STF-KernelSHAP: A Model-Agnostic Space–Time–Frequency Shapley Framework for Physiologically Informed EEG Explainability

Diego Armando Pérez-Rosero *, Andres Camilo Lopez-Boscan, Andrés Marino Álvarez-Meza *, David Augusto Cárdenas-Peña, German Castellanos-Dominguez

Posted Date: 4 June 2026

doi: 10.20944/preprints202606.0375.v1

Keywords: EEG; explainable artificial intelligence; SHAP; space–time–frequency analysis; physiological coherence; motor imagery; ADHD; attribution methods

 doi:10.20944/preprints202606.0375.v1

Article

STF-KernelSHAP: A Model-Agnostic Space–Time–Frequency Shapley Framework for Physiologically Informed EEG Explainability

Diego Armando Pérez-Rosero [1,*], Andres Camilo Lopez-Boscan [1], Andrés Marino Álvarez-Meza [1], David Augusto Cárdenas-Peña [2] and German Castellanos-Dominguez [1]

[1] Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales 170003, Colombia
[2] Automatics Research Group, Universidad Tecnológica de Pereira, Pereira 660003, Colombia
* Correspondence: dieaperezros@unal.edu.co

Abstract

Reliable interpretability remains essential for deploying deep learning models in EEG-based neurotechnology, particularly in brain–computer interfaces and clinical decision-support scenarios. However, existing post-hoc explainable artificial intelligence methods often provide single-domain attribution maps, limiting their ability to jointly characterize spatial, temporal, and spectral EEG dynamics. In addition, perturbation-based strategies may disrupt physiological signal organization, whereas gradient-based methods depend on internal model information and are therefore constrained by the classifier architecture. Here, we introduce STF-KernelSHAP, a model-agnostic space–time–frequency Shapley framework for physiologically coherent EEG explainability. Our approach is threefold: first, EEG trials are decomposed into structured channel–time–frequency cells using segment-wise spectral analysis, enabling multidomain attribution over spatial locations, temporal windows, and frequency bands; second, coalitions are defined over complete channel–time–frequency cells and reconstructed back into the signal domain, promoting physiologically informed perturbations; and third, class-conditional relevance is estimated through a KernelSHAP-based weighted surrogate model using only model outputs, ensuring architecture-independent Shapley estimation. We evaluate STF-KernelSHAP on motor imagery (MI) decoding and Attention-Deficit/Hyperactivity Disorder (ADHD) detection, and compare it against KernelSHAP, LIME, Occlusion, Integrated Gradients, and Grad-CAM++. Fidelity is quantified using Deletion and ROAD, while qualitative analyses examine topographic and frequency-band relevance patterns. Results indicate that STF-KernelSHAP remains visually and functionally comparable to classical XAI methods, while additionally providing window-dependent and frequency-specific explanations. Overall, STF-KernelSHAP offers a physiologically informed and model-agnostic alternative for multidomain EEG interpretability.

Keywords: EEG; explainable artificial intelligence; SHAP; space–time–frequency analysis; physiological coherence; motor imagery; ADHD; attribution methods

1. Introduction

The decoding of oscillatory brain dynamics constitutes a central component of modern neurotechnology, supporting progress in both active neural interfaces and clinical diagnostics [1]. In the context of Brain–Computer Interfaces (BCIs), Motor Imagery (MI) remains a foundational paradigm: the mental simulation of movement modulates sensorimotor rhythms that enable applications in rehabilitation and assistive communication [2]. Beyond BCI, the characterization of intrinsic neural activity plays a relevant role in computational psychiatry, where resting-state spectral markers contribute to the identification of neurodevelopmental alterations such as those associated with Attention-Deficit/Hyperactivity Disorder (ADHD) [3]. To observe these diverse neural processes,

 doi:10.20944/preprints202606.0375.v1

electroencephalography (EEG) has become the predominant non-invasive modality due to its high temporal resolution, portability, and comparatively low acquisition cost—properties that make it suitable for large-scale and real-world deployments [4].

As EEG-based applications expand, the decoding of complex space–time–frequency patterns has increasingly relied on deep learning (DL), which has demonstrated superior performance compared to classical machine learning approaches [5]. However, these performance gains introduce a critical challenge: the interpretability of DL models [6]. Because such models typically operate as opaque systems, it is difficult to ascertain whether their decisions reflect meaningful neurophysiology or spurious correlations—an issue that is particularly consequential in clinical and neurotechnology settings, where transparent and physiologically plausible explanations are essential [7]. These concerns have motivated growing interest in Explainable Artificial Intelligence (XAI), a field focused on generating interpretable accounts of model behavior and on ensuring that predictive performance is accompanied by reliable and human-understandable insights [8].

Despite the increasing adoption of Explainable Artificial Intelligence (XAI) in neurotechnology, reliable interpretability for EEG remains difficult to achieve due to two persistent limitations. The first limitation relates to the lack of multidomain resolution, because EEG signals distribute information jointly across space, time, and frequency, and analyses restricted to a single domain cannot capture the spectral components that underpin core neurophysiological dynamics [9]. This restriction leads to explanatory outputs that overlook oscillatory patterns essential for decoding [10]. The second limitation concerns physiological coherence. Many current methods operate through perturbations or manipulations that ignore the functional organization of EEG signals, thereby disrupting dependencies among neurophysiological features and producing explanations that may be mathematically acceptable but biologically implausible [11]. Addressing these two limitations simultaneously is further constrained by the requirement of architectural independence: while architecture-specific solutions can partially mitigate one of these issues, a general approach should support multidomain explanations through physiologically informed perturbations, without relying on gradients or internal model parameters, thereby enabling applicability across pre-trained black-box models and heterogeneous ensembles [12].

In response to these challenges, the literature has explored several interpretability strategies, each typically targeting one of the aforementioned issues, but rarely both. Among architecture-independent approaches, methods such as LIME, KernelSHAP, and occlusion-based techniques estimate feature relevance through local perturbations, surrogate explanations, or partial masking of the input [13]. Although these methods provide flexibility and compatibility with black-box models, the introduced perturbations may alter the intrinsic temporal, spatial, and spectral dependencies of EEG signals, potentially compromising the physiological coherence of the explanations [14]. To improve multidomain resolution, hybrid techniques combine temporal attribution scores with mechanisms such as attention, explicit time-windowing, or concept-based activation vectors [11]. Although these approaches aim to link temporal activations with spectral modulations, they typically infer frequency information from temporal proxies rather than manipulating the spectral domain directly, which can introduce spectral leakage and limit the ability to attribute decisions to well-defined oscillatory bands [15]. Complementarily, gradient-based methods such as Integrated Gradients and Grad-CAM++ leverage information derived from internal activations and backpropagated gradients to identify salient regions relevant to the model decision; however, this dependence on the architecture and internal properties of the neural network limits their applicability to black-box models and heterogeneous architectures [15].

To address physiological coherence, segmentation-based approaches inspired by superpixels or graph clustering group input features according to structural similarity, aiming to preserve the integrity of the input by treating coherent regions as units of analysis [16]. However, such grouping methods were originally designed for visual data, where spatial adjacency determines structure, and do not naturally extend to EEG, where meaningful dependencies occur across non-adjacent electrodes and cross-frequency interactions [17]. As a result, existing methods provide partial advances toward

 doi:10.20944/preprints202606.0375.v1

either multidomain resolution or physiological coherence, but do not jointly satisfy both requirements under architecture-independent conditions [18].

We propose Space–Time–Frequency KernelSHAPley Attribution (STF-KernelSHAP), a model-agnostic interpretability framework designed to provide multidomain EEG explanations using physiologically informed channel–time–frequency perturbations. The proposed framework addresses three main requirements in EEG explainability:

- Multidomain EEG attribution: To address the lack of multidomain resolution, STF-KernelSHAP decomposes each EEG trial into structured channel–time–frequency cells using segment-wise spectral analysis. This representation allows the attribution process to operate directly over spatial locations, paradigm-specific temporal windows, and physiologically meaningful frequency bands, rather than relying on flattened features or single-domain temporal explanations.
- Physiologically informed perturbation: To promote the functional organization of EEG signals, STF-KernelSHAP defines coalitions over complete channel–time–frequency cells instead of isolated samples. Each coalition is mapped back to the signal domain through spectral reconstruction, allowing perturbations to be applied over physiologically informed EEG components and mitigating the risk of generating biologically implausible signal manipulations.
- Architecture-independent Shapley estimation: To ensure applicability across heterogeneous classifiers, STF-KernelSHAP operates as a black-box explainer that only requires access to model outputs. Class-conditional relevance is estimated through a KernelSHAP-based weighted surrogate model, without using gradients, internal activations, or architecture-specific parameters.

We evaluate STF-KernelSHAP on two EEG classification scenarios with distinct neurophysiological characteristics: motor imagery decoding using the GIGA MI-ME dataset and ADHD detection using a pediatric EEG dataset. The proposed framework is applied to pre-trained EEG classifiers and compared against representative gradient-based and model-agnostic XAI methods, including Integrated Gradients, Grad-CAM++, Occlusion, LIME, and KernelSHAP. Quantitative fidelity is assessed using perturbation-based criteria such as Deletion and ROAD, whereas qualitative analyses examine the physiological plausibility of the resulting topographic and frequency-band attribution patterns. Overall, STF-KernelSHAP provides a unified explanation strategy that preserves standard spatial EEG interpretability while extending it toward structured space–time–frequency analysis.

The remainder of this paper is organized as follows: Section 2 presents the related work. Section 3 introduces the materials and methods. Section 4 describes the experimental set-up. Section 5 presents and discusses the results. Finally, Section 6 provides the concluding remarks.

2. Related Work

2.1. EEG Classification and Decoding Architectures

The evolution of EEG decoding has progressively transitioned from manual feature engineering toward end-to-end learning paradigms, largely driven by the increasing availability of large-scale neurophysiological datasets [19]. Early approaches relied extensively on statistical signal processing techniques to isolate informative components from noise, thereby establishing a rigorous foundation for handcrafted feature extraction prior to the emergence of deep learning methods [20].

In this context, comprehensive reviews have highlighted that deep learning (DL) has become a central technology for EEG decoding across a wide range of applications, including rehabilitation and clinical diagnosis [21]. These data-driven approaches substantially outperform classical feature engineering techniques, such as Common Spatial Patterns (CSP), by automatically learning hierarchical representations directly from raw signals [22]. However, the expressive non-linear nature of these models renders the classification process increasingly opaque. Several CSP variants, including Filter Bank CSP (FBCSP) and L1-norm CSP (L1CSP), were proposed to improve robustness by optimizing frequency selection and reducing artifact sensitivity. Nevertheless, despite these enhancements, traditional approaches remain highly power-dependent and sensitive to noise, reinforcing the intrinsic limitations of handcrafted pipelines [23].

 doi:10.20944/preprints202606.0375.v1

To bridge the gap between classical signal processing and neural representation learning, a class of specialized, neuro-inspired architectures has emerged, aiming to replicate traditional filtering operations within neural layers [24]. These models seek to preserve the interpretability of linear filters while leveraging the expressive power of deep networks, progressively moving away from generic computer vision architectures toward domain-specific designs optimized for time-series data [25].

Within this paradigm, EEGNet constitutes a seminal example, introducing temporal and depth-wise convolutional layers to explicitly model spectral and spatial EEG features in a manner analogous to traditional filter banks [26]. Related architectures, such as ShallowConvNet and DeepConvNet, emphasize compact designs to efficiently capture band-power modulations [27]. Subsequent developments include TCFusionNet, which integrates dilated convolutions to enlarge the receptive field, as well as deep kernel learning approaches such as KREEGNet, which compute functional connectivity through Gaussian kernels [28]. Although these architectures encode spatial and temporal structure by design, they struggle to accommodate the pronounced non-stationarity of EEG signals across subjects and trials without complex adaptive mechanisms, such as deformable convolutions. As a result, their interpretability remains closely tied to specific architectural assumptions and geometric transformations [29].

More recently, EEG decoding has shifted toward modeling global dependencies and complex relational structures [30]. This transition reflects a broader view of the brain not merely as a grid of sensors, but as an interconnected network exhibiting long-range temporal and spatial interactions [31]. Representative examples include CT-Net, which combines convolutional neural networks and Transformers for refined feature extraction, as well as Spatial Graph Neural Networks (SGNNs), which explicitly encode electrode connectivity via graph topology [32]. In parallel, unsupervised architectures such as Deep Belief Networks (DBNs) and Autoencoders have been employed for feature reconstruction and dimensionality reduction [33]. While these models improve multidomain representation learning, their interpretability is typically derived from attention weights or graph pooling scores [34]. In practice, such mechanisms often capture correlation rather than causal relevance, thereby coupling explanations to internal model components instead of offering physiologically grounded insight [35].

2.2. Interpretability and Explainable AI in EEG Analysis

The increasing architectural complexity of EEG decoding models—and the resulting opacity of modern classifiers—has amplified the demand for principled interpretability mechanisms, particularly in high-stakes clinical and BCI scenarios [36]. In this context, and to address the diversity of explanatory requirements, the literature has converged on taxonomies that categorize explainable artificial intelligence (XAI) methods according to their operational principles [37].

A widely adopted taxonomy distinguishes XAI techniques into four categories: example-based, rule-based, hidden semantics, and attribution-based approaches [38]. Example-based methods, such as Influence Functions and prototype learning, explain predictions by identifying influential or representative training instances [39]. Rule-based approaches approximate complex decision boundaries through symbolic logic, often using decision trees or if–then rule sets [40]. Hidden semantics methods analyze internal neuron activations to associate them with abstract concepts, employing techniques such as activation maximization or network dissection [41].

For high-dimensional neurophysiological signals, however, attribution-based methods have emerged as the most relevant category [42]. These techniques explicitly map model predictions back to the input space, assigning quantitative relevance scores to individual features such as time samples or electrodes. Gradient-based approaches, including Saliency Maps and Layer-wise Relevance Propagation (LRP), exploit the differentiable structure of neural networks to trace activation flow from the output to the input [43]. Grad-CAM and its extensions, such as Grad-CAM++ and LayerCAM, have been adapted to one-dimensional biosignals to highlight discriminative temporal or spatial regions [44]. Despite their effectiveness within convolutional architectures, these methods require access to internal gradients and are therefore limited to differentiable models, restricting their applicability to heterogeneous or ensemble-based EEG pipelines [12].

 doi:10.20944/preprints202606.0375.v1

To overcome these architectural constraints, perturbation-based methods have been proposed as model-agnostic alternatives [45]. By treating the classifier as an oracle and observing output variations in response to input perturbations, these approaches estimate feature importance without accessing internal parameters, often drawing on cooperative game theory [46]. SHAP represents a prominent example and has been applied to the interpretation of cognitive and emotional states in BCI systems, alongside LIME [47]. However, these general-purpose explainers typically rely on pointwise perturbations that assume feature independence, leading to physiologically implausible scenarios and a loss of signal coherence due to the violation of inherent dependencies [48].

Recent efforts have attempted to mitigate this limitation by grouping features based on correlation or spatial proximity, thereby preserving structural relationships during the explanation process [49]. CorrSHAP, for instance, groups correlated features to reduce computational cost and improve structural fidelity, while related segmentation and masking strategies have been explored to identify biomarkers in motor learning tasks [16]. Nevertheless, these approaches are largely adapted from computer vision, where dependencies are predominantly local and spatial [50]. As a result, they fail to capture the non-local functional connectivity and cross-frequency interactions characteristic of EEG signals, underscoring the need for domain-specific interpretability strategies [11].

