

MedSpaformer: a Transferable Transformer with Multi-granularity Token Sparsification for Medical Time Series Classification

Jiexia Ye¹, Weiqi Zhang², Ziyue Li³, Jia Li¹*, Fugee Tsung²

¹The Hong Kong University of Science and Technology (Guangzhou), China

²The Hong Kong University of Science and Technology, Hong Kong SAR, China

³ Department of Operations and Technology, and Munich Data Science Institute, and Heilbronn Data Science Center, Technical University of Munich, Germany

jye324@connect.hkust-gz.edu.cn, wzhangcd@connect.ust.hk, ziyue.li@tum.de, jjalee@hkust-gz.edu.cn, season@ust.hk

Abstract

Accurate medical time series (MedTS) classification is essential for effective clinical diagnosis, yet remains challenging due to complex multi-channel temporal dependencies, information redundancy, and label scarcity. While transformer-based models have shown promise in time series analysis, most are designed for forecasting tasks and fail to fully exploit the unique characteristics of MedTS. In this paper, we introduce MedSpaformer, a transformer-based framework tailored for MedTS classification. It incorporates a sparse token-based dual-attention mechanism that enables global context modeling and token sparsification, allowing dynamic feature refinement by focusing on informative tokens while reducing redundancy. This mechanism is integrated into a multi-granularity cross-channel encoding scheme to capture intra- and inter-granularity temporal dependencies and inter-channel correlations, enabling progressive refinement of task-relevant patterns in medical signals. The sparsification design allows our model to flexibly accommodate inputs with variable lengths and channel dimensions. We also introduce an adaptive label encoder to extract label semantics and address cross-dataset label space misalignment. Together, these components enhance the model’s transferability across heterogeneous medical datasets, which helps alleviate the challenge of label scarcity. Our model outperforms 13 baselines across 7 medical datasets in supervised learning. It also excels in few-shot learning and demonstrates zero-shot capability in both in-domain and cross-domain diagnostics. These results highlight MedSpaformer’s robustness and its potential as a unified solution for MedTS classification across diverse settings.

Code — <https://github.com/start2020/MedSpaformer>

Introduction

Medical time series (MedTS) data—such as multi-channel electrocardiograms (ECGs) and electroencephalograms (EEGs)—encode rich temporal dynamics crucial for diagnosing life-threatening conditions like arrhythmias (Wagner et al. 2020) and epilepsy (Shah et al. 2018). Early and accurate classification of these signals enables timely intervention and personalized treatment (Wang et al. 2022). Yet, MedTS data present unique modeling challenges due to their

complex structure and clinical constraints: First, MedTS signals exhibit *complex multi-channel temporal dependencies*. Pathological patterns span diverse time scales—from millisecond-level epileptic spikes to minute-level slow oscillations—and are distributed across multiple sensors (e.g., 19-lead EEGs), requiring simultaneous modeling of both temporal hierarchy and cross-channel interactions (Wang et al. 2024; Tang et al. 2021). Second, MedTS data are often *redundant and noisy*, with repeated or irrelevant segments that dilute discriminative patterns and increase computational overhead (Zhang et al. 2024a). Third, *label scarcity* is pervasive—clinically annotated datasets are limited due to the high cost of expert labeling, particularly for rare disorders (Yang et al. 2025; Li et al. 2024a).

Traditional MedTS approaches rely on shallow statistical features (Rahman et al. 2015; Riaz et al. 2020). Current deep learning models—including RNNs (Salloum and Kuo 2017), CNNs (Lawhern et al. 2018), and GNNs (Tang et al. 2021)—are capable of capturing increasingly complex patterns. Recently, transformer-based models have emerged as powerful sequence learners, particularly in time series forecasting (Wen et al. 2022). However, most of them are not specifically designed for MedTS classification, and thus fall short in addressing its domain-specific challenges. PatchTST (Nie et al. 2023) and Crossformer (Zhang and Yan 2023) capture local patterns through patching but lack multi-scale flexibility due to fixed patch sizes. MTST (Zhang et al. 2024b) and Pathformer (Chen et al. 2024) introduce multi-resolution strategies but are confined to single-channel inputs. In contrast, FEDformer (Zhou et al. 2022) and Autoformer (Wu et al. 2021) enable cross-channel attention but overlook multi-scale structure. Medformer (Wang et al. 2024) unifies these views via multi-granularity cross-channel modeling, but its dense self-attention indiscriminately attends to all tokens, lacking effective suppression of redundant signals. Furthermore, these models’ rigid architectural designs—fixed input lengths and channel configurations—constrain their adaptability to heterogeneous datasets, hindering their potential to mitigate label scarcity through cross-dataset transfer learning.

To bridge these limitations, we propose MedSpaformer, a transferable transformer specifically designed for MedTS classification. First, we design a **Token-Sparse Dual Attention (TSDA)** mechanism for granularity and channel

*Corresponding Author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

modeling. TSDA employs self-attention to model global token interactions, followed by token-sparse attention to compress tokens using a fixed smaller number of domain-guided learnable queries. This sparsification aims to remove redundant information, preserve task-relevant features, and reduce computational cost. We stack multiple TSDA blocks to progressively encode multi-granularity and cross-channel information. They first capture local granular features, then refine inter-granularity dependencies, and finally model cross-channel interactions to integrate complementary information. The sparse encoding of TSDA can transform input sequences of varying lengths into fixed-length, enabling our model to directly process heterogeneous inputs with different sequence length and channel configurations. Further, we design an adaptive label encoder to project label descriptions into a unified latent space to bridge cross-dataset label space mismatches. Together, they allow MedSpaformer to transfer knowledge across datasets with varying lengths, channels, and classes, demonstrating few-shot and zero-shot transferability in diverse clinical applications, mitigating the limited label challenge. Our main contributions are as follows:

- We propose MedSpaformer, a transformer architecture tailored for MedTS classification. By incorporating a token-sparse dual-attention mechanism into a multi-granularity cross-channel encoding framework, MedSpaformer progressively distills informative patterns, reduces redundancy, and effectively models multi-scale temporal dynamics and cross-channel dependencies in medical data.
- MedSpaformer supports input-output heterogeneity via the sparse encoding mechanism and an adaptive label encoder. To the best of our knowledge, it is the first transformer framework enabling cross-task zero-shot transfer in time series classification.
- We conduct extensive experiments on multiple public datasets, achieving state-of-the-art performance in both supervised and few-shot settings. Furthermore, we evaluate our model in in-domain and cross-domain zero-shot scenarios to demonstrate its cross-dataset transferability.

Related Work

Medical Time Series Classification. Medical time series (MedTS) data, such as EEG (Escudero et al. 2006), ECG (PhysioBank 2000), EMG (Xiong et al. 2021), and EOG (Fan et al. 2021), are widely used in disease diagnosis, monitoring, and rehabilitation (Fatourehchi et al. 2007). Traditional methods, such as nearest neighbor classifiers (Rahman et al. 2015), auto-regressive models (Schaffer, Dobbins, and Pearson 2021), and Gaussian mixture models (Vincent, Risser, and Ciuciu 2009), offer simplicity and interpretability but face challenges when dealing with complex, high-dimensional patterns. With the advent of deep learning, models leveraging RNNs (Salloum and Kuo 2017), CNNs (Lawhern et al. 2018), and GNNs (Tang et al. 2021) have dominated MedTS classification. For instance, EEG-Net (Lawhern et al. 2018) uses depthwise separable convolutions to extract EEG features, while GNNs (Tang et al. 2021) enable self-supervised seizure detection. While these

models show promising results in tasks with single-modality medical signals, they often lack generalizability across different medical modalities. For example, a model tailored for ECG (Ding et al. 2025) may not transfer effectively to other types of medical signals such as EEG (Sharma and Meena 2024) or EOG (van Gorp et al. 2024).

Transformer for Time Series. Transformers have significantly advanced time series analysis. Based on tokenization strategies, they can be categorized into single-timestamp (Wu et al. 2021; Zhou et al. 2021), all-timestamp (Liu et al. 2024), and multi-timestamp approaches (Zhang and Yan 2023; Zhang et al. 2024b; Wang et al. 2024), with the latter further divided into single- and multi-granularity methods. Single-timestamp tokenization struggles with capturing coarse-grained patterns, while all-timestamp strategies may overlook fine-grained local details. Single-granularity methods like PatchTST (Nie et al. 2023) and Crossformer (Zhang and Yan 2023) generate fixed-length patches from single-channel sequences, capturing local patterns but falling short in handling multi-scale dynamics. Multi-granularity models such as MTST (Zhang et al. 2024b) and Pathformer (Chen et al. 2024) address this by using varied patch sizes, yet remain limited to single-channel inputs, which may hinder performance for multivariate time series classification. Medformer (Wang et al. 2024) employs multi-granularity encoding and captures low-level channel correlations via cross-channel patching. In contrast, our model derives high-level channel representations through channel-wise multi-granularity encoding. Moreover, Medformer lacks a mechanism for suppressing redundant signals, which we address via token sparsification. Finally, while prior models show limited cross-dataset transferability, our model enables direct transfer across heterogeneous medical datasets.

Methodology

Problem Formulation Consider a medical time series dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^N$ where each signal $\mathbf{X}_i \in \mathbb{R}^{L \times C}$ contains L timestamps across C channels and each label $y_i \in \{1, 2, \dots, M\}$ is described by a text \mathcal{T}_{y_i} . M is the number of classes. Our objective is to learn a framework to align the temporal signal \mathbf{X}_i and its label description \mathcal{T}_{y_i} into a unified D dimension latent space to obtain their representations $\mathbf{h}_i^{(x)} \in \mathbb{R}^D$ and $\mathbf{h}_i^{(y)} \in \mathbb{R}^D$. The framework is optimized by maximizing the similarity between time series-label pairs $(\mathbf{h}_i^{(x)}, \mathbf{h}_i^{(y)})$.

Overview Figure 1 demonstrates our model. In this section, we first introduce the core component—the token-sparse dual attention block (TSDA). TSDA effectively captures global context among tokens, eliminates redundant signals and refines features by token sparsification. Next, we apply TSDA blocks on multi-granularity encoding for intra- and inter-granularity correlation extraction, and on multi-channel encoding for channel correlation integration. Built upon TSDA blocks, our model is inherently agnostic to input length and channel configurations. We also introduce an adaptive label encoder to align heterogeneous label spaces across datasets. These designs enable the model to be trained across diverse datasets and equip it with few-shot/zero-shot

