MULTI-SCALE AND MULTI-MODAL CONTRASTIVE LEARNING NETWORK FOR BIOMEDICAL TIME SERIES

Hongbo Guo[1,2,3], *Xinzi Xu*[1,2], *Hao Wu*[3,4,*], *Guoxing Wang*[1,2,3]

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Key Laboratory for Thin Film and Microfabrication of Ministry of Education,
Shanghai Jiao Tong University, Shanghai, China
[3]Guangdong JiuZhi Technology Co., Ltd, Zhongshan, Guangdong, China
[4]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong, China

ABSTRACT

Multi-modal biomedical time series (MBTS) data offers a holistic view of the physiological state, holding significant importance in various bio-medical applications. Owing to inherent noise and distribution gaps across different modalities, MBTS can be complex to model. Various deep learning models have been developed to learn representations of MBTS but still fall short in robustness due to the ignorance of modal-to-modal variations. This paper presents a multi-scale and multi-modal biomedical time series representation learning (MBSL) network with contrastive learning to migrate these variations. Firstly, MBTS is grouped based on inter-modal distances, then each group with minimum intra-modal variations can be effectively modeled by individual encoders. Besides, to enhance the multi-scale feature extraction (encoder), various patch lengths and mask ratios are designed to generate tokens with semantic information at different scales and diverse contextual perspectives respectively. Finally, cross-modal contrastive learning is proposed to maximize consistency among inter-modal groups, maintaining useful information and eliminating noises. Experiments against four bio-medical applications show that MBSL outperforms state-of-the-art models by 33.9% mean average errors (MAE) in respiration rate, by 13.8% MAE in exercise heart rate, by 1.88% accuracy in human activity recognition, and by 1.14% F1-score in obstructive sleep apnea-hypopnea syndrome.

***Index Terms*—** bio-medical time series, multi-modal, representation learning, contrastive learning

1. INTRODUCTION

Multi-modal biomedical time series (MBTS) data have been widely used for various bio-medical applications, such as combining photoplethysmogram (PPG) and blood oxygen level (SpO_2)[1] to jointly predict sleep apnea-hypopnea syndrome (OSAHS). Deep learning (DL) models especially temporal convolution networks (TCN)[2][3] are widely used for MBTS representation learning due to their good performance and low resource consumption, which uses dilated causal convolutions to capture long-term temporal dependency. To better capture extract multi-scale patterns (MS) in MBTS, MS extraction methods[4][5][6] are designed. Besides, to learn useful representations from MBTS, contrastive learning[3][7][8] is also often used.

However, the performance of existing models is limited due to the ignorance of the distribution gap across modalities and underfitted MS feature extraction modules. Hence, to further improve the performance, this paper presents a multi-scale and multi-modal biomedical time series contrastive learning (MBSL) network which includes the following main parts. 1) **Inter-modal grouping technique to address distribution gap across modalities**.The amplitude range of MBTS is unbounded that may vary significantly across modalities for example SpO_2 values usually vary from 70% to 100%, while the acceleration values vary from 0 to 1000 cm/s^2. MBTS also features diverse structured patterns, such as the seasonally dominant PPG and trend dominant SpO2, suited for frequency and time domain modeling respectively[9]. Therefore, the commonly used parameter-sharing encoder for all modalities[3][8] may limit the capability of the encoders to represent different modalities. An intuitive approach is to customize a specific encoder for each modality to extract heterogeneous features, but it'll be inefficient and lack robustness[10]. To achieve a more capable and robust model, inter-modal grouping (IMG) is proposed to group MBTS into different groups with minimum intra-model distances, thus data in the same group manifest similar distributions, which is simple to represent with a shared encoder. Then distinct encoders are trained for each group to handle multi-modal data distribution gaps. 2) **A lightweight multi-scale data transformation using multi-scale patching and multi-scale masking.** Most MS extraction algorithms primarily rely on multi-branch convolution utilizing receptive fields of varying scales[5][6]. However, it quadratically increases the memory and computing requirements. Some works like MCNN[4] use down-sampling to generate multi-scale data, but it would lose considerable useful information. This paper proposes a lightweight multi-scale data transformation using multi-scale patching and multi-scale masking, where different patch lengths are defined to obtain input tokens with semantically meaningful information at different resolutions, and the different mask ratios are adaptively defined according to the scale of features to be extracted. 3) **Cross-modal contrastive learning.** Inappropriate data augmentation can alter the intrinsic properties of MBTS, leading to the formation of false positive pairs. This issue causes many time series contrastive learning models to encode invariances that might not align with the requirements of downstream tasks..[11] For instance, excessive masking risks obscuring physiologically informative sub-segments. To construct more reasonable contrastive pairs, cross-modal contrastive learning is utilized to learn modal-invariant representations that capture information shared among modalities. The intuition behind the idea is that different modalities commonly reflect the physiological state so the modal-invariant semantic is useful information for BMA. The contributions are summarized as follows:

* Hao Wu is the correspondence author.

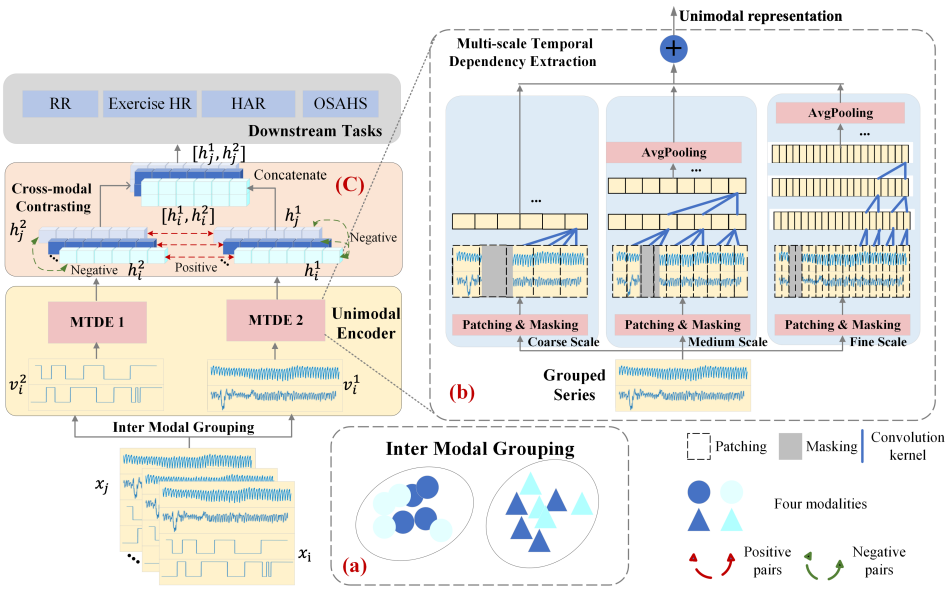

Fig. 1. MBSL architecture. (a) Inter-modal grouping; (b) Multi-scale Temporal Dependency Extraction; (c) Cross-modal Contrasting. x_i represents the i-th input series, v_i^1, and h_i^1 represents the raw data and embedding of the first group of the i-th sample, respectively.

1. We propose multi-scale patching and multi-scale masking for extracting features at various resolutions and adaptively enhancing the capacity of the model, respectively.
2. We propose inter-modal grouping to process multi-modal data and use cross-modal contrastive learning for effective positive pairs construction and multi-modal data fusion.
3. Our model outperforms state-of-the-art (SOTA) methods across four datasets.

2. METHOD

The overall architecture of MBSL is shown in Fig. 1. It includes IMG, MTDE, and cross-modal contrastive loss, which will be briefly introduced below.

2.1. Inter-modal Grouping

As mentioned before, the distribution gap severely impairs the generalization ability of the network. IMG is utilized to address this challenge. Intuitively, MBTS is divided into distinct grouped modalities based on their inter-modal distances. Modalities within each group exhibit similar distributions. Independent encoders are used to model different grouped modalities, respectively. Our approach aligns with the intuition that similar data should be processed similarly.

In detail, we consider a set of M modalities of the data, represented as $X_1, \ldots, X_M$. Each modality is first dimensionally reduced using t-SNE, and then the inter-modal distance is calculated through the Euclidean distance. K group modalities $v_1, \ldots, v_K$ will be obtained. Within a group, the minimum distance between any modality X_i and other modalities X_j ($i \neq j$) needs to be less than a threshold I:

$$D_i = argmin \{d(X_i, X_j\}_{j=1}^{g} < I \quad (1)$$

where g is the number of modalities within a group. K group-specific encoders will be used for the K-grouped modalities.

2.2. Multi-scale data transformation

Recently, TCN has been proven to be a highly effective network architecture for BMA. However, real-world MBTS is very complex, a unified receptive field (RF) within each layer of TCN is often not powerful enough to capture the intricate temporal dynamics. What's more, the effective RFs of the input layers are limited, causing temporal relation loss during feature extraction. To adaptively learn a more comprehensive representation, we designed an MTDE module, which uses MS data transformation (patching and masking) and MS feature extraction encoder.

2.2.1. Multi-scale patching and Multi-scale Masking

Multi-scale patching Different lengths of patches are used to transform the time series from independent points to tokens with semantic information within adaptive subseries, allowing the model to capture features of different scales. In particular, by aggregating more samples, long-term dependence with less bias can be obtained instead of short-term impact. Simultaneously, patching can efficiently alleviate the computational burden caused by the explosive increase in multiple TCN branches by reducing the sequence length from L to $\frac{L}{P}$, where P is the patch length.

Multi-scale Masking Masking random timestamps helps improve the robustness of learned representations by forcing each timestamp to reconstruct itself in distinct contexts, which can be formulated as $x_{mask_i} = x_i * m$, where $m \in 0, 1$. Commonly, the ratio of masked samples to the whole series, denoted as M_R is set to be small to avoid distortion of original data. However, as to different scales of features, the optimized M_R is adaptive based on the experimental results. For instance, for coarser-grained features, larger M_R is preferred so that ample semantic information can still be retained while introducing strong variances, thereby enhancing the robustness. Hence, different M_Rs are selected to adapt to multi-scale features.

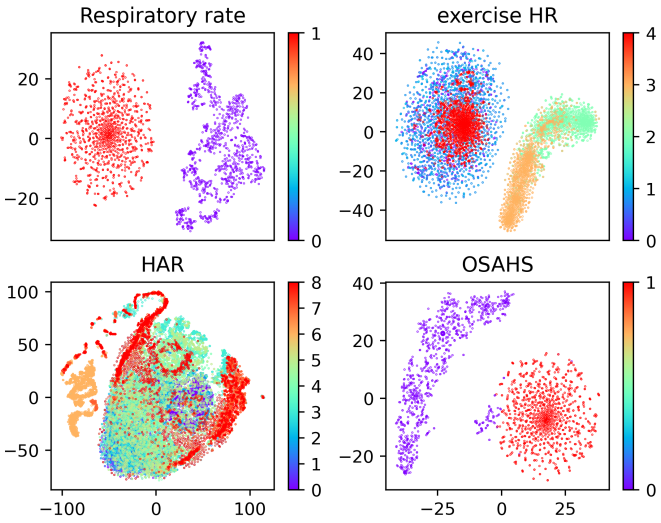

Fig. 2. Distribution gap in multi-modal biomedical time series. Different colors represent different modalities in that dataset.

2.2.2. *Multi-scale feature extraction encoder*

After MS data transformation, sketches of a time series at different scales are generated. Then MS time series pass through parallel TCN encoders with different kernel sizes and layers and then concatenated at the end. Average pooling is used to align sequence lengths to uniform length.

2.3. Cross-modal Contrastive Loss

Previous research on contrastive representation learning mainly tackled the problem of faulty positive pairs and noise in data. In the process of generating positive pairs by data augmentation, incorrect positive samples are also introduced, leading to suboptimal performance.

In this work, the same segments from different modalities are chosen as positive pairs instead of generating them by data augmentation, thus avoiding the risk of introducing faulty positive pairs. Besides, this is beneficial to learn the semantic information shared by modalities because signals from different modalities of the same segments can intuitively reflect the subject's physiological state. Hence, during reducing the loss of contrastive learning, effective information among different modalities of the same segment will be maximized while redundant information across different segments will be ignored.

In particular, we consider all grouped modality pairs (i, j), $i \neq j$, and optimize the sum of a set of pair-wise contrastive losses:

$$\mathcal{L}_{CM} = \sum_{0<i<j\leq M} \mathcal{L}_{V_i,V_j} \tag{2}$$

$$\mathcal{L}_{V_1,V_2} = -\log \frac{\exp\left(\text{sim}\left(h_i^1, h_i^2\right)/\tau\right)}{\sum_{k=1}^{N} \mathbb{1}_{k\neq i} \exp\left(\text{sim}\left(h_i, h_k\right)/\tau\right)} \tag{3}$$

where h_i^1 is the embedding of the first group of the i-th sample, $\text{sim}\left(h_i^1, h_i^2\right) = h_i^1 h_i^2 / \left\|h_i^1\right\| \left\|h_i^2\right\|$, $\mathbb{1}_{k\neq i}$ in 0,1 is an indicator function evaluating to 1 if $k \neq i$ and τ denotes a temperature parameter, N represents the size of batch size.

Table 1. Results of RR detection. W represents the length of signal.

Methods	MAE (W=16)	MAE (W=32)	MAE (W=64)
TFC	4.16 ± 3.31	2.96 ± 3.19	4.54 ± 3.65
TS2vec	3.27 ±3.35	2.56 ± 2.93	2.34 ± 3.01
RespNet[12]	2.45 ± 0.69	2.07 ± 0.98	2.06 ± 1.25
RRWAVE[6]	1.80 ± 0.95	1.62 ± 0.86	1.66 ± 1.01
Ours	**1.28 ± 0.84**	**1.13 ± 0.95**	**0.91 ± 0.81**

Table 2. Results of exercise HR detection.

Methods	RMSE
TFC	31.1
TS2vec	31.2
TST[13]	25.0
InceptionTime[5]	23.9
Ours	**20.6**

3. EXPERIMENTAL RESULT

3.1. Experimental Setup

The size of the hidden layers and output layer is 32 and 64 respectively. The mask ratio, M_R is 0.05, 0.1, and 0.15 and the patch length is $0.04 * fs$, $0.08 * fs$, and $0.16 * fs$ for small, middle, and large scales, respectively, where fs is the sample rate of the dataset. The temperature τ of the contrastive learning is 0.1. The model is optimized using Adam with a batch size of 480 and a learning rate of 0.002.

3.2. Results

The proposed MBSL is applied to MBTS datasets in two major practical tasks including regression and classification. The two latest contrastive learning works, i.e. TFC[8]and TS2Vec[3], are used as baseline models, and we also compare them with SOTA under the corresponding datasets. The t-SNE visualization of these datasets is shown in Fig. 2, which shows that there is a significant distribution gap between different modalities.

3.2.1. *MBTS regression*

We evaluate MBSL on open-source time series regression datasets: Respiratory rate (RR) detection[14], providing 53 8-minute PPG data segment (125 HZ) recordings, and exercise heart rate (HR) detection[15], including 3196 2-channel PPG and 3-axis acceleration sampled at a frequency of 125Hz.

Evaluation method: 1) RR The PPG signal and its Fourier transform in frequency domain signal are used as multivariate biological time series. Leave-one-out cross-validation is applied and the preprocessing method follows [6]. The average mean absolute error (MAE) and the standard deviation (SD) across all patients are used to evaluate the performance of the model [6].

2) Exercise HR. It is divided into two groups, one of which is the two-way PPG signal and one-way acceleration signal, and the other is the two-way acceleration signal. We use root mean squared error (RMSE) to evaluate model performance and divide the dataset into the training set, validation set, and test set as in [13].

Compare with the State-of-arts. Table 1 and table 2 demonstrates our superior performance over existing methods, quantita-

Table 3. Result of HAR and OSAHS

HAR			OSAHS			
Methods	Acc	AUPRC	Methods	F1	Re	Acc
TS2vec	90.57	0.85	TFC	62.30	57.76	82.38
TFC	90.22	0.95	TS2vec	60.57	55.40	81.82
COT[20]	94.05	/	ConCAD[7]	75.70	76.54	87.62
BTSF[11]	94.63	0.99	U-Sleep[1]	76.20	76.49	87.92
Ours	**96.51**	**0.99**	**Ours**	**77.34**	**78.67**	**88.38**

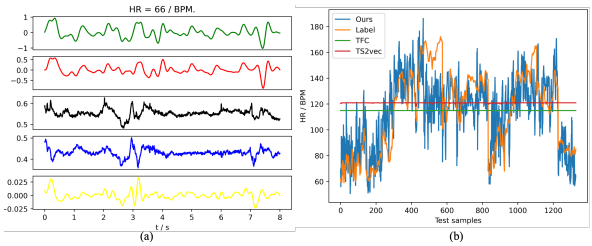

Fig. 3. (a) A sample in the exercise heart rate data set. (b) Prediction results of different models on the whole exercise heart rate testset.

tively, reducing MAE by 33% in RR and RMSE 13.8% in exercise HR.

3.2.2. MBTS classification

We evaluate MBSL on widely-used MBTS classification datasets: HAR[16] including 10299 3-axis accelerometer, 3-axis gyroscope and 3-axis magnetometer, OSAHS[17] including 349032 PPG and SpO_2.

Evaluation Method: 1) HAR. It is divided into two groups, one of which is the 3-axis acceleration, 3-axis gyroscope, and 2-axis magnetometer and the other is the 1-axis magnetometer. The dataset is divided as in [18]. Accuracy (Acc) and the area under the precision-recall curve (AUPRC) are used to evaluate model like [18]. **2) OSAHS.** It contains PPG and SpO_2. The data is prepossessed as in [19] and randomly divided into four parts according to the subject IDs with three parts for training and validation, and the remaining one part for testing. F1-score, recall (Re) and Acc are used to evaluate the performance of the model.

Compare with the State-of-arts. Table 3 shows that our MBSL improves Acc by 1.88% in HAR and increases Acc by 0.54%, Re by 2.18%, and F1-Score by 1.14% in OSAHS.

3.3. Discussion

Multi-scale features extraction ability and noise immunity. As shown in Fig. 3, in exercise HR detection, due to motion artifacts, PPG is severely noisy and the HR varies from 50 beats per minute (BPM) to 150 BPM across different subjects. TFC[8]and TS2Vec[3] collapse under such complex MBTS and the regression results on the entire test dataset are straight lines. InceptionTime[5] achieves a relatively better result thanks to an MS extraction model. Due to additional noise immunity and more powerful MS feature extraction ability, MBSL achieves the best performance.

Table 4. Ablation Experiments on Exercise Heart Rate Dataset.

	RMSE	MAE
MBSL	**20.6**	**15.3**
Inter-modal grouping		
w/o IMG	25.5	20.7
Random Grouping	24.0	24.0
Full Grouping	25.9	21.9
MTDE		
w/o mask	21.9	15.9
Three moderate mask ratios	21.3	16.0
w/o patch	24.1	18.5
Three moderate patch lengths	23.6	18.1
MTDE—>TCN	32.5	26.0
Cross-modal contrastive loss		
Supervised learning	22.2	16.3
Cross-modal contrastive loss —> instance contrastive objective	22.4	16.2

3.4. Ablation experiments

Ablation experiments are conducted on the exercise heart rate dataset because it exists in a very complex scene, which can better prove the effectiveness of our algorithm. We studied the benefits of IMG, the MTDE, and the cross-modal contrastive loss. The result is shown in Table 4. **IMG.** Without IMG, the performance of the model drops a lot due to distribution gaps. We also tried different group strategies. Random grouping means that MBTS are randomly grouped, and the number of groups is the same as IMG. Full grouping means that we apply 5 encoders to the 5 modalities in exercise HR dataset. Both methods have a certain decline, which proves that our grouped MI is a reasonable grouping, which is very important for adapting to new datasets. **MTDE.** Without patching and masking, the MS extraction ability is limited, which results a poor performance. In addition, when three common moderate mask ratios (0.1) and patch lengths (10) are applied, model performance also decreases, which demonstrates the necessity of using data transformation of different strengths for feature extraction at different scales. Surprisingly, when replacing MTDE with TCN, the model performance drops sharply, possibly due to the collapse as TS2Vec.**Cross-modal contrastive loss.** Regardless of removing the cross-modal contrastive loss or replacing the cross-modal contrastive loss with the commonly used instance contrastive loss, the model performance has declined, proving the effectiveness of the cross-modal contrastive loss.

4. CONCLUSION

Deep learning is widely used in BMA. At present, the distribution gap across modalities and the complex dynamics of MBTS are still two major challenges for BMA. This article proposes a multi-scale and multi-modal representation learning network. In particular, the inter-modal grouping is introduced to eliminate the distribution gap. The MTDE is designed based on different patch lengths and mask ratios so that different scales of input perspectives are generated for the encoder. In addition, a cross-modal contrastive loss is used to encode the common semantic information across modalities, avoiding the risk of introducing faulty positive pairs by data augmentation. Evaluated on four datasets, our model outperforms current SOTA models, demonstrating the effectiveness of our approach.

5. REFERENCES

[1] Riku Huttunen, Timo Leppänen, Brett Duce, Erna S. Arnardottir, Sami Nikkonen, Sami Myllymaa, Juha Töyräs, and Henri Korkalainen, "A comparison of signal combinations for deep learning-based simultaneous sleep staging and respiratory event detection," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 5, pp. 1704–1714, 2023.

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271 [cs.LG], 2018, http://arxiv.org/abs/1803.01271.

[3] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu, "Ts2vec: Towards universal representation of time series," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2022.

[4] Zhicheng Cui, Wenlin Chen, and Yixin Chen, "Multi-scale convolutional neural networks for time series classification," arXiv:1603.06995 [cs.CV], 2016, https://arxiv.org/abs/1603.06995.

[5] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean, "InceptionTime: Finding AlexNet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, sep 2020.

[6] Pongpanut Osathitporn, Guntitat Sawadwuthikul, Punnawish Thuwajit, Kawisara Ueafuea, Thee Mateepithaktham, Narin Kunaseth, Tanut Choksatchawathi, Proadpran Punyabukkana, Emmanuel Mignot, and Theerawit Wilaiprasitporn, "Rrwavenet: A compact end-to-end multiscale residual cnn for robust ppg respiratory rate estimation," *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 15943–15952, 2023.

[7] Guanjie Huang and Fenglong Ma, "Concad: Contrastive learning-based cross attention for sleep apnea detection," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021.

[8] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," in *36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[9] Xiyuan Zhang, Xiaoyong Jin, Karthick Gopalswamy, Gaurav Gupta, Youngsuk Park, Xingjian Shi, Hao Wang, Danielle C. Maddix, and Yuyang Wang, "First de-trend then attend: Rethinking attention for time-series forecasting," in *36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[10] Lu Han, Han-Jia Ye, and De-Chuan Zhan, "The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting," 2023.

[11] Ling Yang and Shenda Hong, "Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion," in *Proceedings of the 39 th International Conference on Machine Learning (ICML)*, 2022.

[12] Vignesh Ravichandran, Balamurali Murugesan, Vaishali Balakarthikeyan, Keerthi Ram, S.P. Preejith, Jayaraj Joseph, and Mohanasankar Sivaprakasam, "Respnet: A deep learning model for extraction of respiration from photoplethysmogram," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 5556–5559.

[13] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2021.

[14] Marco A. F. Pimentel, Alistair E. W. Johnson, Peter H. Charlton, Drew Birrenkott, Peter J. Watkinson, Lionel Tarassenko, and David A. