupled window, hop and overlap choice.

On the frequency axis SG is the sparsest method in every one of the seven case studies and is never significantly less faithful than any STFT configuration; on the synthetic setups it also localises the discriminative band best. The resolution trade-off is measurable in the explanations themselves: for the IG and IxG families the window with the best frequency complexity is always the one with the worst time complexity. Across 153 comparisons no window is significantly better than SG on both variants of any metric, and in 117 of them SG is significantly better on both. SG is less sparse in time than a time-optimised window in three case studies. The band-width experiment shows that part of this gap is the $1/\Delta f$ temporal spread of a narrow band, i.e. the trade-off moved and not removed. In every such cell SG remains markedly more faithful on the same axis, so no STFT window is at once sparser and more faithful than SG.

For practitioners, the result means that a joint time-frequency explanation no longer requires an arbitrary choice of analysis window: a single band width, chosen from the physiology or the acoustics of the problem, yields a map that is competitive on both axes. The number of permutations can be fixed by monitoring the stabilisation of infidelity without extra computation.

Code and data availability

The implementation of Spectral Gradients, the model definitions and training scripts of the classifiers in Table A.1, the benchmark and figure-generation code, and the per-sample results are available at `https://github.com/guidogagl/spectralgradients` and archived on Zenodo [47]. Spectral Gradients is also distributed as part of the PhysioEx library [3, 48]. The MIT-BIH Arrhythmia Database [49] and the Sleep-EDF Database Expanded [50] are openly available from PhysioNet, and Google Speech Commands v0.02 [41] from its public distribution; the Montreal Archive of Sleep Studies [42] is available from the MASS consortium under a data-use agreement and cannot be redistributed by the authors. The synthetic datasets are generated by the released code.

CRediT authorship contribution statement

Guido Gagliardi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing, Funding acquisition. **Antonio Luca Alfeo:** Supervision, Writing – review & editing. **Mario G. C. A. Cimino:** Supervision, Writing – review & editing. **Wojciech Samek:** Supervision, Writing – review & editing. **Maarten De Vos:** Supervision, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Funding: This work was supported by the Research Foundation Flanders (FWO) [Strategic Basic Research fellowship 1SH4Z24N and Travel Grant for a Long Research Stay Abroad V458425N, both to G. Gagliardi]; by the Flemish Government under the METHUSALEM programme (METH/26/003, "Methusalem–BioMedAI: Explainable and generative AI for accelerating biomedical discoveries: from genome to function"); and by the Flemish Government AI Research Program. The funders had no role in

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Claude (Anthropic) in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgements

Computational resources were provided by the VSC (Flemish Supercomputer Centre), funded by the FWO and the Flemish Government (Sofia cluster), and by CINECA under the ISCRA initiative (project IscrC_SELFEEG, Leonardo supercomputer).

Table A.1: Architecture and training recipe of the classifiers trained for this study. All are plain 1-D CNNs on the raw waveform; "$a \to b$" lists channel widths, "k" kernel sizes, "s" strides. Splits are train/validation/test fractions; SpeechCommands uses the official split.

Domain	Architecture	Training
Synthetic	3 conv layers, $1 \to 4 \to 8 \to 16$, $k = 100/4/4$, $s = 50/2/2$; ReLU + batch norm after each layer; no pooling; linear head ($48 \to 5$).	Adam, lr 10^{-3}, cross-entropy, 10 epochs, batch 64, 2,000 signals per setup, split 70/15/15, seed 0; best-validation checkpoint.
ECG	4 conv layers, $1 \to 32 \to 64 \to 128 \to 128$, $k = 7/5/3/3$, $s = 1$; ReLU; max-pool (2) after the first three layers, global average pool after the last; linear head ($128 \to 5$); per-beat z-scoring inside the network; no batch norm or dropout.	Adam, lr 10^{-3}, weight decay 10^{-4}, cosine schedule, 50 epochs, batch 128, gradient clipping at 1.0, class rebalancing of N and F beats, split 70/15/15, seed 42.
Speech	4 blocks of two conv layers each, $1 \to 32 \to 64 \to 128 \to 192$, $k = 15/11/9/7$, $s = 1$; batch norm + SiLU; max-pool (4) and dropout (0.10–0.25) per block; global average pool; head $192 \to 256 \to 35$ with dropout 0.35; per-clip z-scoring inside the network.	AdamW, lr $3 \cdot 10^{-4}$, weight decay 10^{-2}, label smoothing 0.1, cosine schedule, 50 epochs, batch 128, gradient clipping at 1.0, waveform augmentation (gain, noise, circular shift), official split, seed 42.