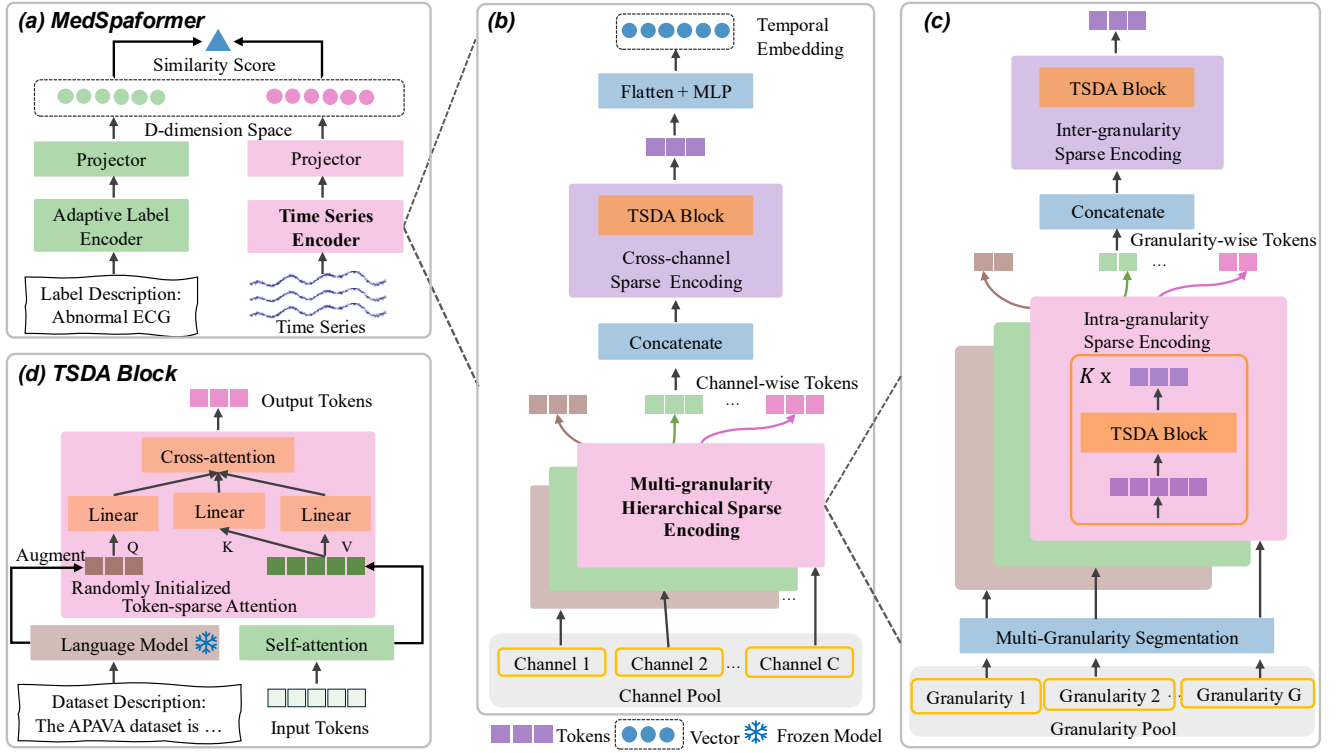


Figure 1: (a) MedSpaformer consists of a time series encoder and a label encoder to map time series and labels into a unified space for optimization. (b) shows the modeling of intra- and inter-channel correlations. (c) illustrates multi-granularity encoding with TSDA blocks that capture intra- and inter-granularity dependencies. (d) The token-sparse dual attention (TSDA) block combines self-attention to model global context and token-sparse attention to focus on informative local patterns.

transferability across different medical applications.

Token-Sparse Dual Attention Block

Inspired by physicians’ two-stage diagnostic process—first holistically contextualizing symptoms, then analyzing specific biomarkers (Hausmann et al. 2016)—we propose the Token-Sparse Dual Attention (TSDA) block, which mirrors this process through global context modeling and dynamic feature refinement using a two-stage attention mechanism.

TSDA first employs self-attention to capture global long-range temporal dependencies, leveraging its proven ability to model pairwise token interactions in sequential data (Chen et al. 2024; Wang et al. 2024). This global modeling capability integrates information across the entire sequence, reinforces inter-token dependencies, and contextualizes local patterns—crucial in medical signals, where waveform anomalies gain meaning only within the broader temporal structure (Wagner et al. 2020). Formally, given an input sequence $\mathbf{H} \in \mathbb{R}^{L \times D}$, the self-attention output is defined as $\mathbf{H}^{\text{self}} \leftarrow \text{Attn}^{\text{self}}(\mathbf{H}, \mathbf{H}, \mathbf{H})$, where $\mathbf{H}^{\text{self}} \in \mathbb{R}^{L \times D}$.

While self-attention captures comprehensive temporal dependencies, medical signals often contain redundant or noisy patterns that obscure diagnostic features. To address this, inspired by Q-Former (Li et al. 2023) which employs learnable queries to extract visual representation most relevant to the text, we design a token-sparse attention layer

that uses learnable queries informed by domain-specific priors to selectively attend to diagnostically salient features—analogue to how physicians narrow their analysis to specific biomarkers with domain knowledge after forming an initial clinical impression. Specifically, we introduce a set of Q randomly initialized learnable query vectors \mathbf{Q} , augmented with domain-specific prior embedding $\mathbf{e}^{\text{prior}}$: $\mathbf{Q}^{\text{aug}} = f(\mathbf{Q}, \mathbf{e}^{\text{prior}})$ where $\mathbf{Q}^{\text{aug}} \in \mathbb{R}^{Q \times D}$. f is the function to fuse queries and priors and concatenation is applied in our experiments. Following (Jin et al. 2023), we utilize a frozen language model to generate domain-specific embedding based on the dataset description: $\mathbf{e}^{\text{prior}} = f_{\text{LM}}(\mathcal{T}^{\text{data}})$. These queries then attend to \mathbf{H}^{self} to generate a sparse token set:

$$\begin{aligned} \mathbf{H}^{\text{sparse}} &\leftarrow \text{Attn}^{\text{sparse}}(\mathbf{Q}^{\text{aug}}, \mathbf{H}^{\text{self}}, \mathbf{H}^{\text{self}}) \\ &= \text{Softmax} \left(\frac{(\mathbf{Q}^{\text{aug}} \mathbf{W}_Q) (\mathbf{H}^{\text{self}} \mathbf{W}_K)^{\top}}{\sqrt{D}} \right) (\mathbf{H}^{\text{self}} \mathbf{W}_V) \end{aligned} \quad (1)$$

where $\mathbf{H}^{\text{sparse}} \in \mathbb{R}^{Q \times D}$ retains Q tokens and $Q \ll L$, aiming to preserve critical features and eliminate irrelevant information, reducing computation. Note that TSDA block transforms variable-length sequences into fixed-length representations via token-sparse attention. This design is input-length-agnostic—its trainable parameters depend solely on the predefined queries number Q and dimension D —en-

abling parameter sharing across inputs of arbitrary lengths and ensuring computational stability.

Multi-granularity Hierarchical Sparse Encoding

Multi-granularity Segmentation. In the multi-granularity encoding module, each channel of the input is independently processed to capture intra-channel distinctive features. To capture intra-channel multi-scale temporal patterns, following (Wang et al. 2024; Zhang et al. 2024b), we partition each channel into multi-granularity segments using varying window sizes $\mathcal{S} = \{s_1, s_2, \dots, s_G\}$ and $|\mathcal{S}| = G$. Each granularity s_i generates a sequence of non-overlapping patches $\{p_1^{(i)}, p_2^{(i)}, \dots\}$, where $p_j^{(i)} \in \mathbb{R}^{s_i}$ represents the j -th patch of granularity i . The number of patches is $L_i = \lceil L/s_i \rceil$ with zero padding to ensure divisibility. These patches are projected into a unified latent space of dimension D via linear transformations to obtain the patch embedding sequence $\mathbf{P}_i = [\hat{p}_1^{(i)}, \hat{p}_2^{(i)}, \dots, \hat{p}_{L_i}^{(i)}] \in \mathbb{R}^{L_i \times D}$. The embeddings of all granularities undergo hierarchical sparse encoding to model both intra-granularity and inter-granularity temporal dependencies.

