HIERARCHICAL TOPOLOGY-AWARE SPARSE ATTENTION FOR MEDICAL TIME-SERIES CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical time series (MedTS) exhibit complementary temporal and inter-channel dependencies that are structured and non-uniform. Existing Transformer-based methods increasingly introduce structured representations and interactions, but generally do not adapt their interaction patterns to the distinct characteristics of temporal and inter-channel dependencies. We propose **TopoAttn**, a sparse attention framework that combines branch-specific interaction structures with hierarchical local-to-global communication. In the channel branch, direct interactions are determined by groups derived from training-set Fast Fourier Transform (FFT)-based spectral similarities, whereas the temporal branch restricts interactions to local temporal neighborhoods; learnable hubs coordinate information across groups. Mixture-of-Experts (MoE)-based temporal embedding and subject-guided enhancement further enrich the complementary branch representations. Under subject-independent five-fold evaluation on five EEG and ECG datasets, TopoAttn achieves the best result in 27 of 30 dataset-metric comparisons among the evaluated methods. Ablation and sensitivity analyses support the benefit of structured sparse interaction and branch-specific grouping beyond sparsity alone. TopoAttn also achieves the best result across all metrics on two human activity recognition datasets, supporting transfer beyond MedTS. The code will be released upon acceptance.

1 INTRODUCTION

Medical time series (MedTS), such as electroencephalography (EEG) and electrocardiography (ECG), provide continuous observations of physiological processes and support applications ranging from disease diagnosis to health monitoring and brain-computer interfaces (Liu et al., 2021; Murat et al., 2020). MedTS exhibit structured dependencies along two complementary dimensions: temporal patterns capture local morphology, rhythmic activity, and evolving physiological events, while inter-channel relationships reflect coordinated responses among sensors or leads (Wang et al., 2024; Guoqi et al., 2026). Effectively modeling both structures is critical for MedTS classification.

A further challenge is inter-subject variability. Because MedTS datasets often contain multiple segments from the same subject, random sample-level splits can place subject-related signals in both training and test sets, leading to overly optimistic performance estimates. Subject-independent evaluation avoids this leakage by keeping subjects disjoint across partitions and provides a stricter assessment of generalization to unseen subjects (Wang et al., 2024; Guoqi et al., 2026).

Transformers (Vaswani et al., 2017) enable global token interaction, but standard self-attention does not explicitly encode MedTS structure. Recent methods introduce more structured communication. For instance, Medformer (Wang et al., 2024) organizes cross-channel patches across temporal granularities, whereas TeCh (Guoqi et al., 2026) uses a common central-token communication scheme for temporal and channel representations. However, neither explicitly tailors token interactions to the distinct structures of temporal and channel dependencies. We use *topology* to denote which token pairs can interact directly and how tokens are grouped; designing dimension-specific interaction topologies for MedTS therefore remains an open problem.

In this work, we approach this problem by modeling MedTS as a *multi-center* structure of locally coherent token groups and propose **TopoAttn**, a topology-aware sparse attention framework. Fig. 1 contrasts representative interaction paradigms for MedTS and highlights TopoAttn's combination

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

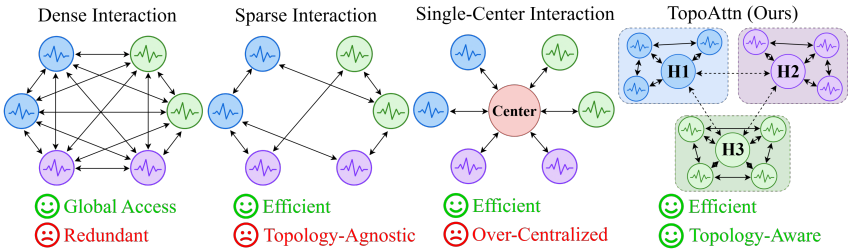

Figure 1: Conceptual comparison of interaction paradigms. Dense interaction offers global access but may introduce redundancy; sparse interaction improves efficiency without MedTS-specific topology; centralized interaction coordinates information through a common center; and TopoAttn combines sparse communication with topology-aware multi-center organization. They are not mutually exclusive and illustrate characteristic interaction patterns rather than individual methods.

of branch-specific topologies with hierarchical local-to-global communication. This hierarchical mechanism is shared across both branches: tokens interact locally, group information is aggregated by hubs, hubs communicate across groups, and the coordinated information is redistributed to the tokens. The topology itself is dimension-specific: channel groups are constructed from training-set Fast Fourier Transform (FFT)-based spectral similarities, whereas temporal interactions follow local temporal neighborhoods and contiguous groups. This design supports fine-grained local interaction and coarse-grained cross-group coordination without dense all-to-all communication. TopoAttn further constructs complementary representations for the two dimensions. The channel branch represents complete channel trajectories, while the temporal branch uses cross-channel multi-timestamp patches with a Mixture-of-Experts (MoE) embedding for heterogeneous local patterns. Subject-guided token enhancement incorporates subject-level context, and adaptive gated fusion combines the channel and temporal representations for classification.

Experiments on MedTS benchmarks show strong overall performance against earlier methods. Ablation and sensitivity analyses support the contributions of TopoAttn's key components and topology construction. Efficiency analysis further shows a favorable performance-efficiency trade-off. Results on human activity recognition (HAR) datasets provide evidence of transfer beyond MedTS.

Our main contributions are summarized as follows. (1) We introduce TopoAttn, a topology-aware sparse attention framework that combines dimension-specific interaction topologies with a shared hierarchical local-to-global communication mechanism, enabling structured within-group interaction and efficient cross-group coordination. (2) We design complementary channel and temporal topology priors based on spectral channel relationships and temporal locality, respectively. Ablation and topology-sensitivity analyses support the benefit of meaningful topology construction beyond sparsity alone. (3) TopoAttn improves over prior methods under subject-independent evaluation across several MedTS datasets and performs strongly on HAR datasets, supporting its effectiveness across modalities and transfer beyond MedTS.

2 Related Work

Earlier MedTS approaches employed handcrafted features (Fahimi et al., 2017; Schmidt et al., 2013), convolutional and recurrent networks (Lawhern et al., 2018; Roy et al., 2019; Salloum & Kuo, 2017), and graph models (Tang et al., 2021; Fan et al., 2025), whereas recent methods increasingly exploit attention-based representations.

Tab. 1 compares Transformer-based time-series methods along representation, interaction, and evaluation dimensions. *Multi-timestamp* (MT) denotes representations that encode multiple observations within a modeling unit, while *cross-channel* (CC) denotes explicit modeling of dependencies across variables or sensors. *Multi-granularity* (MG) captures information at multiple structured levels: MTST (Zhang et al., 2024) uses multiple temporal resolutions, Medformer (Wang et al., 2024) and MedSpaformer (Ye et al., 2026) use multiple temporal granularities with cross-channel modeling, whereas TopoAttn realizes fine-to-coarse structure through local token interactions and group-level hub coordination. The main distinction lies in how interactions are organized. Informer (Zhou et al.,

2021) and Reformer (Kitaev et al., 2020) sparsify or approximate attention primarily for efficiency, while Medformer (Wang et al., 2024) structures communication across temporal granularities. MedSpaformer (Ye et al., 2026) combines global self-attention with learnable-query token-sparse attention to compress representations within multi-granularity cross-channel encoding. TeCh (Guoqi et al., 2026) applies a common core-token aggregation-redistribution mechanism to both temporal and channel representations. TopoAttn instead uses branch-specific topology priors: temporal interactions follow local temporal structure, while channel interactions follow training-set frequency-domain relationships, with hubs enabling cross-group coordination. *Multi-domain* (MD) denotes evaluation on both MedTS and other time-series domains, and *subject-independent* (SI) denotes evaluation on subjects unseen during training (Wang et al., 2024; Guoqi et al., 2026).

Table 1: Comparison of Transformer-based time-series methods. See text for abbreviations and explanations.

Models	MT	CC	MG	Interaction	MD	SI
Autoformer (Wu et al., 2021)		✓		Auto-Correlation + decomposition		
FEDformer (Zhou et al., 2022)		✓		Frequency-enhanced interaction + decomposition		
Informer (Zhou et al., 2021)		✓		ProbSparse attention		
iTransformer (Liu et al., 2024)	✓	✓		Attention across variable tokens		
MTST (Zhang et al., 2024)	✓		✓	Multi-branch attention across temporal resolutions		
Nonformer (Liu et al., 2022)		✓		De-stationary attention		
PatchTST (Nie et al., 2022)	✓			Channel-independent patch attention		
Reformer (Kitaev et al., 2020)		✓		Locality-Sensitive Hashing attention		
Transformer (Vaswani et al., 2017)		✓		Dense self-attention		
Medformer (Wang et al., 2024)	✓	✓	✓	Intra-/inter-granularity attention	✓	✓
MedSpaformer (Ye et al., 2026)	✓	✓	✓	Token-sparse dual attention		
TeCh (Guoqi et al., 2026)	✓	✓		Core-token aggregation-redistribution	✓	✓
TopoAttn (Ours)	✓	✓	✓	Topology-aware sparse attention with branch-specific topologies	✓	✓

3 TopoAttn

Consider an input MedTS sample $\boldsymbol{X} \in \mathbb{R}^{T \times C}$, where T denotes the number of timestamps and C denotes the number of channels. Given a dataset $\mathcal{D} = \{(\boldsymbol{X}_i, y_i)\}_{i=1}^{N}$ with N samples, where $y_i \in \{1, \ldots, K\}$ denotes the class label and K is the number of classes, our objective is to learn a function f_θ that maps each input sample to a vector of class scores $\boldsymbol{o} = f_\theta(\boldsymbol{X}) \in \mathbb{R}^K$. The predicted class is then given by $\hat{y} = \arg\max_{k \in \{1,\ldots,K\}} o_k$. Depending on the dataset, the classes may represent disease categories, disease stages, or other activity categories.

3.1 Rationale and Overview of TopoAttn

Fig. 2 illustrates the overall framework of TopoAttn. We use *topology* to denote the pattern of permitted direct token interactions and token grouping. This design is motivated by the non-uniform structure of MedTS dependencies, which differs across channel and temporal dimensions. EEG electrodes reflect spatially distributed neural activity, while ECG leads observe cardiac activity from different orientations (Sanei & Chambers, 2013; Kayyali et al., 2025), yielding structured inter-channel relationships. Meanwhile, MedTS exhibit both local dynamics and longer-range rhythmic variation, such as EEG frequency-band activity and ECG beat-to-beat variability (Sanei & Chambers, 2013; Electrophysiology, 1996). These properties motivate branch-specific interaction structures, where related tokens interact directly, and cross-group information is coordinated through hubs.

Given an input MedTS sample, *Augmented Patch Embedding* constructs channel-wise and temporal-wise tokens for two parallel branches. *Topology-Aware Sparse Attention* then applies a spectral-similarity topology derived from the training split to the channel branch and a locality-based topology to the temporal branch. Both branches follow the same hierarchical interaction scheme: local token interaction, hub aggregation, inter-hub communication, and redistribution to tokens. *Subject-Guided Token Enhancement* incorporates subject-level context, while the *Adaptive Gated Fusion Classifier* integrates the complementary branch representations for final classification.

3.2 Augmented Patch Embedding

MedTS exhibit complementary structure across channels and time. Channel-centric tokens preserve holistic channel dynamics but may miss local temporal patterns, whereas patch-centric tokens capture local dynamics but may lose global channel context (Liu et al., 2024; Nie et al., 2022). We therefore construct two parallel branches: the channel branch represents each complete channel as one token, while the temporal branch uses non-overlapping cross-channel patches. Building on dual

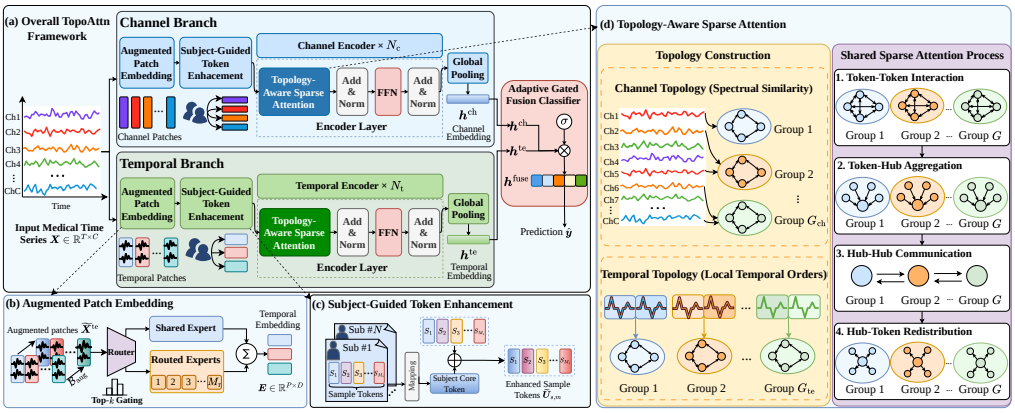

Figure 2: Overview of TopoAttn. (a) Dual-branch architecture; (b) augmented patch embedding; (c) subject-guided token enhancement; and (d) topology-aware sparse attention.

tokenization (Guoqi et al., 2026), TopoAttn further distinguishes the two branches through branch-specific topology construction and hierarchical topology-aware sparse interaction (See Sec. 3.3).

Before tokenization, we independently augment the two branches during training to provide regularization while preserving task-relevant information. Following common time-series augmentation strategies (Chen et al., 2020; Wang et al., 2023), we randomly sample augmentations per branch from the bank in App. A.1: $\widetilde{\boldsymbol{X}}^{\mathrm{ch}} = \mathcal{A}_{\mathrm{ch}}(\boldsymbol{X})$ and $\widetilde{\boldsymbol{X}}^{\mathrm{te}} = \mathcal{A}_{\mathrm{te}}(\boldsymbol{X})$, where $\mathcal{A}_{\mathrm{ch}}, \mathcal{A}_{\mathrm{te}} \overset{\text{ind.}}{\sim} \mathcal{B}_{\mathrm{aug}}$ and $\widetilde{\boldsymbol{X}}^{\mathrm{ch}}, \widetilde{\boldsymbol{X}}^{\mathrm{te}} \in \mathbb{R}^{T\times C}$. These branch-specific views are subsequently converted into channel-wise and temporal-wise tokens.

Channel Embedding. The channel branch treats each sensor channel as a coherent modeling unit. This preserves the complete temporal trace of a channel, rather than dividing it into independent temporal fragments, and provides a channel-specific representation for subsequent inter-channel modeling. Following the channel-centric tokenization of Liu et al. (2024), the complete augmented sequence of each channel is projected into the token space as $\boldsymbol{H}_{c,:} = (\widetilde{\boldsymbol{X}}^{\mathrm{ch}}_{:,c})^{\top}\boldsymbol{W}_{\mathrm{ch}} + \boldsymbol{b}_{\mathrm{ch}} + \boldsymbol{P}^{\mathrm{ch}}_{c,:}$, $c = 1, \ldots, C$. Here, $\boldsymbol{W}_{\mathrm{ch}} \in \mathbb{R}^{T\times D}$ and $\boldsymbol{b}_{\mathrm{ch}} \in \mathbb{R}^{D}$ are learnable projection parameters, while $\boldsymbol{P}^{\mathrm{ch}} \in \mathbb{R}^{C\times D}$ denotes fixed sinusoidal positional encodings over channel-token positions. This yields $\boldsymbol{H} \in \mathbb{R}^{C\times D}$, with one token per channel encoding its behavior over the full input window.

Temporal Embedding. While the channel branch preserves the full temporal trajectory of each channel, it does not explicitly expose local temporal events as separate modeling units. We therefore divide the input into non-overlapping cross-channel patches, each jointly representing a short temporal interval across all channels. Given patch length L, we zero-pad the temporal dimension when necessary and obtain $P = \lceil T/L \rceil$ patches. Let $\overline{\boldsymbol{X}}^{\mathrm{te}} \in \mathbb{R}^{PL\times C}$ denote the padded version of $\widetilde{\boldsymbol{X}}^{\mathrm{te}}$. The p-th patch is $\boldsymbol{Z}_p = \overline{\boldsymbol{X}}^{\mathrm{te}}_{(p-1)L+1:pL,:} \in \mathbb{R}^{L\times C}$, $p = 1, \ldots, P$, with flattened representation $\boldsymbol{z}_p = \mathrm{vec}(\boldsymbol{Z}_p) \in \mathbb{R}^{LC}$. Thus, each token jointly represents L consecutive timestamps across all C channels. To accommodate heterogeneous temporal content, we employ a noisy-gated MoE (Shazeer et al., 2017). Each expert applies $\mathcal{E}^q_m(\boldsymbol{z}_p) = \boldsymbol{z}_p\boldsymbol{W}^q_m + \boldsymbol{b}^q_m \in \mathbb{R}^D$, where $q \in \{s, r\}$ denotes shared and routed experts, respectively, with $\boldsymbol{W}^q_m \in \mathbb{R}^{LC\times D}$ and $\boldsymbol{b}^q_m \in \mathbb{R}^D$. For the M_r routed experts, noisy top-k gating computes routing scores and sparse expert weights, while M_s shared experts capture patterns common across patches. The temporal token is obtained as

$$
\begin{aligned}
&\boldsymbol{h}_p = \mathrm{Linear}_{\mu}(\boldsymbol{z}_p) + \boldsymbol{\epsilon}_p \odot \mathrm{Linear}_{\sigma}(\boldsymbol{z}_p), \qquad \boldsymbol{\epsilon}_p \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{M_r}),\\
&\boldsymbol{g}_p = \mathrm{Softmax}(\mathrm{KeepTopK}(\boldsymbol{h}_p, k)) \in \mathbb{R}^{M_r},\\
&\boldsymbol{E}_{p,:} = \sum_{m=1}^{M_s} \mathcal{E}^s_m(\boldsymbol{z}_p) + \sum_{m=1}^{M_r} g_{p,m}\mathcal{E}^r_m(\boldsymbol{z}_p) + \boldsymbol{P}^{\mathrm{te}}_{p,:}.
\end{aligned}
\tag{1}
$$

Here, $\mathrm{KeepTopK}(\cdot, k)$ retains the k largest routing scores and masks the rest with $-\infty$, yielding sparse expert weights. $\boldsymbol{P}^{\mathrm{te}} \in \mathbb{R}^{P\times D}$ denotes fixed sinusoidal positional encodings. This yields the temporal embedding $\boldsymbol{E} \in \mathbb{R}^{P\times D}$. Consequently, $\boldsymbol{H}$ captures full-window channel behavior,

while $\boldsymbol{E}$ captures heterogeneous local temporal dynamics with cross-channel context. Both are subsequently processed by their branch-specific topology-aware sparse attention modules.

3.3 Topology-Aware Sparse Attention

Dense self-attention treats all token pairs uniformly, although MedTS dependencies are often local or structure-dependent. We therefore introduce *topology-aware sparse attention*, which restricts direct interactions using branch-specific topologies while preserving cross-group communication through learnable hubs. The central design principle is to separate local specialization from global coordination: tokens interact within topology groups, hubs aggregate group information, communicate globally, and redistribute the coordinated information to tokens. This shared hierarchical mechanism is instantiated with different priors for the two branches: a data-driven spectral topology for channels and a locality-based topology for temporal patches, making branch-specific topology construction central to TopoAttn.

Topology-Aware Interaction. Let $\boldsymbol{U} \in \mathbb{R}^{S\times D}$ denote the input tokens, where $S = C$ for the channel branch and $S = P$ for the temporal branch. A token-token mask $\boldsymbol{M}^{\mathrm{tt}} \in \{0,1\}^{S\times S}$ specifies direct interactions, while a token-hub mask $\boldsymbol{M}^{\mathrm{th}} \in \{0,1\}^{S\times G}$ assigns tokens to G topology groups. Masked multi-head attention is then applied according to these topology constraints. Given G learnable group hubs, interaction proceeds in four stages: tokens first interact under $\boldsymbol{M}^{\mathrm{tt}}$; each hub aggregates its assigned tokens according to $\boldsymbol{M}^{\mathrm{th}}$; hubs communicate globally through multi-head attention; and the resulting hub information is redistributed through $\boldsymbol{M}^{\mathrm{th}}$ and combined with the local token features. This separates topology-constrained direct interaction from cross-group coordination through a compact set of hub representations. The resulting interaction block is followed by residual normalization and a feed-forward network. Detailed equations are in App. A.4.2.

Channel Topology. The channel branch models interactions among sensor or lead tokens. Since physiological channels may exhibit distinct yet correlated spectral patterns, fully connected interactions can introduce redundancy. We therefore construct a data-driven topology from frequency-domain similarity. FFT decomposes signals into frequency components, from which relative spectral energies can be computed, revealing periodic and global characteristics that may be obscured in raw waveforms, while similar spectral profiles can reflect similar underlying oscillatory structures (Savvides et al., 2008; Chen et al., 2012). Accordingly, channels are grouped according to their normalized spectral-energy profiles. For each cross-validation fold, the topology is constructed solely from the training subjects and then kept fixed for training, validation, and test sets, thereby preventing information contained in the evaluation partitions from influencing topology construction.

For each training sample and channel, we center the signal, compute its FFT, and summarize the normalized spectral energy over F predefined frequency bins, yielding $\boldsymbol{f}_{n,c} \in \mathbb{R}^F$. We obtain a robust channel prototype using the element-wise median across training samples, $\bar{\boldsymbol{f}}_c = \mathrm{Median}\, n(\boldsymbol{f}n, c) \in \mathbb{R}^F$, and compute pairwise spectral similarity as $s_{ij} = (\bar{\boldsymbol{f}}_i^\top \bar{\boldsymbol{f}}_j)/(\|\bar{\boldsymbol{f}}_i\|_2 \|\bar{\boldsymbol{f}}_j\|_2)$, $i, j = 1, \ldots, C$. We define the channel-distance matrix as $\boldsymbol{\Delta} = [1 - s_{ij}]_{i,j=1}^{C} \in \mathbb{R}^{C\times C}$, then apply average-linkage agglomerative clustering (Moseley & Wang, 2023) until G_{ch} non-overlapping groups $\{\mathcal{G}_g^{\mathrm{ch}}\}_{g=1}^{G_{\mathrm{ch}}}$ remain. The resulting topology masks are $\boldsymbol{M}_{ij}^{\mathrm{tt,ch}} = \mathbb{I}[\exists g : i, j \in \mathcal{G}_g^{\mathrm{ch}}] \in \{0,1\}^{C\times C}$ and $\boldsymbol{M}_{i,g}^{\mathrm{th,ch}} = \mathbb{I}[i \in \mathcal{G}_g^{\mathrm{ch}}] \in \{0,1\}^{C\times G_{\mathrm{ch}}}$. Thus, channels assigned to the same spectral group interact directly, while cross-group information is exchanged through the hub pathway, yielding a data-driven alternative to fixed or fully connected channel organization.

Temporal Topology. The temporal branch contains P ordered patch tokens and uses temporal order as its structural prior. Since neighboring patches correspond to adjacent temporal intervals and often exhibit locally related dynamics (Chen et al., 2024), we consequently restrict direct token interactions to a temporal neighborhood of radius r. We further partition the sequence into G_{te} contiguous, non-overlapping groups $\{\mathcal{G}_g^{\mathrm{te}}\}_{g=1}^{G_{\mathrm{te}}}$, each associated with one temporal hub. The token-token and token-hub masks are defined as

$$
\begin{aligned}
&\boldsymbol{M}_{pq}^{\mathrm{tt,te}} = \mathbb{I}[|p-q| \le r] \in \{0,1\}^{P\times P}, \qquad p, q = 1, \ldots, P, \\
&\mathcal{G}_g^{\mathrm{te}} = \left\{ p \;\middle|\; \left\lfloor \frac{(g-1)P}{G_{\mathrm{te}}} \right\rfloor + 1 \le p \le \left\lfloor \frac{gP}{G_{\mathrm{te}}} \right\rfloor \right\}, \\
&\boldsymbol{M}_{p,g}^{\mathrm{th,te}} = \mathbb{I}\big[p \in \mathcal{G}_g^{\mathrm{te}}\big] \in \{0,1\}^{P\times G_{\mathrm{te}}}, \qquad g = 1, \ldots, G_{\mathrm{te}}.
\end{aligned}
\tag{2}
$$

$\boldsymbol{M}^{\mathrm{tt,te}}$ preserves short-range interactions between neighboring patches, while $\boldsymbol{M}^{\mathrm{th,te}}$ assigns each patch to a contiguous temporal group. Hub aggregation and inter-hub attention then propagate information across distant groups, combining local temporal specialization with long-range coordination.

Overall, both branches share the same hierarchical sparse interaction mechanism but use different structural priors: spectral similarity estimated from the training samples for channel topology, and temporal order for local neighborhoods and groups. This branch-specific topology construction with shared hub-mediated communication enables TopoAttn to model structured dependencies without dense all-to-all interaction.

3.4 Subject-Guided Token Enhancement

A single MedTS segment provides only a partial observation of a subject, whereas samples from the same subject generally share subject-specific physiological characteristics and complementary information (Marcel & Millán, 2007; Israel et al., 2005; Biel et al., 2001). Inspired by the aggregation-redistribution mechanism of (Han et al., 2024), we extend its within-sample token aggregation to same-subject samples. Specifically, samples sharing a subject ID are first aggregated into a subject-level core, which is then redistributed to enhance each sample representation. This mechanism is applied independently to the channel and temporal branches. For subject s with M_s samples, let $\boldsymbol{U}_{s,m} \in \mathbb{R}^{S\times D}$ denote its m-th sample and $\boldsymbol{Z}_{s,m} \in \mathbb{R}^{S\times D_c}$ its projected features, where $D_c = D/4$. The subject-level core is: $\boldsymbol{S}_s = \sum_{m=1}^{M_s} \mathrm{Softmax}_{M_s}\left(\{\boldsymbol{Z}_{s,m}\}_{m=1}^{M_s}\right)_m \odot \boldsymbol{Z}_{s,m} \in \mathbb{R}^{S\times D_c}$. It is then redistributed to each sample as $\widetilde{\boldsymbol{U}}_{s,m} = \mathcal{R}(\boldsymbol{U}_{s,m}, \boldsymbol{S}_s)$, where $\mathcal{R}(\cdot)$ denotes the redistribution operator detailed in App. A.4.2. Unlike a sample-level core, $\boldsymbol{S}_s$ aggregates information across same-subject samples while preserving token positions. It is constructed once per epoch and cached for the corresponding samples. After this initial enhancement, subject IDs are no longer required inside the encoder. Each layer instead applies the same aggregation-redistribution mechanism within each sample before topology-aware sparse attention, progressively refining its global context. Thus, subject-level context is introduced once and subsequently refined through layer-wise sample-level aggregation, jointly capturing subject- and sample-specific information while improving robustness to inter-subject variability.

3.5 Adaptive Gated Fusion Classifier

The channel and temporal branches capture complementary dependencies and may contribute differently across samples and feature dimensions. We therefore employ adaptive gated fusion to dynamically balance their representations before classification. Given the final token embedding $\boldsymbol{U}^{\mathrm{ch}} \in \mathbb{R}^{C\times D}$ and $\boldsymbol{U}^{\mathrm{te}} \in \mathbb{R}^{P\times D}$, we obtain normalized branch-level representations $\boldsymbol{h}^{\mathrm{ch}}, \boldsymbol{h}^{\mathrm{te}} \in \mathbb{R}^{D}$ by mean pooling. Their concatenation, element-wise product, and absolute difference are combined into $\boldsymbol{z}$ to characterize both complementary information and cross-branch agreement. The feature-wise gate and fused representation are: $\boldsymbol{g} = \sigma(\mathrm{MLP}_g(\boldsymbol{z})) \in (0,1)^D, \quad \boldsymbol{h}^{\mathrm{fuse}} = \mathrm{LN}\big(\boldsymbol{g}\odot\boldsymbol{h}^{\mathrm{ch}} + (1-\boldsymbol{g})\odot\boldsymbol{h}^{\mathrm{te}} + \mathrm{MLP}_r([\boldsymbol{h}^{\mathrm{ch}}\|\boldsymbol{h}^{\mathrm{te}}])\big)$. The gate adaptively balances the two branches across feature dimensions, while the residual fusion path preserves complementary information. Finally, $\boldsymbol{h}^{\mathrm{fuse}}$ is linearly projected to K class scores. Detailed fusion operations are provided in App. A.4.2.

3.6 Inference

All topology construction and evaluation are performed independently within each subject-level cross-validation fold. The channel topology is constructed exclusively from the training subjects of each fold and remains fixed for training, validation, and testing. Subject-guided token enhancement is applied separately within each split: subject cores are constructed by aggregating samples belonging to the same subject and are used to enhance the corresponding sample representations. Therefore, no information is shared across training, validation, and test subjects, while subject identity is used only to associate samples from the same subject for subject-level aggregation.

4 EXPERIMENTAL ANALYSIS

Datasets. We compare our TopoAttn with 11 SOTA methods across five MedTS datasets, including three EEG datasets (**APAVA**, **TDBrain**, and **ADFTD**) and two ECG datasets (**PTB** and **PTB-XL**). We evaluate all methods under the *subject-independent* setting, in which the training, validation, and test sets are partitioned according to subject identity. To further examine the generalizability of TopoAttn beyond MedTS, we conduct additional experiments on two human activity recognition (HAR) datasets, namely **FLAAP** and **UCI-HAR**. We follow the preprocessing protocol of Medformer (Wang et al., 2024) for each dataset, which is described in detail in App. A.2 and A.3.