Appendix A. Classifier architectures and analysis configurations

Table A.1 gives the architecture and training recipe of the classifiers trained for this study, and Table A.2 the per-domain analysis configuration of SG and of the STFT baselines.

Table A.2: Per-domain analysis configurations. STFT baselines use three window sizes n_{fft} (time-optimised / balanced / frequency-optimised) with hop = $n_{\text{fft}}/4$. SG uses a single band width Δf, giving b bands over $[0, f_s/2]$ (for ECG the requested 2.5 Hz rounds to one DFT bin, f_s/n = 1.8 Hz), K sampled permutations, and s = 10 integration steps. For ECG and Speech, K = 100 was used, a conservative setting fixed before the convergence analysis of Section 5, which shows the estimate to be on its plateau from $K \approx 50$. Frequency- and time-infidelity ablate the signal in patches of the sizes shown (DFT bins / time samples).

Domain	STFT n_{fft} (t / b / f)	SG Δf (Hz), b	SG K	Patch (freq / time)
Synthetic	32/128/512	2.0, 26	20	10/20
ECG	16/64/128	1.8, 101	100	5/20
Speech	256/1024/4096	50, 161	100	20/500
Sleep-MASS	32/128/512	0.5, 129	20	1/50
Sleep-EDF	32/128/512	0.5, 101	20	1/50

Table A.3: Effect sizes for the decision-relevant comparison: signed Cliff's δ of SG against the best-median STFT rival on each metric (positive favours SG; $|\delta| < 0.147$ is negligible). $*/^{\dagger}$ mark SG significantly better / worse under the corrected standard (Holm-Bonferroni + $|\delta| \geq 0.147$). Localization applies to synthetic domains only. Three decimals are shown because two cells lie within 0.005 of the threshold.

Domain	Freq.Inf	Freq.Cmplx	Time.Inf	Time.Cmplx	Freq.Loc	Time.Loc
Synt-0	-0.070	+0.519*	+0.234*	-0.090	+0.767*	-0.032
Synt-1	+0.013	+0.243*	+0.091	-0.179†	+0.963*	-0.105
Synt-2	-0.072	+0.634*	+0.401*	-0.153†	+0.890*	-0.006
ECG	-0.056	+0.327*	-0.146	+0.366*	—	—
Speech	+0.241*	+0.155*	+0.385*	+0.169*	—	—
Sleep-MASS	-0.077	+0.930*	+0.281*	-0.677†	—	—
Sleep-EDF	+0.046	+0.841*	+0.323*	-0.142	—	—

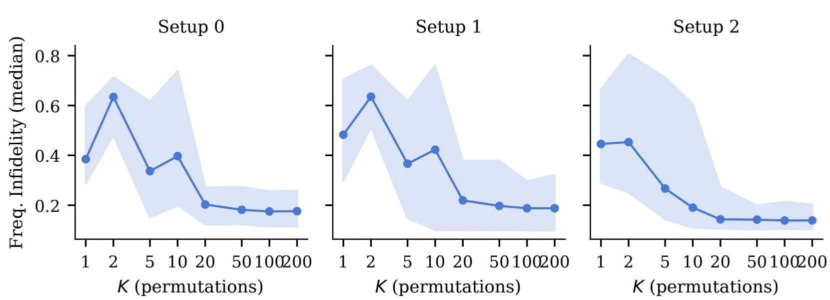

Figure A.1: Median frequency infidelity as a function of the number of permutations K on the three synthetic setups (a fixed 200-sample subset of each, b = 26 bands). Below K = 20, the approximation is unstable; convergence is reached around K=50.

References

[1] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data Mining and Knowledge Discovery 33 (4) (2019) 917–963.

