# Multi-scale Progressive Gated Transformer for Physiological Signal Classification

**Wei Zhou** WZHOU@MY.SWJTU.EDU.CN
*Southwest Jiaotong University*

**Hao Wang** CSHAOWANG@GMAIL.COM
*Nanyang Technological University*

**Yiling Zhang** ZYLSCIENCE@FOXMAIL.COM
*Southwest Jiaotong University*

**Cheng Long** C.LONG@NTU.EDU.SG
*Nanyang Technological University*

**Yan Yang** YYANG@SWJTU.EDU.CN
*Southwest Jiaotong University*

**Dongjie Wang** WANGDONGJIE@KNIGHTS.UCF.EDU
*University of Central Florida*

**Editors:** Emtiyaz Khan and Mehmet Gönen

## Abstract

Physiological signal classification is of great significance for health monitoring and medical diagnosis. Deep learning-based methods (e.g. RNN and CNN) have been used in this domain to obtain reliable predictions. However, the performance of existing methods is constrained by the long-term dependence and irregular vibration of the univariate physiological signal sequence. To overcome these limitations, this paper proposes a Multi-scale Progressive Gated Transformer (MPGT) model to learn multi-scale temporal representations for better physiological signal classification. The key novelties of MPGT are the proposed Multi-scale Temporal Feature extraction (MTF) and Progressive Gated Transformer (PGT). The former adopts coarse- and fine-grained feature extractors to project the input signal data into different temporal granularity embedding spaces and the latter integrates such multi-scale information for data representation. Classification task is then conducted on the learned representations. Experimental results on real-world datasets demonstrate the superiority of the proposed model.

**Keywords:** Physiological signal classification, Time series data, Multi-scale representation, Transformer.

## 1. Introduction

As a type of time series data, physiological signal data are sparkling in the field of intelligent healthcare with the maturity of sensor technology, especially the increasing usage quantity of wearable devices. Knowledge discovery or pattern recognition from physiological signal data is of great significance for both health monitoring and medical aided diagnosis(Zhang et al., 2021). The classification of physiological signal data is indeed a key and worth studying research task, e.g., sleep-stage classification based on EEG, arrhythmia detection

based on ECG, etc. Since the excellent feature representation ability of deep learning, deep learning-based approaches have recently been investigated for physiological signal classification (Rim et al., 2020). For example, following convolutional neural networks (CNNs) in image classification, CNN-based methods (Huy et al., 2019; Michael et al., 2020) have been designed for the sleep stage classification of physiological signals. Inspired by the time unfolding properties of recurrent neural network (RNN) including its variant (such as LSTM and GRU), RNN-based methods (Zhang et al., 2019; Wang and Zhou, 2019) have been used for physiological signal classification to effectively model temporal patterns. In addition, attention-based model (Eldele et al., 2021) has been proposed for physiological signals classification to capture dynamic dependencies of data by using the attention mechanism.

However, there are certain limitations in most of these existing methods for physiological signal classification. As a typical time series data, physiological signal data collected by sensors with a specific frequency not only have temporal correlation, but also have long-term temporal dependence and irregular time series oscillation. For example, a 30-second EEG signal collected at a frequency of 100 Hz (i.e., 100 time steps sampled per second) eventually forms a long-term sequence with 3000 time steps. Fig. 1 shows a case of sampled EEG signal data. In Fig. 1, the upper image is a sequence of 600 time steps in the EEG signal, and the lower image is a sub-sequence of 100 time steps. From the figure, we can observe some temporal characteristics of such physiological signals from the figure: (1) Comparing Sub-sequence 1 with Sub-sequence 2, the temporal dependence of the relatively stable Sub-sequence 1 is easily ignored, due to the irregular change trend of the long sequence; and (2) When Sub-sequence 1 has been amplified, it is clear that there are multi-scale irregular temporal correlations in this sequence. The reason is due to signal value drastically fluctuates with time on different scales. Although existing methods can model nonlinear relationship of physiological signals, most of them cannot fully exploit the above-mentioned complex temporal characteristics of physiological signals.

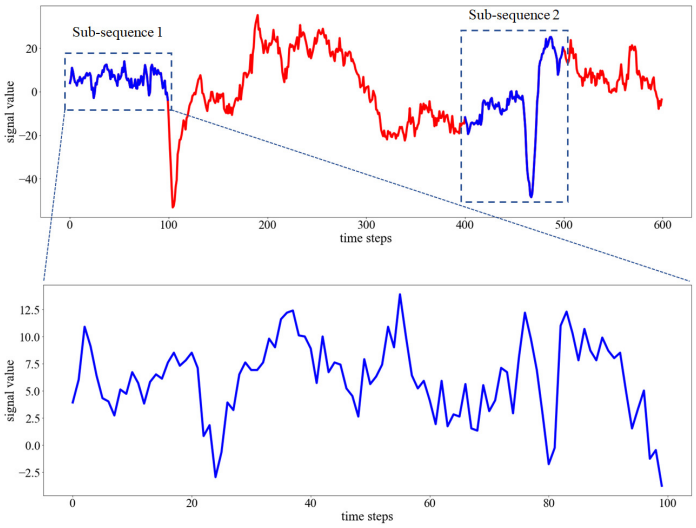


Figure 1: Multi-scale information among EEG signal time series.