SOTA. We compare our method with the following methods: Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), Informer (Zhou et al., 2021), iTransformer (Liu et al., 2024), MTST (Zhang et al., 2024), Nonformer (Liu et al., 2022), PatchTST (Nie et al., 2022), Reformer (Kitaev et al., 2020), Medformer (Wang et al., 2024), TeCh (Guoqi et al., 2026), and the vanilla Transformer (Vaswani et al., 2017). Implementation details for all comparison methods are provided in App. A.4.3. We reproduce the SOTA results using the official implementations and training frameworks provided by Medformer (Wang et al., 2024) and TeCh (Guoqi et al., 2026).

Evaluation metrics. Following the evaluation practice of SOTA MedTS studies, e.g., (Wang et al., 2024; Guoqi et al., 2026), we report six metrics: accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1 score, macro-averaged AUROC, and macro-averaged AUPRC. For each dataset, we perform 5-fold cross-validation using a fixed split seed. Then, we repeat the complete five-fold evaluation for five independent training seeds and report the mean and standard deviation across all resulting runs. For each run, we select the checkpoint with the best F1 score on the validation set and evaluate it on the test set, in line with (Wang et al., 2024; Guoqi et al., 2026). We additionally aggregate predictions across all samples of each held-out subject and report subject-level evaluation results in App. A.9.4.

4.1 MAIN RESULTS

Results on MedTS datasets. As shown in Tab. 2, TopoAttn achieves the best performance on 27 out of 30 evaluation metrics, with an average F1 improvement of 4.48% over TeCh, the strongest SOTA, across five MedTS datasets. The gains are consistent across both EEG and ECG benchmarks, with particularly pronounced improvements on APAVA, where F1 and AUPRC increase by 7.33% and 10.66%, respectively. TopoAttn also exhibits improved stability under cross-validation evaluations, reducing the average standard deviation by 7.44% relative to TeCh. Overall, these gains and lower variability demonstrate that TopoAttn models structured MedTS dependencies more effectively and consistently across datasets and modalities. Statistical significance against TeCh is further examined in App. A.11.

Generalization Beyond MedTS. TopoAttn achieves the best performance across all six metrics on both HAR datasets (see App. Tab. 7). In terms of F1, TopoAttn achieves 81.19% on FLAAP and 97.85% on UCI-HAR, while yielding an average improvement of 0.72% over TeCh across the 12 dataset-metric entries. These consistent gains indicate that the effectiveness of TopoAttn extends beyond MedTS to broader multivariate sensor-classification tasks.

4.2 MODEL ANALYSES

Ablation Study. We evaluate the contribution of the main components of TopoAttn through five ablation settings on the five MedTS datasets. *w/o Topology* replaces topology-aware sparse attention in both branches with dense self-attention; *w/o Channel Topology* and *w/o Temporal Topology* remove the topology constraint from the corresponding branch; *w/o MoE* replaces the MoE-based temporal embedding with a single linear projected embedding; and *w/o Sub* removes subject-guided token enhancement. All remaining components and hyperparameters are kept unchanged, following the implementation details in (App. A.4.1). Tab. 3 shows that the complete TopoAttn consistently achieves the highest accuracy and F1 across all five datasets. Averaged across datasets, removing topology-aware interactions entirely reduces F1 by 2.43%, while removing the channel or temporal topology individually leads to relative decreases of 0.96% and 2.70%, respectively. The channel-topology ablation consistently reduces F1, while the larger drop from removing temporal topology

Table 2: Results on five MedTS datasets. The training, validation, and test sets are distributed based on subject IDs. The best result is marked in **bold red**, and the second best is marked in *blue*.

Dataset	Model	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
ADFTD (3 Classes)	Autoformer	46.66±0.90	45.31±1.12	44.44±1.10	43.50±1.17	63.01±1.11	46.40±1.44
	FEDformer	47.94±0.85	47.05±1.42	45.63±0.71	44.63±0.67	64.77±0.86	48.64±0.93
	Informer	49.58±1.19	48.19±0.85	46.95±1.26	46.72±1.11	65.42±1.16	48.38±1.34
	iTransformer	*51.68±0.43*	48.66±0.24	*48.55±0.36*	*48.19±0.33*	*68.00±0.25*	50.71±0.34
	MTST	50.34±1.20	*49.67±0.87*	47.43±0.54	46.73±0.89	65.78±0.68	49.12±0.43
	Nonformer	49.36±0.50	47.67±0.75	47.10±0.68	46.19±0.52	64.82±0.88	48.55±0.82
	PatchTST	49.63±1.58	48.29±1.39	46.51±1.26	45.86±1.48	64.80±1.03	48.48±1.11
	Reformer	47.00±1.47	46.17±0.77	44.44±1.35	43.04±1.88	64.04±0.84	48.16±0.91
	Transformer	50.51±0.89	48.60±1.04	48.19±0.80	47.29±0.79	66.04±1.07	49.83±0.87
	Medformer	48.84±0.77	45.47±2.82	43.40±0.97	40.22±1.72	64.09±1.07	47.68±1.03
	TeCh	50.71±0.74	48.35±0.61	47.28±0.68	46.58±0.97	66.57±0.63	*50.73±0.58*
	TopoAttn	**54.09±0.99**	**51.78±1.25**	**51.01±1.05**	**50.42±1.03**	**69.35±0.87**	**54.10±1.01**
APAVA (2 Classes)	Autoformer	64.11±1.32	63.31±3.04	65.34±1.59	58.00±1.92	71.90±1.94	70.04±2.18
	FEDformer	65.24±2.07	64.18±1.99	64.74±1.27	57.12±2.25	78.81±1.28	74.08±1.74
	Informer	76.82±1.91	72.78±1.50	76.13±3.23	71.13±2.92	83.22±1.10	76.06±1.22
	iTransformer	78.47±0.55	75.30±0.76	78.66±1.12	73.30±0.72	84.70±1.00	77.87±1.07
	MTST	74.11±2.45	71.51±3.05	72.32±2.06	69.05±2.85	79.41±1.60	74.34±1.67
	Nonformer	74.86±0.81	70.47±0.73	73.82±1.63	69.17±1.20	81.35±1.41	73.87±1.60
	PatchTST	70.59±2.42	68.87±3.49	69.22±1.49	65.04±3.31	77.46±2.63	73.37±2.26
	Reformer	76.30±2.04	72.59±1.53	76.19±1.52	71.09±2.13	82.99±1.90	75.85±2.36
	Transformer	75.91±1.36	71.27±0.90	74.13±2.14	69.89±1.69	81.90±0.83	74.39±1.00
	Medformer	76.08±1.00	71.97±0.65	76.40±0.66	70.93±0.93	82.81±0.50	74.99±0.58
	TeCh	*80.25±1.18*	*78.11±0.72*	*80.22±1.96*	*74.66±1.68*	*88.46±0.55*	*80.84±0.60*
	TopoAttn	**84.08±0.24**	**79.92±0.23**	**85.20±0.47**	**80.13±0.27**	**91.30±0.23**	**89.46±0.70**
TDBrain (2 Classes)	Autoformer	83.10±1.91	82.52±2.11	81.62±1.75	81.34±1.87	89.44±1.32	89.44±1.32
	FEDformer	79.60±1.13	78.97±1.12	77.85±1.04	77.54±1.19	87.80±1.00	87.50±1.08
	Informer	84.91±0.60	84.66±0.43	*84.21±0.78*	83.67±0.72	*93.68±0.35*	92.57±0.37
	iTransformer	74.14±0.61	73.96±0.38	71.46±0.59	71.30±0.71	80.24±0.20	79.11±0.31
	MTST	76.74±0.78	75.62±1.03	74.12±0.96	73.63±1.11	83.51±0.72	81.04±0.43
	Nonformer	84.90±1.32	84.70±1.07	83.87±1.29	83.53±1.36	93.62±0.75	92.38±1.06
	PatchTST	78.03±0.79	77.06±0.93	75.73±0.66	75.49±0.74	85.87±0.39	83.44±0.65
	Reformer	83.46±0.57	83.53±0.47	82.76±0.69	82.11±0.66	92.29±0.35	91.26±0.41
	Transformer	84.17±0.70	83.75±0.75	83.48±0.85	82.85±0.80	93.19±0.70	92.01±0.80
	Medformer	83.48±0.40	83.79±0.31	81.83±0.58	81.87±0.49	92.04±0.46	91.21±0.53
	TeCh	**86.23±0.48**	*85.69±0.50*	**84.87±0.45**	**84.73±0.46**	*93.68±0.30*	*92.75±0.30*
	TopoAttn	*86.14±0.36*	**86.99±0.36**	83.67±1.22	*84.04±1.05*	**93.90±0.32**	**93.27±0.43**
PTB (2 Classes)	Autoformer	85.55±0.46	73.73±0.97	71.32±1.37	71.70±0.90	81.88±1.82	76.29±1.01
	FEDformer	88.20±0.64	80.00±1.26	76.98±2.54	76.90±1.34	88.83±1.35	83.62±1.36
	Informer	89.68±0.37	81.76±0.67	80.98±0.94	80.54±0.72	92.71±0.36	86.22±0.38
	iTransformer	89.36±0.48	80.43±0.65	83.11±0.63	81.05±0.63	92.67±0.61	86.66±0.72
	MTST	89.99±0.33	82.12±0.91	*83.37±0.98*	81.76±0.38	92.88±0.27	86.48±0.61
	Nonformer	88.31±0.82	79.00±1.77	77.65±0.30	77.37±0.91	90.97±0.53	83.58±1.20
	PatchTST	*90.18±0.20*	*82.66±0.32*	82.91±0.69	*81.87±0.41*	93.00±0.25	*87.30±0.32*
	Reformer	89.37±0.41	81.40±0.37	80.30±1.06	79.92±0.81	92.42±0.50	85.56±0.87
	Transformer	88.44±0.36	79.51±0.35	77.52±0.67	77.50±0.52	90.89±0.23	83.85±0.69
	Medformer	89.30±0.28	80.96±0.28	81.43±1.23	80.54±0.66	91.38±0.73	86.50±0.74
	TeCh	89.67±0.58	81.14±0.98	82.92±0.69	81.24±0.73	*93.17±0.72*	87.16±1.09
	TopoAttn	**92.72±0.71**	**86.89±1.42**	**87.37±0.75**	**86.45±1.04**	**96.05±0.28**	**92.82±1.06**
PTB-XL (5 Classes)	Autoformer	60.61±1.41	50.41±1.51	49.77±1.25	48.98±1.33	81.70±0.76	51.45±1.59
	FEDformer	54.59±11.34	53.70±3.91	48.80±6.79	47.10±8.32	82.53±3.66	53.61±5.84
	Informer	71.48±0.21	62.24±0.24	59.24±0.18	60.32±0.16	88.56±0.10	64.40±0.19
	iTransformer	69.48±0.13	59.52±0.24	55.32±0.12	56.80±0.11	87.09±0.11	60.84±0.17
	MTST	72.13±0.20	63.08±0.14	59.82±0.21	61.01±0.18	88.85±0.07	65.35±0.23
	Nonformer	70.69±0.34	61.29±0.39	58.77±0.32	59.67±0.28	88.20±0.16	63.54±0.35
	PatchTST	73.20±0.17	64.70±0.19	60.97±0.22	62.38±0.21	89.66±0.05	67.14±0.07
	Reformer	71.39±0.25	61.80±0.42	59.50±0.23	60.37±0.09	88.57±0.11	64.19±0.25
	Transformer	70.68±0.09	61.37±0.16	58.64±0.45	59.54±0.31	88.21±0.05	63.56±0.15
	Medformer	73.36±0.10	64.87±0.14	61.19±0.15	62.69±0.13	89.93±0.07	67.29±0.14
	TeCh	*73.53±0.16*	*65.33±0.41*	*61.49±0.31*	*62.85±0.17*	*90.16±0.08*	*67.82±0.17*
	TopoAttn	**74.93±0.07**	**67.12±0.23**	**62.05±0.09**	**63.61±0.11**	**91.01±0.05**	**69.51±0.14**

indicates a stronger temporal contribution on these benchmarks. The two effects are not strictly additive as the branches are jointly optimized. Beyond the topology design, replacing the MoE temporal embedding reduces F1 by 1.60%, and removing subject-guided enhancement leads to a 2.48% reduction. Overall, these results show that TopoAttn benefits from the coordinated contributions of topology-aware interactions, heterogeneous temporal representation, and subject-guided enhancement, rather than relying on any single component.

Sensitivity Analysis of Topologies. To evaluate the importance of the proposed topology design, we compare the channel and temporal topologies with alternative constructions while keeping all other model settings unchanged. For the channel branch, we compare the FFT-derived topology

Table 3: Ablation Study. We report Accuracy and F1-Score across five MedTS datasets. The best is highlighted in **bold red**. Full ablation results are in App. A.8.

Variants	ADFTD		APAVA		TDBrain		PTB		PTB-XL	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
w/o Topology	50.66±1.50	47.60±1.62	84.05±0.61	78.88±0.66	84.91±0.20	82.68±1.12	91.99±0.59	85.18±0.87	73.96±0.09	62.39±0.15
w/o Channel Topology	52.78±0.51	49.42±1.22	84.04±0.71	79.96±0.98	85.75±0.27	83.51±1.12	91.85±0.43	84.87±0.63	74.89±0.07	63.52±0.07
w/o Temporal Topology	50.35±1.88	47.44±1.94	82.59±1.10	77.80±1.23	85.08±0.45	82.76±1.51	91.94±0.64	85.30±1.13	73.96±0.08	62.45±0.19
w/o MoE	52.51±1.86	49.21±1.62	81.27±1.69	77.24±1.85	85.68±0.82	83.87±1.34	92.23±0.60	85.74±1.01	73.51±0.23	62.98±0.12
w/o Sub	52.91±0.62	48.86±0.52	81.32±3.37	77.16±3.85	85.55±0.84	82.80±1.89	90.95±0.50	83.29±0.91	74.74±0.11	63.32±0.17
TopoAttn	**54.09**±0.99	**50.42**±1.03	**84.08**±0.24	**80.13**±0.27	**86.14**±0.36	**84.04**±1.05	**92.72**±0.71	**86.45**±1.04	**74.93**±0.07	**63.61**±0.11

with a fixed domain-prior topology and a random topology; for the temporal branch, we compare the locality-based topology with a random topology. Here, *Prior* denotes a fixed domain-informed grouping based on established acquisition conventions, whereas *Random* denotes an unstructured sparse grouping corresponding to the topology-agnostic sparse setting illustrated in Fig. 1. Detailed prior constructions are provided in App. A.5.

Table 4: Topology sensitivity across seven datasets. F1 is reported as mean ± std, with full metrics provided in the appendix A.9.

Dataset	Channel topology			Temporal topology	
	Prior	Random	TopoAttn	Random	TopoAttn
ADFTD	49.91±0.79	49.48±1.36	**50.42**±1.03	46.47±1.45	**50.42**±1.03
APAVA	77.83±0.83	77.73±1.37	**80.13**±0.27	77.54±0.89	**80.13**±0.27
TDBrain	83.40±0.80	82.69±1.26	**84.04**±1.05	82.14±1.65	**84.04**±1.05
PTB	85.77±1.47	84.58±1.53	**86.45**±1.04	85.33±1.21	**86.45**±1.04
PTB-XL	62.80±0.13	62.76±0.20	**63.61**±0.11	62.06±0.12	**63.61**±0.11
FLAAP	79.50±0.33	80.76±0.40	**81.19**±0.27	80.84±0.28	**81.19**±0.27
UCI-HAR	97.65±0.09	97.48±0.13	**97.85**±0.19	97.54±0.06	**97.85**±0.19

As seen in Tab. 4, the proposed topology construction consistently outperforms its alternatives. Across the five MedTS datasets, the FFT-derived channel topology improves F1 by an average of 1.37% over the domain-prior topology and 2.04% over random topology, while the locality-based temporal topology yields a larger 3.59% improvement over random temporal topology. These results suggest that spectral similarity provides a more informative organization of channel relationships, while preserving local temporal order strengthens temporal modeling. The same trend is also observed on HAR datasets, with consistent F1 gains over the alternative topology constructions, indicating that the benefits of structured topology generalize beyond MedTS. Overall, the consistent gains present that TopoAttn benefits not from sparsity alone, but from branch-specific topologies that capture the distinct structures of channel and temporal dependencies. Full results are provided in App. Tabs. 9- 10.

Additional Analyses. Apps. A.6, A.9.3, and A.10 report theoretical complexity, topology robustness, and computational efficiency, respectively. App. **??** further shows that TopoAttn largely retains its performance under strict sample-wise inference on representative EEG and ECG datasets.

5 Conclusion

We introduced TopoAttn, a Transformer designed for medical time-series classification across unseen subjects. TopoAttn represents the input through complementary channel and temporal branches, constructs the channel topology from frequency-domain similarity, and defines temporal structure through local neighborhoods. These branch-specific structures are coordinated through hierarchical sparse interactions mediated by learnable hubs. Combined with an MoE temporal embedding and subject-guided token enhancement, TopoAttn captures structured dependencies across channels and time and generalizes to broader multivariate sensor data. While the current model relies on fixed topologies constructed before training, future work will investigate adaptive topology learning that jointly discovers and updates interaction structures to improve flexibility across subjects, datasets, and physiological conditions.

AI USE STATEMENT

In this work, we used generative AI tools for editing portions of the manuscript to improve grammar, clarity, and readability. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide ingredients for mathematical proofs, assist in writing proofs, propose or refine hypotheses, design or provide feedback on the methodology or experiments, implement methods, translate content, clean or reformat datasets, conduct qualitative or thematic analysis, or interpret results. Any remaining required-disclosure categories were not applicable to this work.

Additionally, we used generative AI tools for editing the paper to improve its readability, which is a recommended disclosure category. All AI-assisted text was reviewed and edited by the authors. The authors independently verified the analyses, results, and claims, and take responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available EEG, ECG, and human-activity datasets collected and released by the original dataset creators. We did not recruit human participants or collect new data. We used the datasets according to their applicable access conditions, licenses, and terms of use, and we cite the original sources throughout the manuscript.

Our experiments operate on the released recordings and do not attempt to identify participants or infer information beyond the prediction tasks defined by the datasets. We report aggregate results and do not release raw recordings or additional personally identifiable information. The datasets may contain population-, site-, or device-specific biases, and therefore the reported performance may not generalize to unseen populations or clinical settings.

Our TopoAttn is presented as a research method and should not be interpreted as a clinically validated diagnostic system or used for medical decision-making without further external validation. Responsibility for participant consent, institutional review, and de-identification rests with the original dataset collection and release procedures; we do not make additional claims beyond the documentation provided by the dataset authors.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our experiments throughout the paper and appendix. The main text describes the TopoAttn architecture, branch-specific topology construction, training protocol, and evaluation procedure. App. A.3 reports dataset statistics and preprocessing details, including the subject-independent five-fold split protocol and fixed split seed. App. A.4 provides the implementation details and hyperparameters for TopoAttn and all comparison methods, with dataset-specific settings summarized in Tab. 6. The complete ablation and topology-sensitivity results are reported in App. A.8 and A.9. All results are averaged over five independent training seeds, and checkpoints are selected according to validation F1 score, as described in the main text. The source code, including the training and evaluation scripts, topology-construction procedures, and configuration files, will be released upon acceptance.

REFERENCES

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, pp. 3–4, 2013.

Lena Biel, Ola Pettersson, Lennart Philipson, and Peter Wide. Ecg analysis: a new approach in human identification. *IEEE transactions on instrumentation and measurement*, 50(3):808–812, 2001.

Gyorgy Buzsaki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304 (5679):1926–1929, 2004.

Ling Chen, Ling-Jun Zou, and Li Tu. A clustering algorithm for multiple data streams based on spectral component similarity. *Information Sciences*, 183(1):35–47, 2012.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 10527–10545, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Sam Dixon, Lina Yao, and Robert Davidson. Modality aware contrastive learning for multimodal human activity recognition. *Concurrency and Computation: Practice and Experience*, 36(16): e8020, 2024.

Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996.

J Escudero, Daniel Abásolo, Roberto Hornero, Pedro Espino, and Miguel López. Analysis of electroencephalograms in alzheimer's disease patients with multiscale entropy. *Physiological measurement*, 27(11):1091–1106, 2006.

Golshan Fahimi, Seyed Mahmoud Tabatabaei, Elnaz Fahimi, and Hamid Rajebi. Index of theta/alpha ratio of the quantitative electroencephalogram in alzheimer's disease: a case-control study. *Acta Medica Iranica*, pp. 502–506, 2017.

Wei Fan, Jingru Fei, Dingyu Guo, Kun Yi, Xiaozhuang Song, Haolong Xiang, Hangting Ye, and Min Li. Towards multi-resolution spatiotemporal graph learning for medical time series classification. In *Proceedings of the ACM on Web Conference 2025*, pp. 5054–5064, 2025.

Yu Guoqi, Juncheng Wang, Chen Yang, Jing Qin, Angelica Aviles-Rivero, and Shujun Wang. Decentralized attention fails centralized signals: Rethinking transformers for medical time series. In *International Conference on Learning Representations*, volume 2026, pp. 108935–108955, 2026.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. Softs: Efficient multivariate time series forecasting with series-core fusion. *Advances in Neural Information Processing Systems*, 37: 64145–64175, 2024.

Steven A Israel, John M Irvine, Andrew Cheng, Mark D Wiederhold, and Brenda K Wiederhold. Ecg to identify individuals. *Pattern recognition*, 38(1):133–142, 2005.

Mohammad Kayyali, Ana Mincholé, Shuang Qian, Alistair Young, Devran Ugurlu, Elliot Fairweather, Steven Niederer, John Whitaker, Martin Bishop, and Pablo Lamata. Anatomical-electrical coupling of cardiac axes: Definitions and population variability for advancing personalised ecg interpretation. *PLoS computational biology*, 21(7):e1013161, 2025.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology: a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology. *Circulation*, 115(10):1306–1324, 2007.

Prabhat Kumar and S Suresh. Flaap: An open human activity recognition (har) dataset for learning and finding the associated activity patterns. *Procedia Computer Science*, 212:64–73, 2022.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ecg diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35: 9881–9893, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Sebastien Marcel and José del R Millán. Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 29(4):743–752, 2007.

Andreas Miltiadous, Katerina D Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G Tsalikakis, Pantelis Angelidis, Markos G Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, et al. A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg. *Data*, 8(6):95, 2023.

Benjamin Moseley and Joshua R Wang. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. *Journal of Machine Learning Research*, 24(1):1–36, 2023.

Fatma Murat, Ozal Yildirim, Muhammed Talo, Ulas Baran Baloglu, Yakup Demir, and U Rajendra Acharya. Application of deep learning techniques for heartbeats detection using ecg signals-analysis and review. *Computers in biology and medicine*, 120:103726, 2020.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Maria E Peltola, Markus Leitinger, Jonathan J Halford, Kollencheri Puthenveettil Vinayan, Katsuhiro Kobayashi, Ronit M Pressler, Ioana Mindruta, Luis Carlos Mayor, Leena Lauronen, and Sándor Beniczky. Routine and sleep eeg: minimum recording standards of the international federation of clinical neurophysiology and the international league against epilepsy. *Epilepsia*, 64 (3):602–618, 2023.

PhysioToolkit PhysioBank. Physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

Ronald Salloum and C-C Jay Kuo. Ecg-based biometrics using recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2062–2066. IEEE, 2017.

Saeid Sanei and Jonathon A. Chambers. *EEG Signal Processing*. John Wiley & Sons, 2013. ISBN 9781118691236. doi: 10.1002/9780470511923.

Alexios Savvides, Vasilis J Promponas, and Konstantinos Fokianos. Clustering of biological time series by cepstral coefficients based distances. *Pattern Recognition*, 41(7):2398–2412, 2008.

Magali T Schmidt, Paulo AM Kanda, Luis FH Basile, Helder Frederico da Silva Lopes, Regina Baratho, Jose LC Demario, Mario S Jorge, Antonio E Nardi, Sergio Machado, Jéssica N Ianof, et al. Index of alpha/theta ratio of the electroencephalogram: a new marker for alzheimer's disease. *Frontiers in aging neuroscience*, 5:60, 2013.

Margitta Seeck, Laurent Koessler, Thomas Bast, Frans Leijten, Christoph Michel, Christoph Baumgartner, Bin He, and Sándor Beniczky. The standardized eeg electrode array of the ifcn. *Clinical neurophysiology*, 128(10):2070–2077, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Siyi Tang, Jared A Dunnmon, Khaled Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel L Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. *arXiv preprint arXiv:2104.08336*, 2021.

Hanneke Van Dijk, Guido Van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde Van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (tdbrain) database. *Scientific data*, 9(1):333, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. *Advances in Neural Information Processing Systems*, 36:55694–55717, 2023.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37:36314–36341, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Jiexia Ye, Weiqi Zhang, Ziyue Li, Jia Li, and Fugee Tsung. Medspaformer: a transferable transformer with multi-granularity token sparsification for medical time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27791–27799, 2026.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting. In *International conference on artificial intelligence and statistics*, pp. 4222–4230. PMLR, 2024.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

A Appendix

This appendix provides supplementary methodological, implementation, and experimental details supporting the main paper, including the training augmentations, seven datasets and preprocessing pipelines with the subject-independent splitting protocol, additional TopoAttn implementation

details on topology-aware interactions, subject-guided token enhancement, adaptive gated fusion, dataset-specific hyperparameters, and domain-prior topology construction, as well as the settings of the comparison methods. We further report the full results on the two HAR datasets, subject-level evaluation, complete ablation results across all metrics, sensitivity analyses of channel and temporal topology construction and critical hyperparameters, the theoretical attention complexity analysis, and statistical significance tests against the strongest competing method.

A.1 Data Augmentation Banks

Following augmentation practices adopted in recent MedTS studies e.g., (Wang et al., 2024; Guoqi et al., 2026), we use a training-time augmentation bank comprising temporal flipping, channel shuffling, temporal masking, frequency masking, jittering, and dropout. We do not optimize the augmentation policy in this work; investigating task- and dataset-specific augmentation strategies is left for future study.

Temporal Flipping. We reverse the MedTS data along the temporal dimension. The probability of applying this augmentation is controlled by a parameter prob, with a default value of 0.5.

Channel Shuffling. We randomly shuffle the order of MedTS channels. The probability of applying channel shuffling is controlled by the parameter prob, also set by default to 0.5.

Temporal Masking. We randomly mask some timestamps across all channels. The proportion of timestamps masked is controlled by the parameter ratio, with a default value of 0.1.

Frequency Masking. First introduced in (Zhang et al., 2022) for contrastive learning, this method involves converting the MedTS data into the frequency domain, randomly masking some frequency bands, and then converting it back. The proportion of frequency bands masked is controlled by the parameter ratio, with a default value of 0.1.

Jittering. Random noise, ranging from 0 to 1, is added to the raw data. The intensity of the noise is adjusted by the parameter scale, which is set by default to 0.1.

Dropout. Similar to the dropout layer in neural networks, this method randomly drops some values. The proportion of values dropped is controlled by the parameter ratio, with a default setting of 0.1.

A.2 The Details of the Used Dataset

We evaluate TopoAttn on seven datasets spanning EEG, ECG, and inertial-sensor-based human activity recognition (HAR):

APAVA (Escudero et al., 2006) is an EEG dataset for binary Alzheimer's disease classification.

TDBrain (Van Dijk et al., 2022) is an EEG dataset used here for binary Parkinson's disease classification.

ADFTD (Miltiadous et al., 2023) is an EEG dataset with three classes: healthy controls, Frontotemporal Dementia, and Alzheimer's disease.

PTB (PhysioBank, 2000) is an ECG dataset used for binary Myocardial Infarction classification.

PTB-XL (Wagner et al., 2020) is a large-scale ECG dataset with five diagnostic classes.

FLAAP (Kumar & Suresh, 2022) is a smartphone-based HAR dataset containing accelerometer and gyroscope recordings for activity recognition.

UCI-HAR (Anguita et al., 2013) is a HAR dataset comprising accelerometer and gyroscope signals collected using waist-mounted smartphones.

Tab. 5 summarizes the main characteristics of the seven datasets, including the number of subjects, samples, classes, channels, timestamps, sampling rate, modality, and dataset size.

Table 5: Summary statistics of the seven datasets used in our experiments.

Dataset	#-Subject	#-Sample	#-Class	#-Channel	#-Timestamps	Sampling Rate	Modality	File Size
ADFTD	88	69,752	3	19	256	256 Hz	EEG	2.52 GB
APAVA	23	5,967	2	16	256	256 Hz	EEG	186 MB
TDBrain	72	6,240	2	33	256	256 Hz	EEG	571 MB
PTB	198	64,356	2	15	300	250 Hz	ECG	2.15 GB
PTB-XL	17,596	191,400	5	12	250	250 Hz	ECG	4.28 GB
FLAAP	8	13,123	10	6	100	100 Hz	IMU	60 MB
UCI-HAR	30	10,299	6	9	128	50 Hz	IMU	91 MB

A.3 Data Preprocessing

We follow the dataset preparation and preprocessing protocol of Medformer (Wang et al., 2024), using the processed datasets released with its official implementation.[1] For all MedTS datasets, we adopt subject-independent 5-fold cross-validation, where subjects and all their corresponding samples are partitioned into five mutually disjoint folds, with three folds used for training, one for validation, and one for testing. The fold partition is generated using a fixed random seed of 42.

APAVA Preprocessing. The **A**lzheimer's **P**atients' **R**elatives **A**ssociation of **Va**lladolid (APAVA) dataset[2] (Escudero et al., 2006) is a public EEG dataset comprising 23 subjects, including 12 patients with Alzheimer's disease and 11 healthy controls. On average, each subject has 30.0 ± 12.5 trials, with each trial containing 5 seconds of EEG recorded over 16 channels and 1,280 timestamps. Each trial is standardized and segmented into nine half-overlapping 1-second samples of 256 timestamps, yielding 5,967 samples in total. Each sample retains the subject ID of its originating subject.

TDBrain Preprocessing. The TDBrain dataset[3] (Van Dijk et al., 2022) is a permission-accessible EEG dataset containing recordings from 1,274 subjects over 33 channels. Each subject has eyes-open and eyes-closed recordings, and the full dataset includes 60 diagnostic labels, with subjects potentially associated with multiple conditions. Following the adopted preprocessing protocol, we use the eyes-closed recordings from a subset of 25 subjects with Parkinson's disease and 47 healthy controls. Each recording is segmented into non-overlapping 1-second samples of 256 timestamps, with incomplete segments discarded, resulting in 6,240 samples. Each sample retains the subject ID of its originating subject.

ADFTD Preprocessing. The **A**lzheimer's **D**isease and **F**ron**T**otemporal **D**ementia (ADFTD) dataset[4] (Miltiadous et al., 2023) is a public EEG dataset comprising 36 patients with Alzheimer's disease (AD), 23 patients with Frontotemporal Dementia (FTD), and 29 healthy controls (HC). Recordings contain 19 channels sampled at 500 Hz. Trial durations are approximately 13.5 minutes for AD subjects (5.1-21.3 minutes), 12.0 minutes for FTD subjects (7.9-16.9 minutes), and 13.8 minutes for HC subjects (12.5-16.5 minutes). A 0.5-45 Hz band-pass filter is applied, after which recordings are downsampled to 256 Hz and segmented into non-overlapping 1-second samples of 256 timestamps. Incomplete segments are discarded, yielding 69,752 samples.

PTB Preprocessing. The PTB dataset[5] (PhysioBank, 2000) is a public ECG dataset comprising recordings from 290 subjects over 15 channels, with eight diagnostic categories. We use a subset of 198 subjects comprising patients with myocardial infarction and healthy controls. Recordings are downsampled to 250 Hz and standardized. The signals are then segmented into individual heartbeats: R-peaks are identified, outlying R-peak intervals are removed, and each heartbeat is extracted relative to its R-peak. Shorter heartbeats are zero-padded to a common length determined by the maximum duration across channels. This procedure yields 64,356 samples.

[1] https://github.com/DL4mHealth/Medformer

[2] https://osf.io/jbysn/

[3] https://brainclinics.com/resources/

[4] https://openneuro.org/datasets/ds004504/versions/1.0.6

[5] https://physionet.org/content/ptbdb/1.0.0/

PTB-XL Preprocessing. The PTB-XL dataset[6] (Wagner et al., 2020) is a large public ECG dataset comprising 18,869 subjects, 12 channels, and five diagnostic classes. Because subjects may have multiple recordings, we discard subjects whose diagnoses vary across trials, leaving 17,596 subjects. PTB-XL provides 10-second recordings sampled at both 100 Hz and 500 Hz; we use the 500 Hz version, downsample it to 250 Hz, and standardize the signals. Each recording is then segmented into non-overlapping 1-second samples of 250 timestamps, with incomplete segments discarded, resulting in 191,400 samples.

A.4 Implementation Details

A.4.1 Implementation Details of TopoAttn

TopoAttn is optimized using Adam with cross-entropy loss, no weight decay, and a fixed learning rate of 1×10^{-4}. Early stopping is based on the validation macro-F1 score, with the dataset-specific maximum number of epochs $E_{\max}$ and patience PT reported in Tab. 6. All multi-head attention modules use eight heads. The feed-forward network in each encoder layer follows a $D \to 2D \to D$ structure with GELU activation. The subject-guided aggregation-redistribution core dimension is fixed to $D_c = D/4$, and the temporal-neighborhood radius is set to $r = 2$. The noisy-gated MoE contains four routed experts and one shared expert, with top-k routing using $k = 2$. Each topology group is associated with one D-dimensional learnable hub, such that G_{ch} and G_{te} also specify the numbers of channel and temporal hubs.

Dataset-specific architectural settings are summarized in Tab. 6. Here, D controls the representation capacity, PL determines the temporal granularity of patch tokens, and N_t and N_c specify the depths of the temporal and channel branches, respectively. We use $B = 256$ for PTB-XL to improve training efficiency on the larger dataset. All dataset-specific settings are selected using **validation performance only**; the test sets are not used for hyperparameter selection. Therefore, the selected hyperparameter values may differ across datasets.

For the FFT-derived channel topology, we use $F = 8$ logarithmically spaced frequency bins over $[f_{\min}, f_{\max}]$, keeping F fixed across all datasets to provide a compact spectral representation without dataset-specific tuning of the frequency resolution. Logarithmic spacing provides finer resolution at lower frequencies while retaining coverage of higher-frequency components, and is particularly consistent with the approximately logarithmic organization of electrophysiological oscillations (Buzsaki & Draguhn, 2004; Peltola et al., 2023). We set $f_{\min} = 0.5$ Hz and choose $f_{\max}$ according to the effective spectral bandwidth and sampling characteristics of each dataset. In particular, the 0.5-70 Hz range used for the EEG datasets is consistent with established clinical EEG recording recommendations (Peltola et al., 2023), while the ECG ranges follow standard considerations on diagnostically relevant ECG bandwidth and sampling requirements (Kligfield et al., 2007). For the inertial HAR datasets, the upper bound is constrained by the corresponding sampling bandwidth. The resulting dataset-specific ranges R_{FFT} are summarized in Tab. 6. The bin boundaries are defined as

$$b_i = f_{\min} \left(\frac{f_{\max}}{f_{\min}} \right)^{i/F}, \qquad i = 0, \ldots, F. \tag{3}$$

For each channel, the signal is centered before computing the real-valued Fourier spectrum. The spectral power is integrated within each frequency bin and normalized across bins to form an F-dimensional spectral-energy representation, which is subsequently used for channel-topology construction.

We adopt subject-independent 5-fold cross-validation with split seed 42 and repeat the complete 5-fold evaluation for five training seeds (42-46), yielding 25 runs per configuration for reporting mean and standard deviation. Sensitivity to the key topology parameters G_{ch}, G_{te}, and r is further examined on APAVA in App. A.9.3. All experiments are conducted on a server with eight NVIDIA Tesla V100 GPUs.

[6] https://physionet.org/content/ptb-xl/1.0.3/

Table 6: Dataset-specific implementation settings of TopoAttn. B, D, and PL denote the batch size, representation dimension, and patch length; N_t and N_c denote the numbers of temporal and channel encoder layers; and G_{ch} and G_{te} denote the corresponding topology group numbers. $E_{\max}$ and PT denote the maximum training epochs and early-stopping patience, while R_{FFT} specifies the FFT frequency range (Hz) used for channel topology construction.

Dataset	B	D	PL	N_t	N_c	G_{ch}	G_{te}	$E_{\max}$	PT	R_{FFT}
ADFTD	128	128	4	6	6	5	8	6	6	0.5-70
APAVA	128	256	2	6	6	5	4	40	20	0.5-70
TDBrain	128	128	2	3	1	8	8	60	60	0.5-45
PTB	128	256	6	3	3	8	5	80	40	0.5-100
PTB-XL	256	128	5	5	1	4	8	20	20	0.5-45
FLAAP	128	512	2	6	1	2	4	100	40	0.5-45
UCI-HAR	128	256	6	5	6	5	4	60	40	0.5-25

A.4.2 Architecture Details of TopoAttn

Topology-Aware Interaction Details. The four stages described in Sec. 3.3 are implemented as

$$\begin{aligned}
&\boldsymbol{U}^{\mathrm{tt}} = \mathrm{MHA}(\boldsymbol{U}; \boldsymbol{M}^{\mathrm{tt}}),\\
&\boldsymbol{R}^{\mathrm{agg}} = \mathrm{Attn}(\boldsymbol{R}^0\boldsymbol{W}_Q^h, \boldsymbol{U}^{\mathrm{tt}}\boldsymbol{W}_K^h, \boldsymbol{U}^{\mathrm{tt}}\boldsymbol{W}_V^h; (\boldsymbol{M}^{\mathrm{th}})^\top),\\
&\boldsymbol{R}^{\mathrm{g}} = \mathrm{MHA}(\boldsymbol{R}^{\mathrm{agg}}),\\
&\boldsymbol{U}^{\mathrm{out}} = \boldsymbol{U}^{\mathrm{tt}} + \mathrm{Attn}(\boldsymbol{U}^{\mathrm{tt}}\boldsymbol{W}_Q^r, \boldsymbol{R}^{\mathrm{g}}\boldsymbol{W}_K^r, \boldsymbol{R}^{\mathrm{g}}\boldsymbol{W}_V^r; \boldsymbol{M}^{\mathrm{th}})\boldsymbol{W}_O^r.
\end{aligned} \tag{4}$$

Here, $\boldsymbol{R}^0 \in \mathbb{R}^{G\times D}$ denotes the learnable group hubs and $\boldsymbol{W}$ the corresponding projection matrices. First, $\boldsymbol{M}^{\mathrm{tt}}$ restricts direct interaction to topology-related tokens, producing $\boldsymbol{U}^{\mathrm{tt}}$. Each hub then aggregates only the tokens assigned to its group through $(\boldsymbol{M}^{\mathrm{th}})^\top$, after which global multi-head attention exchanges information among the resulting hubs. Finally, the coordinated hub representations are redistributed to their associated tokens through $\boldsymbol{M}^{\mathrm{th}}$ and added residually to the local token features.

Subject-Guided Token Enhancement Details. We detail the aggregation-redistribution operator used in Sec. 3.4. Following SOFTS (Han et al., 2024), given $\boldsymbol{U} \in \mathbb{R}^{S\times D}$, the input tokens are projected into a compact core space with $D_c = D/4$, aggregated into a sample-level core, and redistributed through a residual projection:

$$\begin{aligned}
&\boldsymbol{Z} = \mathrm{Linear}_2(\mathrm{GELU}(\mathrm{Linear}_1(\boldsymbol{U}))) \in \mathbb{R}^{S\times D_c},\\
&\boldsymbol{c} = \sum_{i=1}^{S} \mathrm{Softmax}_S(\boldsymbol{Z})_i \odot \boldsymbol{Z}_i \in \mathbb{R}^{D_c},\\
&\boldsymbol{C} = \mathrm{Repeat}(\boldsymbol{c}, S) \in \mathbb{R}^{S\times D_c},\\
&\mathcal{R}(\boldsymbol{U}, \boldsymbol{C}) = \mathrm{LN}(\boldsymbol{U} + \mathrm{Linear}_4(\mathrm{GELU}(\mathrm{Linear}_3([\boldsymbol{U}\|\boldsymbol{C}])))).
\end{aligned} \tag{5}$$

Here, $\boldsymbol{Z}$ denotes the projected token features, $\boldsymbol{c}$ their aggregated core, and $\boldsymbol{C}$ repeats this core across the S token positions for redistribution; $\|$ and $\odot$ denote concatenation and element-wise multiplication. For subject-guided enhancement, we extend the initial aggregation from tokens within one sample to samples sharing the same subject ID, while preserving corresponding token positions, as defined in Eq. 5. The resulting subject core is constructed once per epoch and cached for redistribution to the corresponding samples. Subject IDs are thereafter no longer required: each encoder layer applies Eq. 5 within each sample before topology-aware sparse attention. The procedure is applied independently to the channel and temporal branches.

Adaptive Gated Fusion Details. The fusion module described in Sec. 3.5 first obtains branch-level representations by mean pooling and normalization of the final channel and temporal token features:

$$
\begin{aligned}
&\boldsymbol{h}^{\mathrm{ch}} = \mathrm{LN}_{\mathrm{ch}}\left(\frac{1}{C}\sum_{i=1}^{C}\boldsymbol{U}^{\mathrm{ch}}_{i,:}\right), \qquad \boldsymbol{h}^{\mathrm{te}} = \mathrm{LN}_{\mathrm{te}}\left(\frac{1}{P}\sum_{i=1}^{P}\boldsymbol{U}^{\mathrm{te}}_{i,:}\right),\\
&\boldsymbol{z} = [\boldsymbol{h}^{\mathrm{ch}} \| \boldsymbol{h}^{\mathrm{te}} \| (\boldsymbol{h}^{\mathrm{ch}} \odot \boldsymbol{h}^{\mathrm{te}}) \| |\boldsymbol{h}^{\mathrm{ch}} - \boldsymbol{h}^{\mathrm{te}}|] \in \mathbb{R}^{4D},\\
&\boldsymbol{g} = \sigma(\mathrm{Linear}_{g,2}(\mathrm{GELU}(\mathrm{Linear}_{g,1}(\boldsymbol{z})))) \in (0,1)^{D},\\
&\boldsymbol{h}^{\mathrm{gate}} = \boldsymbol{g} \odot \boldsymbol{h}^{\mathrm{ch}} + (1-\boldsymbol{g}) \odot \boldsymbol{h}^{\mathrm{te}},\\
&\boldsymbol{h}^{\mathrm{fuse}} = \mathrm{LN}\big(\boldsymbol{h}^{\mathrm{gate}} + \mathrm{Linear}_{r,2}\big(\mathrm{GELU}\big(\mathrm{Linear}_{r,1}([\boldsymbol{h}^{\mathrm{ch}} \| \boldsymbol{h}^{\mathrm{te}}])\big)\big)\big).
\end{aligned}
\tag{6}
$$

The product and absolute-difference terms in $\boldsymbol{z}$ encode cross-branch agreement and discrepancy, respectively, while $\boldsymbol{g}$ provides feature-wise adaptive weighting. The residual path from $[\boldsymbol{h}^{\mathrm{ch}} \| \boldsymbol{h}^{\mathrm{te}}]$ preserves complementary information beyond gated averaging. Finally, the fused representation is projected to class scores as $\boldsymbol{o} = \mathrm{Linear}_p(\boldsymbol{h}^{\mathrm{fuse}}) \in \mathbb{R}^K$, with the predicted class given by $\hat{y} = \arg\max_k o_k$.