Despite the breadth of existing methods, a persistent gap remains in consistently addressing the multidimensional nature of EEG, where spatial, temporal, and spectral components are intrinsically coupled [51]. Current approaches often impose a trade-off between physiological fidelity and architectural flexibility, motivating the search for unified frameworks that reconcile both aspects [52]. Explanation quality varies substantially across models and samples, and hybrid solutions frequently revert to temporal proxies that introduce spectral leakage or lack unified quantitative measures [53]. Consequently, increasing model complexity may enhance performance while simultaneously degrading interpretability [54]. This unresolved tension motivates the development of model-agnostic approaches that respect the tri-domain organization of EEG, namely space, time, and frequency, without relying on arbitrary segmentation or architectural constraints, thereby justifying the proposed STF-KernelSHAP framework.

3. Materials and Methods

3.1. Tested Datasets

To assess the robustness and versatility of the proposed STF-KernelSHAP framework, experiments were conducted on two datasets drawn from distinct neurophysiological domains. The first dataset corresponds to a classical Brain–Computer Interface (BCI) paradigm based on voluntary motor control and high-density EEG recordings, whereas the second targets the clinical characterization of a neurodevelopmental disorder using a standard clinical montage. This dual selection enables a comprehensive evaluation of the interpretability of STF-KernelSHAP across heterogeneous acquisition systems, sampling rates, and cognitive states.

For the evaluation of motor imagery (MI), we employed the GIGAScience dataset [55]. The original dataset comprises EEG recordings from 52 healthy subjects; however, participants identified as 29 and 34 were excluded due to data inconsistencies, yielding a final cohort of 50 subjects. EEG signals were acquired using a 64-channel Ag/AgCl active electrode system (BioSemi ActiveTwo), arranged according to the international 10–10 system (Figure 1), with a sampling rate of 512 Hz. Signal acquisition and experimental control were managed through the BCI2000 platform, which also delivered visual cues instructing left- or right-hand motor imagery tasks. To ensure adherence to the MI paradigm and exclude overt motor execution, electromyographic (EMG) signals were recorded concurrently, allowing verification of the absence of actual physical movement during imagery intervals.

 doi:10.20944/preprints202606.0375.v1

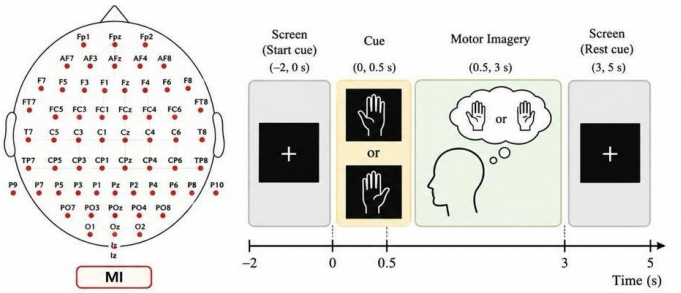

Figure 1. Experimental protocol and acquisition configuration of the GIGAScience dataset for MI–EEG classification. **Right:** Temporal structure of a single trial, where a visual cue prepares the subject, followed by a three-second interval dedicated to the imagination of left- or right-hand movement. **Left:** Spatial arrangement of EEG electrodes, illustrating the sequential channel layout from left frontal regions toward posterior areas and back along the central axis, in accordance with the international 10–10 system.

Complementing the analysis of oscillatory motor tasks, we extended the evaluation to a clinical cognitive context by incorporating a publicly available EEG dataset from IEEE DataPort [56]. This dataset comprises recordings from children aged 7 to 12 years, divided into a group diagnosed with Attention Deficit Hyperactivity Disorder (ADHD) and a healthy control group. Diagnoses were established by an experienced psychiatrist according to DSM-IV criteria; participants in the ADHD group had received Ritalin treatment for a period not exceeding six months, whereas control subjects had no history of psychiatric or neurological disorders. To ensure class balance in the classification tasks, one subject from the ADHD group was randomly excluded, resulting in a final balanced cohort of 120 participants (60 ADHD and 60 controls).

In this clinical setting, EEG signals were acquired at a sampling frequency of 128 Hz using a standard 19-electrode montage based on the international 10–20 system, referenced to the earlobes A1 and A2. The spatial configuration of the electrodes is depicted in Figure 2. The experimental protocol evaluated sustained visual attention, a cognitive function commonly impaired in ADHD, by means of a perceptual counting task involving sequential visual stimuli. For the subsequent analysis, the continuous EEG recordings were segmented into 4-second epochs (512 samples) with a 50% overlap. Each epoch was treated as an independent instance, while strictly preserving subject identity in order to prevent data leakage across validation folds.

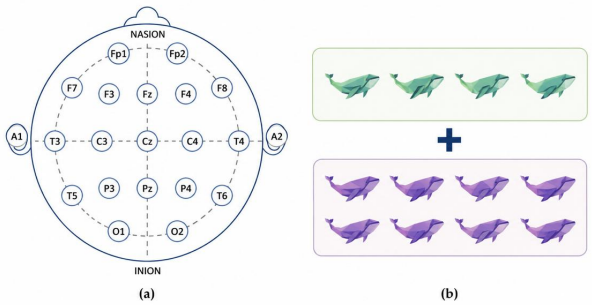

Figure 2. EEG channel configuration of the ADHD dataset, acquired using a standard 19-channel montage based on the international 10–20 system.

 doi:10.20944/preprints202606.0375.v1

3.2. Classification Deep Learning Models

Given an EEG trial represented as $\boldsymbol{X}_n = \{\boldsymbol{x}_{\check{c}} \in \mathbb{R}^T : \check{c} \in \check{C}\}$, and its corresponding ground-truth label $\boldsymbol{y}_n \in \{0,1\}^C$, where $\check{C}$ denotes the number of electrodes, T the number of temporal samples, and C the number of classes, a deep learning classifier is employed to map the resulting multichannel signal into a vector of class probabilities.

Specifically, a classifier $\mathcal{F} : \mathbb{R}^{\check{C}\times T} \to [0,1]^C$ produces a predicted class-probability vector $\hat{\boldsymbol{y}}_n \in [0,1]^C$, which is defined through a composition of successive nonlinear transformations as follows:

$$\hat{\boldsymbol{y}}_n = \mathcal{F}(\boldsymbol{X}_n) = (\check{f}_L \circ \check{f}_{L-1} \circ \cdots \circ \check{f}_1)(\boldsymbol{X}_n), \tag{1}$$

where $\check{f}_l(\cdot)$ denotes the l-th layer of the classifier for $l \in \{1,\ldots,L\}$, and $\circ$ represents the composition operator. The output vector satisfies the probability simplex constraints, namely $\sum_{c=1}^{C} \hat{y}_{n,c} = 1$ with each component $\hat{y}_{n,c} \in \hat{\boldsymbol{y}}_n$.

For multi-class classification, the model output $\hat{\boldsymbol{y}}$ is interpreted as a categorical probability distribution over the C classes. Accordingly, the training objective is formulated using the cross-entropy loss function:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log\left(\hat{y}_{n,c}\right), \tag{2}$$

Within this general learning formulation, the classifier $\mathcal{F}$ is not restricted to a particular EEG decoding architecture. Instead, it provides the predictive function whose decisions are subsequently explained by post-hoc attribution methods. In this work, different deep learning models are used to evaluate whether the proposed STF-KernelSHAP framework can generate consistent explanations across architectures with distinct spatial, temporal, and spectral representation mechanisms. Therefore, classification performance is considered as the predictive basis for the interpretability analysis, rather than as the main methodological contribution of the study.

3.3. SHAP Fundamentals

Due to the complex and highly non-linear composition of layers in $\mathcal{F}$, as described in Eq. 1, the resulting model operates effectively as a black box, which precludes direct inspection of its internal decision mechanisms. This intrinsic lack of transparency motivates the use of post-hoc interpretability methods. In this context, additive feature attribution approaches are particularly relevant, as they seek to locally approximate the behavior of the classifier by means of a linear surrogate defined over a simplified and interpretable representation of the input [57].

Let $\mathcal{M} = \{\mathcal{M}_1,\ldots,\mathcal{M}_M\}$ denote the set of interpretable features, where $M = \check{C} \times T$ corresponds to the total number of interpretable components. For a given sample n, the simplified input representation is denoted by $\boldsymbol{z}_n \in \{0,1\}^M$, whose binary entries indicate the presence or absence of each interpretable feature. To incorporate the bias term within a unified linear formulation, the extended representation $\tilde{\boldsymbol{z}}_n = [1, \boldsymbol{z}_n^\top]^\top \in \mathbb{R}^{M+1}$ is introduced.

Within this framework, the surrogate output across all C classes is denoted by $\check{\boldsymbol{y}}_n \in \mathbb{R}^C$ and is expressed as

$$\check{\boldsymbol{y}}_n = \boldsymbol{\Phi}_n \tilde{\boldsymbol{z}}_n, \tag{3}$$

where $\boldsymbol{\Phi}_n \in \mathbb{R}^{C\times(M+1)}$ denotes the matrix of class-specific attribution coefficients. The c-th row of $\boldsymbol{\Phi}_n$, given by $\boldsymbol{\phi}_{n,c}^\top = [\phi_{n,c,0}, \phi_{n,c,1}, \ldots, \phi_{n,c,M}]$, collects the base value together with the feature attributions associated with class c. The base value is defined as $\phi_{n,c,0} = \mathcal{F}(\boldsymbol{X}_{n,\varnothing})^\top \boldsymbol{y}_n$, where $\boldsymbol{y}_n$ acts as a selector for the class of interest.

To connect the surrogate model with the original input space $\mathbb{R}^{\check{C}\times T}$, the simplified representation $\boldsymbol{z}_n$ is mapped back through a reconstruction function $\mathcal{H}_{\boldsymbol{X}} : \{0,1\}^M \to \mathbb{R}^{\check{C}\times T}$. This mapping satisfies $\boldsymbol{X}_n = \mathcal{H}_{\boldsymbol{X}}(\mathbf{1})$, where $\mathbf{1}$ denotes the all-ones vector. Under this construction, the surrogate output satisfies $\check{\boldsymbol{y}}_n \approx \mathcal{F}(\mathcal{H}_{\boldsymbol{X}}(\boldsymbol{z}_n)) \quad \text{for } \boldsymbol{z}_n \approx \mathbf{1}$, thereby ensuring local fidelity in the neighborhood of the original input.

 doi:10.20944/preprints202606.0375.v1

Having defined the additive surrogate model, the remaining task is to determine the attribution coefficients in $\mathbf{\Phi}_n$ in a principled and theoretically grounded manner. This is achieved by adopting Shapley values, which provide a canonical mechanism for quantifying the contribution of each interpretable feature to the model output [58].

For a given feature $m \in \mathcal{M}$, let $\mathcal{S}_m = \{S_i\}_{i=1}^{2^{M-1}}$, with $S_i \subseteq \mathcal{M} \setminus \{m\}$, denote the set of all coalitions that exclude feature m. Each coalition S_i defines a reference context against which the marginal contribution of feature m is evaluated. For any $S_i \in \mathcal{S}_m$, the reconstructed input $\boldsymbol{X}_n^{S_i} = \mathcal{H}_{\boldsymbol{X}}(\boldsymbol{z}_n^{S_i})$ corresponds to the original input $\boldsymbol{X}_n$ restricted to the features in S_i, while all remaining components are replaced by the reference values specified by $\mathcal{H}_{\boldsymbol{X}}$.

Within this formulation, the Shapley value associated with feature m and class c is defined as

$$\phi_{n,c,m} = \sum_{S_i \in \mathcal{S}_m} \Pi(S_i, M) \left(\mathcal{F}(\boldsymbol{X}_n^{S_i \cup \{m\}}) - \mathcal{F}(\boldsymbol{X}_n^{S_i}) \right)^\top \boldsymbol{y}_n, \tag{4}$$

where the weighting function $\Pi(S_i, M) = \frac{|S_i|!\left(M-|S_i|-1\right)!}{M!}$ depends solely on the coalition cardinality and the total number of interpretable features.

While the exact Shapley formulation provides a rigorous attribution criterion, it becomes computationally intractable when the number of interpretable features M is large, due to the exhaustive enumeration of 2^{M-1} possible coalitions. KernelSHAP addresses this limitation by introducing a finite-sampling approximation based on a set of sampled coalitions [46] $\mathcal{Z}_n = \{\boldsymbol{z}_n^i\}_{i=1}^{N_s} \subset \{0,1\}^M$,where N_s denotes the number of coalitions independently and identically distributed (i.i.d.) according to a prescribed distribution over $\{0,1\}^M$. This strategy enables the approximation of the expectations implicit in the theoretical Shapley definition using a finite number of classifier evaluations. Feature attributions are then obtained by fitting the additive surrogate model through a weighted least-squares procedure,

$$\begin{aligned} \boldsymbol{\phi}_{n,c}^* = \arg\min_{\boldsymbol{\phi}_{n,c}} & \sum_{i=1}^{N_s} \Pi\left(\boldsymbol{z}_n^i\right) \left(\mathcal{F}\left(\mathcal{H}_{\boldsymbol{X}}\left(\boldsymbol{z}_n^i\right)\right)^\top \boldsymbol{y}_n - \boldsymbol{\phi}_{n,c}^\top \tilde{\boldsymbol{z}}_n^i \right)^2 \\ \text{s.t.} \quad & \boldsymbol{\phi}_{n,c}^\top \tilde{\mathbf{0}} = \mathcal{F}(\mathcal{H}_{\boldsymbol{X}}(\mathbf{0}))^\top \boldsymbol{y}_n, \\ & \boldsymbol{\phi}_{n,c}^\top \tilde{\mathbf{1}} = \mathcal{F}(\mathcal{H}_{\boldsymbol{X}}(\mathbf{1}))^\top \boldsymbol{y}_n. \end{aligned} \tag{5}$$

Here, $\tilde{\boldsymbol{z}}_n^i = [1, (\boldsymbol{z}_n^i)^\top]^\top \in \mathbb{R}^{M+1}$ explicitly incorporates the bias term. The constraints are enforced via the extended representations of the empty coalition $\tilde{\mathbf{0}} = [1, \mathbf{0}^\top]^\top$ and the full coalition $\tilde{\mathbf{1}} = [1, \mathbf{1}^\top]^\top$, thereby ensuring exactness at both boundary cases. The weighting function $\Pi(\boldsymbol{z}_n^i) = \frac{M-1}{\binom{M}{|\boldsymbol{z}_n^i|}\,|\boldsymbol{z}_n^i|\,(M-|\boldsymbol{z}_n^i|)}$ corresponds to the Shapley kernel and depends exclusively on the coalition cardinality.

From this perspective, KernelSHAP provides a classical additive attribution formulation in which the Shapley values are estimated through a weighted linear surrogate fitted over sampled coalitions. When the weighted design matrix is well conditioned, Eq. (5) can be solved using standard weighted least-squares solvers, whereas regularized variants may be adopted to improve numerical stability in ill-conditioned settings. In practical implementations, the boundary constraints associated with the empty and full coalitions are commonly enforced through anchor coalitions with large weights, ensuring local accuracy at both reference cases. Thus, KernelSHAP offers a general model-agnostic basis for estimating feature contributions, which is subsequently adapted in this work to structured EEG channel–time–frequency coalitions.

The relevance of Shapley values in the context of model interpretability stems from their axiomatic foundation. These properties establish formal requirements for additive feature attributions and justify the use of Shapley values as a principled mechanism for assigning relevance to individual interpretable components. In particular, the following properties are central to the SHAP formulation [59].