To tackle aforementioned limitations and challenges, we propose a novel deep learning model in this paper, called Multi-scale Progressive Gated Transformer (MPGT), which can progressively learn data representation for physiological signals from different temporal scales. Specifically speaking, we first construct a Multi-scale Temporal Features extraction (MTF) module to focus on the raw signal sequence with different temporal ranges (i.e. coarse- and fine-grained). Unlike existing methods, our MTF utilizes convolution kernels of

different sizes to extract coarse- and fine-grained features of sequences and then map them into embedding spaces of different granularities. Besides, MTF adopts pooling operations to retain key time points and those trend changing points. The proposed multi-scale extraction provides a comprehensive encoding of long sequence at multiple scales, enabling the model to effectively capture long-term dependencies. In addition, on the basis of long sequence processing, the irregular temporal relations of physiological signals at different scales is further explored. We then propose a novel Progressive Gated Transformer (PGT) module. PGT exploits two convolution-transformer encoders to capture temporal dependencies from multiple embeddings. One of the two encoders is to learn fine-grained embedding by capturing local fluctuations of signal sequence and another encoder is to learn coarse-grained embedding by mining global variation trend of signal sequence. The two encoders are bridged by a gated unit. In practice, the gated unit integrates fine-grained features into coarse-grained features to progressively fuse different grained temporal information to enhance data representation. The learned temporal representations are then fed into a classifier module to produce the final classification results.

Our main contributions in this paper can be summarized as follows:

- We design a Multi-scale Temporal Feature extraction (MTF) to approach the complex long-term temporal dependencies of physiological signal data. MTF projects physiological sequences into embedding spaces of different granularities to capture multi-scale temporal dependencies of the time series data.
- We propose a novel Progressive Gated Transformer (PGT) to mine and fuse irregular multi-scale temporal information for physiological sequences. Each layer of PGT consists of two convolution-transformer encoders and a gated unit, which can mine local and global multi-scale temporal dependencies and progressively fuse multi-scale information to enhance data representation.
- We conduct extensive experiments on three real-world datasets. Experimental results demonstrate the superiority of our proposed model. In addition, the visualization analysis is performed to verify the effectiveness of the proposed model.

## 2. Related Work

### 2.1. Physiological signal classification

As the maturity of sensor technology (e.g. the popularity of wearable devices) and the rapid development of machine learning, physiological signal classification has attracted more and more attention in recent years. Yang et al. (2018) designed a principal component analysis network (PCANet) for feature extraction, and then exploited a linear support vector machine (SVM) for ECG signal classification. Tuncer et al. (2019) combined discrete wavelet transform (DWT) with a novel ingredient called 1-dimensional hexadecimal local pattern (1D-HLP) for arrhythmia detection in electrocardiogram signals. However, these traditional machine learning methods require hand-crafted feature extraction. In addition, most of them cannot fully tackle and explore the complex nonlinear relationship of physiological signal time series.

Since the success of deep learning applied in other research domains (Fawaz et al., 2019), deep learning-based techniques have also boosted significant improvements in physiological signal classification. For instances, Huy et al. (2019) proposed a joint classification-and-prediction framework based on CNNs for automatic sleep stage classification, and subsequently designed a simple yet efficient CNN architecture to power the framework. Michael et al. (2020) designed a deep CNN architecture for automated sleep stage classification on EEG and EOG signals. Saeed et al. (2020) adopted a novel architecture consisting of wavelet transform and multiple LSTM recurrent neural networks for ECG signals classification on wearable devices. To improve the generalization of the model, Lin et al. (2021) presented a Multi-Branch Network to separate the background features and task features of EEG signals for emotion recognition to achieve better model performance. In addition, Eldele et al. (2021) presented a novel multi-resolution CNN and multi-head attention based deep learning architecture for sleep stages classification on single channel EEG signals. Jia et al. (2020) proposed a novel spatial-spectral-temporal attention based 3D dense network for emotion recognition from EEG signals. Although deep learning-based methods effectively extract the nonlinear relationship of physiological signals, they ignore the long-term dependencies and irregular temporal correlations of physiological signals, resulting in insufficient exploration of the complex temporal characteristics for physiological time series.

### 2.2. Transformer networks

Transformer (Ashish et al., 2017) based on self-attention mechanism has shown great power in many fileds (Wen et al., 2022). For example, Zhou et al. (2021) studied the long-sequence time-series forecasting problem and designed a solid Transformer-based model, called Informer, to predict long sequences. George et al. (2021) presented a novel framework based on Transformer encoder to learn the effective representation of multivariate time series. Kolesnikov et al. (2021) split an image into a sequence of patches and designed a novel model based on Transformer encoder for image classification. Chen et al. (2021a) proposed a dual-branch transformer to learn multi-scale features from image patches of different sizes to produce stronger image features for classification. Wu et al. (2021) introduced the convolution into Vision Transformer architecture to further improve the performance on image recognition tasks in performance and efficiency. Chen et al. (2021b) designed the Spatial-Temporal Transformer Networks consisting of the spatial-temporal Transformer encoder and the temporal Transformer to model the spatial and temporal dependencies for trajectories prediction. Xu et al. (2021) investigated a Spatial-Temporal Transformer Networks to learn dynamical directed spatial dependencies and long-range temporal dependencies for long-term traffic prediction.

Motivated by the powerful ability of Transformer for data representation, in this paper, we explore Transformer and propose a Multi-scale Progressive Gated Transformer model to learn temporal features at different scales for physiological signal classification.

## 3. Method

### 3.1. Problem Formulation

Given a raw physiological signal time series $\mathbf{S} = \{s_1, s_2, ..., s_i, ..., s_t\} \in \mathbb{R}^t$, $t$ denotes the number of time steps and $s_i$ is the signal value at the $i$-th time step. Our goal is to learn an end-to-end deep neural network model to classify such a physiological signal time series, formulated as follows:

$$\mathbf{Y} = \mathcal{F}(\mathbf{S}|\Theta) \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^C$ is the final classification output distribution, $C$ denotes the number of class, $\mathcal{F}$ denotes the proposed neural network model, and $\Theta$ is the parameters of the model.

