Self-Supervised Framework Based on Subject-Wise Clustering for Human Subject Time Series Data

Eunseon Seong*, Harim Lee*, Dong-Kyu Chae

Department of Artificial Intelligence, Hanyang University, Seoul, South Korea
{emilyseong, hrimlee, dongkyu}@hanyang.ac.kr

Abstract

With the widespread adoption of IoT, wearable devices, and sensors, time series data from human subjects are significantly increasing in the healthcare domain. Due to the laborious nature of manual annotation in time series data and the requirement for human experts, self-supervised learning methods are attempted to alleviate the limited label situations. While existing self-supervised methods have been successful to achieve comparable performance to the fully supervised methods, there are still some limitations that need to be addressed, considering the nature of time series data from *human subjects*: In real-world clinical settings, data labels (e.g., sleep stages) are usually annotated by *subject-level*, and there is a substantial variation in patterns between subjects. Thus, a model should be designed to deal with not only the label scarcity but also subject-wise nature of data to ensure high performance in real-world scenarios. To mitigate these issues, we propose a novel self-supervised learning framework for human subject time series data: **S**ubject-**A**ware **T**ime **S**eries **C**lustering (SA-TSC). In the unsupervised representation learning phase, SA-TSC adopts a subject-wise learning strategy rather than instance-wise learning which randomly samples data instances from different subjects within the batch during training. Specifically, we generate subject-graphs with our graph construction method based on Gumbel-Softmax and perform graph spectral clustering on each subject-graph. In addition, we utilize graph neural networks to capture dependencies between channels and design our own graph learning module motivated from self-supervised loss. Experimental results show the outstanding performance of our SA-TSC with the limited & subject-wise label setting, leading to its high applicability to the healthcare industry. The code is available at: https://github.com/DILAB-HYU/SA-TSC

Introduction

Time series data plays an important role in representing various real-life domains and situations (Lee, Park, and Chae 2023; Kim and Chae 2023). However, unlike images or texts, time series data is less human-friendly, making manual annotation significantly time-consuming, laborious, and requires human experts (Eldele et al. 2021). Therefore, designing a machine learning framework that considers the limited labeled data on time series has gained great attention from researchers recently. In this context, Self-Supervised Learning (SSL) methods have been widely adopted.

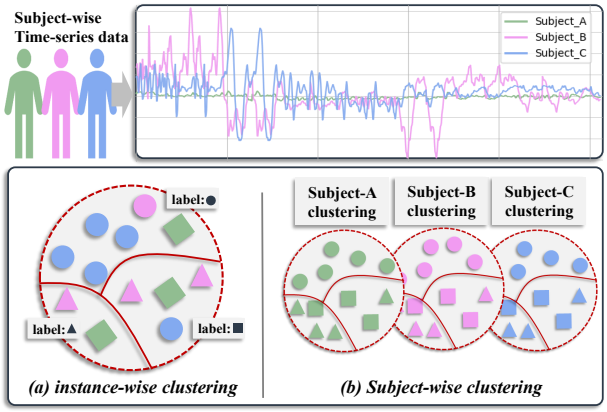

Figure 1: The necessity of subject-wise learning

Especially, as time-series data from human subjects are increasing due to the widespread adoption of wearable devices and sensors (Mohammadi et al. 2018), self-supervised models such as TS-TCC (Eldele et al. 2021) and CA-TCC (Eldele et al. 2022) have gained prominence in areas like human activity and sleep data research. Furthermore, some studies (e.g., SleepDPC (Xiao et al. 2021), CoSleep (Ye et al. 2021), SSLAPP (Lee, Seong, and Chae 2022)) specifically focus on the sleep staging task. Self-supervised learning methods first learn a meaningful representation of data in an unsupervised manner and then fine-tune the model using only a few labels in its downstream task. Remarkably, these strategies exhibit comparable performance to the supervised methods even in the limited labeled settings.

Despite the significant achievements of such self-supervised methods on time series data, there still exist limitations that need to be addressed specifically for *time series data from human subjects*[1] which are vastly collected nowa-

*Both authors contributed equally to this research.

[1]In a clinical setting, time series data is collected from various *subjects* with multiple data instances (e.g., EEG, EOG, EMG), with the collection/annotation process being conducted on a subject-by-subject basis. Here, subjects typically refer to 'patients'. Given the constraints and complexities of such studies, it is often expected that only a very limited number of subjects can be recruited.

days in various domains such as healthcare (Chung et al. 2022; Kwon et al. 2021). Considering that data labels such as sleep stages are typically annotated at the subject level, and there exists a significant variation in patterns among different subjects, we claim that the learning strategy needs to be re-designed into a subject-oriented manner. In the upper part of Figure 1, we represent instances from the Human Activity Recognition dataset, each with identical labels but from different human subjects. This reveals that even with consistent labels, time series patterns can be distinct depending on individuals. We thus emphasize the need for special consideration to this domain, designating it as '**human subject time series data**'. Previous studies focus on the general nature of time series, overlooking the specific characteristics inherent to human subject time series data. These studies have focused on individual data instance units, without the consideration of *subject* where such instances belong to.

Considering these inherent individual characteristics in human subject time series data, we propose a subject-wise clustering strategy for representation learning. Fig 1. (a) and (b) compare instance-wise clustering and subject-wise clustering for representation learning. The instance-wise method, adopted from prior works, performs representation learning in the full data space which leads to incorporating inter-subject variance. In contrast, our proposed subject-wise clustering strategy considers the subject characteristics and reduces inter-subject variance for representation learning.

Furthermore, we argue that for optimal performance in real-world scenarios, the model must address not only label scarcity but also subject-wise labels, since real-world data annotations are typically performed subject-wise, not instance-wise. Hence, we assert the importance of verifying the model performance in limited & subject-wise label setting configuration to confirm its applicability.

From this perspective, we propose **S**ubject-**A**ware **T**ime **S**eries **C**lustering (SA-TSC): a novel self-supervised framework for human subject time series data based on a subject-aware learning strategy. Our subject-wise learning is expected to enable the model to extract representations within the subject context by reducing *inter-subject variance*, which time series data from human subjects may have due to individual human characteristics. This helps the model learn discriminative representation in the subject space (the embedding space of each subject graph representation obtained by subject-wise learning; the space for learning robust representations to inter-subject variance), which we expect results in an outstanding performance in the limited & subject-wise label setting.