A.4.3 Implementation Details of SOTA

We evaluate TopoAttn against 11 Transformer-based methods using the unified experimental framework provided by Medformer (Wang et al., 2024), enabling consistent preprocessing, training, and evaluation across methods. The compared methods are Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), Informer (Zhou et al., 2021), iTransformer (Liu et al., 2024), MTST (Zhang et al., 2024), Nonformer (Liu et al., 2022), PatchTST (Nie et al., 2022), Reformer (Kitaev et al., 2020), Medformer (Wang et al., 2024), TeCh (Guoqi et al., 2026), and the vanilla Transformer (Vaswani et al., 2017).

Autoformer. Autoformer (Wu et al., 2021) replaces standard self-attention with an auto-correlation mechanism for modeling periodic dependencies in time-series forecasting. It additionally employs series decomposition to separate trend and seasonal components.

FEDformer. FEDformer (Zhou et al., 2022) incorporates frequency-domain representations through Fourier-based frequency-enhanced blocks and attention, together with time-series decomposition for modeling complex temporal patterns.

Informer. Informer (Zhou et al., 2021) introduces ProbSparse self-attention to reduce the computational and memory costs of long-sequence modeling and employs a generative decoding strategy for time-series forecasting.

iTransformer. iTransformer (Liu et al., 2024) inverts the conventional time-series tokenization scheme by representing each variable through its complete temporal sequence as a token and applying attention across variable tokens.

MTST. MTST (Zhang et al., 2024) models time series at multiple temporal resolutions by processing patch tokens with different patch lengths in parallel, enabling the model to capture temporal patterns at different scales.

Nonformer. Nonformer (Liu et al., 2022) addresses non-stationarity in time series through destationary attention together with normalization and denormalization mechanisms designed to mitigate over-stationarization.

PatchTST. PatchTST (Nie et al., 2022) divides each univariate channel into temporal patches and represents each patch as a token, enabling Transformer-based modeling over longer temporal contexts.

Reformer. Reformer (Kitaev et al., 2020) replaces standard dot-product attention with locality-sensitive hashing attention and employs reversible residual layers to reduce the computational and memory requirements of Transformer models.

Transformer. The vanilla Transformer (Vaswani et al., 2017) employs standard dense self-attention. In our time-series setting, each timestamp across all channels is embedded as an input token.

Medformer. Medformer (Wang et al., 2024) is a Transformer architecture designed for medical time-series classification. It constructs cross-channel patch representations at multiple temporal granularities to model temporal and inter-channel dependencies at different scales.

TeCh. TeCh (Guoqi et al., 2026) is a MedTS classification framework that replaces pairwise token attention with an aggregation-redistribution mechanism in which information is aggregated into a central representation and subsequently redistributed to individual tokens. It applies complementary temporal and channel tokenization to model temporal and inter-channel dependencies.

A.5 Domain-Prior Channel Topology

As a domain-informed alternative to the proposed FFT-derived channel topology, we construct a fixed prior topology for the channel-topology sensitivity analysis in Tab. 4. This comparison evaluates whether the performance of TopoAttn depends specifically on its data-driven spectral grouping or can also be achieved using established anatomical, physiological, or acquisition-based channel organization.

For EEG, electrodes are grouped by scalp region following the standard 10-20/10-10 organization (Seeck et al., 2017). ADFTD uses five groups: frontal {Fp1, Fp2, F7, F3, Fz, F4, F8}, central {C3, Cz, C4}, temporal {T3, T4, T5, T6}, parietal {P3, Pz, P4}, and occipital {O1, O2}. APAVA follows the same five-region partition using its available channels: frontal {Fp1, Fp2, F7, F3, F4, F8}, central {C3, C4}, temporal {T3, T4, T5, T6}, parietal {P3, P4}, and occipital {O1, O2}. TDBrain uses six groups: frontal {Fp1, Fp2, F7, F3, Fz, F4, F8, FC3, FCz, FC4}, central {C3, Cz, C4, CP3, CPz, CP4}, temporal {T7, T8}, parietal {P7, P3, Pz, P4, P8}, occipital {O1, Oz, O2}, and auxiliary {VPVA, VNVB, HPHL, HNHR, Erbs, OrbOcc, Mass}.

For ECG, we follow the conventional lead organization (Kligfield et al., 2007). PTB-XL uses two groups: limb leads {I, II, III, aVR, aVL, aVF} and precordial leads {V1-V6}, while PTB additionally includes a third vectorcardiographic group {VX, VY, VZ}.

For HAR, channels are grouped according to sensor modality and acquisition semantics (Kumar & Suresh, 2022; Dixon et al., 2024). FLAAP uses two groups: accelerometer {AccX, AccY, AccZ} and gyroscope {GyroX, GyroY, GyroZ}. UCI-HAR uses three groups: body acceleration {BodyAccX, BodyAccY, BodyAccZ}, body gyroscope {BodyGyroX, BodyGyroY, BodyGyroZ}, and total acceleration {TotalAccX, TotalAccY, TotalAccZ}. Channels within each prior-defined group interact directly and share one hub, yielding a fixed domain-informed topology for comparison with the proposed data-driven spectral topology.

A.6 Theoretical Attention Complexity Analysis

We analyze the theoretical attention cost of TopoAttn in terms of the query-key interactions defined by its sparse attention pattern, omitting the common feature-dimension factor. Let C and P denote the numbers of channel and temporal tokens, and G_{ch} and G_{te} the corresponding numbers of topology groups. Dense attention over the two branches requires $\mathcal{O}(C^2 + P^2)$ interactions.

For the channel branch, direct attention is restricted within groups. If group g contains C_g channels, with $\sum_{g=1}^{G_{\mathrm{ch}}} C_g = C$, its token-token cost becomes $\sum_g C_g^2 < C^2$ for multiple non-empty groups. When group sizes are comparable, $C_g \approx C/G_{\mathrm{ch}}$, yielding

$$\sum_{g=1}^{G_{\mathrm{ch}}} C_g^2 \approx G_{\mathrm{ch}} \left(\frac{C}{G_{\mathrm{ch}}} \right)^2 = \frac{C^2}{G_{\mathrm{ch}}}, \tag{7}$$

so the channel token-token term is reduced from $\mathcal{O}(C^2)$ to approximately $\mathcal{O}(C^2/G_{\mathrm{ch}})$. Token-hub communication and hub-hub attention further contribute $\mathcal{O}(C)$ and $\mathcal{O}(G_{\mathrm{ch}}^2)$.

For the temporal branch, each patch attends to at most $2r + 1$ neighboring patches, reducing the token-token term from $\mathcal{O}(P^2)$ to $\mathcal{O}(Pr)$. In our settings, r is small (typically 1-3) and satisfies $r \ll P$; token-hub and hub-hub operations additionally contribute $\mathcal{O}(P)$ and $\mathcal{O}(G_{\mathrm{te}}^2)$.

The general TopoAttn attention cost and its balanced-group approximation are therefore

$$\begin{aligned}\mathcal{T}_{\text{Topo}} &= \mathcal{O}\left(\sum_{g=1}^{G_{\text{ch}}} C_g^2 + Pr + C + P + G_{\text{ch}}^2 + G_{\text{te}}^2\right), \\ &\approx \mathcal{O}\left(\frac{C^2}{G_{\text{ch}}} + Pr + C + P + G_{\text{ch}}^2 + G_{\text{te}}^2\right),\end{aligned} \tag{8}$$

compared with $\mathcal{T}_{\text{dense}} = \mathcal{O}(C^2 + P^2)$. In the considered MedTS settings, $C \ll P$, while G_{ch} and G_{te} are small (typically 1-8). For a fixed sensor/lead configuration and fixed topology hyperparameters, C, G_{ch}, G_{te}, and r are constant with respect to P. Hence,

$$\mathcal{T}_{\text{Topo}} = \mathcal{O}(Pr + P) = \mathcal{O}(P), \qquad \mathcal{T}_{\text{dense}} = \mathcal{O}(P^2). \tag{9}$$

Therefore, with respect to increasing temporal resolution, TopoAttn reduces the dominant attention cost from quadratic to linear scaling while retaining complementary channel and temporal representations through within-group interactions and compact hub-mediated coordination.

A.7 Results on two Human Activity Recognition Datasets

Tab. 7 evaluates the generalization ability of TopoAttn beyond medical time series on two human activity recognition datasets. TopoAttn achieves the best performance across all six metrics on both FLAAP and UCI-HAR, demonstrating consistent improvements in classification performance and ranking quality. On FLAAP, TopoAttn outperforms the second-best method, TeCh, by 0.84%, 0.62%, 0.71%, and 0.83% in accuracy, precision, recall, and F1 score, respectively. It also improves AUROC and AUPRC by 0.15% and 1.42%, respectively. On UCI-HAR, TopoAttn surpasses the second-best method, iTransformer, by 0.56% in accuracy and approximately 0.61% in precision, recall, and F1 score. These results indicate that TopoAttn can effectively exploit inter-channel relationships and remains competitive when applied to general multivariate time-series classification tasks beyond the medical domain.

Table 7: Results on two HAR datasets. To evaluate the performance of our method on general time series, we test it on two HAR datasets: FLAAP and UCI-HAR, which exhibit potential channel correlations inherently. The best result is marked in **bold red**, and the second best is marked in *blue*.

Dataset	Model	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
FLAAP (10 Classes)	**Autoformer**	40.27±0.11	41.40±1.49	39.65±0.07	34.51±0.10	74.81±0.20	37.78±0.36
	FEDformer	65.69±9.54	65.95±8.61	65.51±9.31	64.97±9.94	92.63±4.59	69.10±10.58
	Informer	73.54±0.24	73.96±0.38	73.39±0.21	73.30±0.29	95.85±0.12	77.86±0.50
	iTransformer	75.67±0.20	75.83±0.29	75.49±0.18	75.44±0.22	96.49±0.06	80.97±0.21
	MTST	73.04±0.20	73.64±0.16	73.06±0.26	73.00±0.23	95.12±0.10	76.02±0.38
	Nonformer	75.57±0.40	76.35±0.35	75.58±0.44	75.49±0.39	96.45±0.06	80.26±0.35
	PatchTST	56.23±0.07	56.39±0.14	55.68±0.06	55.76±0.10	88.80±0.10	59.27±0.27
	Reformer	71.45±0.60	71.81±0.41	71.19±0.74	71.09±0.65	95.22±0.14	74.66±0.50
	Transformer	76.40±0.66	77.02±0.55	76.52±0.60	76.36±0.63	96.65±0.08	81.31±0.53
	Medformer	75.90±0.46	76.52±0.39	75.89±0.47	75.63±0.47	97.08±0.11	81.94±0.63
	TeCh	*80.75±0.15*	*80.70±0.20*	*80.72±0.25*	*80.52±0.21*	*97.69±0.03*	*86.46±0.17*
	TopoAttn	**81.43±0.30**	**81.20±0.26**	**81.29±0.27**	**81.19±0.27**	**97.84±0.02**	**87.69±0.19**
UCI-HAR (6 Classes)	**Autoformer**	49.84±1.20	61.67±2.67	53.68±1.06	42.59±1.69	89.22±0.48	69.76±0.98
	FEDformer	89.58±12.20	91.11±9.69	90.27±11.21	89.56±12.63	97.97±3.49	93.30±10.85
	Informer	96.30±0.50	96.47±0.47	96.45±0.49	96.45±0.49	99.77±0.04	98.99±0.18
	iTransformer	*97.16±0.17*	*97.27±0.17*	*97.27±0.17*	*97.26±0.17*	*99.87±0.01*	*99.44±0.02*
	MTST	96.37±0.32	96.43±0.30	96.45±0.32	96.42±0.31	99.80±0.03	99.11±0.13
	Nonformer	96.55±0.26	96.77±0.24	96.76±0.24	96.75±0.24	99.75±0.02	98.84±0.09
	PatchTST	88.90±0.12	89.65±0.13	89.61±0.11	89.58±0.11	98.66±0.03	94.63±0.13
	Reformer	95.82±0.26	96.04±0.26	96.05±0.24	96.01±0.25	99.71±0.04	98.67±0.18
	Transformer	96.33±0.19	96.59±0.20	96.55±0.18	96.54±0.18	99.74±0.02	98.84±0.10
	Medformer	95.56±0.11	95.87±0.10	95.85±0.11	95.84±0.10	99.60±0.02	98.18±0.10
	TeCh	96.38±0.18	96.60±0.17	96.61±0.17	96.58±0.17	99.74±0.02	98.85±0.08
	TopoAttn	**97.70±0.20**	**97.86±0.20**	**97.86±0.18**	**97.85±0.19**	**99.89±0.02**	**99.49±0.07**

A.8 Full Ablation Results

For concision, the main text reports ablation results on five representative datasets, while the complete results across all seven datasets are provided in Tab. 8.

Tab. 8 reports the complete ablation results across all seven datasets. The full TopoAttn model achieves the best accuracy and F1 score on every dataset, indicating that the final performance re-

Table 8: Full Ablation Study. We report Accuracy and F1-Score across all seven datasets. The full TopoAttn model is highlighted in **bold red**.

Variants	ADFTD		APAVA		TDBrain		PTB		PTB-XL		FLAAP		UCI-HAR	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
w/o Topology	50.66±1.50	47.60±1.62	84.05±0.61	78.88±0.66	84.91±0.20	82.68±1.12	91.99±0.59	85.18±0.87	73.96±0.09	62.39±0.15	80.55±0.96	80.36±0.92	97.56±0.03	97.72±0.02
w/o Channel Topology	52.78±0.51	49.42±1.22	84.04±0.71	79.96±0.98	85.75±0.27	83.51±1.12	91.85±0.43	84.87±0.63	74.89±0.07	63.52±0.07	80.84±0.26	80.60±0.23	97.65±0.07	97.80±0.07
w/o Temporal Topology	50.35±1.88	47.44±1.94	82.59±1.10	77.80±1.23	85.08±0.45	82.76±1.51	91.94±0.64	85.30±1.13	73.96±0.08	62.45±0.19	81.09±0.97	80.92±0.92	97.41±0.32	97.56±0.30
w/o MoE	52.51±1.86	49.21±1.62	81.27±1.69	77.24±1.85	85.68±0.82	83.87±1.34	92.23±0.60	85.74±1.01	73.51±0.23	62.98±0.12	79.92±0.30	79.71±0.31	97.09±0.14	97.26±0.13
w/o Sub	52.91±0.62	48.86±0.52	81.32±3.37	77.16±3.85	85.55±0.84	82.80±1.89	90.95±0.50	83.29±0.91	74.74±0.11	63.32±0.17	81.12±0.48	80.92±0.47	97.55±0.37	97.70±0.35
TopoAttn	**54.09**±0.99	**50.42**±1.03	**84.08**±0.24	**80.13**±0.27	**86.14**±0.36	**84.04**±1.05	**92.72**±0.71	**86.45**±1.04	**74.93**±0.07	**63.61**±0.11	**81.43**±0.30	**81.19**±0.27	**97.70**±0.20	**97.85**±0.19

sults from the joint contribution of topology-aware attention, branch-specific topology construction, MoE-based temporal embedding, and subject-guided enhancement.