– Property 1: Missingness. If a feature is absent from the simplified representation, i.e., $z_{n,m} = 0$, then it contributes nothing to the explanation, which implies $\phi_{n,c,m} = 0$.

 doi:10.20944/preprints202606.0375.v1

– Property 2: Local accuracy. The explanation model must recover the model output for the original input, such that the sum of the base value and all feature attributions equals the target model score:

$$\mathcal{F}(\boldsymbol{X}_n)^\top \boldsymbol{y}_n = \phi_{n,c,0} + \sum_{m=1}^{M} \phi_{n,c,m}.$$

– Property 3: Consistency. Let $\mathcal{F}$ and $\mathcal{F}'$ denote two predictive models with corresponding attribution coefficients $\phi_{n,c,m}$ and $\phi'_{n,c,m}$. If, for all coalitions $S_i \in \mathcal{S}_m$, the marginal contribution of feature m under $\mathcal{F}'$ is greater than or equal to that under $\mathcal{F}$, namely,

$$\mathcal{F}'(\boldsymbol{X}_n^{S_i\cup\{m\}})^\top \boldsymbol{y}_n - \mathcal{F}'(\boldsymbol{X}_n^{S_i})^\top \boldsymbol{y}_n \;\geq\; \mathcal{F}(\boldsymbol{X}_n^{S_i\cup\{m\}})^\top \boldsymbol{y}_n - \mathcal{F}(\boldsymbol{X}_n^{S_i})^\top \boldsymbol{y}_n, \tag{6}$$

then the attribution assigned to feature m cannot decrease, i.e., $\phi'_{n,c,m} \geq \phi_{n,c,m}$.

These axioms are not merely desirable qualitative properties; rather, they uniquely characterize the Shapley value solution within the class of additive feature attribution methods, as formalized by the following theorem [60].

Theorem 1. Within the class of additive feature attribution methods, the Shapley value formulation in Eq. 4 is the unique solution satisfying the standard SHAP axioms of local accuracy, missingness, and consistency.

3.4. EEG-Driven Multidomain Shapley Attribution Framework

We instantiate the general SHAP framework introduced above for structured multichannel EEG signals by redefining the interpretable feature space over channel-wise time–frequency cells and by introducing a reconstruction operator that maps each sampled coalition back to the original signal domain.

Let $\boldsymbol{X}_n \in \mathbb{R}^{\check{C}\times T}$ denote the input signal for sample n. We first project $\boldsymbol{X}_n$ onto the time–frequency domain through a deterministic transformation $\mathcal{T} : \mathbb{R}^{\check{C}\times T} \to \mathbb{C}^{\check{C}\times\check{T}\times K}$, yielding $\check{\boldsymbol{X}}_n = \mathcal{T}(\boldsymbol{X}_n)$, where $\check{T}$ denotes the number of temporal windows and K denotes the number of discrete spectral components.

The time–frequency plane $\{1,\ldots,\check{T}\} \times \{1,\ldots,K\}$ is partitioned into Q non-overlapping cells $\{\mathcal{G}_q\}_{q=1}^{Q}$, defined as disjoint window–band regions that jointly cover the plane, with $Q = (\#\,\text{windows}) \times (\#\,\text{bands})$. Accordingly, the interpretable feature space is defined as $\check{M} = \check{C} \times Q$, so that each interpretable feature corresponds to one time–frequency cell within one channel. Coalitions are encoded by binary vectors $\check{\boldsymbol{z}}_n^i \in \{0,1\}^{\check{M}}$, collected in the finite set $\check{\mathcal{Z}}_n = \{\check{\boldsymbol{z}}_n^i\}_{i=1}^{N_s}$.

To evaluate each coalition in the original input space, we define the composite reconstruction mapping

$$\check{\mathcal{H}}_{\boldsymbol{X}} = \mathcal{T}^{-1} \circ \mathcal{H}_{\mathrm{TF}} \circ \mathcal{H}_Q. \tag{7}$$

Here, $\mathcal{H}_Q : \{0,1\}^{\check{M}} \to \mathbb{R}^{\check{C}\times Q}$ maps a binary coalition to a channel–cell representation by activating or deactivating complete time–frequency cells within each channel. The operator $\mathcal{H}_{\mathrm{TF}} : \mathbb{R}^{\check{C}\times Q} \to \mathbb{C}^{\check{C}\times\check{T}\times K}$ expands this representation over the indices defined by $\mathcal{G}_q$, and $\mathcal{T}^{-1} : \mathbb{C}^{\check{C}\times\check{T}\times K} \to \mathbb{R}^{\check{C}\times T}$ maps the reconstructed representation back to the temporal domain. In this way, each coalition S_i induces a valid realization

$$\boldsymbol{X}_n^{S_i} = \check{\mathcal{H}}_{\boldsymbol{X}}(\check{\boldsymbol{z}}_n^i), \tag{8}$$

which can be directly evaluated by the classifier.

 doi:10.20944/preprints202606.0375.v1

Under this structured coalition space, the attribution coefficients are estimated through the KernelSHAP weighted least-squares formulation adapted to the proposed EEG representation. For class c, this estimator is given by

$$\begin{aligned} \boldsymbol{\phi}^{*}_{n,c} = \arg\min_{\boldsymbol{\phi}_{n,c}} \sum_{i=1}^{N_s} \Pi\left(\check{\mathbf{z}}^{i}_{n}\right) \left(\mathcal{F}\left(\check{\mathcal{H}}_{\mathbf{X}}\left(\check{\mathbf{z}}^{i}_{n}\right)\right)^{\top} \boldsymbol{y}_n - \boldsymbol{\phi}^{\top}_{n,c} \hat{\mathbf{z}}^{i}_{n} \right)^2 \\ \text{s.t.} \quad \boldsymbol{\phi}^{\top}_{n,c} \hat{\mathbf{0}} = \mathcal{F}\left(\check{\mathcal{H}}_{\mathbf{X}}(\check{\mathbf{0}})\right)^{\top} \boldsymbol{y}_n, \\ \boldsymbol{\phi}^{\top}_{n,c} \hat{\mathbf{1}} = \mathcal{F}\left(\check{\mathcal{H}}_{\mathbf{X}}(\check{\mathbf{1}})\right)^{\top} \boldsymbol{y}_n. \end{aligned} \tag{9}$$

where $\hat{\mathbf{z}}^{i}_{n} = [1, (\check{\mathbf{z}}^{i}_{n})^{\top}]^{\top} \in \mathbb{R}^{\check{M}+1}$, $\hat{\mathbf{0}} = [1, \check{\mathbf{0}}^{\top}]^{\top}$, and $\hat{\mathbf{1}} = [1, \check{\mathbf{1}}^{\top}]^{\top}$. The Shapley kernel is preserved from the general formulation, with the argument now defined over the proposed channel–cell coalition space.

Finally, the resulting attributions are organized into the tensor $\boldsymbol{\Phi}_n \in \mathbb{R}^{C \times (\check{C}Q+1)}$, which aggregates, for each class, the contributions of all time–frequency cells across all channels, together with the bias term. The complete workflow of the proposed EEG-driven multidomain Shapley attribution strategy is summarized in Figure 3.

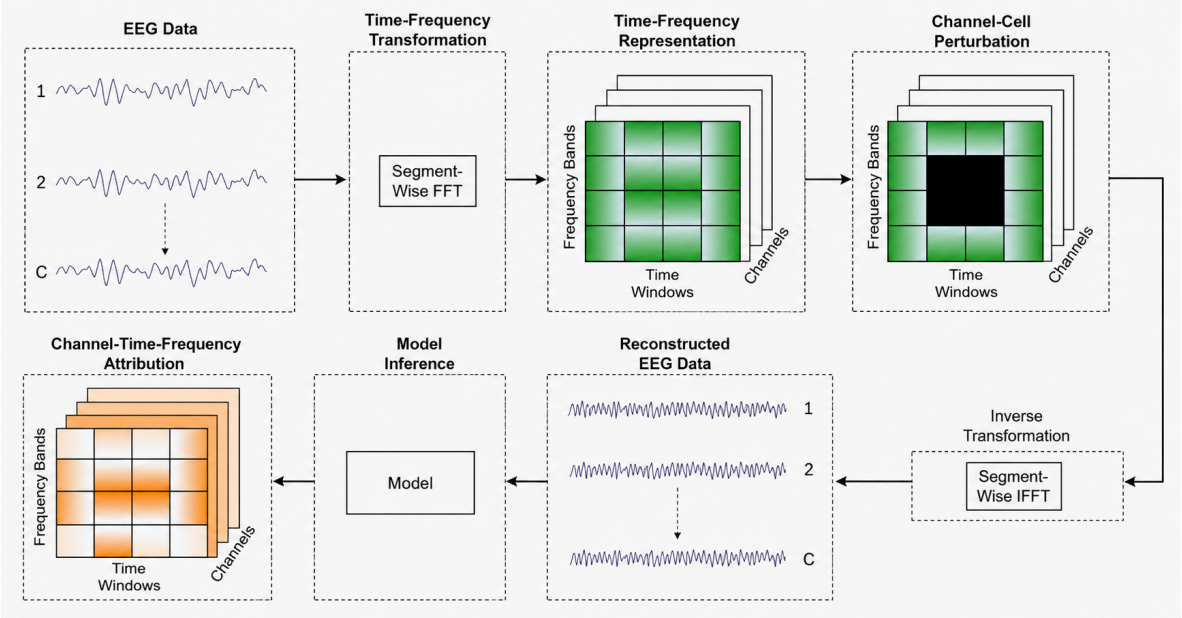

Figure 3. Overview of the proposed EEG-driven multidomain Shapley attribution framework.

4. Experimental Set-Up

4.1. Assessment and Method Comparison

The experimental evaluation was structured to compare, under a common protocol, both the predictive performance of the EEG classifiers and the fidelity of the explanations generated by the considered XAI strategies. Accordingly, this subsection first presents the classification models employed and then describes the performance metrics, explanation methods, and perturbation-based fidelity criteria.

First, EEGNet was considered, a compact convolutional architecture specifically designed for EEG signal classification [61]. Likewise, ShallowConvNet was included as a shallow convolutional network aimed at capturing relevant spatio–temporal patterns in EEG signals [62]. Finally, T-GARNet was evaluated as an architecture that integrates temporal encoding and kernelized representations for EEG classification [63].

 doi:10.20944/preprints202606.0375.v1

Once the classifiers were established, predictive performance was quantified using the following metrics:

- Accuracy (ACC): measures the proportion of correctly classified trials with respect to the total number of evaluated samples:

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

where TP and TN denote true positives and true negatives, whereas FP and FN correspond to false positives and false negatives.

- Area under the ROC curve (AUC): quantifies the discriminative capability of the classifier by integrating the relationship between the true positive rate and the false positive rate across different thresholds:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}\Big(\mathrm{FPR}^{-1}(\nu)\Big)\, d\nu, \tag{11}$$

where TPR represents the true positive rate, FPR the false positive rate, and $\nu \in [0,1]$ is an auxiliary integration variable.

- Cohen's kappa coefficient (κ): measures the agreement between the predicted labels and the ground-truth labels, correcting for the agreement expected by chance:

$$\kappa = \frac{\varrho_o - \varrho_e}{1 - \varrho_e}, \tag{12}$$

where ϱ_o is the observed agreement and ϱ_e is the agreement expected by chance.

Subsequently, from the trained and evaluated models, local explanations were generated to identify the signal regions contributing to the decision of each classifier. In the following formulations, $\boldsymbol{y}_n$ is used as a class-selector vector to extract the scalar model response associated with the class of interest. Its practical definition depends on the analysis: predicted classes are used for perturbation-based fidelity assessment, whereas ground-truth labels are used for class-wise topographic interpretation. To ensure a homogeneous comparison, all XAI strategies were applied to the same EEG trial $\boldsymbol{X}_n$, the same target class selected by $\boldsymbol{y}_n$, and the same scalar output $\mathcal{F}(\boldsymbol{X}_n)^\top \boldsymbol{y}_n$. Under this configuration, each method produces an attribution $\boldsymbol{\phi}_{n,c}^{(\cdot)}$, which is then used to assess explanation fidelity through controlled perturbations of the signal.

The considered XAI strategies are described below:

- KernelSHAP [64]: estimates the contribution of each interpretable feature by fitting a weighted additive surrogate model over perturbed coalitions of the input. In this work, KernelSHAP follows the constrained weighted least-squares formulation previously defined in Eq. (5).
- LIME [65]: fits an interpretable surrogate model in the neighborhood of the explained trial. To this end, let $\boldsymbol{z} \in \{0,1\}^M$ be the interpretable representation associated with a perturbation of $\boldsymbol{X}_n$. In the linear formulation employed in this work, such a local surrogate is expressed as

$$\breve{g}_{\boldsymbol{\beta}_{n,c}}(\boldsymbol{z}) = \beta_{n,c,0} + \sum_{m=1}^{M} \beta_{n,c,m} z_m. \tag{13}$$

The surrogate coefficients are estimated as

$$\boldsymbol{\beta}_{n,c}^{\mathrm{LIME}} = \arg\min_{\boldsymbol{\beta}_{n,c}} \mathcal{L}_{\mathrm{loc}}\Big(\mathcal{F}, \breve{g}_{\boldsymbol{\beta}_{n,c}}, \pi_{\boldsymbol{X}_n}\Big) + \Omega(\boldsymbol{\beta}_{n,c}), \tag{14}$$

 doi:10.20944/preprints202606.0375.v1

where $\mathcal{L}_{\text{loc}}$ measures the local discrepancy between $\mathcal{F}$ and the surrogate, $\pi_{\boldsymbol{X}_n}$ weights the perturbations according to their proximity to $\boldsymbol{X}_n$, and Ω regulates the model complexity. Consequently, the optimal surrogate is determined by $\boldsymbol{\beta}_{n,c}^{\text{LIME}}$, while the attributions are defined as

$$\boldsymbol{\phi}_{n,c}^{\text{LIME}} = \left[\beta_{n,c,1}^{\text{LIME}}, \ldots, \beta_{n,c,M}^{\text{LIME}}\right]^\top. \tag{15}$$

- Integrated Gradients [66]: computes attributions by integrating the gradients of the output associated with the target class along a continuous path between a reference $\boldsymbol{X}_B$ and the trial $\boldsymbol{X}_n$:

$$\boldsymbol{\phi}_{n,c}^{\text{IG}} = (\boldsymbol{X}_n - \boldsymbol{X}_B) \odot \int_0^1 \nabla_{\boldsymbol{X}} \left[\mathcal{F}(\boldsymbol{X}_B + \eta(\boldsymbol{X}_n - \boldsymbol{X}_B))^\top \boldsymbol{y}_n\right] d\eta, \tag{16}$$

  where $\eta \in [0,1]$ is the interpolation parameter and $\odot$ represents the Hadamard product.
- Occlusion [42]: estimates the relevance of an input region by replacing it with a reference and quantifying the induced change in the target-class score. Let $\mathcal{R} = \{\check{r}_1, \ldots, \check{r}_R\}$ be the set of occlusion regions. For a region $\check{r} \in \mathcal{R}$, let $\boldsymbol{X}_n^{\check{r}}$ denote the perturbed version of $\boldsymbol{X}_n$, in which only the region $\check{r}$ is replaced by the reference. In this case, the regional attribution is defined as

$$\phi_{n,c}^{\text{Occ}}(\check{r}) = \mathcal{F}(\boldsymbol{X}_n)^\top \boldsymbol{y}_n - \mathcal{F}\left(\boldsymbol{X}_n^{\check{r}}\right)^\top \boldsymbol{y}_n. \tag{17}$$

  Therefore, the complete occlusion-based explanation is given by

$$\boldsymbol{\phi}_{n,c}^{\text{Occ}} = \left[\phi_{n,c}^{\text{Occ}}(\check{r}_1), \ldots, \phi_{n,c}^{\text{Occ}}(\check{r}_R)\right]^\top. \tag{18}$$

- Grad-CAM++ [67]: obtains a relevance map from the activations of an internal convolutional layer:

$$\boldsymbol{\phi}_{n,c}^{\text{GC++}} = \text{ReLU}\left(\sum_{\bar{k}=1}^{\bar{K}} \omega_{n,c,\bar{k}}^{\text{GC++}} \boldsymbol{B}_n^{\bar{k}}\right), \tag{19}$$

  where $\boldsymbol{B}_n^{\bar{k}}$ is the $\bar{k}$-th activation map of the selected layer, $\omega_{n,c,\bar{k}}^{\text{GC++}}$ represents its weight associated with the target class, and $\bar{K}$ is the number of activation maps considered.