Technically, SA-TSC is composed of a subject-wise clustering framework for representation learning and a fine-tuning model with few subject-level labels. The clustering framework is composed of representation extraction from each data instance within a subject, followed by representation graphs of each subject for spectral clustering. We design spatial graph neural networks to capture multi-channel dependencies and use positive and negative pairs for spatial graph representation learning. Here, we employ a positive pairwise loss with a graph augmentation technique that leverages cross-domain information integrating both time domain and frequency domain data. Furthermore, we propose an agreement loss with explicit positive and negative pairs within the batch. Then, to enable clustering for each subject, we embed data within the same subject into a graph structure and perform graph spectral clustering. To ensure that the subject graph is well-constructed, we propose a graph construction method utilizing Gumbel-Softmax on the Gaussian kernel adjacency matrix. Experiments on three real-world datasets show that our SA-TSC outperforms the previous self-supervised methods with few subject-level labels given in the fine-tuning stage. Our extensive ablation study highlights the effectiveness of each proposed idea.

Related Work

In recent years, self-supervised learning (Ericsson et al. 2022; Chen et al. 2020; Chen and He 2021; Gao, Yao, and Chen 2021) has been widely introduced to solve limited labeled data issues. The self-supervised method learns representation in an unsupervised manner and fine-tunes the model in a downstream task using few labels. While many of the studies focus on image and text data, there are some studies focusing on the healthcare domain. Self-supervised learning in the healthcare domain focuses on extracting well-learned representation of time series data since healthcare data shows time series patterns collected from human bio-signals. Among them, contrastive learning approach which employs positive and negative pair is one of the most popular methods for representation learning. SleepDPC (Xiao et al. 2021) utilizes segment-based contrasting on an auto-regressive model to learn time-based representation from sleep signals. CoSleep (Ye et al. 2021) proposes a co-training mechanism of multi-view (time view and frequency view), and applies contrastive learning to top-K positive pairs based on the similarity between the two views. SleepDPC and CoSleep are both for sleep dataset which includes various bio-signals such as electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), etc. TS-TCC (Eldele et al. 2021) learns discriminative representation by temporal and contextual contrasting. CA-TCC (Eldele et al. 2022) adopts a class-aware contrasting strategy on TS-TCC. TS-TCC and CA-TCC are developed for general time series data and they conduct experiments also on human activity data and sleep data. While previous approaches learn representations in an instance-wise manner, our method is specifically designed to learn class discriminative representations in the subject space, which is better suited for real-world scenarios.

Motivation

Time series data from human subjects has the potential to include individual characteristics leading to distinct patterns between subjects, which we refer to as inter-subject variance.

Figure 2 illustrates the distinct time series patterns between human subjects in sleep staging. For example, subjects with insomnia exhibit a higher frequency of the 'awake' stage than healthy patients. To account for such

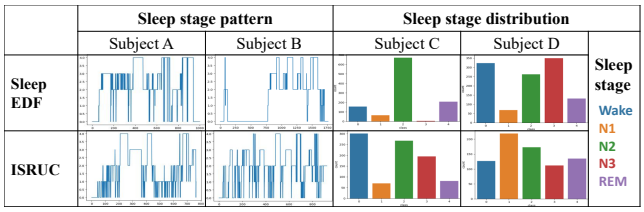

Figure 2: Different time series patterns between individuals.

Notation	Description
$\mathbf{X}$	time series input data
$\mathbf{H}$	Spatial representation of $\mathbf{X}$
$\mathbf{S}$	Class assignment of $\mathbf{X}$
$\mathcal{G}$	Spatial Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$
$\mathcal{V}$	Multi-channel node set
$\mathcal{E}$	Edge connection between node set $\mathcal{V}$
$\mathbf{A}$	Adjacency matrix of $\mathcal{G}$
$\tilde{\mathcal{G}}$	Positive pair $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \mathcal{E}, \mathbf{A})$
$\tilde{\mathcal{V}}$	Multi-channel cross-domain node set
$\mathbf{G}$	Subject Graph
$\mathbf{W}$	Adjacency matrix of $\mathbf{G}$

Table 1: List of notations defined in the paper.

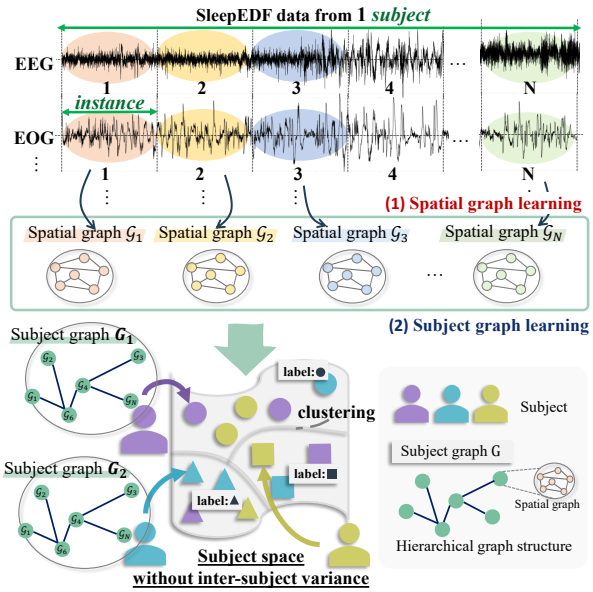

Figure 3: (1) The spatial graph $\mathcal{G}$ acting as a feature extractor to capture the *intrinsic relationship* among multi-channel sensor data. (2) The subject graph $\mathbf{G}$ for subject-wise representation learning avoiding inter-subject variance.

inter-subject variance within a dataset, we design the subject-wise learning strategy. Its main idea is to form batches with only instances belonging to the same subject, whereas prior works form batches with randomly sampled instances from various subjects. In addition, in real-world scenarios, labels are very scarce and likely exist on a per-subject (patient) basis since annotation is performed in a per-subject manner. For example, sleep experts label each epoch (30 seconds) into a sleep stage considering the comprehensive trend of the sleep cycle during the patient's entire sleep recordings. However, prior works do not address this real-world scenario and just randomly sample instances from various subjects for model finetuning. We argue that the finetuning setting must be re-designed to be subject-oriented, to demonstrate the model's generalization to new subjects given limited & subject-wise labeled data for real-world application.

Proposed Method

We propose a graph-based time series clustering framework for subject-wise representation learning, which we refer to as SA-TSC. The overall architecture of SA-TSC is illustrated in Figure 4. We describe our proposed method across two primary sections. In the first section, we define and outline the model structure, and in the second section, we delve into the specifics of the training process.

Model Statement

Our model is composed of two distinct modules: 1) **Spatial Graph**, which is designed to capture multi-channel dependencies of individual data instance units, and 2) **Subject Graph**, in which individual data instances are represented in a subject-wise graph structure. Our full framework can be viewed as a hierarchical graph structure: each subject graph contains nodes of data instances, and each data instance node represents a graph of multi-channel time series data, where each node corresponds to an individual channel. Figure 3 illustrates the concept of the spatial graph and the subject graph. The notations used in the paper are summarized in Table 1.

1) Spatial Graph We employ graph structure to characterize multi-channel data. Specifically, we represent multi-channel time series data as an undirected unweighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ denotes the set of multi-channel nodes (e.g., EEG Fpz-Cz, EEG Pz-Oz, EOG horizontal, EMG submental). A graph $\mathcal{G}$ has $m = |\mathcal{V}|$ nodes and each node $\{v_1 \dots v_m\} \in \mathcal{V}$ consists of node feature $\mathbf{X} \in \mathbb{R}^{m \times t}$, where t is the length of each data. The edge set $\mathcal{E}$ represents the relationship between the channels based on topological geometry. The connection between channels are represented by the adjacency matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$, where $\mathbf{A}_{i,j} = 1$ iff$(i, j) \in \mathcal{E}$. Here, since each dataset has a different topological characteristic, we carefully design a spatial graph based on the specific characteristics of each dataset. For instance, in the case of the activity recognition dataset which comprises 3 triaxial sensors, we establish connections for each sensor (e.g., x, y, z axis of accelerometer) and each axis (e.g., x-axis of accelerometer, gyroscope)

Spatial representation $\mathbf{H} \in \mathbb{R}^{m \times d}$ is extracted from graph $\mathcal{G}$ through the utilization of a spatial GNN. This spatial GNN employs a 1D convolutional layer (CONV) to encode sequential patterns of the time series data, and a Message Passing (MP) layer to embed channel-wise relationships and capture multi-channel dependencies.

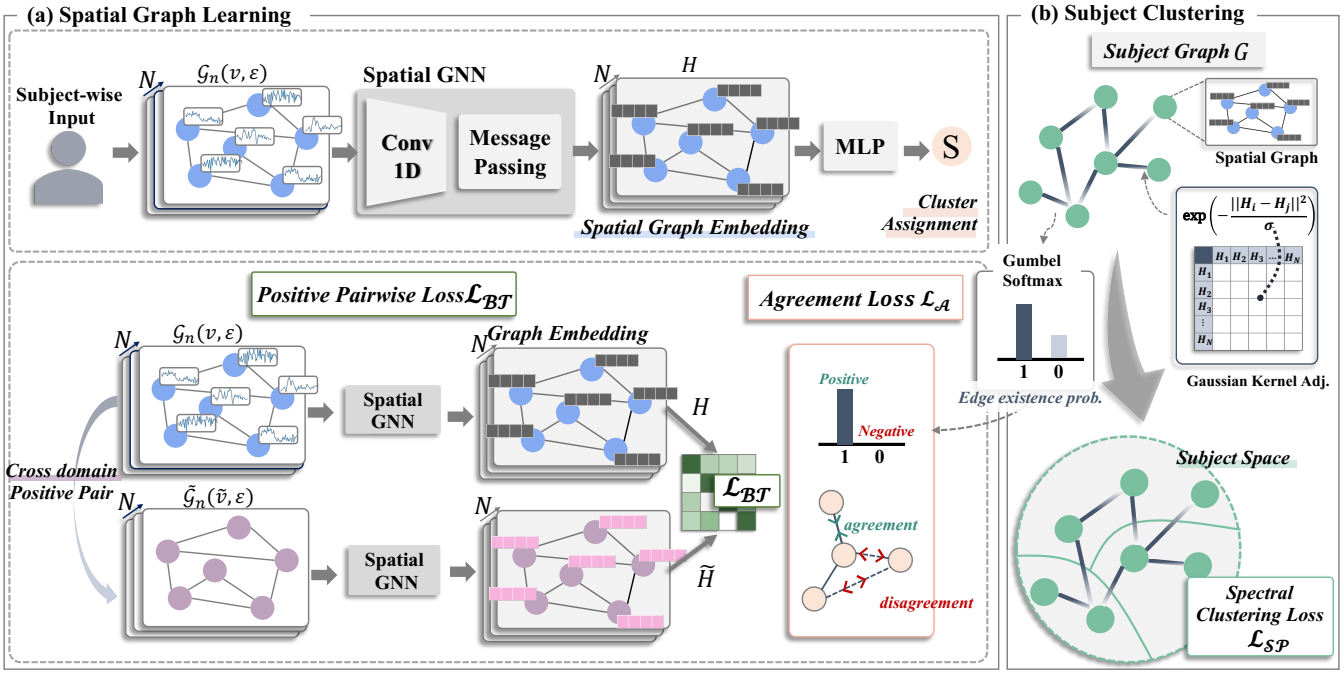

Figure 4: The overall framework of SA-TSC, composed of two components: 1) Spatial representation learning to extract multi-channel time series representation and 2) subject-wise clustering which performs graph spectral clustering on the subject space. Note that the inputs to the model are individual subject's data.

$$\mathbf{X}' = \text{CONV}(\mathbf{X}; \Theta_{\text{CONV}}) \tag{1}$$

$$\mathbf{H} = \text{MP}(\mathbf{X}', \mathbf{A}; \Theta_{\text{MP}}) \tag{2}$$

We use multi-layer perceptron (MLP) on node representation $\mathbf{H}$ to derive cluster assignment $\mathbf{S}$:

$$\mathbf{S} = \text{MLP}(\mathbf{H}; \Theta_{\text{MLP}}) \tag{3}$$

2) Subject Graph In order to perform effective representation learning in the subject space, we introduce a subject-wise graph. The construction of the subject-graph $\mathbf{G}$ is specifically oriented to differentiate subjects, preventing data intermixing from distinct subjects, so that the model ensures learning discriminative representations within the designated subject space.