[2] H. Phan, K. Mikkelsen, Automatic sleep staging of eeg signals: recent development, challenges, and future directions, Physiological Measurement 43 (4) (2022) 04TR01.

[3] G. Gagliardi, L. Alfeo, M. G. Cimino, G. Valenza, M. De Vos, Physioex, a new python library for explainable sleep staging through deep learning, Physiological Measurement (2025).

[4] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature communications 10 (1) (2019) 1096.

[5] A. Theissler, F. Spinnato, U. Schlegel, R. Guidotti, Explainable AI for time series classification: A review, taxonomy and research directions, IEEE Access 10 (2022) 100700–100724.

[6] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: International Conference on Learning Representations (ICLR), Workshop Track, 2014, arXiv:1312.6034.

[7] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences, arXiv preprint arXiv:1605.01713 (2016).

[8] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (7) (2015) e0130140.

[9] Y. Li, H. Liang, L. Zheng, WB-LRP: Layer-wise relevance propagation with weight-dependent baseline, Pattern Recognition 158 (2025) 110956. `doi:10.1016/j.patcog.2024.110956`.

[10] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[11] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, S.-I. Lee, Improving performance of deep learning models with axiomatic attribution priors and expected gradients, Nature machine intelligence 3 (7) (2021) 620–631.

[12] T. D. T. Kien, N. T. Tam, N. H. Son, K. Than, D. A. Nguyen, Enhancing visual feature attribution via weighted integrated gradients, Pattern Recognition Letters 207 (2026) 68–74. `doi:10.1016/j.patrec.2026.06.009`.

[13] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[14] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[15] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications, Proceedings of the IEEE 109 (3) (2021) 247–278.

[16] C. Iber, A. A. of Sleep Medicine, The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, American Academy of Sleep Medicine, 2007.
URL `https://books.google.be/books?id=jnsEMwAACAAJ`

[17] R. S. Rosenberg, S. Van Hout, The american academy of sleep medicine interscorer reliability program: sleep stage scoring, Journal of clinical sleep medicine 9 (1) (2013) 81–87.

[18] L. R. Rabiner, R. W. Schafer, Theory and Applications of Digital Speech Processing, Pearson, Upper Saddle River, NJ, 2011.

[19] G. D. Clifford, F. Azuaje, P. E. McSharry, Advanced Methods and Tools for ECG Data Analysis, Artech House, Norwood, MA, 2006.

[20] R. Younis, A. Hakmeh, Z. Ahmadi, MTS2Graph: Interpretable multivariate time series classification with temporal evolving graphs, Pattern Recognition 152 (2024) 110486. `doi:10.1016/j.patcog.2024.110486`.

[21] D. Lee, S. Malacarne, E. Aune, Explainable time series anomaly detection using masked latent generative modeling, Pattern Recognition 156 (2024) 110826. `doi:10.1016/j.patcog.2024.110826`.

[22] W. Jeong, G. Shin, J. Lee, JoCE: Joint counterfactual explanations for interpretable time series anomaly detection, Pattern Recognition 177 (2026) 113324. `doi:10.1016/j.patcog.2026.113324`.

[23] J. Vielhaben, S. Lapuschkin, G. Montavon, W. Samek, Explainable ai for time series via virtual inspection layers, Pattern Recognition 150 (2024) 110309.

[24] S. Rezaei, A. Halev, X. Liu, On the necessity of multi-domain explanation: An uncertainty principle approach for deep time series models, in: International Joint Conference on Neural Networks (IJCNN), IEEE, 2025, arXiv:2506.03267.

[25] D. Mercier, A. Dengel, S. Ahmed, FREQuency ATTribution: Benchmarking frequency-based occlusion for time series data, Applied Intelligence 56 (4) (2026) 108. `doi:10.1007/s10489-026-07116-1`.

[26] L. S. Shapley, A value for *n*-person games, in: H. W. Kuhn, A. W. Tucker (Eds.), Contributions to the Theory of Games II, no. 28 in Annals of Mathematics Studies, Princeton University Press, 1953, pp. 307–317.

[27] D. Gabor, Theory of communication, Journal of the Institution of Electrical Engineers—Part III: Radio and Communication Engineering 93 (26) (1946) 429–457. doi:10.1049/ji-3-2.1946.0074.