Intra-Granularity Hierarchical Sparse Encoding.

Each granularity is first processed independently to capture granularity-specific temporal dynamics. Inspired by the human cognitive process of analyzing time series signals (e.g., ECG waveforms) (Wagenmakers, Farrell, and Ratcliff 2005)—which involves iteratively filtering noise, aggregating local patterns, and distilling global semantics—we integrate K TSDA blocks for intra-granularity processing to enable hierarchical feature refinement. The forward process of k -th TSDA block is formulated as follows:

$$\mathbf{H}_k = \text{TSDA}_k(\mathbf{H}_{k-1}; \Theta_k, O_k) \quad (2)$$

where \mathbf{H}_{k-1} is the input token sequence and $\mathbf{H}_0 = \mathbf{P}_i$. $\mathbf{H}_k \in \mathbb{R}^{O_k \times D}$ is the output token sequence. Θ_k denotes trainable parameters and O_k is the critical hyperparameter controlling token compression and $O_k < O_{k-1}$. The output of the whole hierarchical TSDA processing is denoted as $\mathbf{H}^{\text{intra}} \in \mathbb{R}^{O_K \times D} = \mathbf{H}_K$, a granularity-wise representation for subsequent inter-granularity correlation modeling.

Inter-Granularity Sparse Encoding. Intra-granularity encoding has learned diverse high-level temporal features in different granularities, which are subsequently concatenated into a token sequence, denoted as $\mathbf{H}_S^{\text{intra}} \in \mathbb{R}^{(G \cdot O^K) \times D} = [\mathbf{H}_{s_1}^{\text{intra}}; \mathbf{H}_{s_2}^{\text{intra}}; \dots; \mathbf{H}_{s_G}^{\text{intra}}]$. A single TSDA block then models inter-granularity relationships as follows:

$$\mathbf{H}^{\text{inter}} \in \mathbb{R}^{O^{\text{inter}} \times D} = \text{TSDA}(\mathbf{H}_S^{\text{intra}}; \Theta, O^{\text{inter}}) \quad (3)$$

where $O^{\text{inter}} \ll G \cdot O^K$. The self-attention of TSDA block serves to establish a global context among the tokens of different granularities. This operation allows the model to understand the overall structure and interconnections among the different granularity tokens. Then the token-sparse attention compresses the information from the different granularities into a more manageable and focused representation to refine the information as well as reduce computation. Additionally, different datasets may favor distinct granularities, and the domain knowledge-based learnable queries can

guide the selection of optimal granularities for each dataset to enhance the model’s generalization.

Cross-Channel Sparse Encoding

The output of inter-granularity encoding, $\mathbf{H}^{\text{inter}}$, serves as high-level semantic tokens for each channel. By concatenating all channel representations, we obtain the channel embedding matrix $\mathbf{H}^C = [\mathbf{H}_1^{\text{inter}}; \mathbf{H}_2^{\text{inter}}; \dots; \mathbf{H}_C^{\text{inter}}] \in \mathbb{R}^{(C \cdot O^{\text{inter}}) \times D}$, which is then passed through a TSDA block to model and enhance inter-channel dependencies as follows:

$$\mathbf{H}_C^{\text{self}} \in \mathbb{R}^{(C \cdot O^{\text{inter}}) \times D} \leftarrow \text{Attn}^{\text{self}}(\mathbf{H}^C, \mathbf{H}^C, \mathbf{H}^C) \quad (4)$$

$$\mathbf{H}_C^{\text{sparse}} \in \mathbb{R}^{U \times D} \leftarrow \text{Attn}^{\text{sparse}}(\mathbf{Q}_C^{\text{aug}}, \mathbf{H}_C^{\text{self}}, \mathbf{H}_C^{\text{self}}) \quad (5)$$

The first self-attention layer in Equation 4 computes dense pairwise correlations across all channels, establishing a global context that captures both complementary relationships (e.g., spatially distant EEG channels jointly detecting propagating epileptic spikes). Leveraging the comprehensive context from the previous layer, the second token-sparse attention layer in Equation 5 distills the C channel tokens into U ($U < C \cdot O^{\text{inter}}$) task-specific prototypes through learnable, domain-informed queries $\mathbf{Q}_C^{\text{aug}}$, aiming to filtering out irrelevant noise (e.g., overlapping functionalities among biosensors). The output of TSDA block is flattened and projected into a D -dimensional space to generate the final temporal embedding: $\mathbf{h}_i^{(x)} \in \mathbb{R}^D = \text{MLP}(\text{Flatten}(\mathbf{H}_C^{\text{sparse}}))$. Note that this module’s trainable parameters are independent of channel number C , allowing deployment across heterogeneous datasets with varying channel counts (e.g., 6-channel ICU monitors vs. 12-channel wearable arrays) without architectural adaptation. This refinement amplifies critical channel interactions while suppressing noise.

Adaptive Label Encoder

Traditional classification models rely on one-hot embedding for label representation, struggling to adapt to heterogeneous label spaces or generalize to unseen classes, limiting their cross-dataset transferability. Recent advances attempt to mitigate this challenge: ZeroG (Li et al. 2024b) constructs a unified cross-dataset label space via pre-trained language model (LM) for graph classification. UniTS (Gao et al. 2024) introduces trainable CLS tokens as label embeddings to support different time series classification task adaptation. Akata et al. (Akata et al. 2016) utilize attribute embeddings as priors and update label embeddings for image classification through labeled training data. Inspired by these works, we propose an adaptive label encoder designed to enhance the model’s cross-dataset transferability and generalization capabilities. A subsequent learnable projector dynamically refines the label embeddings, mapping them to a unified D -dimensional space shared with the time series embeddings. The formula is as follows:

$$\mathbf{h}_i^{(y)} \in \mathbb{R}^D = \mathbf{W}_1 \cdot (\text{ReLU}(\mathbf{W}_2 \cdot f_{\text{LM}}(\mathcal{T}_{y_i}) + b)) \quad (6)$$

where f_{LM} refers to a frozen language model, and \mathcal{T}_{y_i} denotes the textual description of label y_i . $\mathbf{h}_i^{(y)}$ represents the adaptive label embedding.

Loss Function: A cross-entropy loss is utilized for training, formulated as follows:

$$\mathcal{L} = - \sum_{i=1}^N \log \frac{\exp \left(\text{sim} \left(\mathbf{h}_i^{(x)}, \mathbf{h}_i^{(y)} \right) \right)}{\sum_{j=1}^M \exp \left(\text{sim} \left(\mathbf{h}_i^{(x)}, \mathbf{h}_j^{(y)} \right) \right)} \quad (7)$$

where $\text{sim}(\cdot)$ is a function to measure the similarity between temporal embedding and class embedding. During the inference stage, the class with the highest similarity score is predicted as label of the medical signal, formalized as:

$$y'_i = \text{argmax}_j \left(\text{sim} \left(\mathbf{h}_i^{(x)}, \mathbf{h}_j^{(y)} \right) \mid j \in \{1, \dots, M\} \right) \quad (8)$$

where y'_i is the predicted label for sample \mathbf{X}_i . We employ the dot product as the function $\text{sim}(\cdot)$.

Datasets	# Samples	# Channels	# Steps
APAVA (2-Classes)	5,967	16	256
ADFTD (3-Classes)	69,752	19	256
TUSZ (2-Classes)	22,040	19	6,000
TUSZ (4-Classes)	2,891	19	6,000
PTB (2-Classes)	64,356	15	300
PTB-XL (4-Classes)	17,110	12	1,000
PTB-XL (5-Classes)	17,110	12	1,000

Table 1: Statistics of datasets.

Experiments

We evaluate our model’s efficacy and in-domain/cross-domain transferability on 7 real-world datasets from 3 medical domains with 13 baselines.

Experimental Setup

Datasets. We select datasets from three medical domains. **(1) Alzheimer’s Disease:** APAVA (Escudero et al. 2006) and ADFTD (Miltiadous et al. 2023) are two EEG datasets for Alzheimer’s Disease classification. **(2) Epilepsy:** TUSZ v1.5.2 (Shah et al. 2018) is a large-scale corpus of EEG signals for Epilepsy. It offers two label sets: a coarse-grained label set (2 Classes) that distinguishes between seizure and non-seizure signals, and a fine-grained label set (4 Classes) that categorizes seizures into four types. **(3) Heart Disease:** PTB (PhysioBank 2000) and PTB-XL (Wagner et al. 2020) are two large-scale ECG databases for heart disease diagnosis. PTB-XL also provides two label sets: PTB-XL (4 Classes) with coarse-grained labels, and PTB-XL (5 Classes) with fine-grained labels. Table 1 provides brief information about the processed datasets.

Baselines. We compare our model with a diverse set of baselines categorized as follows: **(1) Non-Transformer Models:** DLinear (Zeng et al. 2023), MultiRocket (Tan et al. 2022), LightTS (Zhang et al. 2022), TimesNet (Wu

et al. 2022). **(2) Non-Multi-granularity Transformer-based Models:** PatchTST (Nie et al. 2023), Autoformer (Wu et al. 2021), Crossformer (Zhang and Yan 2023), ETSformer (Woo et al. 2022), FEDformer (Zhou et al. 2022), Informer (Zhou et al. 2021). **(3) Multi-granularity Transformer-based Models:** PathFormer (Chen et al. 2024), Medformer (Wang et al. 2024), MTST (Zhang et al. 2024b).

Implementation Details. Following (Wang et al. 2024), we macro-averaged F1, macro-averaged AUROC, macro-averaged AUPRC and accuracy as metrics. We save the model with the best F1 score on the validation set and evaluate it on the test set. Due to data imbalance, we use the F1 score in the main paper to better reflect model performance. We find a set of hyper-parameters that perform optimally for most datasets through fine-tuning: multi-granularity window $\mathcal{S} = [25, 50, 100, 150]$; TSDA blocks $K = 3$ with hierarchical token list $\{O_1, \dots, O_K\} = [128, 64, 32]$; inter-granularity encoding output tokens $O^{\text{inter}} = 10$ and cross-channel encoding output tokens $U = 5$; hidden dimension $D = 128$. We adopt ClinicalBERT (Wang et al. 2023) as the frozen LM. The batch size is set to 128 for the ADFTD and PTB datasets, while 32 for the remaining datasets. We employ the AdamW optimizer and the Cosine scheduler for learning rate decay. The training session is conducted with ten random seeds (41–50) on fixed training, validation, and test sets to compute the mean and standard deviation of model performance. Each training process runs for up to 60 epochs, with early stopping if there is no improvement in F1 score of validation set for 7 consecutive epochs. All experiments are conducted using the PyTorch framework on NVIDIA A6000 (48GB) GPU.

Supervised Learning

Key findings in Table 2 include: (1) MultiRocket, DLinear, and LightTS perform poorly due to their simplified architectures, which struggle with complex temporal dependencies. (2) Transformer-based models generally surpass traditional methods, underscoring the effectiveness of self-attention mechanism. (3) Five suboptimal performances in the multi-granularity models highlight the effectiveness of the multi-granularity mechanism in leveraging the contributions of different granularities to decision-making. (4) MedSpaformer outperforms all datasets, demonstrating its strong generalization capability. By emphasizing useful multi-granularity tokens and progressively discarding redundant information, it extracts higher-level channel interactions, enhancing performance compared to Medformer.

Few-shot Learning

Few-shot learning addresses the label scarcity challenge by transferring knowledge from source domain with ample labeled data to target domain with limited labels. In this section, we pre-train all models on the source dataset, then fine-tune them on the target dataset under $\{5, 10, 20, 30, 40, 50\}$ -shot settings. Since our baselines have fixed input dimensions, direct transfer between heterogeneous datasets is infeasible. We select source-target dataset pair with the

Datasets Model	APAVA (2-Classes)	ADFTD (3-Classes)	TUSZ (2-Classes)	TUSZ (4-Classes)	PTB-XL (4-Classes)	PTB-XL (5-Classes)	PTB (2-Classes)
MultiRocket	0.511±0.015	0.342±0.004	0.629±0.029	0.733±0.013	0.224±0.018	0.275±0.016	0.572±0.006
Dlinear	0.482±0.012	0.295±0.002	0.645±0.022	0.736±0.016	0.239±0.013	0.251±0.001	0.599±0.008
LightTS	0.526±0.023	0.378±0.021	0.695±0.003	0.843±0.014	0.469±0.009	0.431±0.011	0.735±0.015
TimesNet	0.703±0.019	<u>0.463±0.024</u>	0.764±0.017	0.849±0.012	0.485±0.006	0.526±0.022	0.781±0.027
PatchTST	0.565±0.011	0.453±0.015	0.746±0.005	0.855±0.022	0.566±0.003	0.509±0.024	0.762±0.024
Autoformer	0.715±0.021	0.435±0.011	0.725±0.018	0.802±0.021	0.432±0.019	0.493±0.012	0.635±0.014
Crossformer	0.691±0.009	0.428±0.014	0.741±0.021	0.837±0.011	0.548±0.017	0.485±0.007	0.742±0.013
ETSformer	0.652±0.022	0.451±0.011	0.811±0.024	0.834±0.015	0.508±0.013	0.439±0.009	0.803±0.021
FEDformer	<u>0.742±0.018</u>	0.432±0.019	0.718±0.012	0.803±0.024	0.528±0.023	0.527±0.015	0.686±0.012
Informer	0.676±0.014	0.461±0.008	0.772±0.011	0.848±0.003	0.448±0.025	0.463±0.006	0.728±0.015
PathFormer	0.674±0.016	0.415±0.009	0.713±0.004	0.794±0.022	0.503±0.012	0.482±0.002	0.618±0.018
Medformer	0.711±0.017	0.459±0.013	<u>0.821±0.015</u>	0.839±0.008	<u>0.575±0.014</u>	0.519±0.005	<u>0.814±0.003</u>
MTST	0.637±0.013	0.427±0.012	0.765±0.002	<u>0.855±0.009</u>	0.546±0.011	<u>0.532±0.003</u>	0.711±0.002
MedSpaformer	0.821±0.014	0.468±0.012	0.852±0.007	0.901±0.011	0.583±0.014	0.562±0.009	0.843±0.014

Table 2: Supervised Learning in F1 score The best results are highlighted in bold, and the second-best results are underlined.

Zero-shot Experiments		Test Datasets						
		Alzheimer’s Disease		Epilepsy		Heart Disease		
		APAVA (2-Classes)	ADFTD (4-Classes)	TUSZ (2-Classes)	TUSZ (4-Classes)	PTB-XL (4-Classes)	PTB-XL (5-Classes)	PTB (2-Classes)
Pre-training Domains	Alzheimer’s Disease	0.533±0.045	0.291±0.062	0.474±0.048	0.481±0.030	0.285±0.054	0.194±0.025	0.422±0.099
	Epilepsy	0.407±0.015	0.273±0.037	0.470±0.056	0.515±0.032	0.238±0.034	0.173±0.011	0.381±0.063
	Heart Disease	0.413±0.024	0.305±0.049	0.517±0.068	0.506±0.011	0.271±0.079	0.230±0.034	0.452±0.055
Supervised/ Few-shot Experiments	Dlinear (50-shot)	N/A	N/A	0.434±0.023	0.478±0.025	0.187±0.028	0.163±0.047	N/A
	MedSpaformer (5-shot)	N/A	N/A	0.546±0.017	0.577±0.012	0.253±0.045	0.259±0.009	N/A
	Dlinear (Supervised)	0.482±0.012	0.295±0.002	0.645±0.022	0.736±0.016	0.239±0.013	0.251±0.001	0.599±0.008

Table 3: Zero-shot Learning in F1 score: In-domain (gray background) versus cross-domain experiments, with comparisons to few-shot and supervised learning.

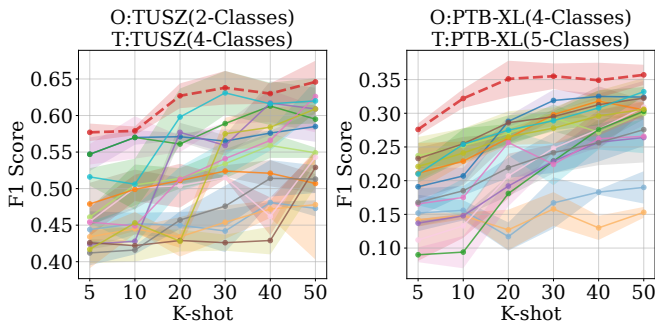


Figure 2: Few-shot results under different shots on two experiments. *O* refers to source domain while *T* refers to target domain.

same input length and channel counts, namely PTB-XL (4-Classes) and PTB-XL (5-Classes), TUSZ (2-Classes) and TUSZ (4-Classes). After pre-training, we freeze the model backbone and train a task-specific classification head for fine-tuning. Figure 2 shows that (1) The performance of

nearly all the models increases when the shots increases. And our model has the best performance on almost all the shots, demonstrating its robustness in transferability. (2) The transformer-based models usually have better performance than non-transformer models, which is consistent with the supervised model performance. (3) The performance gap between our model and baselines is more pronounced in PTB-XL than in TUSZ, showing its superior few-shot learning capacity in PTB-XL.

Zero-shot Learning

To evaluate the zero-shot transferability of our model, we conduct in-domain and cross-domain experiments. In-domain experiments transfer knowledge between datasets within the same domain. For example, when the target dataset is APAVA, we use the remaining datasets in "Alzheimer’s Disease" domain, specifically ADFTD, as sources. In contrast, cross-domain experiments involve pre-training our model on all datasets of the source domain and evaluating it on the test dataset from the target domain. Since our baselines lack zero-shot capability, we provide few-shot and supervised learning results for comparison. In Table 3,

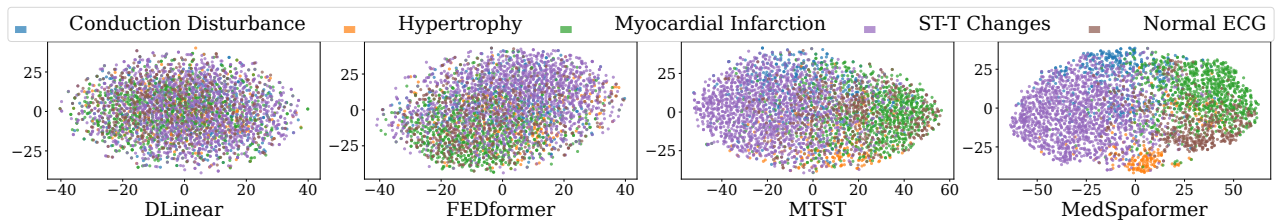


Figure 3: Embedding visualization of PTB-XL(5-Classes) on representative models, with colors indicating class labels.

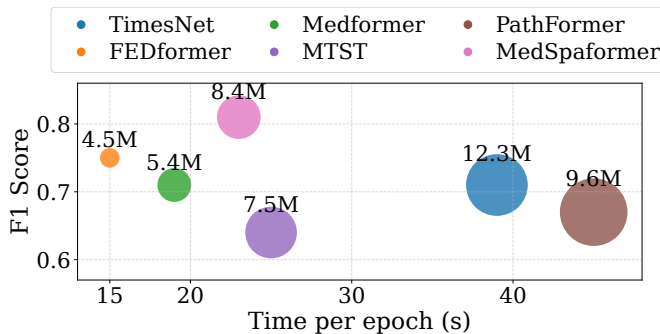


Figure 4: Efficiency comparison on the APAVA dataset (2 Classes). M (million) serves as unit for trainable parameters.

(1) There are four best performances in in-domain experiments and only three in cross-domain experiments, indicating that in-domain transfer exhibits stronger zero-shot performance than cross-domain. (2) Our model’s zero-shot performance excels DLinear under 50 shots. (3) Our model’s best zero-shot performance on APAVA, ADFTD, and PTB-XL (4-Classes) surpasses DLinear under supervised learning, likely due to the extensive pre-training data that captures a wider range of temporal patterns.

Model	APAVA (2-Classes)	TUSZ (2-Classes)	PTB (2-Classes)
W/O Multi-Granularity	0.727±0.012	0.771±0.003	0.753±0.008
W/O Channel Attention	0.766±0.009	0.794±0.005	0.787±0.006
W/O Sparse Attention	0.752±0.008	0.788±0.004	0.767±0.003
W/O Label Encoder	0.796±0.019	0.828±0.007	0.820±0.015
MedSpaformer	0.821±0.014	0.852±0.007	0.843±0.014

Table 4: Ablation Study in F1 score on three datasets

More Experiments

Ablation Study. To assess the impact of critical modules in our model, we perform ablation studies in four configurations. “W/O Multi-Granularity” uses single-granularity {25} to replace multi-granularity. “W/O Channel Attention” replaces cross-channel encoding with simple concatenation of all channel representations. “W/O Sparse Attention” substitutes token-sparse attention with self-attention. “W/O Label Encoder” uses one-hot encoding for ground truth. In Ta-

ble 4, Multi-Granularity makes the most significant contribution, improving performance by approximately 7% on average of mean performance. Sparse attention follows with an enhancement of about 6%. Channel attention contributes nearly 5% as well, while the Label Encoder provides an additional improvement of around 2%. These results underscore the efficacy of our proposed mechanisms.

Efficiency Analysis. In Figure 4, we compare the efficiency of our model against representative baselines on APAVA dataset, including the time required to train one epoch, F1 score, and trainable parameters. FEDformer is the fastest and most lightweight model, ranking second in performance. Medformer is the second fastest and second smallest, with the third-best performance. MedSpaformer ranks third in training time and has 8.4 million parameters, smaller than TimesNet and PathFormer. While it sacrifices some training speed compared to FEDformer and Medformer, it achieves a significantly higher F1 score. Overall, MedSpaformer presents a balanced option in the trade-off between efficiency and effectiveness. Its relatively high performance with a reasonable trainable parameter count and training time makes it a viable choice.

Visualization. To better visualize the learned representations from supervised learning, we use t-SNE (Maaten and Hinton 2008) to project the representations of representative models on the PTB-XL (5-Classes) dataset into a 2D space in Figure 3. DLinear struggles to distinguish with different classes. FEDformer demonstrates better class separation, particularly for the dominant class, ST-T. MTST further improves by identifying the second-largest class, Myocardial Infarction, but fails in the small classes. In contrast, our model provides better discrimination among all classes.

Conclusion

We propose MedSpaformer, a novel transformer-based model tailored for medical time series classification. By incorporating token-sparse dual-attention mechanism into multi-granularity cross-channel encoding, MedSpaformer effectively captures both intra- and inter-channel dependencies as well as multi-scale temporal patterns critical to medical signals. The combination of sparse encoding and an adaptive label encoder enables MedSpaformer to process heterogeneous datasets with few-shot and zero-shot transferability. Extensive experiments validate its superiority, robustness, and adaptability to improve diagnostic performance in medical contexts.

Acknowledgements

This work is funded by National Natural Science Foundation of China Grant No. 72371217, the Guangzhou Industrial Informatics and Intelligence Key Laboratory No. 2024A03J0628, the Nansha Key Area Science and Technology Project No. 2023ZD003, and Project No. 2021JC02X191.