Replacing topology-aware sparse attention with dense self-attention (**w/o Topology**) reduces the average accuracy and F1 score across the seven datasets from 81.58 and 77.67 to 80.53 and 76.40, respectively. This corresponds to relative reductions of 1.30% in accuracy and 1.63% in F1. The degradation is particularly pronounced on ADFTD, where F1 decreases by 5.59%, and on TDBrain, PTB, and PTB-XL, where relative reductions of more than 1% in F1 are also observed. Because this variant retains the same attention projections and changes primarily the interaction pattern from structured sparse attention to dense attention, the comparison supports the usefulness of structured sparse communication.

The two branch-specific topology ablations further demonstrate that the channel and temporal topologies contribute complementary information. Removing the channel topology while retaining the temporal topology (**w/o Channel Topology**) reduces the average F1 score by 0.74% across all seven datasets. The largest dataset-specific reductions occur on PTB and ADFTD, with decreases of 1.83% and 1.98%, respectively. Removing the temporal topology (**w/o Temporal Topology**) produces a larger average F1 reduction of 1.74% across the seven datasets and a reduction of 2.44% when considering only the five MedTS datasets. The largest decreases are observed on ADFTD and APAVA, where F1 drops by 5.91% and 2.91%, respectively. This difference suggests that the temporal topology is particularly important for preserving local temporal interactions, while the channel topology provides complementary benefits for modeling cross-channel relationships.

The MoE-based temporal embedding also makes a substantial contribution. Removing the MoE module (**w/o MoE**) lowers the average F1 score by 1.41% across the seven datasets. Its effect is especially clear on APAVA and FLAAP, where F1 decreases by 3.61% and 1.82%, respectively. This pattern is consistent with the role of the MoE module in providing different embedding transformations for heterogeneous temporal patches.

Removing subject-guided token enhancement (**w/o Sub**) reduces the average F1 score by 1.77% across the seven datasets. The largest reductions are observed on PTB and APAVA, with decreases of 3.66% and 3.71%, respectively. These results suggest that subject-level contextual information is particularly useful for datasets with substantial inter-subject variability. At the same time, the smaller changes on FLAAP and UCI-HAR indicate that the contribution of this module depends on the dataset and the degree of subject-specific variation.

Overall, the ablation results show that no single component fully explains the performance of TopoAttn. The topology-aware attention mechanism provides structured local and cross-group communication, the branch-specific topologies encode complementary channel and temporal dependencies, the MoE module models heterogeneous temporal patterns, and subject-guided enhancement supplies additional subject-level context. Their combination consistently yields the strongest performance across both medical and general time-series datasets.

A.9 Full Sensitivity Analysis Results

A.9.1 Sensitivity Analysis: Channel Topology

Tab. 9 presents the complete comparison between the proposed FFT-derived channel topology and two alternatives: a domain-prior topology based on predefined channel groupings and a randomly

grouped topology. TopoAttn achieves the best mean performance in all 42 dataset-metric comparisons across the seven datasets, including accuracy, precision, recall, F1 score, AUROC, and AUPRC.

The advantage of the FFT-derived topology is particularly clear on the MedTS datasets. Averaged across the five MedTS datasets, TopoAttn achieves an F1 score of 72.93, compared with 71.94 for the domain-prior topology and 71.45 for the random topology. Thus, it provides relative improvements of 1.37% and 2.07% over the two alternatives, respectively. The improvements are especially pronounced on APAVA, where TopoAttn improves F1 by 2.96% and 3.09% over the domain-prior and random topologies, respectively. It also consistently improves F1 on ADFTD, TDBrain, PTB, and PTB-XL. The gains extend beyond F1: TopoAttn is superior in every reported metric on each of the five MedTS datasets.

The comparison with the domain-prior topology is important because it evaluates whether manually specified anatomical or sensor-based groupings are sufficient for channel interaction modeling. Although such prior groupings can provide meaningful structure, they do not consistently identify channels with similar signal dynamics. In contrast, TopoAttn constructs channel groups from training-set FFT-based spectral similarities, allowing channels with related frequency characteristics to interact directly within the same group, while the hub pathway provides communication across groups. The results therefore suggest that data-derived spectral relationships provide a more informative organization of channel interactions than fixed medical groupings.

The random topology provides a complementary control for testing whether structured grouping is beneficial beyond imposing a grouping pattern alone. Although random grouping is competitive on some datasets and metrics, TopoAttn remains consistently superior. This indicates that the performance gains do not arise merely from restricting attention to groups, but from forming groups according to meaningful channel relationships. The results on the HAR datasets support the same conclusion beyond the medical domain. TopoAttn improves over the domain-prior and random topologies in all six metrics on both FLAAP and UCI-HAR, although the margins are smaller on UCI-HAR, where all methods already achieve near-saturated performance.

Overall, these results provide evidence that the channel topology should reflect the statistical relationships among channels. In TopoAttn, FFT-derived spectral similarity supplies a data-driven channel prior, while the hierarchical hub mechanism preserves communication between otherwise separate groups. Together, these design choices yield more reliable channel interactions than either predefined medical grouping or arbitrary random grouping.

A.9.2 Sensitivity Analysis: Temporal Topology

Tab. 10 presents the complete comparison between TopoAttn and its variant with a randomly constructed temporal topology. TopoAttn achieves the best result in all 42 dataset-metric comparisons across the seven datasets. In particular, it consistently improves accuracy, precision, recall, F1 score, AUROC, and AUPRC, indicating that its advantage is not restricted to a single evaluation criterion.

On the five MedTS datasets, the improvements are particularly pronounced on ADFTD, where TopoAttn surpasses the random variant by 9.14%, 7.96%, 8.72%, 8.50%, 6.58%, and 10.52% in accuracy, precision, recall, F1 score, AUROC, and AUPRC, respectively. APAVA shows the largest AUPRC improvement, with a gain of 8.91%. Consistent improvements are also observed on TDBrain, PTB, and PTB-XL. Averaged across these five datasets, TopoAttn improves accuracy, precision, recall, F1 score, AUROC, and AUPRC by approximately 2.50%, 2.12%, 2.59%, 3.14%, 1.98%, and 4.08%, respectively.

The advantage extends to the HAR datasets, although the margins are smaller. On FLAAP and UCI-HAR, TopoAttn improves all six metrics over the random variant, with average relative gains of approximately 0.39% in accuracy, 0.36% in precision, 0.36% in recall, 0.37% in F1 score, 0.05% in AUROC, and 0.27% in AUPRC. Overall, these results suggest that an order-aware temporal topology is more effective than an unstructured random topology when the remaining TopoAttn architecture and training configuration are held fixed.

Table 9: Full results for sensitivity analysis of domain-prior and random channel topology construction. We compare TopoAttn with its variants using channel topologies constructed from domain prior knowledge and random grouping across all seven datasets. The best is highlighted in **bold red**.

Dataset	Setting	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
ADFTD	Prior	53.50±0.71	51.52±1.04	50.66±0.82	49.91±0.79	68.98±0.55	53.65±0.58
	Random	52.96±1.07	50.79±1.55	50.14±1.34	49.48±1.36	68.52±1.20	52.92±1.35
	TopoAttn	**54.09±0.99**	**51.78±1.25**	**51.01±1.05**	**50.42±1.03**	**69.35±0.87**	**54.10±1.01**
APAVA	Prior	82.13±0.59	78.86±0.25	82.85±1.12	77.83±0.83	90.14±0.51	88.17±1.14
	Random	82.03±1.08	78.61±0.87	82.47±1.17	77.73±1.37	90.06±0.87	87.97±1.37
	TopoAttn	**84.08±0.24**	**79.92±0.23**	**85.20±0.47**	**80.13±0.27**	**91.30±0.23**	**89.46±0.70**
TDBrain	Prior	85.10±0.73	85.09±0.64	83.52±0.95	83.40±0.80	93.07±0.30	92.20±0.46
	Random	85.33±0.93	86.35±1.01	82.34±1.33	82.69±1.26	93.30±0.49	92.64±0.58
	TopoAttn	**86.14±0.36**	**86.99±0.36**	**83.67±1.22**	**84.04±1.05**	**93.90±0.32**	**93.27±0.43**
PTB	Prior	92.26±0.73	86.07±1.22	86.78±1.77	85.77±1.47	96.03±1.24	92.47±1.35
	Random	91.65±0.95	85.06±1.62	85.48±1.59	84.58±1.53	96.04±0.80	92.34±1.24
	TopoAttn	**92.72±0.71**	**86.89±1.42**	**87.37±0.75**	**86.45±1.04**	**96.05±0.28**	**92.82±1.06**
PTB-XL	Prior	73.42±0.12	64.88±0.17	61.77±0.14	62.80±0.13	90.24±0.08	67.80±0.15
	Random	73.40±0.21	64.75±0.35	61.76±0.34	62.76±0.20	90.15±0.16	67.66±0.25
	TopoAttn	**74.93±0.07**	**67.12±0.23**	**62.05±0.09**	**63.61±0.11**	**91.01±0.05**	**69.51±0.14**
FLAAP	Prior	79.70±0.35	79.62±0.29	79.59±0.35	79.50±0.33	97.48±0.08	85.83±0.42
	Random	80.97±0.38	80.80±0.39	80.84±0.40	80.76±0.40	97.71±0.07	87.20±0.31
	TopoAttn	**81.43±0.30**	**81.20±0.26**	**81.29±0.27**	**81.19±0.27**	**97.84±0.02**	**87.69±0.19**
UCI-HAR	Prior	97.50±0.09	97.67±0.09	97.66±0.09	97.65±0.09	99.86±0.01	99.36±0.04
	Random	97.31±0.14	97.49±0.12	97.50±0.12	97.48±0.13	99.86±0.01	99.37±0.06
	TopoAttn	**97.70±0.20**	**97.86±0.20**	**97.86±0.18**	**97.85±0.19**	**99.89±0.02**	**99.49±0.07**

Table 10: Full results for sensitivity analysis of random temporal topology construction. We compare TopoAttn with its variant using randomly constructed temporal topology across all seven datasets. The best is highlighted in **bold red**.

Dataset	Setting	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
ADFTD	Random	49.56±1.47	47.96±1.20	46.92±1.44	46.47±1.45	65.07±1.43	48.95±1.32
	TopoAttn	**54.09±0.99**	**51.78±1.25**	**51.01±1.05**	**50.42±1.03**	**69.35±0.87**	**54.10±1.01**
APAVA	Random	82.35±0.72	79.87±0.55	83.83±0.92	77.54±0.89	89.08±0.78	82.14±1.05
	TopoAttn	**84.08±0.24**	**79.92±0.23**	**85.20±0.47**	**80.13±0.27**	**91.30±0.23**	**89.46±0.70**
TDBrain	Random	84.74±0.70	85.25±0.51	82.02±1.71	82.14±1.65	92.99±0.57	92.26±0.66
	TopoAttn	**86.14±0.36**	**86.99±0.36**	**83.67±1.22**	**84.04±1.05**	**93.90±0.32**	**93.27±0.43**
PTB	Random	91.91±0.74	85.72±1.19	86.86±1.25	85.33±1.21	95.76±0.35	92.40±0.83
	TopoAttn	**92.72±0.71**	**86.89±1.42**	**87.37±0.75**	**86.45±1.04**	**96.05±0.28**	**92.82±1.06**
PTB-XL	Random	73.85±0.10	66.15±0.21	60.36±0.10	62.06±0.12	90.14±0.06	67.76±0.14
	TopoAttn	**74.93±0.07**	**67.12±0.23**	**62.05±0.09**	**63.61±0.11**	**91.01±0.05**	**69.51±0.14**
FLAAP	Random	81.05±0.29	80.86±0.27	80.95±0.29	80.84±0.28	97.76±0.04	87.24±0.19
	TopoAttn	**81.43±0.30**	**81.20±0.26**	**81.29±0.27**	**81.19±0.27**	**97.84±0.02**	**87.69±0.19**
UCI-HAR	Random	97.38±0.06	97.55±0.06	97.55±0.06	97.54±0.06	99.87±0.01	99.43±0.03
	TopoAttn	**97.70±0.20**	**97.86±0.20**	**97.86±0.18**	**97.85±0.19**	**99.89±0.02**	**99.49±0.07**

A.9.3 Hyperparameter Sensitivity Results

Although the dataset-specific topology group numbers G_{ch} and G_{te} are selected based on validation performance within the training folds, and $r = 2$ is used as the default temporal-neighborhood radius, we additionally evaluate alternative values of all three parameters on APAVA to assess whether TopoAttn's performance depends on a narrowly tuned topology configuration.

We examine the sensitivity of TopoAttn to three critical hyperparameters, G_{ch}, G_{te}, and r, on APAVA. These parameters directly govern topology construction by determining the numbers of channel and temporal groups and the scope of local temporal interactions. We vary one parameter at a time while keeping all remaining settings fixed to the main configuration reported in Tab. 6. As shown in Tab. 11, the default setting ($G_{\mathrm{ch}} = 5$, $G_{\mathrm{te}} = 4$, $r = 2$) provides the strongest overall per-

formance: $G_{\mathrm{ch}} = 5$ and $r = 2$ achieve the best results on five of six metrics, while $G_{\mathrm{te}} = 4$ achieves the best results across all six metrics. Performance varies moderately around these settings, indicating that the topology configuration affects classification quality while the selected values provide a favorable balance for APAVA.

Importantly, TopoAttn remains strong even when these topology hyperparameters are moved away from their selected values. Compared with TeCh, the best SOTA on APAVA in Tab. 2, TopoAttn achieves higher performance in 71 of the 72 metric comparisons across the 12 non-default configurations. The only exception is Precision at $G_{\mathrm{ch}} = 6$ (77.62% vs. 78.11%). Moreover, the lowest F1 score observed across all evaluated configurations is 76.67%, which remains above TeCh's 74.66%. These results show that the advantage of TopoAttn is not confined to the selected topology configuration and remains robust across reasonable variations in G_{ch}, G_{te}, and r.

Table 11: Sensitivity analysis of key topology hyperparameters on APAVA. We vary one parameter at a time while keeping all others fixed to the default configuration ($G_{\mathrm{ch}} = 5$, $G_{\mathrm{te}} = 4$, $r = 2$). Results are reported as mean±std over 25 runs (5 folds × 5 seeds). † denotes the default setting. The best is highlighted in **bold red**.

Parameter	Value	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
G_{ch}	3	83.47±1.26	**80.04±0.61**	84.01±1.92	78.93±1.85	91.17±0.49	88.94±0.78
	4	83.24±0.54	79.60±0.48	83.55±0.68	78.95±0.63	90.84±0.93	88.50±1.30
	5†	**84.08±0.24**	79.92±0.23	**85.20±0.47**	**80.13±0.27**	**91.30±0.23**	**89.46±0.70**
	6	81.20±1.78	77.62±1.29	81.76±1.89	76.67±2.19	89.46±0.83	87.20±1.29
	7	82.54±1.31	78.77±1.00	83.79±1.40	78.46±1.38	90.06±1.09	88.05±0.88
G_{te}	2	82.48±1.72	78.79±1.22	83.24±2.16	78.24±2.27	89.73±0.27	87.84±0.50
	3	82.30±2.41	78.74±1.61	83.17±2.24	78.10±2.76	89.30±2.40	87.32±2.32
	4†	**84.08±0.24**	**79.92±0.23**	**85.20±0.47**	**80.13±0.27**	**91.30±0.23**	**89.46±0.70**
	5	81.47±3.12	78.35±2.13	82.68±2.73	76.87±3.68	89.41±2.41	86.99±3.23
	6	83.08±1.98	79.12±1.84	83.70±2.20	79.04±2.46	90.32±0.95	88.56±1.40
r	1	83.73±0.47	**79.96±0.48**	84.55±0.08	79.45±0.50	90.67±0.33	88.41±0.77
	2†	**84.08±0.24**	79.92±0.23	**85.20±0.47**	**80.13±0.27**	**91.30±0.23**	**89.46±0.70**
	3	82.62±0.98	78.96±0.73	83.53±1.13	78.27±1.27	90.69±0.69	88.45±1.35
	4	81.48±1.49	78.17±1.22	81.53±1.70	76.67±1.92	90.49±0.64	87.77±1.39
	5	83.30±1.41	79.11±1.03	83.40±1.68	78.80±1.68	91.19±0.47	88.40±0.93

A.9.4 Subject-Level Evaluation

To complement the sample-level evaluation in the main paper, we further assess diagnostic performance at the subject level under the same subject-independent 5-fold protocol. For each held-out subject, we average the predicted class probabilities across all of its samples and assign the class with the highest aggregated probability. Each training seed thus yields one complete out-of-fold prediction for each subject, and we report the mean and standard deviation over five training seeds. This evaluation directly examines whether the evidence captured from individual time-series samples remains consistent when making a single prediction for each unseen subject.

As shown in Tab. 12, TopoAttn achieves the best performance on 23 of the 30 evaluation metrics. TopoAttn obtains the highest AUROC and AUPRC on all five datasets, demonstrating consistently strong subject-level discrimination and class ranking after probability aggregation. The persistence of these gains after combining multiple samples from the same subject suggests that the advantage of TopoAttn is not driven by isolated or particularly easy segments, but reflects more consistent subject-level predictive evidence. Although TeCh achieves higher hard-decision metrics on TDBrain, TopoAttn still attains the best AUROC and AUPRC, indicating that its discriminative scores remain strong even when the final argmax decisions differ. Overall, the subject-level results confirm that the improvements observed in the main experiments largely transfer from individual samples to the clinically relevant task of classifying unseen subjects.

A.10 Efficiency Analysis

Fig. 3 compares F1-Score, inference time, and forward memory overhead on APAVA. TopoAttn achieves the highest F1-Score while maintaining competitive computational and memory efficiency.

Table 12: **Subject-level evaluation on five medical time-series datasets.** For each held-out subject, predicted class probabilities are averaged across all samples to obtain a single subject-level prediction. Results are reported as mean±std over five complete 5-fold out-of-fold evaluations with training seeds 42-46. The best and second-best results are highlighted in **bold red** and *blue*, respectively.

Dataset	Model	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
ADFTD	**Medformer**	51.59±1.36	43.96±5.33	46.48±1.66	42.62±2.97	68.42±1.24	52.20±1.14
	TeCh	*59.55±1.70*	*56.75±1.88*	*56.65±1.43*	*55.75±1.72*	*74.06±1.01*	*62.53±1.37*
	TopoAttn	**60.68±2.11**	**57.28±2.70**	**58.02±2.07**	**56.54±2.24**	**76.22±0.96**	**64.58±1.47**
APAVA	**Medformer**	81.74±3.25	81.84±3.22	81.67±3.40	81.66±3.32	86.52±1.55	87.48±1.59
	TeCh	*82.61±2.75*	*82.65±2.76*	*82.65±2.76*	*82.59±2.75*	*90.00±0.88*	*90.97±0.89*
	TopoAttn	**89.57±2.13**	**89.62±2.04**	**89.62±2.04**	**89.56±2.12**	**94.39±0.77**	**94.97±0.73**
TDBrain	**Medformer**	87.78±0.56	87.42±0.48	*85.21±0.80*	*86.13±0.69*	96.14±0.53	95.62±0.54
	TeCh	**92.22±1.88**	**91.76±1.56**	**91.05±2.78**	**91.32±2.26**	*96.56±0.30*	*96.36±0.30*
	TopoAttn	*88.06±1.11*	*88.96±0.56*	84.67±2.35	86.10±1.71	**98.11±0.46**	**97.66±0.50**
PTB	**Medformer**	89.29±0.59	*88.04±1.11*	83.06±1.27	85.07±0.92	94.31±0.57	*93.41±0.39*
	TeCh	**90.20±0.82**	**88.94±1.17**	*84.69±1.50*	*86.48±1.22*	*94.50±0.55*	93.32±0.56
	TopoAttn	*90.10±0.76*	87.77±1.31	**85.91±0.89**	**86.75±0.92**	**95.16±0.42**	**93.80±0.42**
PTB-XL	**Medformer**	75.46±0.12	67.84±0.27	62.45±0.23	64.53±0.23	91.02±0.09	69.36±0.14
	TeCh	*75.76±0.21*	*68.22±0.42*	**62.92±0.09**	*64.73±0.08*	*91.33±0.13*	*70.08±0.26*
	TopoAttn	**76.72±0.09**	**70.16±0.25**	*62.84±0.17*	**64.82±0.19**	**92.16±0.03**	**71.70±0.08**

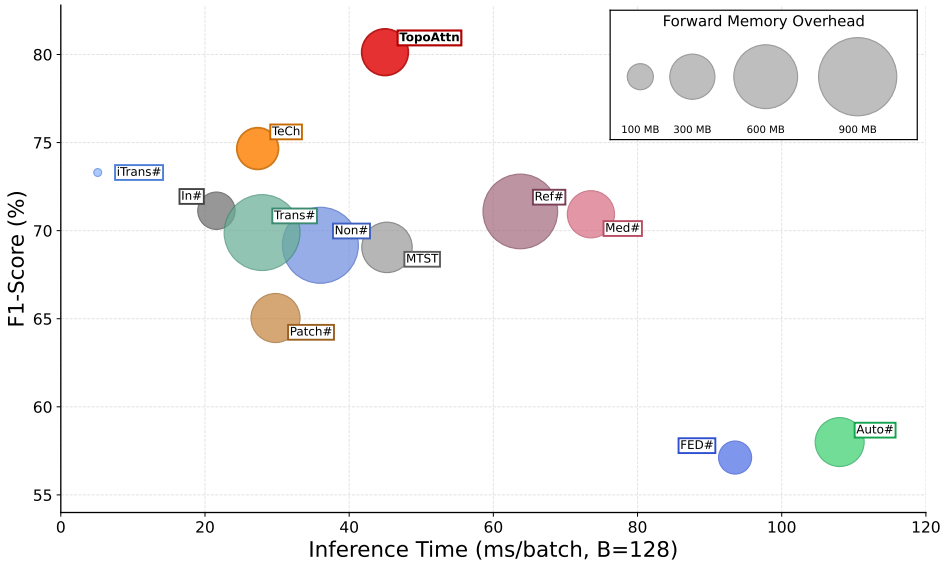

Figure 3: Comparison of F1 scores and computational efficiency across different models. Bubble size denotes forward memory overhead, i.e., the additional GPU memory incurred during inference. The batch size is $B = 128$, and # abbreviates *former* to save space.

Compared with Medformer, it improves F1 by 12.97% while reducing inference time by 38.84% and memory overhead by 2.44%, demonstrating a favorable performance-efficiency trade-off. Although TeCh is more lightweight, TopoAttn further improves F1 by 7.33%. Overall, TopoAttn achieves a strong balance between predictive performance and computational efficiency.

A.11 Statistical Significance Analysis

To assess whether the performance differences between TopoAttn and the strongest competing method, TeCh, are statistically significant, we perform t-tests for each of the five MedTS datasets and six evaluation metrics. Tab. 13 summarizes the direction and significance of the differences. TopoAttn is significantly superior to TeCh for nearly all metrics and datasets. The main exception

is TDBrain, where the difference in accuracy is not significant, while TeCh remains significantly better in recall and F1 score. Conversely, TopoAttn is significantly better in precision, AUROC, and AUPRC on TDBrain. These results support the consistency of the improvements reported in the main results while also identifying the metrics for which TopoAttn does not provide a statistically reliable advantage.

Table 13: Statistical comparison between TopoAttn and TeCh. Symbols indicate whether TopoAttn is significantly higher ($\uparrow^*$), significantly lower ($\downarrow^*$), or not significantly different (ns) according to a two-sided t-test. Differences are considered statistically significant only when $p < 0.05$; results with $p \geq 0.05$, including $p = 0.05$, are marked as not significant.

Dataset	Accuracy	Precision	Recall	F1	AUROC	AUPRC
ADFTD	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$
APAVA	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$
TDBrain	ns	$\uparrow^*$	$\downarrow^*$	$\downarrow^*$	$\uparrow^*$	$\uparrow^*$
PTB	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$
PTB-XL	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$	$\uparrow^*$

A.12 Sensitivity to Same-Subject Aggregation at Inference

All experiments in this work use subject-independent data splits: the training, validation, and test sets contain disjoint subjects. Therefore, no subject is shared across these splits. This property is unchanged in the analysis presented in this section.

A separate consideration arises at inference time because an unseen test subject may contribute multiple recordings. Under the default TopoAttn inference procedure, subject-guided token enhancement can aggregate representations from multiple unlabeled test recordings belonging to the same unseen subject when constructing the subject-level core. Thus, although the evaluation remains subject-independent, the prediction of an individual test sample can benefit from information contained in other test samples from the same subject. This does not introduce information sharing across the training, validation, and test splits, but it raises a distinct question about how much the model benefits from same-subject multi-sample information at inference.

To isolate this effect, we additionally evaluate an independent per-sample inference setting on APAVA and PTB-XL. In this setting, each test recording is processed without accessing any other recording from the same test subject. The subject-level representation used to enhance a test sample is therefore constructed exclusively from that sample. No test labels, predictions, or representations from other test samples are used. All other aspects of the experimental protocol remain unchanged, including the subject-independent splits, five-fold cross-validation, five training seeds (42-46), model architecture, trained model parameters, channel and temporal topologies, and other hyperparameter settings.

We report this comparison on APAVA and PTB-XL to cover two complementary MedTS settings. APAVA is an EEG dataset with 23 subjects, whereas PTB-XL is an ECG dataset with 17,596 subjects. The purpose of this analysis is not to compare the two datasets directly, but to examine whether TopoAttn remains effective when same-subject multi-sample aggregation is unavailable at inference.

As shown in Table 14, removing same-subject multi-sample aggregation has almost no effect on APAVA. Accuracy changes from 84.08% to 84.06%, F1 from 80.13% to 80.11%, while AUROC and AUPRC are unchanged at the reported precision. This indicates that, on APAVA, TopoAttn's predictions do not materially depend on access to other recordings from the same unseen test subject. On PTB-XL, independent per-sample inference produces a larger but still moderate reduction. Accuracy decreases from 74.93% to 73.33%, F1 from 63.61% to 61.85%, AUROC from 91.01% to 90.23%, and AUPRC from 69.51% to 67.44%. Thus, same-subject aggregation provides additional predictive information on PTB-XL, while the model retains most of its performance when each test recording is processed independently.

Overall, this analysis separates two distinct aspects of the evaluation. Subject-independent evaluation concerns whether subjects are disjoint across training, validation, and test sets; this condition holds in both inference settings considered here. The comparison in Table 14 instead measures

Table 14: Sensitivity to same-subject multi-sample aggregation at **inference**. All settings use the same subject-independent train/validation/test splits with disjoint subjects. "Same-subject aggregation" denotes the default inference procedure, in which representations from multiple unlabeled test recordings of the same unseen subject can contribute to the subject-level core. "Independent per-sample" processes each test recording without using any other recording from that subject. Δ denotes independent per-sample minus same-subject aggregation performance. All values are percentages.

Dataset	Inference setting	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
APAVA	Same-subject aggregation	$84.08_{\pm 0.24}$	$79.92_{\pm 0.23}$	$85.20_{\pm 0.47}$	$80.13_{\pm 0.27}$	$91.30_{\pm 0.23}$	$89.46_{\pm 0.70}$
	Independent per-sample	$84.06_{\pm 0.23}$	$79.91_{\pm 0.23}$	$85.19_{\pm 0.46}$	$80.11_{\pm 0.26}$	$91.30_{\pm 0.22}$	$89.46_{\pm 0.69}$
	Δ	-0.02	-0.01	-0.01	-0.02	0.00	0.00
PTB-XL	Same-subject aggregation	$74.93_{\pm 0.07}$	$67.12_{\pm 0.23}$	$62.05_{\pm 0.09}$	$63.61_{\pm 0.11}$	$91.01_{\pm 0.05}$	$69.51_{\pm 0.14}$
	Independent per-sample	$73.33_{\pm 0.34}$	$65.21_{\pm 0.59}$	$60.09_{\pm 0.48}$	$61.85_{\pm 0.42}$	$90.23_{\pm 0.23}$	$67.44_{\pm 0.54}$
	Δ	-1.60	-1.91	-1.96	-1.76	-0.78	-2.07

the effect of allowing information to be aggregated across multiple unlabeled recordings belonging to the same unseen test subject. Accordingly, the observed differences quantify the contribution of same-subject multi-sample information at inference rather than any change in the subject-independent evaluation protocol. The channel and temporal topology components are identical in both settings and are evaluated separately in the topology ablation experiments.