Graph construction is crucial when embedding real-world data into a graph structure as emphasized by (Qiao et al. 2018). Since the model performance highly depends on the quality of constructed graph, graph construction should be performed carefully. Within this context, we utilize the spatial representation $\mathbf{H}$ to represent individual data instances in subject-graph $\mathbf{G}$.

To be more specific, within total N data instances from individual subjects, we construct a subject-wise graph $\mathbf{G} = (\mathbf{H}, \mathbf{W})$ as a weighted graph with the spatial representation node feature $\mathbf{H} \in \mathbb{R}^{B \times d}$, where B is batch size controlling the size of the graph. The adjacency matrix $\mathbf{W}$ represents the edge weight between nodes. In detail, N/B graphs are derived for each subject, which may differ between subjects depending on N because each subject can have different total instances.

Here, we propose a graph construction method based on Gumbel-Softmax and adjacency normalization in order to construct a more representative graph structure. First, we construct the initial adjacency matrix W_{ij} using the Gaussian kernel (Babaud et al. 1986), a non-linear function of the Euclidean distance. $W_{ij} = \exp\left(-\frac{||\mathbf{H}_i - \mathbf{H}_j||^2}{\sigma}\right)$ represents the edge weight between node $\mathbf{H}_i$ and $\mathbf{H}_j$, where σ is the length scale parameter for the Gaussian kernel. Then, Gumbel-Softmax distribution is computed to determine the probability of edge existence. Let π^1_{ij} be the initial logit of edge existence derived from W_{ij}, and $\pi^0_{ij} = 1 - \pi^1_{ij}$ be its opposite probability. The probability y^1_{ij} of edge existence is calculated by:

$$y^1_{ij} = \frac{\exp((\log(\pi^1_{ij}) + g^1_{ij})/\tau)}{\Sigma^1_{p=0}\exp((\log(\pi^p_{ij}) + g^p_{ij})/\tau)}, \tag{4}$$

where g^p are samples drawn from $Gumbel(0, 1)$. Finally, we generate $\mathbf{W}_{ij}$ by incorporating both the similarity and the probability of edge existence:

$$\mathbf{W}_{ij} = W_{ij} \otimes y^1_{ij}, \tag{5}$$

where $\otimes$ stands for element-wise multiplication.

We execute spectral clustering on individual subject-wise graphs $\mathbf{G}$, which we provide detailed information in the 'training procedure' section.

Training Procedure

In the following section, we detail the training procedure to optimize our model defined in the previous section. The training process pursues two key objectives. First, the spatial Graph Neural Network (GNN) (Scarselli et al. 2008) is optimized to effectively represent individual multi-channel time series data instances, which is achieved through a self-supervised loss with positive and negative pairs. Secondly, we apply spectral clustering on the subject graph $\mathbf{G}$ to generate a more comprehensive latent representation in the subject space.

Spatial Graph Learning Module We design our spatial graph learning module to capture class-discriminant representations in unlabeled graphs motivated by self-supervised learning. The loss term for spatial graph learning is composed of two objectives $\mathcal{L}_{\mathcal{BT}}$, positive pairwise loss and $\mathcal{L}_{\mathcal{A}}$, agreement loss.

First, for **positive pairwise loss $\mathcal{L}_{\mathcal{BT}}$,** we apply a simple but efficient graph augmentation technique for positive pair generation. Based on our initial spatial graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ we generate $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \mathcal{E}, \mathbf{A})$ by utilizing 'cross-domain' (e.g., frequency domain for time domain, and vice versa) knowledge to integrate both time and frequency domain knowledge of time series data. The node $\tilde{\mathcal{V}} = \{\tilde{v}_1, \ldots, \tilde{v}_m\}$ is generated by $\tilde{v}_m = f(\mathbf{X}_m)$, where $f(\cdot)$ transforms the input data into the cross-domain view.

Based on the spatial representation $\mathbf{H}$ and $\tilde{\mathbf{H}}$ from $\mathcal{G}$ and $\tilde{\mathcal{G}}$, we apply the positive pairwise loss for spatial representation learning inspired by (Zbontar et al. 2021):

$$\begin{aligned} \mathbf{H} &= \text{MP}(\text{CONV}(\mathbf{X}, \mathbf{A}; \Theta_{\text{CONV}}); \Theta_{\text{MP}}) \\ \tilde{\mathbf{H}} &= \text{MP}(\text{CONV}(f(\mathbf{X}), \mathbf{A}; \Theta_{\text{CONV}}); \Theta_{\text{MP}}) \end{aligned} \tag{6}$$

Then, we compute the cross-correlation matrix $\mathbf{C} \in \mathbb{R}^{m \times m}$, which represents the cross-correlation between the spatial representation features:

$$\mathbf{C}_{ij} = \frac{\Sigma_m z_{m,i} \Sigma_m \tilde{z}_{m,j}}{\sqrt{\Sigma_m (z_{m,i})^2} \sqrt{\Sigma_m (\tilde{z}_{m,i})^2}} \tag{7}$$

where z and $\tilde{z}$ are the normalized vectors of $\mathbf{H}$ and $\tilde{\mathbf{H}}$, respectively. The normalization sets the mean to zero and the standard deviation to one.

Finally, the positive pairwise loss $\mathcal{L}_{\mathcal{BT}}$ is composed of the two terms: one for making the spatial representation invariant to the augmentation and the other for reducing redundancy between its features:

$$\mathcal{L}_{\mathcal{BT}} = \sum_i (1 - \mathbf{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} {\mathbf{C}_{ij}}^2 \tag{8}$$

where λ is a positive constant controlling the importance between loss terms.

Following the positive pairwise loss, the **agreement loss** $\mathcal{L}_{\mathcal{A}}$ is introduced to enhance the prior loss with explicitly defined positive and negative pairs. While augmentation is applied previously for positive pair generation, here we leverage the utilization of variable y^1_{ij} derived by Eq. (4) to define contrastive pairs. Based on the probability of edge existence y^1_{ij} between data points, we assign a binary value $\mathbf{y}_{ij} \in \{0, 1\}$ to specify positive and negative pairs.

$$\mathbf{y}_{ij} = \begin{cases} 1, & \text{if } y^1_{ij} \geq 0.5 \\ 0, & \text{if } y^1_{ij} < 0.5 \end{cases} \tag{9}$$

Based on Eq. (9), if $\mathbf{y}_{ij}$ is 1, indicating that the probability of edge existence is greater than its non-existence, $\mathbf{y}_{ij}$ is 1, $\mathbf{H}_i$ and $\mathbf{H}_j$ are assigned as positive pairs; otherwise, negative pairs.

To achieve the fundamental essence of the agreement loss, which aims to maximize the alignment among positive pairs while simultaneously mitigating the alignment among negative pairs, the agreement loss $\mathcal{L}_{\mathcal{A}}$ is derived by,

$$\mathcal{L}_{\mathcal{A}} = \sum_i \frac{\Sigma_j \mathbb{1}_1(\mathbf{y}_{ij})(\mathbf{S}_i - \mathbf{S}_j)(sg(\mathbf{H}_i) - sg(\mathbf{H}_j))}{\Sigma_j \mathbb{1}_0(\mathbf{y}_{ij})(\mathbf{S}_i - \mathbf{S}_j)(sg(\mathbf{H}_i) - sg(\mathbf{H}_j))}. \tag{10}$$

Here, sg denotes for 'stop-gradient' operator which restricts gradient flow, thereby maintaining the model parameter from being updated. We apply stop-gradient on spatial representation $\mathbf{H}$ so that the model enforces the learning of cluster assignment $\mathbf{S}$ based on $\mathbf{H}$.

Finally, the spatial graph module is optimized by a combined loss function given by,

$$\mathcal{L} = \mathcal{L}_{\mathcal{BT}} + \lambda_{agree} \cdot \mathcal{L}_{\mathcal{A}}, \tag{11}$$

Algorithm 1: Training procedure for SA-TSC.

Input: Time series input data $X = \{X\}_{i=1}^N$, Adjacency matrix $\mathbf{A}$, Message passing network MP, 1D Convolutional layer CONV, multi-layer perceptron MLP, agreement loss weight λ_{agree}, spectral clustering loss weight $\lambda_{\mathcal{SP}}$, batch size B

for #epoch **do**
  {# Spatial Graph Learning}
  for sampled batch $\{X_i\}_{i=1}^B$ **do**
    Construct positive pair:
    $\mathbf{H}_i = \text{MP}(\text{CONV}(\mathbf{X}_i, \mathbf{A}; \Theta_{\text{CONV}}); \Theta_{\text{MP}})$
    $\tilde{\mathbf{H}}_i = \text{MP}(\text{CONV}(f(\mathbf{X}_i), \mathbf{A}; \Theta_{\text{CONV}}); \Theta_{\text{MP}})$
    Compute Positive Pairwise loss $\mathcal{L}_{BT}$ (Eq.8)
  end for
  Compute edge existence (Eq.4)
  for sampled batch $\{X_i\}_{i=1}^B$ **do**
    Specify positive and negative pairs (Eq.9)
    Compute cluster assignment:
    $\mathbf{S}_i = \text{MLP}(\mathbf{H}_i; \Theta_{\text{MLP}})$
    Compute Agreement loss $\mathcal{L}_A$ (Eq.10)
  end for
  {# Subject Graph Learning}
  Given $\mathbf{H} = \{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_B\}$,
  Compute adjacency matrix $\mathbf{W}$ (Eq.5)
  Construct Subject graph $\mathbf{G} = (\mathbf{H}, \mathbf{W})$
  Compute spectral clustering loss $\mathcal{L}_{\mathcal{SP}}$ (Eq. 12)
  if epoch $\leq t$ **then**
    $\mathcal{L} = \mathcal{L}_{\mathcal{BT}} + \lambda_{agree} \cdot \mathcal{L}_{\mathcal{A}}$
  else
    $\mathcal{L} = \mathcal{L}_{\mathcal{BT}} + \lambda_{agree} \cdot \mathcal{L}_{\mathcal{A}} + \lambda_{spectral} \cdot \mathcal{L}_{\mathcal{SP}}$
  end if
end for

where hyperparameter λ_{agree} controls the importance of the agreement loss term.

Subject Graph Clustering Module In order to perform our learning strategy in the subject space, we adopt spectral clustering approach, which is known to be effective in finding distinguishable cluster in the subspace (Von Luxburg 2007; Von Luxburg et al. 2010). We apply the spectral clustering loss $\mathcal{L}_{\mathcal{SP}}$ based on cluster assignment $\mathbf{S}$ so that the inter-cluster distance is maximized and the intra-cluster distance is minimized. Designed from the minCUT problem (Bianchi, Grattarola, and Alippi 2020), $\mathcal{L}_{\mathcal{SP}}$ is defined as:

$$\mathcal{L}_{\mathcal{SP}} = -\frac{Tr(\mathbf{S}^T\mathbf{W}\mathbf{S})}{Tr(\mathbf{S}^T\mathbf{D}\mathbf{S})} + \left\| \frac{\mathbf{S}^T\mathbf{S}}{\|\mathbf{S}^T\mathbf{S}\|_F} - \frac{\mathbf{I}_K}{\sqrt{K}} \right\|_F, \quad (12)$$

where $\mathbf{D}$ is the degree of $\mathbf{G}$ and $\|\cdot\|_F$ stands for the Frobenius norm.

Finally, our final loss $\mathcal{L}$ is minimized with additional hyperparameter $\lambda_{spectral}$ to control the importance of the spectral loss:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\mathcal{BT}} + \lambda_{agree} \cdot \mathcal{L}_{\mathcal{A}} & \text{, if epoch} \leq t \\ \mathcal{L}_{\mathcal{BT}} + \lambda_{agree} \cdot \mathcal{L}_{\mathcal{A}} + \lambda_{spectral} \cdot \mathcal{L}_{\mathcal{SP}} & \text{, otherwise} \end{cases} \quad (13)$$

In the early training phase, the model is optimized solely by the spatial graph learning module, and after a specific epoch t, it is jointly optimized with the spectral clustering loss $\mathcal{L}_{\mathcal{SP}}$. This strategy ensures that a well-learned spatial graph representation is embedded into the subject graph, and the clustering module enforces the model to extract discriminative representation in the subject space considering individual characteristics.

Experimental Setup

Datasets We evaluate our model on three datasets: UCI Human Activity Recognition (UCI-**HAR**) (Anguita et al. 2013), **ISRUC** (Khalighi et al. 2016), and **SleepEDF** (Goldberger et al. 2000). UCI-HAR contains sensor data from 30 subjects performing 6 activities: walking, walking upstairs, downstairs, sitting, standing, and laying, collected from a smartphone consisting of total 9 channels (triaxial acceleration, estimated body acceleration, and angular velocity). ISRUC and Sleep-EDF are polysomnographic (PSG) sleep datasets which we classify into 5 classes: Wake, N1, N2, N3, and REM. ISRUC is consisted of 10 subjects with 6 brain signals (EEG C3-A2, C4-A1, F3-A2, F4-A1, O1-A2, O2-A1), and SleepEDF contains 20 subject with total 39 night recordings from 4 channels (EEG Fpz—Cz, EEG Pz-Oz, EOG, and EMG). Table 2 summarizes the statistics of the datasets used in our paper.

Evaluation Scenarios Each full dataset is split into 80%/20% designated as training and testing datasets respectively. Then, we first perform *unsupervised* representation learning with the full training dataset, and then conduct fine-tuning by using the labeled instances belonging to *the same one or two subjects* by randomly choosing '*1 subject*' or '*2 subjects*' labels from all the subjects involved. For the classification task, the cluster assignment $\mathbf{S}$ from spectral clustering is used as the class prediction label. We report classification accuracy (**ACC**) and **F1**-score for each dataset. Standard normalization is applied to each channel for preprocessing.

Dataset	# Recordings	# Instances	# Classes	# Channels
HAR	30	10,299	6	9
ISRUC	10	8,589	5	6
SleepEDF	39	42,308	5	4

Table 2: Statistics of the datasets used in the experiment. In HAR and ISRUC datasets, the number of subjects is equal to '# Recordings', and in SleepEDF, it is 20 (19 subjects with 2 recordings each, and the rest 1 subject with 1 recording). '# Instances' stands for the number of total instances which eventually represents the total number of nodes that construct the subject-graphs. '# Channels' denotes the number of nodes in the spatial graph.

Baselines We compare the performance of our SA-TSC with several baselines: 1) **Random Initialization**, 2) **CoSleep** (Ye et al. 2021), 3) Time Series Representation Learning via Temporal and Contextual Contrasting (**TS-TCC**) (Eldele et al. 2021), and 4) Class-Aware TS-TCC (**CA-TCC**) (Eldele et al. 2022). The random initialization baseline trains a linear classifier on a frozen and randomly initialized weighted model, representing the lower bound of our learning strategy. Among the baselines, CoSleep is tested on ISRUC and SleepEDF, as its performance is not on a par with the HAR dataset because it is specifically designed for sleep datasets. We fine-tune and compare all the methods using the aforementioned *'1 subject'* or *'2 subjects'* labels.

Results and Analysis

Performance Comparisons

In Table 3, we compare the accuracy and F1-score of our SA-TSC and the baselines on each dataset. SA-TSC outperforms the baselines in the majority of evaluations. Although it slightly trails the baseline on the ISRUC *'2 subjects'* setting, it demonstrates higher stability with a smaller standard deviation. The overall result shows the superiority of our model in the limited & subject-level label setting, which is closer to a real-world scenario. The outperformance is because SA-TSC learns class-discriminative representations in the subject space based on subject-wise learning. In contrast, the baselines learn instance-specific representation, which turns out to be less effective when the subject-level labels are scarce and not easily transferable to other subjects.

Representation Visualization

We represent the t-SNE (Van der Maaten and Hinton 2008) visualization of the learned representations obtained by SA-TSC, TS-TCC, and CA-TCC in Figure 5. Specifically, we plot the representation output of the SleepEDF dataset, with distinct colors signifying the respective ground-truth classes. The result clearly reveals the superiority of SA-TSC, showing a more distinguishable representation.

# Sub	Dataset	HAR		ISRUC		SleepEDF	
		ACC	F1	ACC	F1	ACC	F1
1sub	Random Init	45.90 ± 0.09	34.80 ± 0.02	30.90 ± 0.08	16.00 ± 0.10	55.20 ± 0.02	29.40 ± 0.05
	CoSleep	-	-	45.30 ± 1.60	37.40 ± 5.10	60.00 ± 5.40	45.70 ± 9.40
	TS-TCC	85.07 ± 0.77	84.80 ± 0.87	61.50 ± 0.68	57.96 ± 0.80	70.90 ± 0.52	63.93 ± 0.88
	CA-TCC	87.14 ± 0.63	86.97 ± 0.71	69.27 ± 0.74	65.90 ± 0.70	70.51 ± 0.99	63.20 ± 1.60
	Ours	**91.32 ± 0.90**	**91.17 ± 0.90**	**69.63 ± 0.36**	**67.13 ± 2.43**	**73.11 ± 0.41**	**65.43 ± 0.45**
2sub	Random Init	49.60 ± 0.10	40.10 ± 0.30	31.50 ± 0.10	16.60 ± 0.10	55.60 ± 0.48	30.60 ± 0.14
	CoSleep	-	-	51.20 ± 0.30	49.20 ± 2.10	57.50 ± 4.60	47.20 ± 6.40
	TS-TCC	83.56 ± 0.82	82.73 ± 0.97	64.44 ± 0.99	61.91 ± 1.06	74.27 ± 0.95	66.81 ± 0.71
	CA-TCC	83.85 ± 1.66	82.72 ± 2.19	**73.03 ± 0.96**	**69.94 ± 0.38**	74.92 ± 0.47	65.51 ± 0.34
	Ours	**93.11 ± 1.82**	**93.18 ± 1.90**	72.69 ± 0.41	69.54 ± 0.10	**76.01 ± 0.01**	**69.72 ± 0.02**

Table 3: Performance comparison of SA-TSC and baseline methods on HAR, ISRUC, and SleepEDF. The accuracy and F1 score of the finetuned model are reported for both the '1 subject' and '2 subjects' finetuning scenarios. All models were executed five times, and the results present the mean values and standard deviation.

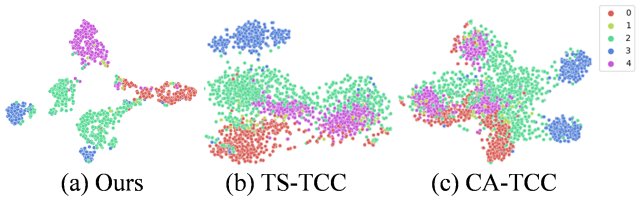

Figure 5: t-SNE visualization of the representations obtained by each model. 0: Wake, 1: N1, 2: N2, 3: N3, and 4: REM.

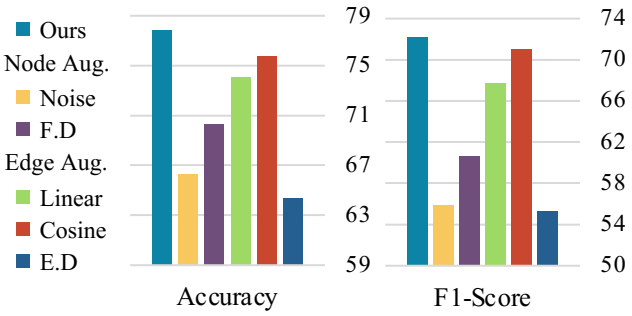

Figure 6: Results on various graph-augmentations.

Ablation Study

To show the effectiveness of the strategies taken in SA-TSC, we conduct two ablation studies. We test the efficacy of our augmentation technique in the spatial GNN module, as well as the individual effectiveness of each module in SA-TSC. Due to the space limitation, we report the results on SleepEDF under the *'2 subject'* fine-tuning setting. The other cases show very similar trend.

Graph Augmentation We conduct a comparative analysis with other widely-used graph augmentation methods (Marrium and Mahmood 2022; Liu et al. 2022; Zhao et al. 2022).

- **Node Feature Augmentation**: Add noise (Noise), which adds Gaussian noise to node features; Feature Drop (FD), which randomly drops node features;
- **Edge Augmentation**: Linear kernel, which constructs edges based on the linear kernel; Cosine kernel, which constructs edges by using the cosine kernel; Edge Drop (ED), which randomly drops edges.

Figure 6 shows the compared result and proves that our graph augmentation technique which integrates cross-domain information from both time and frequency domains is simple yet efficient in time series data.

Model Analysis Table 4 shows the ablation study on SA-TSC, which evaluates the performance impact of each specific loss component, by sequentially adding individual modules. Three model variants are compared, differing based on their loss components for model optimization. Additionally, we also examine the variations of graph construction on our full framework. 'Gaussian' refers to the initial graph construction based on the Gaussian Kernel and 'Gumbel' denotes the adjacency matrix which represents the edge existence. 'Gaussian & Gumbel' stands for our construction method. The result shows the effectiveness of our representation learning module and the graph construction method.

Component	ACC	F1
Positive Pairwise Loss $\mathcal{L}_{\mathcal{BT}}$	69.68	63.11
+ Agreement Loss $\mathcal{L}_{\mathcal{A}}$	73.21	67.03
+ Clustering Loss $\mathcal{L}_{\mathcal{SP}}$		
Gaussian	70.11	63.69
Gumbel	71.32	65.67
Gaussian & Gumbel	**75.24**	**66.82**

Table 4: Ablation analysis of Component Contributions

Impact of the 'stop-gradient' Operator Table 5 illustrates the impact of the 'stop-gradient' operator introduced in Eq. (10). This operator acts to ensure that during the optimization, only $\mathbf{S}$ can be updated and $\mathbf{H}$ remains fixed and is not updated through backpropagation. As a result, the model

is forced to learn the cluster assignment $\mathbf{S}$ based on $\mathbf{H}$.

	HAR		ISRUC		SleepEDF	
	ACC	F1	ACC	F1	ACC	F1
w/ stop-grad	**96.24**	**96.37**	**72.69**	**69.54**	**75.91**	**70.77**
w/o stop-grad	93.73	94.06	66.32	63.87	72.75	65.44

Table 5: Performance according to the 'stop-gradient' operator.

Hyperparamter Effects

In this section, we analyze the influence of the hyperparameters on the model's performance. The reported results correspond to the evaluation under '2 subjects' fine-tuning scenario.

Loss Term Weight Table 6 reports the results on varying hyperparameters λ_{agree}, $\lambda_{spectral}$ which controls the importance of each loss term. In the table, only specific values of the hyperparameters are listed; any value not explicitly mentioned is set as the default value 1.

Hyperparameter		HAR		ISRUC		SleepEDF	
		ACC	F1	ACC	F1	ACC	F1
λ_{agree}	0.25	94.55	94.66	72.56	69.34	74.28	69.08
	0.50	86.27	85.82	71.42	68.73	65.30	54.42
	0.75	95.13	95.24	69.84	67.48	73.69	67.31
$\lambda_{spectral}$	0.25	90.17	89.84	70.91	67.85	73.54	65.09
	0.50	94.07	94.20	75.39	72.17	75.66	69.64
	0.75	93.83	93.95	67.68	65.21	68.64	58.41

Table 6: Performance variations based on loss term weights λ_{agree}, $\lambda_{spectral}$.

Length Scale Parameter of Gaussian Kernel Table 7 illustrates how variations in σ influence the graph sparsity and the resultant performance measures. This relationship emphasizes that careful selection of σ can lead to more meaningful results in graph-based learning.