References

- Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2016. Label-Embedding for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7): 1425–1438.
- Chen, P.; ZHANG, Y.; Cheng, Y.; Shu, Y.; Wang, Y.; Wen, Q.; Yang, B.; and Guo, C. 2024. Pathformer: Multi-scale Transformers with Adaptive Pathways for Time Series Forecasting. In *International Conference on Learning Representations*.
- Ding, C.; Yao, T.; Wu, C.; and Ni, J. 2025. Advances in deep learning for personalized ECG diagnostics: A systematic review addressing inter-patient variability and generalization constraints. *Biosensors and Bioelectronics*, 271: 117073.
- Escudero, J.; Abásolo, D.; Hornero, R.; Espino, P.; and López, M. 2006. Analysis of electroencephalograms in Alzheimer’s disease patients with multiscale entropy. *Physiological measurement*, 27(11): 1091.
- Fan, J.; Sun, C.; Long, M.; Chen, C.; and Chen, W. 2021. EOGNET: A Novel Deep Learning Model for Sleep Stage Classification Based on Single-Channel EOG Signal. *Frontiers in Neuroscience*.
- Fatourechi, M.; Bashashati, A.; Ward, R. K.; and Birch, G. E. 2007. EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical neurophysiology*, 118(3): 480–494.
- Gao, S.; Koker, T.; Queen, O.; Hartvigsen, T.; Tsiligkaridis, T.; and Zitnik, M. 2024. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37: 140589–140631.
- Hausmann, D.; Zulian, C.; Battegay, E.; and Zimmerli, L. 2016. Tracing the decision-making process of physicians with a Decision Process Matrix. *BMC medical informatics and decision making*, 16(1): 133.
- Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; et al. 2023. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *The Twelfth International Conference on Learning Representations*.
- Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEG-Net: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces. *Journal of Neural Engineering*, 056013.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International conference on machine learning*, volume 202, 19730–19742.
- Li, J.; Liu, C.; Cheng, S.; Arcucci, R.; and Hong, S. 2024a. Frozen language model helps ECG zero-shot learning. In *Medical Imaging with Deep Learning*, 402–415. PMLR.
- Li, Y.; Wang, P.; Li, Z.; Yu, J. X.; and Li, J. 2024b. ZeroG: Investigating Cross-dataset Zero-shot Transferability in Graphs. *KDD*.
- Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *International Conference on Learning Representations*.
- Maaten, L. v. d.; and Hinton, G. E. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.
- Miltiadous, A.; Tzamourta, K. D.; Afrantou, T.; Ioannidis, P.; Grigoriadis, N.; Tsalikakis, D. G.; Angelidis, P.; Tsipouras, M. G.; Glavas, E.; Giannakeas, N.; et al. 2023. A dataset of scalp EEG recordings of Alzheimer’s disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6): 95.
- Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*.
- PhysioBank, P. 2000. Physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220.
- Rahman, S. A.; Huang, Y.; Claassen, J.; Heintzman, N.; and Kleinberg, S. 2015. Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. *Journal of biomedical informatics*, 58: 198–207.
- Riaz, F.; Rehman, S.; Ajmal, M.; Hafiz, R.; Hassan, A.; Aljohani, N. R.; Nawaz, R.; Young, R.; and Coimbra, M. 2020. Gaussian mixture model based probabilistic modeling of images for medical image segmentation. *IEEE Access*, 8: 16846–16856.
- Salloum, R.; and Kuo, C.-C. J. 2017. ECG-based biometrics using recurrent neural networks. In *ICASSP*, 2062–2066.
- Schaffer, A. L.; Dobbins, T. A.; and Pearson, S.-A. 2021. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. *BMC Medical Research Methodology*.
- Shah, V.; Von Weltin, E.; Lopez, S.; McHugh, J. R.; Veloso, L.; Golmohammadi, M.; Obeid, I.; and Picone, J. 2018. The temple university hospital seizure detection corpus. *Frontiers in neuroinformatics*, 12: 83.
- Sharma, R.; and Meena, H. K. 2024. Emerging trends in EEG signal processing: A systematic review. *SN Computer Science*, 5(4): 415.
- Tan, C. W.; Dempster, A.; Bergmeir, C.; and Webb, G. I. 2022. MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, 36(5): 1623–1646.
- Tang, S.; Dunnmon, J. A.; Saab, K.; Zhang, X.; Huang, Q.; Dubost, F.; Rubin, D.; and Lee-Messer, C. 2021. Self-Supervised Graph Neural Networks for Improved Electroen-

- cephalographic Seizure Analysis. In *International Conference on Learning Representations*.
- van Gorp, H.; van Gilst, M. M.; Overeem, S.; Dujardin, S.; Pijpers, A.; van Wetten, B.; Fonseca, P.; and van Sloun, R. J. 2024. Single-channel EOG sleep staging on a heterogeneous cohort of subjects with sleep disorders. *Physiological measurement*, 45(5): 055007.
- Vincent, T.; Risser, L.; and Ciuciu, P. 2009. Spatially adaptive mixture modeling for analysis of fMRI time series. *NeuroImage*, 47: S167.
- Wagenmakers, E.-J.; Farrell, S.; and Ratcliff, R. 2005. Human cognition and a pile of sand: a discussion on serial correlations and self-organized criticality. *Journal of experimental psychology: General*, 134(1): 108.
- Wagner, P.; Strodthoff, N.; Boussejot, R.-D.; Kreiseler, D.; Lunze, F. L.; Samek, W.; and Schaeffter, T. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data*, 7(1): 1–15.
- Wang, G.; Liu, X.; Ying, Z.; Yang, G.; Chen, Z.; Liu, Z.; Zhang, M.; Yan, H.; Lu, Y.; Gao, Y.; et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10): 2633–2642.
- Wang, W. K.; Chen, I.; Hershkovich, L.; Yang, J.; Shetty, A.; Singh, G.; Jiang, Y.; Kotla, A.; Shang, J. Z.; Yerrabelli, R.; Roghanizad, A. R.; Shandhi, M. M. H.; and Dunn, J. 2022. A Systematic Review of Time Series Classification Techniques Used in Biomedical Applications. *Sensors*, 22(20): 8016.
- Wang, Y.; Huang, N.; Li, T.; Yan, Y.; and Zhang, X. 2024. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37: 36314–36341.
- Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; and Sun, L. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.
- Woo, S.; Qin, Y.; Arik, S. O.; and Pfister, T. 2022. ETSformer: Exponential smoothing transformers for time-series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 22898–22909.
- Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*.
- Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34: 22419–22430.
- Xiong, D.; Zhang, D.; Zhao, X.; and Zhao, Y. 2021. Deep Learning for EMG-based Human-Machine Interaction: A Review. *IEEE/CAA Journal of Automatica Sinica*, 512–533.
- Yang, F.; Li, X.; Wang, B.; Zhang, T.; Yu, X.; Yi, X.; and Zhu, R. 2025. MMSeg: A novel multi-task learning framework for class imbalance and label scarcity in medical image segmentation. *Knowl. Based Syst.*, 309: 112835.
- Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.
- Zhang, S.; Lian, C.; Xu, B.; Su, Y.; and Alhudhaif, A. 2024a. 12-Lead ECG signal classification for detecting ECG arrhythmia via an information bottleneck-based multi-scale network. *Information Sciences*, 662: 120239.
- Zhang, T.; Zhang, Y.; Cao, W.; Bian, J.; Yi, X.; Zheng, S.; and Li, J. 2022. Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures. *arXiv preprint arXiv:2207.01186*.
- Zhang, Y.; Ma, L.; Pal, S.; Zhang, Y.; and Coates, M. 2024b. Multi-resolution time-series transformer for long-term forecasting. In *International Conference on Artificial Intelligence and Statistics*, 4222–4230. PMLR.
- Zhang, Y.; and Yan, J. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*.
- Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Liu, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.
- Zhou, T.; Ma, Z.; Wen, Q.; Yi, X.; and Sun, L. 2022. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, 27268–27286.