Finally, explanation fidelity was evaluated using MoRF Deletion and ROAD. For each trial $\boldsymbol{X}_n$, a retention mask is defined as

$$\boldsymbol{z}_{n,c}^{\rho} = \left[z_{n,c,1}^{\rho}, \ldots, z_{n,c,M}^{\rho}\right]^\top \in \{0,1\}^M, \tag{20}$$

obtained by deactivating a fraction $\rho \in [0,1]$ of the features with the highest attribution in $\boldsymbol{\phi}_{n,c}^{(\cdot)}$, following the Most Relevant First criterion. Thus, $z_{n,c,m}^{\rho} = 0$ indicates the removal of feature m, whereas $z_{n,c,m}^{\rho} = 1$ indicates its preservation.

In MoRF Deletion, the perturbed trial and its deletion curve are defined as

$$\boldsymbol{X}_{n,\text{MoRF}}^{\rho} = \mathcal{H}_{\boldsymbol{X}}\left(\boldsymbol{z}_{n,c}^{\rho}\right), \qquad D_{\text{MoRF}}(\rho) = \mathcal{F}\left(\boldsymbol{X}_{n,\text{MoRF}}^{\rho}\right)^\top \boldsymbol{y}_n. \tag{21}$$

A rapid decrease in $D_{\text{MoRF}}(\rho)$ indicates that the removed features exert a relevant influence on the classifier decision.

In turn, in ROAD, the perturbation is performed using an explicit reference $\bar{\boldsymbol{X}}_n$, and the post-removal performance is computed as

$$\boldsymbol{X}_{n,\text{ROAD}}^{\rho} = \bar{\mathcal{H}}_{\boldsymbol{X}}\left(\boldsymbol{z}_{n,c}^{\rho}, \bar{\boldsymbol{X}}_n\right), \qquad \text{ROAD}(\rho) = \mathcal{Q}\left(\mathcal{F}\left(\boldsymbol{X}_{n,\text{ROAD}}^{\rho}\right), \boldsymbol{y}_n\right). \tag{22}$$

Here, $\mathcal{Q}(\cdot,\cdot)$ denotes a generic predictive evaluation criterion.

 doi:10.20944/preprints202606.0375.v1

4.2. Training Details

To ensure a fair and reproducible evaluation, all models were trained using stratified five-fold cross-validation, adapted to the structure of each database. In MI, EEG signals were filtered using a fifth-order Butterworth bandpass filter between 4 and 40 Hz, downsampled from 512 Hz to 128 Hz, and subsequently segmented into two temporal partitions: 0–7 s and 2.5–5 s. For this database, stratification was performed at the sample level independently for each subject. In ADHD, the signals were sampled at 128 Hz, a single 0–4 s window was used, and a notch filter was applied to suppress the 50 Hz power-line component. Unlike MI, the partitioning was carried out at the subject level to avoid information leakage between EEG segments from the same participant, assigning a single class label per subject. In both databases, each fold included an external training–test split with an 80%–20% ratio, while the training set was further divided into training and validation subsets, reserving 20% through a stratified random split. The average spectral behavior of both EEG databases after preprocessing is summarized in Figure 4.

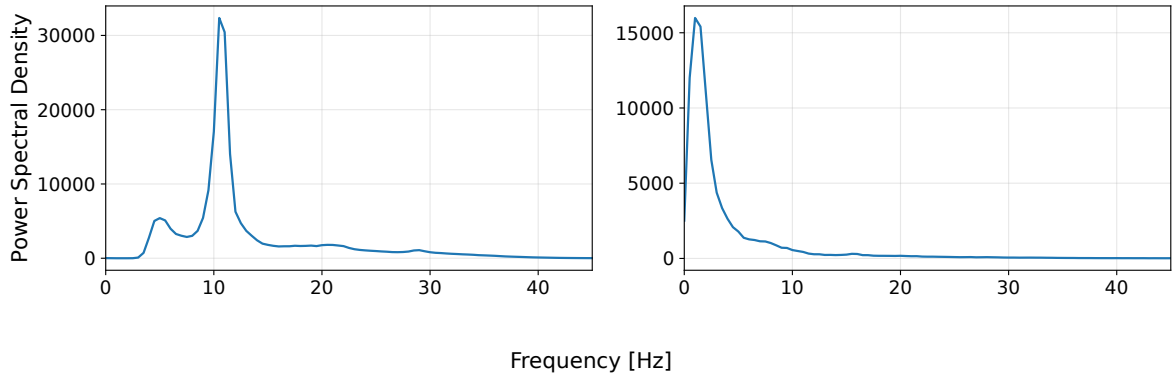

Figure 4. Average power spectral density of the EEG signals. **Left:** Mean frequency spectrum for the MI dataset. **Right:** Mean frequency spectrum for the ADHD dataset.

Based on these partitions, training was implemented in TensorFlow using the Adam optimizer. All models were trained for a maximum of 100 epochs, with a batch size of 16 and a fixed seed of 42. To control overfitting, early stopping was applied to the validation loss, with a patience of 25 epochs, a minimum change of 10^{-4}, and restoration of the best weights. Additionally, the learning rate was automatically reduced when the validation loss stopped improving, using a reduction factor of 0.5, a patience of 10 epochs, and a minimum value of 10^{-6}. The main loss function was normalized binary cross-entropy, and performance was monitored using binary accuracy and AUC. In T-GARNet, this function was complemented with a kernel-based Rényi entropy regularization term, incorporated as a second auxiliary output during training.

Hyperparameter optimization was performed with Optuna, using GPSampler with seed 42, local JournalStorage, and MedianPruner with five warm-up steps. In each trial, the objective was to maximize the mean validation accuracy computed across the five folds, retaining only the weights associated with the best trial. The learning rate was explored as a continuous variable on a logarithmic scale within the interval $(10^{-5}, 10^{-3})$. For the three architectures, the dropout rate and norm constraint were explored as continuous variables in $(0.1, 0.75)$, with a step of 0.05. In EEGNet, the number of temporal filters was selected as an integer variable in $[4, 32]$, with a step of 4; the depth multiplier in $[1, 4]$; the number of separable filters was defined as the product of both; and the temporal kernel length was selected from $\{16, 32, 64, 96, 128\}$. In ShallowConvNet, the number of filters, temporal kernel length, and pooling size were selected as integer variables in $[8, 64]$, $[8, 128]$, and $[16, 64]$, respectively, all with a step of 8; whereas the pooling stride was explored in $[2, 32]$, with a step of 2, constrained not to exceed the pooling size. In T-GARNet, the number of convolutional filters and the number of attention heads were selected as integer variables in $[2, 8]$ and $[1, 5]$, respectively; the Gaussian kernel standard deviation was explored as a continuous variable in $(1, 20)$; the intermediate dimension of the

 doi:10.20944/preprints202606.0375.v1

Transformer block was selected from $\{16, 32, 64, 128\}$; and the relative weight of the classification loss was explored as a continuous variable in $(0.1, 0.9)$, while the weight associated with Rényi entropy regularization was defined as its complement.

From the trained models, post-hoc explanations were computed on the test-set samples to obtain relevance maps compatible with the input structure of each EEG signal. KernelSHAP was applied to the flattened representation of the signal, using a reference set adaptively defined as the minimum between 100 samples and 5% of the training set, ensuring at least 8 samples; for attribution estimation, 500 coalition samples were used, with regularization limited to 200 features. Similarly, LIME was applied to the flattened signal, using as reference the minimum between 200 samples and 10% of the training set, with a minimum of 30 samples; additionally, 1000 local perturbations were generated, and up to 200 relevant features were retained. In turn, Occlusion was implemented through channel-wise temporal perturbations, replacing 1 s windows with a 0.25 s stride by an average reference computed from the training set. Integrated Gradients used the stratified average of the training set as baseline and approximated the integral using 50 interpolation steps. Finally, Grad-CAM++ was applied to a compatible convolutional layer of each model, and the resulting relevance map was resized to the input shape to facilitate comparison with the remaining strategies. Specifically, Conv2D_1 was explicitly selected for ShallowConvNet, whereas the last convolutional layer was used for the remaining architectures.

For STF-KernelSHAP, each EEG signal was transformed using a segmented FFT with nfft $= 512$, and coalitions were defined over the channel–time–frequency cells $\mathcal{G}_q$ established in Section 3.4. In MI, the considered bands were $(4, 8)$, $(8, 13)$, $(13, 30)$, and $(30, 40)$ Hz. For the full 0–7 s window, the temporal regions were $(0, 2)$, $(2, 2.5)$, $(2.5, 5)$, and $(5, 7)$ s, whereas for the 2.5–5 s window, the entire available temporal interval was used. Analogously, in ADHD, the full duration of each EEG segment was employed, and the $(0.5, 4)$ Hz band was additionally included. In all cases, 500 coalition samples were used, with regularization limited to 200 features. Moreover, the baseline was fixed as null in the time–frequency domain, such that $\check{\mathcal{H}}_{\boldsymbol{X}}(\check{\mathbf{0}})$ represents the signal reconstruction when the spectro-temporal content of the cells is suppressed. This choice enables quantifying the contribution of each cell $\mathcal{G}_q$ with respect to a reference state without active content in the corresponding region.

To ensure methodological consistency and avoid information leakage, all methods requiring a reference used only information from the training set. Likewise, the reference-set sizes, number of perturbations, coalition samples, and regularization parameters were fixed by considering a balance between explanation stability and computational cost. Finally, the target-class source was defined according to the subsequent analysis. For perturbation-based fidelity analyses, including Deletion and ROAD, model predictions were used as the target class to evaluate whether the regions identified as relevant supported the decision made by the classifier; under this configuration, KernelSHAP, STF-KernelSHAP, LIME, and Occlusion were computed on probabilities, whereas Integrated Gradients and Grad-CAM++ were computed on logits to reduce saturation. In contrast, for scalp topographic maps, attributions were computed with respect to the ground-truth label, enabling the analysis of the spatial distribution of relevance associated with each real class; in this case, KernelSHAP, STF-KernelSHAP, Occlusion, Integrated Gradients, and Grad-CAM++ were computed on logits, whereas LIME was kept on probabilities, since its local formulation corresponds to a classification problem and the use of logits would shift its interpretation toward local regression.

All experiments were executed in Python v3.12.12 for model training and Python v3.12.13 for post-hoc interpretability analysis. Model training was performed in a cloud-based Kaggle environment under a 64-bit Ubuntu 22.04.5 LTS system, using GPU acceleration when available. The computational setup included an Intel Xeon CPU @ 2.00 GHz with 4 logical cores, 31 GB of RAM, and two NVIDIA Tesla T4 GPUs with 15 GB of VRAM each, together with CUDA v13.0, CUDA compilation tools v12.8, and NVIDIA driver v580.105.08. Subsequently, the post-hoc interpretability analyses were conducted in Google Colaboratory, also under a 64-bit Ubuntu 22.04.5 LTS system and using GPU acceleration when available; in this case, the setup included an Intel Xeon CPU @ 2.00 GHz with 2 logical cores, an

 doi:10.20944/preprints202606.0375.v1

NVIDIA Tesla T4 GPU with 15 GB of VRAM, approximately 12.7 GB of system RAM, CUDA v13.0, CUDA compilation tools v12.8, and NVIDIA driver v580.82.07. The main libraries used throughout the complete workflow were NumPy v2.0.2, SciPy v1.16.3, scikit-learn v1.6.1, TensorFlow/Keras v2.19.0/v3.10.0 for model training and v2.20.0/v3.13.2 for interpretability analysis, KerasNLP v0.21.1 for model training and v0.26.0 for interpretability analysis, SHAP v0.50.0 for model training and v0.51.0 for interpretability analysis, LIME v0.2.0.1, Optuna v4.8.0, and tf-keras-vis v0.8.7. To ensure reproducibility, all source code, scripts, and configuration files will be publicly available at: https://github.com/Daprosero/STF-KernelSHAP.The complete experimental workflow, is summarized in Figure 5.

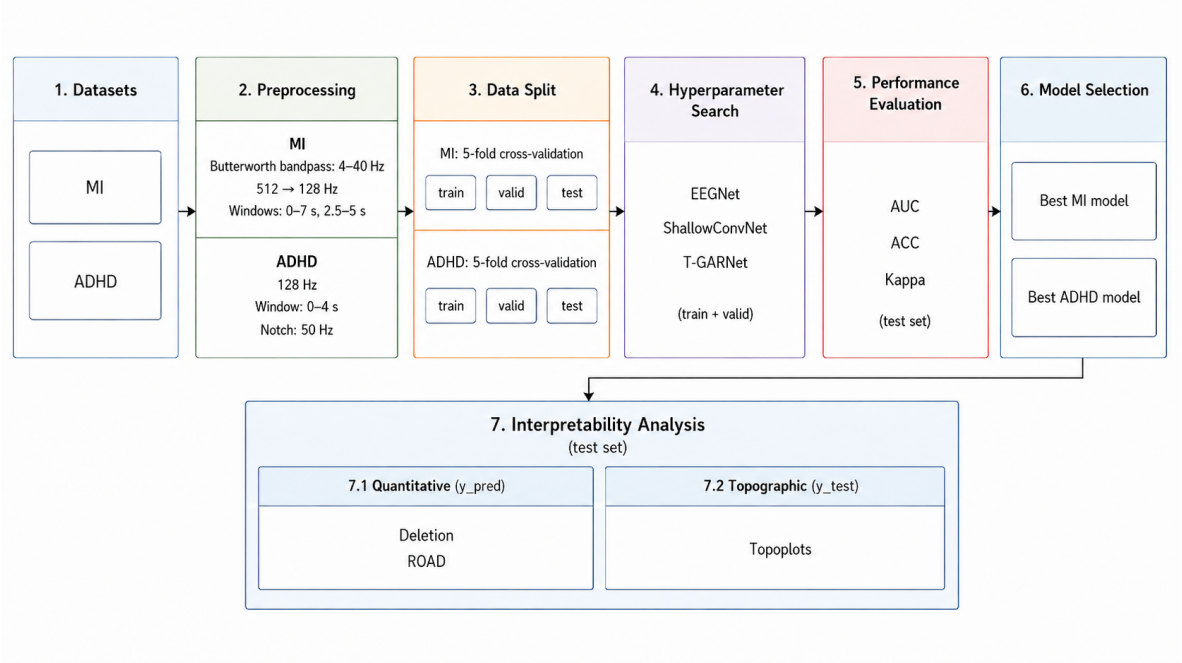

Figure 5. Experimental workflow for model training, selection, performance evaluation, and post-hoc interpretability analysis in the MI and ADHD datasets.

5. Results and Discussion

5.1. Space–Time–Frequency Attribution Analysis in Motor Imagery

The classification results for the 0–7 s window reveal a marked inter-subject variability, as shown in Figure 6. Overall, EEGNet and ShallowConvNet achieve the highest accuracy and AUC values, whereas T-GARNet exhibits a lower performance under this configuration. In particular, ShallowConvNet was selected as the baseline model for the interpretability analysis due to its superior overall performance and relative stability across the evaluated metrics. However, the subject-wise analysis indicates that the classification problem is not determined solely by the model architecture, but also by the individual heterogeneity of EEG responses. In this context, subject 14 belongs to the group of subjects with favorable performance, whereas subject 12 is located in a lower-accuracy region. This contrast defines two complementary analysis scenarios: one in which the classifier learns a sufficiently stable discriminative representation, and another in which the model decision is less reliable.

 doi:10.20944/preprints202606.0375.v1

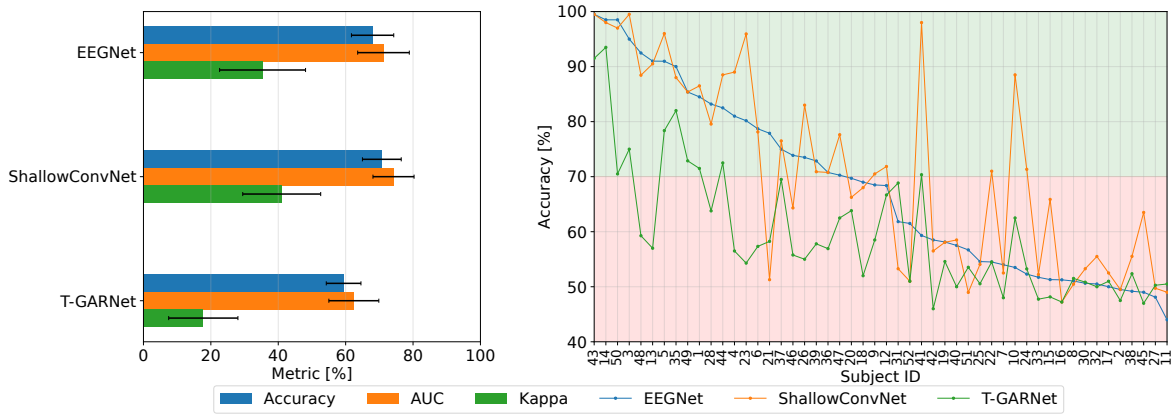

Figure 6. Classification performance obtained on the MI dataset using the 0–7 s temporal window. **Left:** Global summary of the evaluated metrics across the considered deep learning models. **Right:** Subject-wise accuracy distribution, highlighting inter-subject variability in classification performance.

Figure 7 allows us to assess whether STF-KernelSHAP remains comparable to conventional XAI methods when its attributions are projected onto the spatial domain. For subject 14, the evaluated strategies show contributions concentrated over central and centro-lateral regions, which are consistent with the expected sensorimotor activity in hand motor imagery tasks. This behavior is relevant because motor imagery activity is commonly reflected in modulations of mu/alpha and beta rhythms over sensorimotor areas, particularly around central electrodes such as C3, Cz, and C4. In this scenario, STF-KernelSHAP produces an aggregated spatial map that preserves a topographic organization comparable to KernelSHAP, Occlusion, and Integrated Gradients, indicating that the proposed strategy does not lose the ability to provide a standard spatial interpretation. In contrast, for subject 12, the maps are less focalized and exhibit lower anatomical coherence across methods. This loss of sensorimotor reference is consistent with the reduced performance observed for this subject and suggests that, when the classifier decision boundary is weak, the spatial attributions produced by all methods tend to become less informative.

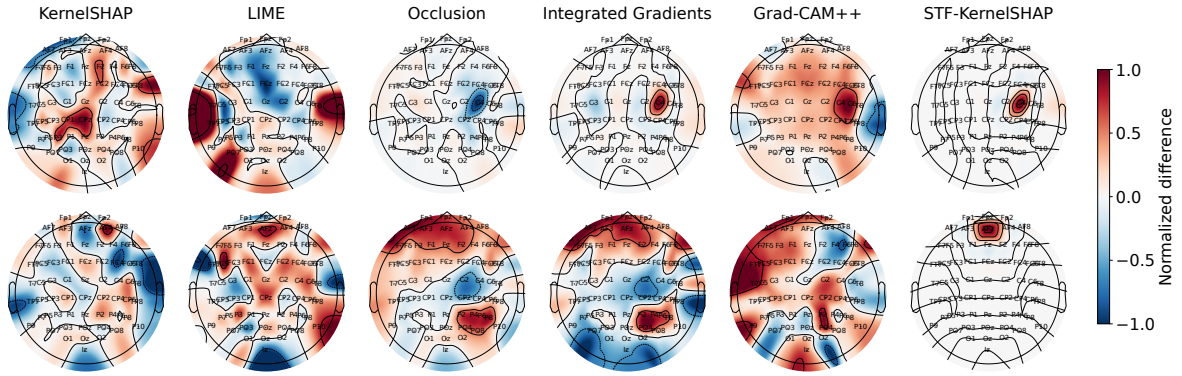

Figure 7. Spatial distribution of the normalized attribution differences obtained on the MI dataset using the 0–7 s temporal window. **Top:** Subject 14 evaluated in fold 4. **Bottom:** Subject 12 evaluated in fold 5.

Figure 8 highlights the distinctive contribution of STF-KernelSHAP. Unlike conventional methods, which provide an aggregated spatial attribution, the proposed strategy preserves the space–time–frequency structure of relevance. For subject 14, the most informative contribution is concentrated within the 2.5–5 s window and the alpha band. This finding is consistent with the experimental protocol described in Figure 1, where the 2.5–5 s interval corresponds to the effective motor imagery period. Moreover, the alpha band matches the mu/alpha range associated with sensorimotor modulation during imagined movement. Therefore, the explanation not only identifies a brain region compatible with the task, but also localizes the contribution within the expected temporal interval and spectral band. This point is central to the proposed approach: STF-KernelSHAP is not limited to competing

 doi:10.20944/preprints202606.0375.v1

with XAI methods through a global topographic map, but also enables verification of whether the model decision relies on physiologically plausible components of the EEG signal.

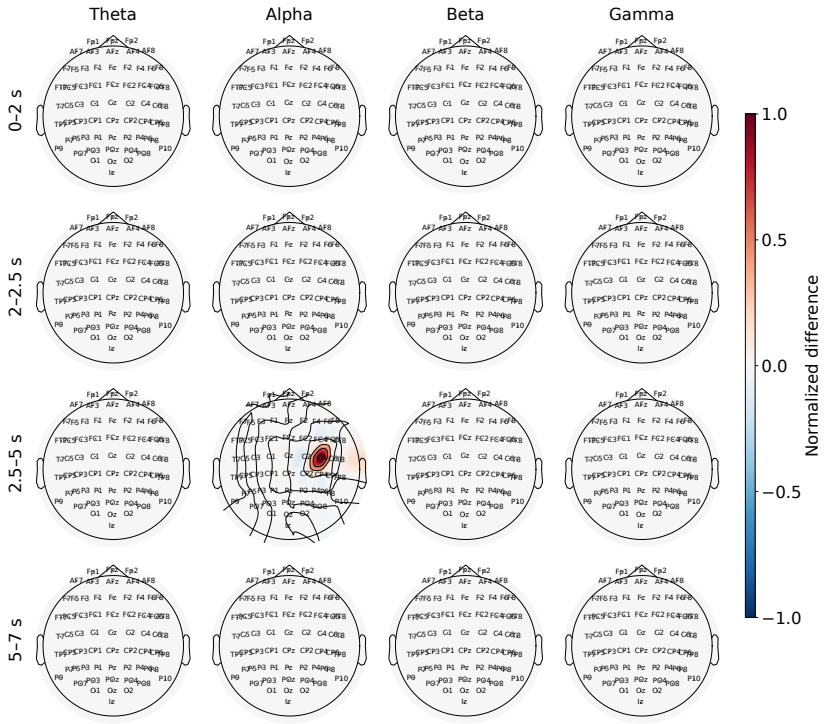

Figure 8. Time–frequency spatial attribution maps obtained with the proposed STF-KernelSHAP strategy on the MI dataset using the 0–7 s temporal window for subject 14 in fold 4. **Rows:** Temporal segments used to localize the contribution of the EEG signal over time. **Columns:** Frequency bands used to characterize the spectral contribution of EEG activity.

Figure 9 presents a more critical case. For subject 12, the space–time–frequency representation does not clearly reproduce a focalized sensorimotor activation pattern, which is consistent with the previously observed low classifier performance. The reduced spatial focalization is therefore coherent with the weaker predictive behavior of the model for this subject. The absence of a well-defined sensorimotor topography indicates that the model did not learn a robust spatial representation for this case. Nevertheless, STF-KernelSHAP still identifies contributions within the 2.5–5 s window and the alpha band, both associated with the effective motor imagery period and the expected sensorimotor modulation. Thus, although the spatial evidence is degraded, the proposed decomposition preserves a temporal–spectral interpretation of the decision process that is not accessible from aggregated XAI maps.

Figure 10 complements the visual analysis through a perturbation-based fidelity assessment. For subject 14, the Deletion MoRF and ROAD curves show that removing relevant regions modifies the classifier response, indicating that the attributions capture components functionally related to the model decision. In this case, STF-KernelSHAP exhibits competitive behavior compared with conventional strategies, confirming that the space–time–frequency decomposition does not compromise functional fidelity. For subject 12, the curves are less stable and less conclusive, which is expected given the lower model performance. In this scenario, perturbing supposedly relevant regions does not produce an ordered degradation pattern, because the initial classifier decision is already less reliable.

 doi:10.20944/preprints202606.0375.v1

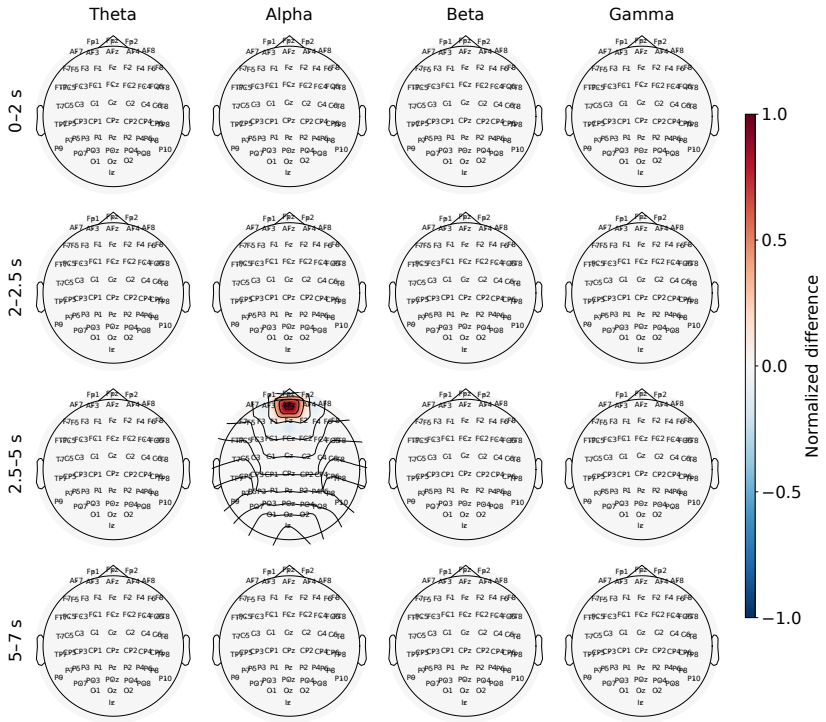

Figure 9. Space–time–frequency attribution maps obtained with the proposed STF-KernelSHAP strategy on the MI dataset using the 0–7 s temporal window for subject 12 in fold 5. **Rows:** Temporal segments used to localize the contribution of the EEG signal over time. **Columns:** Frequency bands used to characterize the spectral contribution of the EEG activity.

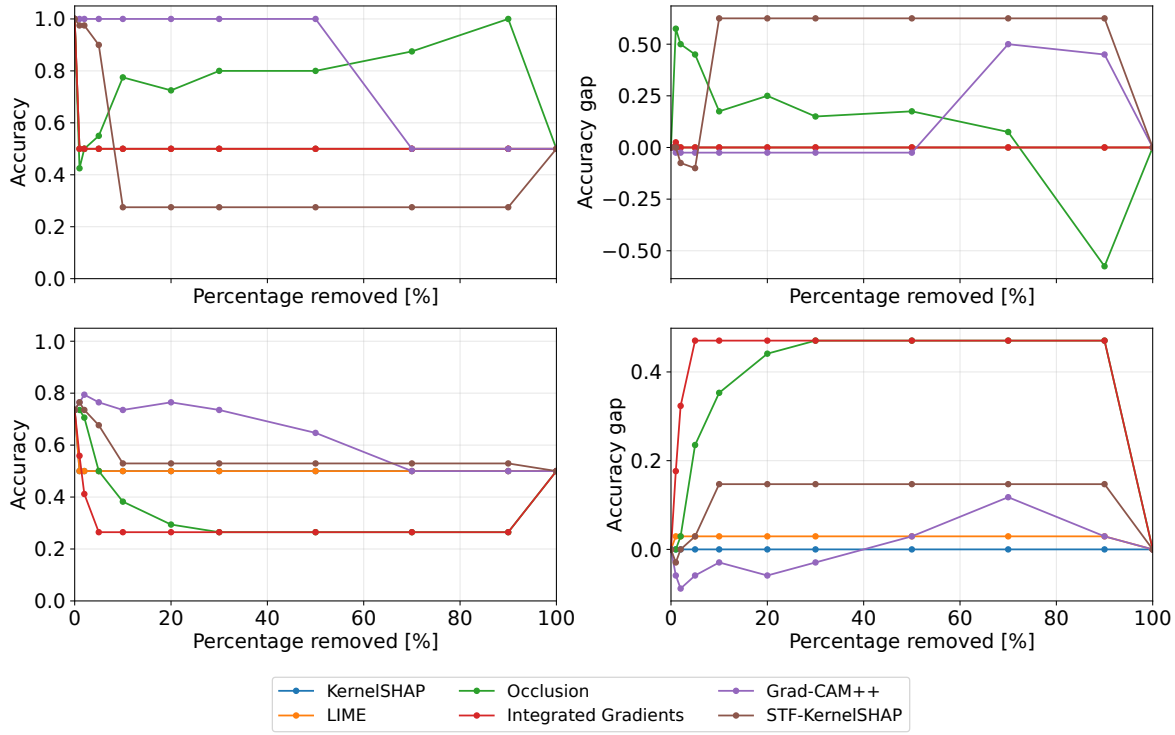

Figure 10. Perturbation-based fidelity analysis obtained on the MI dataset using the 0–7 s temporal window. **Top:** Subject 14 evaluated in fold 4. **Bottom:** Subject 12 evaluated in fold 5. **Left:** Deletion MoRF. **Right:** ROAD.

By restricting the analysis to the 2.5–5 s window, the results in Figure 11 directly assess the temporal interval most closely associated with the execution of the motor imagery paradigm. Under this condition, subject 43 represents the best-performing scenario, whereas subject 12 maintains low

 doi:10.20944/preprints202606.0375.v1

performance, thereby enabling a renewed contrast between a case with reliable decisions and another with lower predictive stability. The relative improvement observed in the best-performing subjects suggests that focusing the analysis on the effective MI window reduces the influence of less informative trial segments, although it does not eliminate the inter-subject variability that characterizes EEG signals.

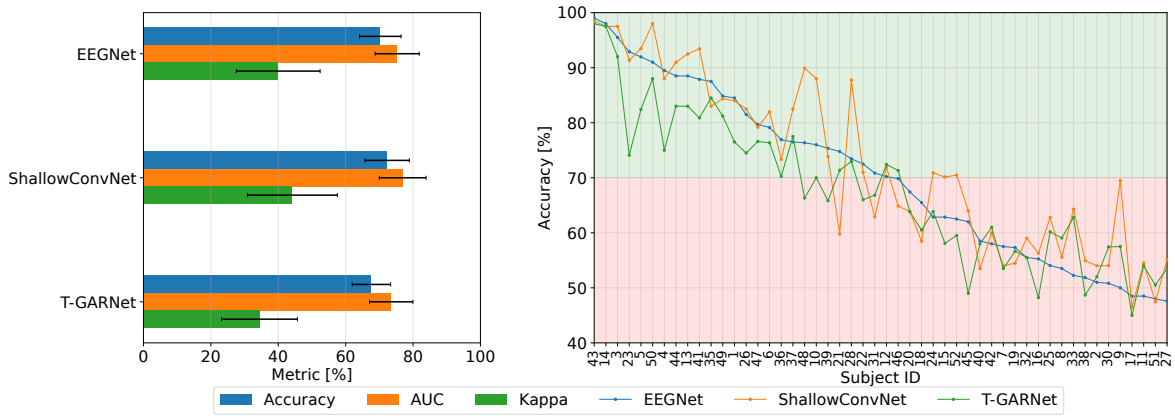

Figure 11. Classification performance obtained on the MI dataset using the 2.5–5 s temporal window. **Left:** Global summary of the evaluated metrics across the considered deep learning models. **Right:** Subject-wise accuracy distribution, highlighting inter-subject variability in classification performance.

Figure 12 compares the spatial distribution of attribution differences obtained by the XAI methods within the 2.5–5 s window. For subject 43, the maps exhibit more defined and localized patterns than those observed for subject 12, in agreement with the higher classification performance. In particular, several methods concentrate relevance over central and centro-lateral regions, which is consistent with the involvement of sensorimotor areas during motor imagery tasks, where mu/alpha and beta rhythms are commonly modulated over the sensorimotor cortex. In this scenario, STF-KernelSHAP preserves a topographic representation comparable to conventional XAI strategies, since it allows spatial contribution regions to be inspected within the same analysis domain. Conversely, for subject 12, the attributions are more scattered and less consistent across methods, which is coherent with a less stable classifier decision.

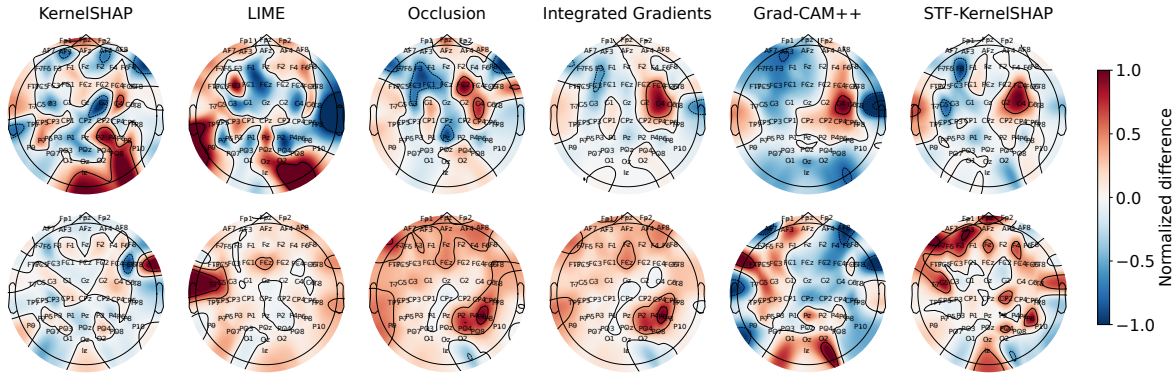

Figure 12. Spatial distribution of the normalized attribution differences obtained on the MI dataset using the 2.5–5 s temporal window. **Top:** Subject 43 evaluated in fold 1. **Bottom:** Subject 12 evaluated in fold 3.

Figure 13 confirms the central contribution of the proposed strategy. For subject 43, STF-KernelSHAP concentrates relevant contributions in the alpha and beta bands, consistently with the expected sensorimotor modulation during motor imagery tasks. This interpretation is more informative than the aggregated spatial map, as it verifies that the model decision is not only localized in plausible regions but is also supported by spectral components consistent with the task. For subject 12, the spatial organization is less clear, in agreement with the low classification performance; nevertheless,

 doi:10.20944/preprints202606.0375.v1

the attribution still enables band-wise inspection of the contribution, preserving a spectral reading that conventional XAI methods do not directly provide.

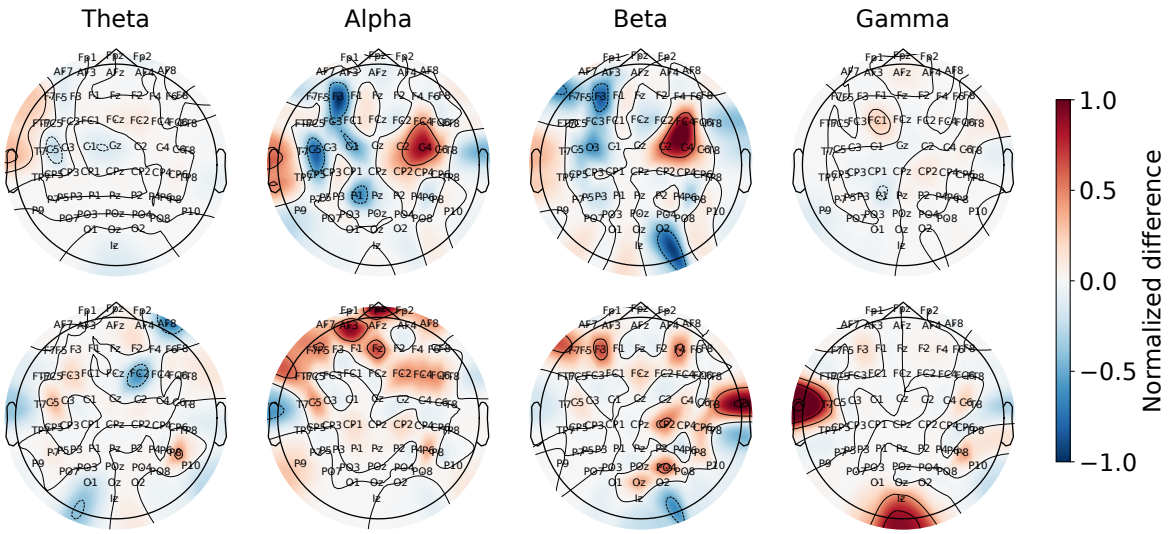

Figure 13. Frequency-band spatial attribution maps obtained with the proposed STF-KernelSHAP strategy on the MI dataset using the 2.5–5 s temporal window. **Top:** Subject 43 evaluated in fold 1. **Bottom:** Subject 12 evaluated in fold 3. **Columns:** Frequency bands used to characterize the spectral contribution of the EEG activity.

The fidelity assessment in Figure 14 reinforces this interpretation. For subject 43, the Deletion MoRF and ROAD curves show a more sensitive response to the perturbation of relevant regions, supporting the functional relationship between the attributions and the classifier decision. For subject 12, the curves are less stable and less conclusive, which is consistent with the lower predictive quality of the model. Thus, the comparison between both subjects indicates that the fidelity of the explanations depends on the classifier performance, whereas the STF decomposition preserves a temporal–spectral interpretation even when the spatial evidence becomes weaker.

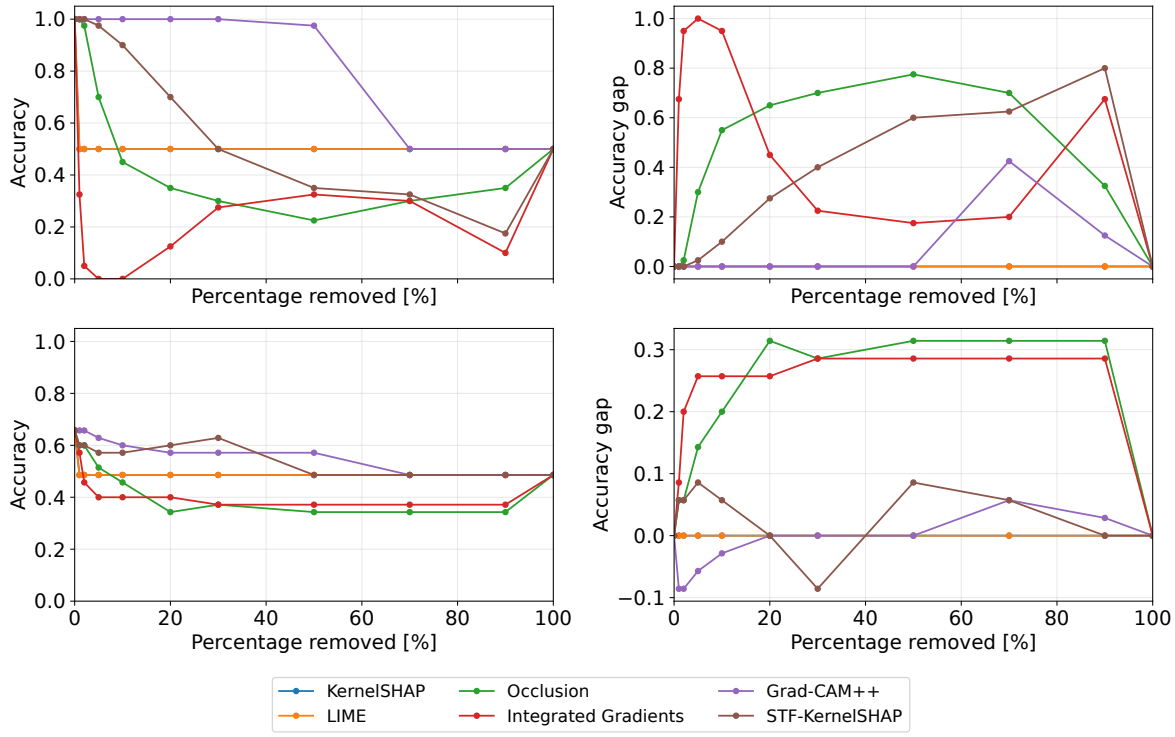

Figure 14. Perturbation-based fidelity analysis obtained on the MI dataset using the 2.5–5 s temporal window. **Top:** Subject 43 evaluated in fold 1. **Bottom:** Subject 12 evaluated in fold 3. **Left:** Deletion MoRF. **Right:** ROAD.

 doi:10.20944/preprints202606.0375.v1

Overall, the comparison between the 0–7 s and 2.5–5 s windows shows that STF-KernelSHAP can be evaluated under the same spatial conditions as conventional XAI methods, while also verifying whether the explanation is coherent with the experimental structure of the paradigm. In the full window, the strategy identifies the contribution within the motor imagery interval; when the analysis is restricted to 2.5–5 s, the explanation concentrates on the alpha and beta bands, which are the expected spectral components for MI. This behavior is consistent with the quantitative fidelity results reported in Table 1, where STF-KernelSHAP remains competitive with conventional post-hoc strategies and achieves the highest ROAD-AUC in the 0–7 s window, suggesting that the most relevant components identified by the method preserve the model decision when physiologically meaningful information is retained. Although Integrated Gradients and Occlusion obtain lower Deletion-AUC values in some settings, their explanations remain defined over the original input space and do not explicitly disentangle the temporal and spectral structure of the MI paradigm. Therefore, the main advantage of STF-KernelSHAP is not merely its ability to produce maps comparable to KernelSHAP, LIME, Occlusion, Integrated Gradients, or Grad-CAM++, but rather its capacity to balance quantitative fidelity with structured neurophysiological interpretability across three complementary levels: spatial localization, temporal window, and spectral band.

Table 1. Quantitative XAI fidelity results for the MI dataset. Methods are ordered according to their global mean rank across Deletion-AUC and ROAD-AUC. Lower Deletion-AUC and higher ROAD-AUC indicate better fidelity.

Method	**Deletion-AUC** $\downarrow$		**ROAD-AUC** $\uparrow$	
	0–7	**2.5–5**	**0–7**	**2.5–5**
Integrated Gradients	$\mathbf{0.291 \pm 0.400}$	$\mathbf{0.207 \pm 0.231}$	0.431 ± 0.466	0.597 ± 0.285
Occlusion	0.470 ± 0.389	0.310 ± 0.212	0.443 ± 0.427	$\mathbf{0.639 \pm 0.202}$
STF-KernelSHAP	0.420 ± 0.378	0.624 ± 0.339	$\mathbf{0.460 \pm 0.369}$	0.262 ± 0.273
KernelSHAP	0.543 ± 0.499	0.602 ± 0.490	0.000 ± 0.000	-0.000 ± 0.002
LIME	0.548 ± 0.495	0.610 ± 0.478	0.009 ± 0.059	0.007 ± 0.059
Grad-CAM++	0.752 ± 0.270	0.789 ± 0.265	0.090 ± 0.166	0.063 ± 0.108

5.2. Space–Frequency Attribution Analysis in ADHD

The classification results obtained for the ADHD database are presented in Figure 15. Unlike the motor imagery analysis, this database is not associated with an event-segmented experimental protocol, but rather with 4 s EEG windows extracted under a more general recording condition. Therefore, the interpretation does not aim to localize a specific phase of a paradigm, but instead to identify spatial and spectral patterns that are relevant for class discrimination. In this scenario, T-GARNet achieves competitive performance compared with the evaluated architectures, which justifies its use as the base model to analyze whether STF-KernelSHAP can be coupled with an architecture different from those employed in MI.

Figure 16 shows the spatial distribution of attribution differences obtained by the XAI methods in ADHD. In contrast to MI, a focalized activation over a specific sensorimotor region is not expected, since the task is not linked to a motor event or to a cognitively bounded window defined by the protocol. Instead, the maps exhibit distributed contributions over frontal, central, and posterior regions, which is consistent with the more global nature of EEG alterations reported in ADHD. In this context, STF-KernelSHAP preserves topographic comparability with KernelSHAP, LIME, Occlusion, Integrated Gradients, and Grad-CAM++, by producing an aggregated spatial map within the same domain of analysis.

 doi:10.20944/preprints202606.0375.v1

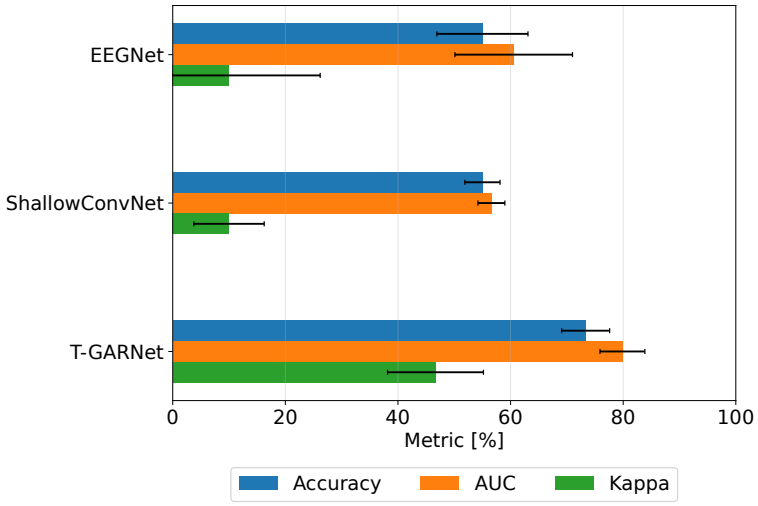

Figure 15. Classification performance obtained on the ADHD dataset.

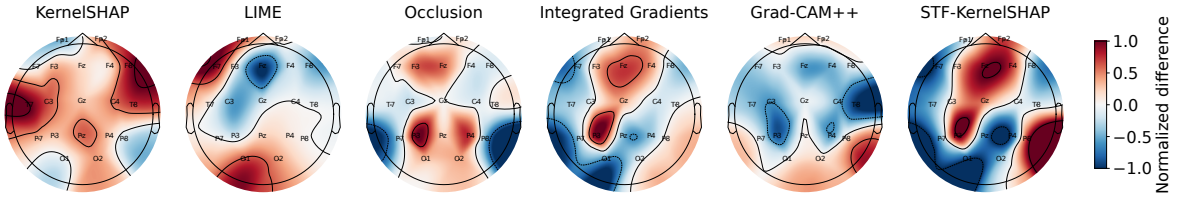

Figure 16. Spatial distribution of the normalized attribution differences obtained on the ADHD dataset.

Figure 17 shows that, because a single 4 s EEG window is available, the STF-KernelSHAP interpretation is mainly concentrated in the spectral domain. This reading is consistent with Figure 4, where the power spectral density is concentrated at low frequencies and progressively decreases toward higher frequencies. Accordingly, the attributions exhibit more defined patterns in delta, theta, alpha, and beta, whereas in gamma the contribution is attenuated and the spatial structure becomes less evident. This correspondence indicates that STF-KernelSHAP captures dominant spectral components of the signal, rather than producing arbitrarily distributed relevance. Moreover, the involvement of low and intermediate frequency bands is compatible with ADHD studies reporting EEG alterations in theta, alpha, and beta, although with sufficient heterogeneity to avoid interpreting a single band or spectral ratio as a universal diagnostic marker.

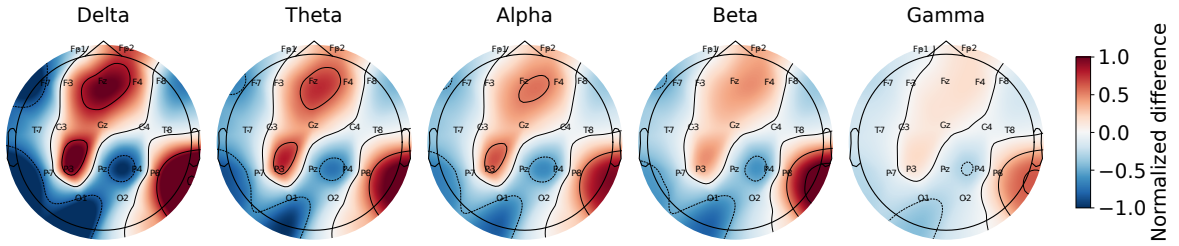

Figure 17. Frequency-band spatial attribution maps obtained with the proposed STF-KernelSHAP strategy on the ADHD dataset. **Columns:** Frequency bands used to characterize the spectral contribution of EEG activity.

Figure 18 complements the analysis through perturbation-based fidelity metrics. In Deletion MoRF, Integrated Gradients exhibits a marked performance drop when the most relevant regions are removed, indicating a strong functional relationship between its attributions and the model decision. STF-KernelSHAP also shows a progressive degradation in accuracy, although less abruptly, which is consistent with an explanation constructed from more structured space–frequency blocks. In ROAD, Integrated Gradients and Occlusion produce larger accuracy gaps, whereas STF-KernelSHAP maintains an intermediate response. This suggests that the proposed strategy preserves functional fidelity,

 doi:10.20944/preprints202606.0375.v1

although its main objective is not to maximize the pointwise performance drop under perturbation, but to provide an explanation organized by regions and frequency bands.

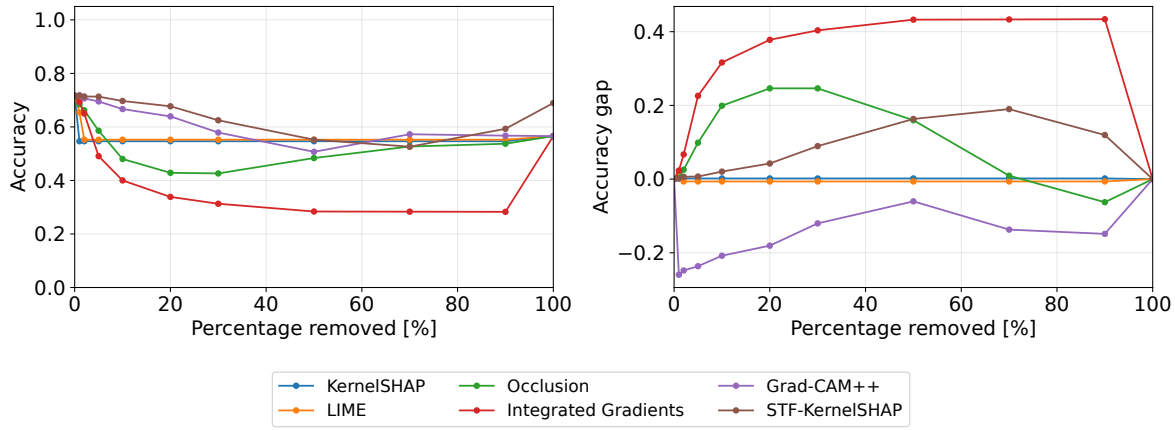

Figure 18. Perturbation-based fidelity analysis obtained on the ADHD dataset using fold 5. **Left:** Deletion MoRF. **Right:** ROAD.

Overall, the ADHD results extend the scope of the analysis conducted in MI. Whereas in motor imagery STF-KernelSHAP enabled verification of the correspondence between the attribution, the effective window of the paradigm, and the expected sensorimotor bands, in ADHD the strategy operates under a different condition: a single 4 s temporal window and a task without an explicit experimental event. Even so, the method preserves the ability to generate spatial maps comparable with conventional XAI approaches and further decomposes the explanation into EEG bands that are relevant for neurophysiological analysis. This behavior is consistent with the quantitative fidelity results reported in Table 2, where Integrated Gradients achieves the strongest overall fidelity, while STF-KernelSHAP remains competitive with perturbation- and Shapley-based methods, outperforming KernelSHAP, LIME, and Grad-CAM++ in both Deletion-AUC and ROAD-AUC. Therefore, the ADHD results indicate that STF-KernelSHAP should not be interpreted solely as the best-performing fidelity method, but as a model-agnostic and structured XAI strategy that can be coupled with different deep architectures and EEG scenarios, balancing quantitative fidelity with spatial and spectral interpretability.

Table 2. Quantitative XAI fidelity results for the TDAH dataset. Methods are ordered according to their global mean rank across Deletion-AUC and ROAD-AUC. Lower Deletion-AUC and higher ROAD-AUC indicate better fidelity.

Method	Deletion-AUC ↓	ROAD-AUC ↑
Integrated Gradients	$\mathbf{0.107 \pm 0.084}$	$\mathbf{0.881 \pm 0.083}$
Occlusion	0.496 ± 0.285	0.326 ± 0.361
STF-KernelSHAP	0.646 ± 0.317	0.331 ± 0.295
KernelSHAP	0.690 ± 0.337	-0.025 ± 0.277
LIME	0.721 ± 0.310	-0.039 ± 0.293
Grad-CAM++	0.729 ± 0.224	-0.286 ± 0.428

5.3. Limitations

Although STF-KernelSHAP provides a structured and model-agnostic strategy for EEG interpretability, some limitations should be acknowledged. The reliability of the explanations depends on the predictive quality of the underlying classifier. When the model exhibits poor or unstable performance, the resulting attribution maps may also become inconsistent, since the learned decision function may not encode robust neurophysiological patterns. This aspect is particularly relevant in

 doi:10.20944/preprints202606.0375.v1

EEG analysis, where inter-subject variability, non-stationarity, and low signal-to-noise ratios affect both classification and interpretation.

The proposed framework also depends on a predefined partition of the signal into temporal windows and frequency bands. While this design enables physiologically coherent perturbations over complete channel–time–frequency cells, it may overlook subject-specific rhythms, transient spectral events, or non-stationary responses that do not match the selected segmentation. Therefore, the resolution of the explanation is partly constrained by the prior definition of the time–frequency grid.

Another relevant aspect concerns the use of a reference baseline in the time–frequency domain to replace absent components during coalition reconstruction. Although this strategy is more structured than direct pointwise masking, the estimated marginal contributions may still depend on the selected baseline. In particular, zero-valued spectral references may not always represent physiologically plausible counterfactual states.

Finally, although the proposed channel–time–frequency grouping reduces the limitations of conventional KernelSHAP, the coalition sampling process still follows the standard approximation based on independently sampled binary masks. This formulation does not explicitly model statistical dependencies among electrodes, temporal windows, and frequency bands. Consequently, some sampled coalitions may remain only partially consistent with the dependency structure of EEG signals.

6. Conclusions

This work introduced STF-KernelSHAP, an architecture-independent explainability framework for deep EEG classifiers. The proposed method represents each EEG trial through structured channel–time–frequency cells and estimates class-conditional relevance using a Shapley value-based perturbation strategy. Unlike conventional post-hoc XAI methods, which commonly rely on flattened inputs, sample-wise perturbations, or architecture-dependent activations, STF-KernelSHAP preserves the multidomain organization of EEG signals while remaining applicable to black-box classifiers.

The experimental results showed that STF-KernelSHAP provides competitive explanatory fidelity with respect to representative XAI methods, including KernelSHAP, LIME, Occlusion, Integrated Gradients, and Grad-CAM++. In the motor imagery scenario, the proposed framework enabled the identification of spatial, temporal, and spectral patterns associated with discriminative sensorimotor activity. In the ADHD detection scenario, it extended the analysis to a different EEG montage, temporal configuration, and neurophysiological condition, supporting its applicability across heterogeneous EEG classification settings.

A central contribution of the proposed framework is that it extends standard EEG topographic interpretation toward structured multidomain analysis. Thus, relevance can be examined not only at the electrode level, but also across task-related temporal windows and frequency bands. This property is particularly relevant in EEG decoding, where discriminative information commonly emerges from the interaction between spatial distributions, oscillatory activity, and temporal dynamics.

The analysis also showed that post-hoc interpretability should be examined together with predictive performance. In high-performing cases, attribution maps were more stable and more consistent with expected neurophysiological patterns, whereas low-performing cases produced less coherent explanations across XAI methods. This finding suggests that unreliable attribution patterns may reflect weak predictive representations rather than limitations of the explanation method alone.

Overall, STF-KernelSHAP offers a unified and architecture-independent alternative for interpreting EEG classifiers. By combining structured time–frequency perturbation, signal-domain reconstruction, and Shapley-based attribution, the framework provides physiologically informed explanations without requiring gradients, internal activations, or model-specific modifications.

Future work will focus on adaptive time–frequency partitions, physiologically informed baselines, and more efficient coalition sampling strategies [68]. A promising direction consists of developing a variational extension of STF-KernelSHAP, in which the coalition distribution is learned from data rather than imposed through independent binary sampling [69]. This extension could better capture

 doi:10.20944/preprints202606.0375.v1

dependencies among electrodes, temporal windows, and frequency bands. In addition, relevance-guided interpolation could be explored to restrict perturbation trajectories to the most informative channel–time–frequency components, with the aim of improving attribution stability [70]. Further validation under subject-independent protocols, cross-dataset transfer, larger EEG cohorts, and expert-based neurophysiological assessment will also be necessary [71].

Author Contributions: Conceptualization, D.A.P.-R.,A.C.L.-B.,A.M.Á.-M.,D.A.C.-P and G.C.-D.; data curation, D.A.P.-R.; methodology, D.A.P.-R.,A.C.L.-B., A.M.Á.-M.; project administration, A.M.Á.-M.,D.A.C.-P; supervision, A.M.Á.-M.,D.A.C.-P and G.C.-D.; resources, D.A.P.-R.,A.C.L.-B and A.M.Á.-M. All authors have read and agreed to the published version of this manuscript.

Funding: Authors gratefully acknowledge support from the program: 'Alianza científica con enfoque comunitario para mitigar brechas de atención y manejo de trastornos mentales relacionados con impulsividad en Colombia (ACEMATE)-91908',This research was supported by the project: 'Sistema multimodal apoyado en juegos serios orientado a la evaluación e intervención neurocognitiva personalizada en trastornos de impulsividad asociados a TDAH como soporte a la intervención presencial y remota en entornos clínicos, educativos y comunitarios-790-2023,' funded by the Colombian Ministry of Science, Technology and Innovation (Minciencias). Also, A.M. Alvarez gives thanks to the project: 'NatureTunes: Inteligencia artificial para el monitoreo de paisajes sonoros y visuales como fomento al aviturismo en el departamento de Caldas', Hermes-63421, funded by Universidad Nacional de Colombia.

Data Availability Statement: Data available upon reasonable request via email.

Acknowledgments: The authors gratefully acknowledge the Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED), through Red 225RT0169, for its academic and collaborative support.

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

References

1. Rohan, N.R.; Vigneswaran, C.; Ghosh, S.; Rajendran, K.; Gaurav, A.; Chakravarthy, V.S. Deep oscillatory neural network. *Scientific Reports* **2025**, *15*, 40968. https://doi.org/10.1038/s41598-025-24837-4.
2. He, X.; et al. Electroencephalographic Motor Imagery Brain Connectivity Analysis for BCI: A Review. *Neural Computation* **2016**, *28*, 999–1021. "BCI based on EEG sensorimotor rhythms is known as motor imagery (MI) ... a type of endogenous EEG-based BCI".
3. Hurjui, I.A.; Hurjui, R.M.; Hurjui, L.L.; Serban, I.L.; Dobrin, I.; Apostu, M.; Dobrin, R.P. Biomarkers and Neuropsychological Tools in Attention-Deficit/Hyperactivity Disorder: From Subjectivity to Precision Diagnosis. *Medicina* **2025**, *61*. https://doi.org/10.3390/medicina61071211.
4. Ramadan, R.A.; Altamimi, A.B. Unraveling the potential of brain-computer interface technology in medical diagnostics and rehabilitation: A comprehensive literature review. *Health and Technology* **2024**, *14*, 263–276.
5. Wang, X.; et al. An in-depth survey on Deep Learning-based Motor Imagery EEG classification. *Neurocomputing* **2024**. "... the rapid advancement of deep learning ... successful application in motor imagery EEG classification".
6. Huang, G.; Li, Y.; Jameel, S.; Long, Y.; Papanastasiou, G. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *Computational and Structural Biotechnology Journal* **2024**, *24*, 362–373. https://doi.org/https://doi.org/10.1016/j.csbj.2024.05.004.
7. Mayor Torres, J.M.; Medina-DeVilliers, S.; Clarkson, T.; Lerner, M.D.; Riccardi, G. Evaluation of interpretability for deep learning algorithms in EEG emotion recognition: A case study in autism. *Artificial Intelligence in Medicine* **2023**, *143*, 102545. https://doi.org/https://doi.org/10.1016/j.artmed.2023.102545.
8. Li, X.; Xiong, H.; Li, X.; Wu, X.; Zhang, X.; Liu, J.; Bian, J.; Dou, D. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems* **2022**, *64*, 3197–3234.
9. Angkan, P.; Jalali, A.; Hungler, P.; Etemad, A. Multi-Domain EEG Representation Learning with Orthogonal Mapping and Attention-Based Fusion for Cognitive Load Classification, 2025, [arXiv:cs.HC/2511.12394]. https://doi.org/10.48550/arXiv.2511.12394.
10. Liu, Z.; Fan, K.; Gu, Q.; Ruan, Y. Channel-Dependent Multilayer EEG Time-Frequency Representations Combined with Transfer Learning-Based Deep CNN Framework for Few-Channel MI EEG Classification. *Bioengineering* **2025**, *12*. https://doi.org/10.3390/bioengineering12060645.

 doi:10.20944/preprints202606.0375.v1

11. Shawly, T.; Alsheikhy, A.A. Eeg-based detection of epileptic seizures in patients with disabilities using a novel attention-driven deep learning framework with SHAP interpretability. *Egyptian Informatics Journal* **2025**, *31*, 100734. https://doi.org/https://doi.org/10.1016/j.eij.2025.100734.
12. given i=S, given=Sophia, f.; given i=M, given=Merle, f.; given i=T, given=Thomas, f.; given i=M, given=Martin, f.; given i=B, given=Benjamin, f. SHAP value-based ERP analysis (SHERPA): Increasing the sensitivity of EEG signals with explainable AI methods. *56*, 6067–6081. https://doi.org/10.3758/s13428-023-02335-7.
13. Niu, Y.; Chen, X.; Fan, J.; Liu, C.; Fang, M.; Liu, Z.; Meng, X.; Liu, Y.; Lu, L.; Fan, H. Explainable machine learning model based on EEG, ECG, and clinical features for predicting neurological outcomes in cardiac arrest patient. *Scientific Reports* **2025**, *15*, 11498.
14. Zhou, X.; Liu, C.; Wang, Z.; Zhai, L.; Jia, Z.; Guan, C.; Liu, Y. Interpretable and robust ai in eeg systems: A survey. *arXiv preprint arXiv:2304.10755* **2023**.
15. Raab, D.; Theissler, A.; Spiliopoulou, M. XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series. *Neural Computing and Applications* **2023**, *35*, 10051–10068.
16. given i=V, given=Vahidin, f.; given i=A, given=Amar, f.; given i=S, given=Senka, f.; given i=V, given=Vahidin, f.; given i=A, given=Amar, f.; given i=S, given=Senka, f. *Superpixel Correlation for Explainable Image Classification*; pp. 27–44. https://doi.org/10.1007/978-3-032-08330-2_2.
17. Gallego-Molina, N.J.; Ortiz, A.; Arco, J.E.; Martinez-Murcia, F.J.; Woo, W.L. Unraveling brain synchronisation dynamics by explainable neural networks using EEG signals: Application to dyslexia diagnosis. *Interdisciplinary Sciences: Computational Life Sciences* **2024**, *16*, 1005–1018.
18. Presacan, O.; Ojha, J.; Yazidi, A.; Monteiro, E.; Lind, P.G. A Comprehensive Review of Explainable AI in Deep Learning Algorithms for EEG Analysis. *ACM Transactions on Computing for Healthcare* **2025**.
19. Ma, W.; Zheng, Y.; Li, T.; Li, Z.; Li, Y.; Wang, L. A comprehensive review of deep learning in EEG-based emotion recognition: classifications, trends, and practical implications. *PeerJ Computer Science* **2024**, *10*, e2065.
20. Zhang, S.; Zhu, Z.; Zhang, B.; Feng, B.; Yu, T.; Li, Z.; Zhang, Z.; Huang, G.; Liang, Z. Overall optimization of CSP based on ensemble learning for motor imagery EEG decoding. *Biomedical signal processing and control* **2022**, *77*, 103825.
21. Elashmawi, W.H.; Ayman, A.; Antoun, M.; Mohamed, H.; Mohamed, S.E.; Amr, H.; Talaat, Y.; Ali, A. A Comprehensive Review on Brain–Computer Interface (BCI)-Based Machine and Deep Learning Algorithms for Stroke Rehabilitation. *Applied Sciences* **2024**, *14*. https://doi.org/10.3390/app14146347.
22. Miao, Y.; Jin, J.; Daly, I.; Zuo, C.; Wang, X.; Cichocki, A.; Jung, T.P. Learning Common Time-Frequency-Spatial Patterns for Motor Imagery Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2021**, *PP*, 1–1. https://doi.org/10.1109/TNSRE.2021.3071140.
23. Saha, P.K.; Rahman, M.A.; Alam, M.K.; Ferdowsi, A.; Mollah, M.N. Common spatial pattern in frequency domain for feature extraction and classification of multichannel EEG signals. *SN Computer Science* **2021**, *2*, 149.
24. Liu, K.; Yang, M.; Yu, Z.; Wang, G.; Wu, W. FBMSNet: A filter-bank multi-scale convolutional neural network for EEG-based motor imagery decoding. *IEEE Transactions on Biomedical Engineering* **2022**, *70*, 436–445.
25. Hong, X.; Du, C.; He, H. Adaptive Domain Alignment Neural Networks for Cross-Domain EEG Emotion Recognition. *IEEE Transactions on Affective Computing* **2024**.
26. 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. *Journal of Neural Engineering* **2018**, *15*, 056013. https://doi.org/10.1088/1741-2552/aace8c.
27. Kim, S.J.; Lee, D.H.; Lee, S.W. Rethinking CNN Architecture for Enhancing Decoding Performance of Motor Imagery-based EEG Signals. *IEEE Access* **2022**.
28. Tobón-Henao, M.; Álvarez Meza, A.M.; Castellanos-Dominguez, C.G. Kernel-Based Regularized EEGNet Using Centered Alignment and Gaussian Connectivity for Motor Imagery Discrimination. *Computers* **2023**, *12*. https://doi.org/10.3390/computers12070145.
29. Luo, J.; Wang, Y.; Xia, S.; Lu, N.; Ren, X.; Shi, Z.; Hei, X. A shallow mirror transformer for subject-independent motor imagery BCI. *Computers in Biology and Medicine* **2023**, *164*, 107254.
30. Xiao, T.; Wang, Z.; Zhang, Y.; Wang, S.; Feng, H.; Zhao, Y.; et al. Self-supervised learning with attention mechanism for EEG-based seizure detection. *Biomedical Signal Processing and Control* **2024**, *87*, 105464.
31. Xie, J.; Zhang, J.; Sun, J.; Ma, Z.; Qin, L.; Li, G.; Zhou, H.; Zhan, Y. A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2022**, *30*, 2126–2136.

 doi:10.20944/preprints202606.0375.v1

32. Liao, L.; Lu, J.; Wang, L.; Zhang, Y.; Gao, D.; Wang, M. CT-Net: an interpretable CNN-Transformer fusion network for fNIRS classification. *Medical & Biological Engineering & Computing* **2024**, *62*, 3233–3247. https://doi.org/10.1007/s11517-024-03138-4.
33. Mirzaei, S.; Ghasemi, P. EEG motor imagery classification using dynamic connectivity patterns and convolutional autoencoder. *Biomedical Signal Processing and Control* **2021**, *68*, 102584.
34. Khare, S.K.; Acharya, U.R. An explainable and interpretable model for attention deficit hyperactivity disorder in children using EEG signals. *Computers in biology and medicine* **2023**, *155*, 106676.
35. Rahman, A.U.; Tubaishat, A.; Al-Obeidat, F.; Halim, Z.; Tahir, M.; Qayum, F. Extended ICA and M-CSP with BiLSTM towards improved classification of EEG signals. *Soft Computing* **2022**, *26*, 10687–10698.
36. Bang, J.S.; Lee, S.W. Interpretable convolutional neural networks for subject-independent motor imagery classification. In Proceedings of the 2022 10th International Winter Conference on Brain-Computer Interface (BCI). IEEE, 2022, pp. 1–5.
37. Schwalbe, G.; Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* **2024**, *38*, 3043–3101.
38. Zhang, Y.; Tino, P.; Leonardis, A.; Tang, K. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* **2021**, *5*, 726–742. https://doi.org/10.1109/tetci.2021.3100641.
39. Koh, P.W.; Liang, P. Understanding Black-box Predictions via Influence Functions, 2020, [arXiv:stat.ML/1703.04730].
40. Averkin, A.; Yarushev, S. Review of research in the field of developing methods to extract rules from artificial neural networks. *Journal of Computer and Systems Sciences International* **2021**, *60*, 966–980.
41. Olah, C.; Mordvintsev, A.; Schubert, L. Feature Visualization. *Distill* **2017**, *2*. https://doi.org/10.23915/distill.00007.
42. Sujatha Ravindran, A.; Contreras-Vidal, J. An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth. *Scientific Reports* **2023**, *13*, 17709.
43. Nielsen, I.E.; Dera, D.; Rasool, G.; Ramachandran, R.P.; Bouaynaya, N.C. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine* **2022**, *39*, 73–84.
44. Jiang, P.T.; Zhang, C.B.; Hou, Q.; Cheng, M.M.; Wei, Y. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing* **2021**, *30*, 5875–5888.
45. Zafar, M.R.; Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction* **2021**, *3*, 525–541.
46. Chen, H.; Covert, I.C.; Lundberg, S.M.; Lee, S.I. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence* **2023**, *5*, 590–601.
47. Sharma, N.; Bollu, T.R., Explainable AI Methods for Interpreting Emotions in Brain–Computer Interface EEG Data. In *Discovering the Frontiers of Human-Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control*; Vinjamuri, R., Ed.; Springer Nature Switzerland: Cham, 2024; pp. 419–436. https://doi.org/10.1007/978-3-031-66656-8_18.
48. given i=V, given=Viswan, f.; given i=N, given=Noushath, f.; given i=M, given=Mufti, f. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. *11*, 10. https://doi.org/10.1186/s40708-024-00222-1.
49. Chen, H.; Lundberg, S.M.; Lee, S.I. Explaining a series of models by propagating Shapley values. *Nature communications* **2022**, *13*, 4512.
50. Subudhi, S.; Patro, R.N.; Biswal, P.K.; Dell'Acqua, F. A survey on superpixel segmentation as a preprocessing step in hyperspectral image analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2021**, *14*, 5015–5035.
51. Cao, N.; Wen, X.; Hao, Y.; Cao, R.; Gao, C.; Cao, R. A Lightweight End-to-End Three-domain Feature Fusion Network for Motor Imagery Decoding. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2024, pp. 1830–1837.
52. Li, H.; Chen, Y.; Wang, Y.; Ni, W.; Zhang, H. Foundation models for cross-domain eeg analysis application: A survey. *arXiv preprint arXiv:2508.15716* **2025**.
53. Cui, J.; Yuan, L.; Wang, Z.; Li, R.; Jiang, T. Towards best practice of interpreting deep learning models for EEG-based brain computer interfaces. *Frontiers in Computational Neuroscience* **2023**, *17*. https://doi.org/10.3389/fncom.2023.1232925.
54. Abibullaev, B.; Keutayeva, A.; Zollanvari, A. Deep learning in EEG-based BCIs: a comprehensive review of transformer models, advantages, challenges, and applications. *IEEE Access* **2023**.

 doi:10.20944/preprints202606.0375.v1

55. Cho, H.; Ahn, M.; Ahn, S.; Kwon, M.; Jun, S.C. EEG datasets for motor imagery brain–computer interface. *GigaScience* **2017**, *6*, gix034, [https://academic.oup.com/gigascience/article-pdf/6/7/gix034/25515099/gix034.pdf]. https://doi.org/10.1093/gigascience/gix034.
56. Nasrabadi, A.M.; Allahverdy, A.; Samavati, M.; Mohammadi, M.R. EEG Data for ADHD/Control Children, 2020. https://doi.org/10.21227/rzfh-zn36.
57. Cremades, A.; Hoyas, S.; Vinuesa, R. Additive-feature-attribution methods: A review on explainable artificial intelligence for fluid dynamics and heat transfer. *International Journal of Heat and Fluid Flow* **2025**, *112*, 109662.
58. Li, M.; Sun, H.; Huang, Y.; Chen, H. Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems* **2024**, *4*, 2.
59. Rozemberczki, B.; Watson, L.; Bayer, P.; Yang, H.T.; Kiss, O.; Nilsson, S.; Sarkar, R. The shapley value in machine learning. In Proceedings of the The 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5572–5579.
60. Olsen, L.H.; Glad, I.K.; Jullum, M.; Aas, K. Using Shapley values and variational autoencoders to explain predictive models with dependent mixed features. *Journal of machine learning research* **2022**, *23*, 1–51.
61. Liu, B.; Chang, H.; Peng, K.; Wang, X. An end-to-end depression recognition method based on EEGNet. *Frontiers in Psychiatry* **2022**, *13*, 864393.
62. Zhang, H.; Xie, J.; Liu, K.; Liu, Y.; Dong, W.; Xu, G. Time frequency transform kernel enhanced ShallowConvNet for auditory selective attention decoding with steady state motion auditory evoked potential. *Biomedical Signal Processing and Control* **2026**, *119*, 109736.
63. Salazar-Dubois, D.V.; Álvarez-Meza, A.M.; Castellanos-Dominguez, G. T-GARNet: A Transformer and Multi-Scale Gaussian Kernel Connectivity Network with Alpha-Rényi Regularization for EEG-Based ADHD Detection. *Mathematics* **2025**, *13*, 4026.
64. Roshan, K.; Zafar, A. Using kernel shap xai method to optimize the network anomaly detection model. In Proceedings of the 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2022, pp. 74–80.
65. Raptis, S.; Ilioudis, C.; Theodorou, K. From pixels to prognosis: unveiling radiomics models with SHAP and LIME for enhanced interpretability. *Biomedical Physics & Engineering Express* **2024**, *10*, 035016.
66. Lundstrom, D.D.; Huang, T.; Razaviyayn, M. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In Proceedings of the International Conference on Machine Learning. PMLR, 2022, pp. 14485–14508.
67. Tripathi, S.; Arya, N.; Kaur, S.; Gupta, T.; Gupta, E.; et al. Grad-CAM++ Enhanced Hybrid CNN-Random Forest Model for Accurate and Transparent Brain Tumor Detection. In Proceedings of the 2025 5th International Conference on Intelligent Technologies (CONIT). IEEE, 2025, pp. 1–6.
68. Saranya, S.; Menaka, R. An explainable machine learning network for classification of autism spectrum disorder using optimal frequency band identification from brain EEG. *IEEE Access* **2025**.
69. Xiao, C.; Dou, J.; Lin, Z.; Ke, Z.; Hou, L. From points to coalitions: Hierarchical contrastive shapley values for prioritizing data samples. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 15995–16003.
70. Reuter, A.; Thielmann, A.; Saefken, B. Neural additive image model: Interpretation through interpolation. *arXiv preprint arXiv:2405.02295* **2024**.
71. Lasfar, R.; Tóth, G. The difference of model robustness assessment using cross-validation and bootstrap methods. *Journal of Chemometrics* **2024**, *38*, e3530.