		σ			
		25	50	75	100
HAR	Sparsity	0.51	0.75	0.94	0.94
	ACC	90.75	95.42	93.01	96.24
	F1	90.50	95.62	92.99	96.37
ISRUC	Sparsity	0.67	0.67	0.67	0.84
	ACC	65.13	64.62	67.91	72.69
	F1	63.22	63.25	62.73	69.54
SleepEDF	Sparsity	0.67	0.67	0.67	0.84
	ACC	75.61	77.86	75.42	75.91
	F1	69.83	72.24	68.87	70.77

Table 7: Performance according to Gaussian kernel length scale parameter σ.

Discussion

Social Impact Time series data from human subjects is increasing significantly from wearable devices and sensors to effectively record human behaviors. However, according to the difference between subjects, inter-subject variance is inevitable since humans are diverse in many aspects (e.g., age, gender, body composition). Our work, which successfully demonstrated its effectiveness on the limited & subject-wise label situation, will be applicable in various real-world healthcare problems, such as EEG-based tasks (e.g., sleep staging, seizure detection, emotion recognition), and multi-channel bio-signal data analysis from wearable sensors or IoT devices (e.g., activity recognition, stress detection). We believe that real-world data is much more scarce and difficult to annotate, making our solution important in tackling practical problems.

Limitations Our full framework is based on graph structure, effectively representing multi-channel dependencies and performing spectral clustering. However, several challenges related to the graph structure leave some discussions. First, although our graph construction method is significant with promising results, the parameter setting that controls the graph sparsity is vital and needs careful consideration. Secondly, as the number of nodes in the graph expands, the computation cost becomes expensive. Future work includes the development of robust models to address graph sparsity and methods that optimize computational efficiency.

Conclusions

Our work is motivated by a typical clinical setting where time series data is collected from various *subjects* but annotated on a subject-by-subject basis, and very few subjects can be recruited in practice. We propose SA-TSC, a self-supervised framework for human subject time series data in a subject-oriented manner. SA-TSC comprises a spatial graph to consider multi-channel dependencies and a subject graph to conduct the learning strategy in the subject space. We apply positive pairwise loss and agreement loss for the spatial graph learning module and employ spectral clustering for subject-wise clustering in the subject space. Qualitative results demonstrate the superiority of our proposed framework and also show the effectiveness of each module on our framework with the ablation studies.

Acknowledgments

This work was partly supported by (1) the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. NRF-2021M3E5D2A01021156), (2) Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01373,Artificial Intelligence Graduate School Program(Hanyang University)) and (3) the National Research Foundation of Korea(NRF) grant funded by the Korea government(*MSIT) (No.2018R1A5A7059549). *Ministry of Science and ICT

References

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J. L.; et al. 2013. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, 3.

Babaud, J.; Witkin, A. P.; Baudin, M.; and Duda, R. O. 1986. Uniqueness of the Gaussian kernel for scale-space filtering. *IEEE transactions on pattern analysis and machine intelligence*, (1): 26–33.

Bianchi, F. M.; Grattarola, D.; and Alippi, C. 2020. Spectral clustering with graph neural networks for graph pooling. In *International conference on machine learning*, 874–883. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Chung, S.; Jeong, C. Y.; Lim, J. M.; Lim, J.; Noh, K. J.; Kim, G.; and Jeong, H. 2022. Real-world multimodal lifelog dataset for human behavior study. *ETRI Journal*, 44(3): 426–437.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2021. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C.-K.; Li, X.; and Guan, C. 2022. Self-supervised contrastive representation learning for semi-supervised time-series classification. *arXiv preprint arXiv:2208.06616*.

Ericsson, L.; Gouk, H.; Loy, C. C.; and Hospedales, T. M. 2022. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3): 42–62.

Gao, T.; Yao, X.; and Chen, D. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220.

Khalighi, S.; Sousa, T.; Santos, J. M.; and Nunes, U. 2016. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124: 180–192.

Kim, S.; and Chae, D.-K. 2023. Look Ahead: Improving the Accuracy of Time-Series Forecasting by Previewing Future Time Features. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2134–2138.

Kwon, H.; Kim, H. H.; An, J.; Lee, J.-H.; and Park, Y. R. 2021. Lifelog data-based prediction model of digital health care app customer churn: retrospective observational study. *Journal of Medical Internet Research*, 23(1): e22184.

Lee, H.; Seong, E.; and Chae, D.-K. 2022. Self-supervised learning with attention-based latent signal augmentation for sleep staging with limited labeled data. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 3868–3876.

Lee, J.; Park, B.; and Chae, D.-K. 2023. DuoGAT: Dual Time-oriented Graph Attention Networks for Accurate, Efficient and Explainable Anomaly Detection on Time-series. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1188–1197.

Liu, Y.; Jin, M.; Pan, S.; Zhou, C.; Zheng, Y.; Xia, F.; and Philip, S. Y. 2022. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 5879–5900.

Marrium, M.; and Mahmood, A. 2022. Data Augmentation for Graph Data: Recent Advancements. *arXiv preprint arXiv:2208.11973*.

Mohammadi, M.; Al-Fuqaha, A.; Sorour, S.; and Guizani, M. 2018. Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials*, 20(4): 2923–2960.

Qiao, L.; Zhang, L.; Chen, S.; and Shen, D. 2018. Data-driven graph construction and graph learning: A review. *Neurocomputing*, 312: 336–351.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17: 395–416.

Von Luxburg, U.; et al. 2010. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3): 235–274.

Xiao, Q.; Wang, J.; Ye, J.; Zhang, H.; Bu, Y.; Zhang, Y.; and Wu, H. 2021. Self-supervised learning for sleep stage classification with predictive and discriminative contrastive coding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1290–1294. IEEE.

Ye, J.; Xiao, Q.; Wang, J.; Zhang, H.; Deng, J.; and Lin, Y. 2021. Cosleep: A multi-view representation learning framework for self-supervised learning of sleep stage classification. *IEEE Signal Processing Letters*, 29: 189–193.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 12310–12320. PMLR.

Zhao, T.; Jin, W.; Liu, Y.; Wang, Y.; Liu, G.; Günnemann, S.; Shah, N.; and Jiang, M. 2022. Graph data augmentation for graph machine learning: A survey. *arXiv preprint arXiv:2202.08871*.