### 3.2. Framework Overview

The overall architecture of the proposed MPGT model is shown in Fig. 2. The model mainly consists of three modules: (1) Multi-scale Temporal Feature extraction (MTF) module, which embeds the raw sequences at different temporal scales to obtain multiple embeddings; (2) Progressive Gated Transformer (PGT) module, which learns informative knowledge from multi-scale embeddings and progressively integrates them; and (3) Classifier module that employs a fully connected layer and a softmax function to calculate the probability distribution of each class for data.

### 3.3. Multi-scale Temporal Feature extraction

Physiological signal time series have a rich temporal structure over multiple time scales. We also note that multi-scale learning has shown promising performance in other domains (Wang et al., 2019; Qian et al., 2020). Therefore, we motivate and propose a Multi-scale Temporal Feature extraction (MTF) module. As shown in Fig. 2, the MTF module (the bottom part in Fig. 2) adopts coarse- and fine-grained feature extractors. Both two feature extractors use the same network structure to encode the temporal information at different scales into multiple embeddings.

We denote that the extracted features are different as the used convolution kernels are with different scales. Convolution with large kernel size can capture relatively holistic features while it may miss detailed (i.e., fine-grained) features. On the contrary, convolution with small kernel size can capture fine-grained features more effectively. In our model, the convolution with small kernel size is used to extract those high-frequency temporal features of physiological sequences to produce fine-grained embeddings. The convolution with large kernel size is utilized to extract those low-frequency temporal features of physiological sequences to generate coarse-grained embeddings. We also apply pooling techniques to preserve information at important time points and trends in sequences with different scales. Next, we take fine-grained feature extractor as an example. The fine-grained feature extractor is a network stacked by a Conv1D(64, 50, 6), a Maxpooling, a Conv1D(128, 8, 1) and an Avgpooling. The Conv1D (64, 50, 6) refers to using 1D convolution layer with 64 filters, a kernel size of 50 and a stride of 6. Maxpooling is the max pooling layer. Avgpooling is the average pooling layer. Given the original physiological signal sequence $\mathbf{S} \in \mathbb{R}^t$, the process

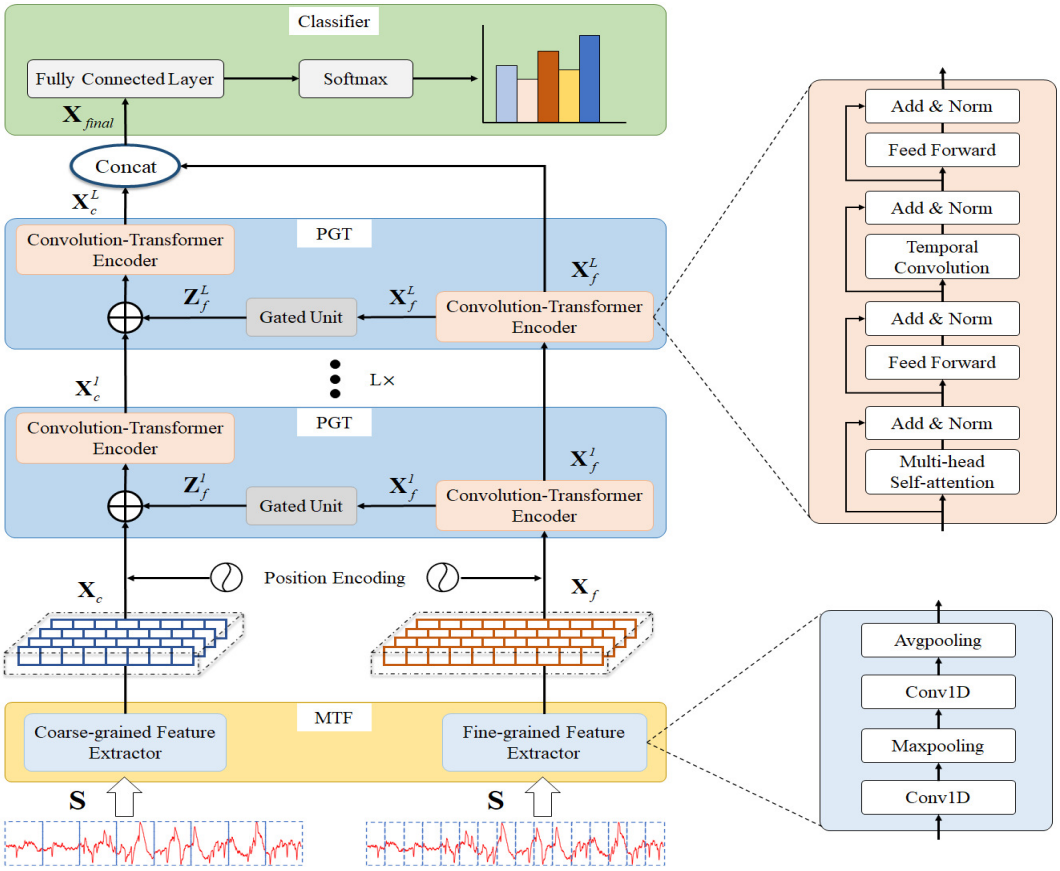


Figure 2: The overall architecture of Multi-scale Progressive Gated Transformer (MPGT)

of feature extraction can be formulated as:

$$\begin{aligned}
\mathbf{X}_1 &= GELU(\mathbf{W}_1 * \mathbf{S} + b_1) \\
\mathbf{X}_2 &= Maxpooling(\mathbf{X}_1) \\
\mathbf{X}_3 &= GELU(\mathbf{W}_3 * \mathbf{X}_2 + b_3) \\
\mathbf{X}_f &= Avgpooling(\mathbf{X}_3)
\end{aligned} \tag{2}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ represent convolutional kernels, $b_1$ and $b_2$ are their biases, and the symbol $*$ denotes convolution operation. $\mathbf{X}_f \in \mathbb{R}^{n \times d_f}$ is the extracted fine-grained embedding, where $n$ is the number of features, and $d_f$ is the dimension of features. Activation function use Gaussian Error Linear Unit (GELU).