[28] U. Bhatt, A. Weller, J. M. F. Moura, Evaluating and aggregating feature-based model explanations, in: Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), 2020, pp. 3016–3022. doi:10.24963/ijcai.2020/417.

[29] M. Bello, R. Amador, M.-M. García, J. Del Ser, P. Mesejo, Ó. Cordón, The level of strength of an explanation: A quantitative evaluation technique for post-hoc XAI methods, Pattern Recognition 161 (2025) 111221. doi:10.1016/j.patcog.2024.111221.

[30] R. P. Bora, P. Terhörst, R. Veldhuis, R. Ramachandra, K. Raja, FRIES: Framework for inconsistency estimation of saliency metrics, Pattern Recognition 171 (2026) 112136. doi:10.1016/j.patcog.2025.112136.

[31] H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, The Annals of Mathematical Statistics 18 (1) (1947) 50–60. doi:10.1214/aoms/1177730491.

[32] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics 6 (2) (1979) 65–70.

[33] N. Cliff, Dominance statistics: Ordinal analyses to answer ordinal questions, Psychological Bulletin 114 (3) (1993) 494–509.

[34] J. Romano, J. D. Kromrey, J. Coraggio, J. Skowronek, Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys?, in: Annual meeting of the Florida Association of Institutional Research, 2006, pp. 1–33.

[35] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, P. Ravikumar, On the (in)fidelity and sensitivity of explanations, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.

[36] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim, A benchmark for interpretability methods in deep neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.

[37] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, E. Kasneci, A consistent and efficient evaluation strategy for attribution methods, in: Proceedings of the 39th International Conference on Machine Learning (ICML), Vol. 162 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 18770–18795.

[38] A. V. Oppenheim, R. W. Schafer, Discrete-Time Signal Processing, 3rd Edition, Prentice Hall, 2010.

[39] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, S. Lapuschkin, Towards best practice in explaining neural network decisions with lrp, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[40] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000 (June 13)) e215–e220, circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[41] P. Warden, [dataset] speech commands: A dataset for limited-vocabulary speech recognition (v0.02), arXiv preprint arXiv:1804.03209; data: `http://download.tensorflow.org/data/speech_commands_v0.02.tar.gz`, accessed 2025 (2018). `doi:10.48550/arXiv.1804.03209`.

[42] C. O'Reilly, N. Gosselin, J. Carrier, T. Nielsen, Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research, Journal of Sleep Research 23 (6) (2014) 628–635. `doi:10.1111/jsr.12169`.

[43] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, J. J. Oberye, Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg, IEEE Transactions on Biomedical Engineering 47 (9) (2000) 1185–1194.

[44] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, A. Gramfort, A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series, IEEE Transactions on Neural Systems and Rehabilitation Engineering 26 (4) (2018) 758–769. `doi:10.1109/TNSRE.2018.2813138`.

[45] O. Tsinalis, P. M. Matthews, Y. Guo, S. Zafeiriou, Automatic sleep stage scoring with single-channel EEG using convolutional neural networks, arXiv preprint arXiv:1610.01683 (2016).

[46] K. Kesgin, H. Jörntell, Super-resolution time-frequency decomposition with hyperlets for neural spike analysis, Pattern Recognition 172 (2026) 112598. `doi:10.1016/j.patcog.2025.112598`.

[47] G. Gagliardi, [software] guidogagl/spectralgradients: Spectral gradients benchmark (v1.0.0), Zenodo (2026). `doi:10.5281/zenodo.22964216`.

[48] G. Gagliardi, [software] PhysioEx: explainable deep learning for physiological signals (v2.0.1), PyPI / GitHub, `https://github.com/guidogagl/physioex`, accessed 25 September 2026 (2026).

[49] G. B. Moody, R. G. Mark, [dataset] MIT-BIH arrhythmia database (version 1.0.0), PhysioNet (2005). `doi:10.13026/C2F305`.

[50] B. Kemp, [dataset] sleep-EDF database expanded (version 1.0.0), PhysioNet (2018). `doi:10.13026/C2X676`.