Similarly, coarse-grained embedding $\mathbf{X}_c \in \mathbb{R}^{n \times d_c}$ is achieved by sampling low-frequency features through a coarse-grained feature extractor, which is composed of a Conv1D(64, 400, 50), a Maxpooling, a Conv1D(128, 7, 1) and an Avgpooling. In addition, to inject the temporal-position information of sequence, we add relative positional encodings to the embeddings of different granularity respectively, which is helpful in extrapolating to long sequences (Dai et al., 2019).

### 3.4. Progressive Gated Transformer

There are irregular and drastic oscillations in physiological signals at different scales. In order to effectively characterize the irregular temporal dependencies of physiological time series, we propose a Progressive Gated Transformer, as shown in the middle part of Fig. 2. Our PGT module utilizes two convolution-transformer encoders to learn coarse- and fine-grained information respectively in multiple embeddings. Meanwhile, we design a gated unit between the two convolution-transformer encoders to progressively aggregate information of different granularities. In practice, we stack multiple PGT modules to learn more informative knowledge at different scales.

The convolution-transformer encoder, which introduces convolution operations on the basis of the transformer encoder architecture, aims to capture the irregular vibration relationship of physiological signal. It is composed of a multi-head self-attention layer, a temporal convolution network module and two feed forward network modules, each of which is set with residual connection and LayerNorm to help build deep model.

**Multi-head Self-Attention (MSA).** Self-attention mechanism is a core of Transformer, which can effectively capture the interaction within the sequence to learn crucial context information. Given an input sequence data $\mathbf{X} \in \mathbb{R}^{n \times d}$, Self-attention mechanism first linearly projects the data into three vectors, i.e., the quires $\mathbf{Q} = \mathbf{W}_q\mathbf{X} \in \mathbb{R}^{n \times d_k}$, the keys $\mathbf{K} = \mathbf{W}_k\mathbf{X} \in \mathbb{R}^{n \times d_k}$ and the values $\mathbf{V} = \mathbf{W}_v\mathbf{X} \in \mathbb{R}^{n \times d_v}$, where $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are the learnable parameters and $n$ is the number of input features. $d$, $d_k$, $d_v$ denote the dimensions of inputs, keys (or queries) and values, respectively. Then, the single head self-attention operation is conducted on query-key-value triplet $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$:

$$\mathbf{H} = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{3}$$

For multi-head self-attention, the queries, keys and values are split to $p$ parts i.e, $[\mathbf{Q}_1, ..., \mathbf{Q}_p]$, $[\mathbf{K}_1, ..., \mathbf{K}_p]$ and $[\mathbf{V}_1, ..., \mathbf{V}_p]$. They are performed the attention function in parallel and the output features of each head are concatenated to obtain the final output:

$$MSA(\mathbf{X}) = Concat(\mathbf{H}_1, ..., \mathbf{H}_p)\mathbf{W}_o \tag{4}$$

where $Concat()$ is the concatenation operation, and $\mathbf{W}_o$ is the linear projection parameters.

**Temporal Convolution Network (TCN).** In the convolution-transformer encoder, MSA plays a leading role in capturing global context information and TCN can extract local features. Inspired by (Gulati et al., 2020; Yan et al., 2021), we employ a temporal convolution network to plug convolution operation into the structure of transformer. As shown in Fig. 3, the TCN consists of three 1D convolution layers (Conv1D), two batch normalization layers (BN) that are deployed following the convolution to aid training deep model. In addition, to improve the ability to model signal sequences, the Gated Linear Units (GLU) are equipped behind the first two convolution layers to replace the sigmoid activation function.

**Feed Forward Network (FFN).** Feed forward network built with two fully connected layers and $Swish$ activation function is applied subsequently after the multi-head self-attention layer and temporal convolution network for feature transformation and nonlinearity. The FFN layer is formulated as below:

$$FFN(\mathbf{X}) = \mathbf{W}_2(Swish(\mathbf{W}_1\mathbf{X} + b_1)) + b_2 \tag{5}$$

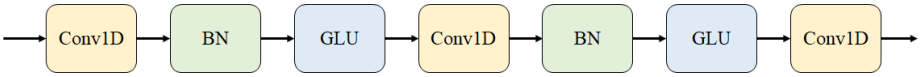


Figure 3: The structure of Temporal Convolution Network

In each Progressive Gated Transformer (PGT), there are two branches in which one branch operates across the coarse-grained temporal embeddings and another one performs on fine-grained temporal embeddings. For the fine-grained embedding $\mathbf{X}_f$, we utilize a convolution-transformer encoder to explore the temporal relation of local irregular fluctuations in physiological signal. The learning process of each PGT is formulated as follows:

$$\begin{aligned}
\mathbf{X}'_f &= MSA(LN(\mathbf{X}_f)) + \mathbf{X}_f \\
\mathbf{X}''_f &= FFN(LN(\mathbf{X}'_f)) + \mathbf{X}'_f \\
\mathbf{X}'''_f &= TCN(LN(\mathbf{X}''_f)) + \mathbf{X}''_f \\
\mathbf{X}^1_f &= FFN(LN(\mathbf{X}'''_f)) + \mathbf{X}'''_f
\end{aligned} \tag{6}$$

where $LN()$ is the LayerNorm used for stable training and faster convergence, $\mathbf{X}^1_f$ is the output of fine-grained features in the first-layer PGT. More generally, we denote that $\mathbf{X}^i_f$ is the output of fine-grained features in the $i$-th layer PGT.

As previously analyzed, there are some subsequences with relatively stable trends but containing important information in the long-term physiological signal time series. If the coarse- and fine-grained features are learned separately, some important information may be missing. Therefore, we consider fusing the learned fine-grained features into coarse-grained features, which helps to amplify the effects of those subsequences with smooth trends. The most straightforward method is to simply add the fine- and coarse-grained features. However, the direct addition of the features of different scales easily leads to information redundancy, which degrades the performance. We thus design a gated unit whose pipeline is illustrated in Fig. 4. Firstly, the learned fine-grained feature $\mathbf{X}^1_f$ is projected to the coarse-grained feature space by two mappings. Meanwhile, the *tanh* and *sigmoid* functions are adopted to guarantee the nonlinear dependence of features and control which information needs to be fused. Secondly, the learned attention vector $\mathbf{A}_s$ is used to update the fine-grained features by element-wise product. The calculation process can be expressed as the following formulas:

$$\begin{aligned}
\mathbf{A}_t &= tanh(\mathbf{W}_t\mathbf{X}^1_f + b_t) \\
\mathbf{A}_s &= \sigma(\mathbf{W}_s\mathbf{A}_t + b_s) \\
\mathbf{Z}^1_f &= \mathbf{A}_s \otimes (\mathbf{W}_{xz}\mathbf{X}^1_f + b_{xz})
\end{aligned} \tag{7}$$

where $\mathbf{Z}^1_f$ is the updated fine-grained feature in the first layer PGT. $\otimes$ denotes the element-wise product between matrices.

Then, the updated fine-grained feature is integrated into coarse-grained embedding $\mathbf{X}_c$ by element-wise addition. Next, the fused coarse-grained representation is fed into another convolution-transformer encoder to learn the global variation trend of the sequence. The

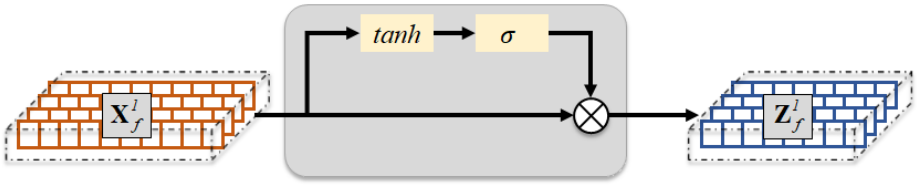


Figure 4: The pipeline of the gated unit

process can be formulated as follows:

$$\begin{aligned}
\mathbf{X}_c &= \mathbf{X}_c \oplus \mathbf{Z}_f^1 \\
\mathbf{X}_c' &= MSA(LN(\mathbf{X}_c)) + \mathbf{X}_c \\
\mathbf{X}_c'' &= FFN(LN(\mathbf{X}_c')) + \mathbf{X}_c' \\
\mathbf{X}_c''' &= TCN(LN(\mathbf{X}_c'')) + \mathbf{X}_c'' \\
\mathbf{X}_c^1 &= FFN(LN(\mathbf{X}_c''')) + \mathbf{X}_c'''
\end{aligned} \tag{8}$$

where $\oplus$ is the element-wise addition between matrices, $\mathbf{X}_c^1$ is the output of coarse-grained features in the first-layer PGT.

Finally, to better classify physiological time series, we concatenate the coarse-grained features focusing on sequence variation trends and the fine-grained features concentrating on time point fluctuations for physiological signal representation enhancement. The final representation is denoted as $\mathbf{X}_{final} = [\mathbf{X}_c^L, \mathbf{X}_f^L]$, where $\mathbf{X}_c^L$ and $\mathbf{X}_f^L$ are the output of coarse- and fine-grained temporal features in the last PGT module.

### 3.5. Classification

In the end, we construct a classifier using a fully connected layer with a softmax activation function (see the top part in Fig. 2). The classifier takes the final learned representation $\mathbf{X}_{final}$ as input to calculate the probability distribution of the class and outputs the classification result. The proposed model is optimized by minimizing the cross-entropy loss between the true class distribution and the predicted class distribution, formulated as:

$$Loss = \sum_{i=1}^{C} \mathbf{y}_i \log(\hat{\mathbf{y}}_i) \tag{9}$$

where $C$ is the number of classes, $\mathbf{y}_i$ is the true class distribution and $\hat{\mathbf{y}}_i$ is the predicted class distribution.

## 4. Experiments

### 4.1. Datasets and Evaluation Metrics

To verify the validity of the proposed model, three real-world physiological signal datasets are used to evaluate the performance of our model. The lengths of EEG and ECG samples are 3000 and 256 time steps, respectively. Table 1 shows the summary of the three datasets. Below is a brief introduce to each dataset.

Table 1: Summary of Three Physiological Signal Datasets

| Dataset | Signal Type | Train Samples | Test Samples | Classes |
|---|---|---|---|---|
| Sleep-EDF-20 | EEG | 33846 | 8462 | 5 (8285/2804/17799/5703/7717) |
| Sleep-EDF-78 | EEG | 156383 | 39096 | 5 (65951/21522/69132/13039/25835) |
| MIT-BIH | ECG | 22827 | 4526 | 6 (11167/6731/3612/2228/787/2828) |

**Sleep-EDF-20**[1]: Sleep-EDF-20 collects the whole-night polysmnographic (PSG) sleep recordings files from 20 subjects. Each PSG file contains two EEG channels (from Fpz-Cz and Pz-Oz electrode locations), one EOG channel (horizontal) and one EMG channel (from submental chin). Following previous studies (Eldele et al., 2021), we select the EEG channel from Fpz-Cz with a sampling rate of 100 Hz as the raw single-channel signal time series. The data is divided into five patterns: W, N1, N2, N3 and REM, which are corresponding to the wake stage, three types of non-rapid eye movement, and rapid eye movement stages.

**Sleep-EDF-78**[2]: Sleep-EDF-78 collects the whole-night polysmnographic (PSG) sleep recordings files from 78 subjects. The sampling settings are the same as Sleep-EDF-20.

**MTI-BIH**[3]: MIT-BIH dataset is used for arrhythmia detection and contains 48 records of individuals from different genders and ages. Each record is a 30-minute long ECG recording of heartbeat signals, with a sampling frequency of 360 Hz. In this work, we sample the raw single-channel arrhythmia signal time series with six classes, i.e., N (normal beat), A (Atrial premature beat), Ventricular (Premature ventricular contraction), F (Fusion of ventricular and normal beat), Paced (Paced beat) and Noise (no kind of heart beat).

In this paper, we utilize Accuracy (ACC) and F1-score to evaluate the performance of each model. Given the True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN), ACC and F1-score are defined as follows:

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

$$F1\text{-score} = \frac{2 \times P \times R}{P + R} \tag{11}$$

where $P$ is the $Precision = \frac{TP}{TP+FP}$, $R$ is the $Recall = \frac{TP}{TP+FN}$.

### 4.2. Baselines and Experimental Setup

In our experiments, we compare the proposed model MPGT with eight baselines, including CNN(Lecun et al., 2015), LSTM(Hochreiter and Schmidhuber, 1997) and MCNN(Cui et al., 2016) that are neural network-based methods, Transformer(Ashish et al., 2017), Informer(Zhou et al., 2021) and GTN(Liu et al., 2021) that are self-attention based methods. In addition, we compare two specific physiological signal classification methods, AttnSleep (Eldele et al., 2021) and HADM (Amin et al., 2021), which are applied to sleep staging and arrhythmia detection respectively.

1. https://www.physionet.org/content/sleep-edf/1.0.0/
2. https://www.physionet.org/content/sleep-edfx/1.0.0/
3. https://www.physionet.org/content/mitdb/1.0.0/

For Transformer and Informer, we first utilize two convolution layers to encode features in the raw sequence, then use the Transformer encoder to learn temporal dependencies, and finally employ fully connected layer to obtain classification results. GTN uses two tower structure to model channel-wise and step-wise correlations. We then use it to learn temporal dependencies at different scales for comparison. We built our models on Ubuntu 16.04 using PyTorch and trained all models on a NVIDA TITAN XP 32GB GPU. All experiments were run ten times and their averages were recorded to reduce the effects of experimental randomness.

### 4.3. Performance Comparison

In this section, we evaluate the proposed MPGT and the aforementioned eight baseline methods on three real-world datasets. The comparison results in terms of ACC and F1-score are shown in Table 2. From the results, we make the following observations:

1. The performance of the models with multi-scales is superior to those models without multi-scales. In particular, MCNN achieves better performance than CNN and LSTM. GTN and MPGT outperform Transformer and Informer. It shows that multi-scale learning can capture richer features of long sequences.

2. Self-attention based models (i.e., Transformer, Informer, GTN and AttnSleep) outperform those CNN and RNN based models. This indicates that self-attention mechanism can better learn irregular long-term dependencies of sequences by exploring correlations within the data.

3. Our MPGT has better performance than GTN in terms of ACC and F1-score. The reason is due to MPGT progressively integrates features from different scales to obtain more robust representations.

4. Compared with methods targeting specific physiological signals (i.e., AttnSleep and HADM), MPGT achieves better performance. This is due to the progressive fusion of multi-scale features during learning process.

In conclusion, our MPGT outperforms these state-of-the-art methods comprehensively. The results indicate that our proposed model can capture more informative features and obtain more discriminative representations for physiological time series classification.

### 4.4. Ablation Study

To demonstrate the effectiveness of the design of each module in our model, we conduct ablation experiments with several variant models. The variants of MPGT are set as follows:

- **MPT**: We remove the gated unit in MPGT, which directly adds fine-grained features into coarse-grained features.

- **MPT-1**: The gated unit is removed and the fine-grained features are not added into coarse-grained features (they are just concatenated at the last layer).

- **MPGT-1**: Instead of concatenating the features of different granularity, we directly use the learned coarse-grained feature representation for classification.

Table 2: Performance comparison results of the proposed model and baselines

| Method \ Dataset | Sleep-EDF-20 | | Sleep-EDF-78 | | MIT-BIH | |
|---|---|---|---|---|---|---|
| | F1-score | ACC | F1-score | ACC | F1-score | ACC |
| CNN | 72.13 | 76.38 | 70.13 | 77.16 | 84.17 | 86.75 |
| LSTM | 72.35 | 75.23 | 69.66 | 76.89 | 85.23 | 87.32 |
| MCNN | 73.95 | 80.04 | 71.65 | 78.37 | 85.73 | 90.11 |
| Transformer | 77.52 | 84.01 | 74.65 | 80.52 | 89.67 | 95.89 |
| Informer | 77.09 | 83.99 | 74.35 | 82.35 | 89.16 | 96.13 |
| GTN | 77.84 | 84.30 | 74.84 | 81.14 | 90.20 | 96.32 |
| AttnSleep | 78.01 | 84.41 | 75.06 | 81.31 | – | – |
| HADM | – | – | – | – | 90.72 | 96.53 |
| MPGT | **80.50** | **86.95** | **77.82** | **83.43** | **92.86** | **98.92** |

- **MPGT-T**: In this variant, the convolution-transformer encoder is replaced by the conventional Transformer encoder.
- **MPGT-C**: We replaces our convolution-transformer encoder with Conformer (Gulati et al., 2020), which is also convolution-augmented transformer structure.

Fig. 5 illustrates the classification performance of different variant models. From the figure, we can see that: (1) The ablation of the gated unit from MPGT results that MPT produces redundancy in the process of progressively fusing information of different granularities, which makes its performance significantly inferior to MPGT; (2) Although MPT-1 is slightly better than MPT, it still shows poor performance. This is because it does not sufficiently learn effective information at different scales. (3) In the process of progressive learning, MPGT-1 has fused features of different granularity, which can effectively avoid some important subsequences being ignored, even if their change trends are stable. Note that only using coarse-grained representation to classify may cause some critical time points in the sequence to be lost, which also degrades the model performance. (4) Compared with the conventional Transformer encoder, the convolution-enhanced Transformer structure can learn more informative knowledge. This results that MPGT and MPGT-C outperform MPGT-T; and (5) MPGT-C is competitive with MPGT, because they are both the convolution-Transformer based architecture.

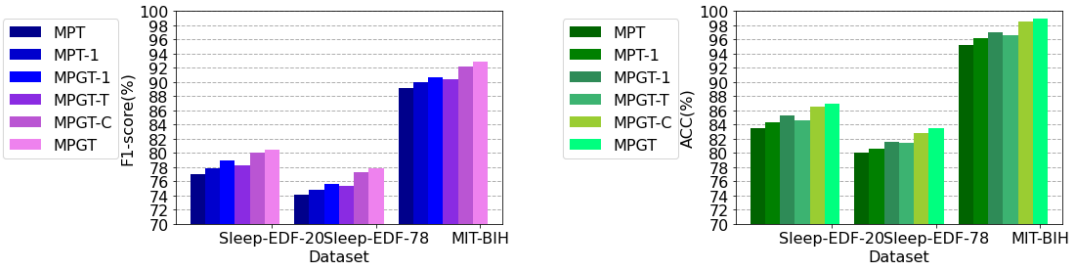


Figure 5: Evaluation of different network modules on all datasets. (Best viewed in color)

### 4.5. Parameters Analysis

In our model, there are mainly two hyperparameters, namely the number of heads $p$ in the multi-head self-attention mechanism and the depth of PGT modules $L$. We set $p$ to $[2, 4, 6, 8, 10]$ and $L$ to $[1, 2, 3, 4, 5]$ respectively to evaluate the performance of the MPGT model. The parameter study results are shown in Fig. 6. From the results, we can observe that our model is not sensitive to the number of heads. In our work, we set the number of heads as 4 because the model performs slightly better when the number of heads is 4. In addition, we can see that as the number of modules increases, the model performance initially improves before it arrives at a peak (say, the number of PGT modules increases to 3). After the peak, the model performance begins to decline and then stabilize. Therefore, in our experiments, the number of PGT modules is set to 3.

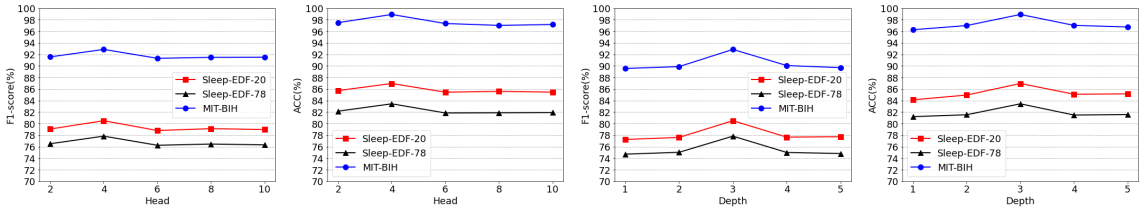


Figure 6: Evaluation in terms of using different numbers of heads and depth

### 4.6. Visualization Results

Fig. 7 shows the visualization results of attention maps at different granularities on Sleep-EDF-20 dataset, where the coordinates denote the sequence lengths of the encoded coarse- and fine-grained features, respectively. As can be seen from the figure, the coarse-grained attention map is relatively sparser. This phenomenon indicates that the learned coarse-grained attention only concentrates on the global change trend of the sequence, and gives more weights to a few crucial subsequences. While the fine-grained attention map is more intensive, this result shows that the learned fine-grained attention tends to focus on the details of the sequence and assign more weights to significant time points in the sequence.

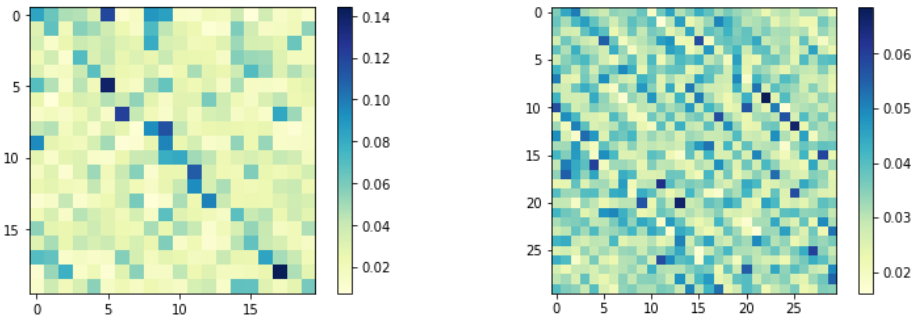

Figure 7: Visualization of coarse-grained attention (*left*) and fine-grained attention (*right*) at different granularity. (Best viewed in color)

## 5. Conclusion

In this paper, we proposed a novel Multi-scale Progressive Gated Transformer model for physiological signal classification. To capture more sufficient temporal dependencies, we first designed a multi-scale temporal feature extraction module to embed physiological signal sequences into different embedding spaces. We then constructed a progressive gated transformer to learn temporal correlation at different scales of physiological serial sequence in a bridge manner. Extensive experiments on three real-world datasets show the effectiveness of the proposed model. In future work, we plan to improve our model for multivariate or multi-channel time series data, or even spatial-temporal data.

## References

Ullah Amin, Rehman Sadaqat ur, Tu Shanshan, Mehmood Raja Majid, Fawad, and Ehatisham ul-haq Muhammad. A hybrid deep cnn model for abnormal arrhythmia detection based on cardiac ecg signal. *Sensors*, 21(3), 2021.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, and Polosukhin Illia. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 1–11, 2017.

Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 357–366, 2021a.

Weihuang Chen, Fangfang Wang, and Hongbin Sun. S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving. In *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157, pages 454–469, 2021b.

Zhicheng Cui, Wenlin Chen, and Yixin Chen. Multi-scale convolutional neural networks for time series classification. *arXiv preprint arXiv:1603.06995*, pages 1735–1780, 2016. doi: arxiv.org/abs/1603.06995.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.

Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33:917–963, 2019.

Zerveas George, Jayaraman Srideepika, Patel Dhaval, Bhamidipaty Anuradha, and Eickhoff Carsten. A transformer-based framework for multivariate time series representation

learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 2114–2124, 2021.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020*, pages 5036–5040, 2020.

Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Phan Huy, Andreotti Fernando, Cooray Navin, Chén Oliver Y, and De Vos Maarten. Joint classification and prediction cnn framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, 66(5):1285–1296, 2019.

Ziyu Jia, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, and Jing Wang. Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 2909–2917, 2020.

Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

Guang Lin, Li Zhu, Bin Ren, Yiteng Hu, and Jianhai Zhang. Multi-branch network for cross-subject eeg-based emotion recognition. In *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157, pages 705–720, 2021.

Minghao Liu, Shengqi Ren, Siyuan Ma, Jiahui Jiao, Yizhou Chen, Zhiguang Wang, and Wei Song. Gated transformer networks for multivariate time series classification. *arXiv preprint arXiv:2103.14438*, 2021. doi: arxiv.org/abs/2103.14438.

Sokolovsky Michael, Guerrero Francisco, Paisarnsrisomsuk Sarun, Ruiz Carolina, and Alvarez Sergio A. Deep learning for automated feature discovery and classification of sleep stages. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6): 1835–1845, 2020.

Xuelin Qian, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Leader-based multi-scale attention deep architecture for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):371–385, 2020.

Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. Deep learning in physiological signal data: A survey. *Sensors*, 20(4), 2020. doi: 10.3390/s20040969.

Saadatnejad Saeed, Oveisi Mohammadhosein, and Hashemi Matin. Lstm-based ecg classification for continuous monitoring on personal wearable devices. *IEEE Journal of Biomedical and Health Informatics*, 24(2):515–523, 2020.

Turker Tuncer, Sengul Dogan, Paweł Pławiak, and U. Rajendra Acharya. Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ecg signals. *Knowledge-Based Systems*, 186, 2019.

Ludi Wang and Xiaoguang Zhou. Detection of congestive heart failure based on lstm-based deep network via short-term rr intervals. *Sensors*, 19(7), 2019. doi: 10.3390/s19071502.

Ning Wang, Jingyuan Li, Lefei Zhang, and Bo Du. Musical: Multi-scale image contextual attention learning for inpainting. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3748–3754, 2019.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022. doi: 10.48550/arXiv.2202.07125.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31, 2021.

Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv.2001.02908*, 2021. doi: 10.48550/arXiv.2001.02908.

Haotian Yan, Zhe Li, Weijian Li, Changhu Wang, Ming Wu, and Chuang Zhang. Contnet: Why not use convolution and transformer at the same time? *arXiv preprint arXiv:2104.13497*, 2021. doi: 10.48550/arXiv.2104.13497.

Weiyi Yang, Yujuan Si, Di Wang, and Buhao Guo. Automatic recognition of arrhythmia based on principal component analysis network and linear support vector machine. *Computers in Biology and Medicine*, 101:22–32, 2018.

Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Transactions on Cybernetics*, 49 (3):839–847, 2019.

Xiaobo Zhang, Yan Yang, Tianrui Li, Yiling Zhang, Hao Wang, and Hamido Fujita. Cmc: A consensus multi-view clustering model for predicting alzheimer's disease progression. *Computer Methods and Programs in Biomedicine*, 199, 2021.

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*, pages 11106–11115, 2021.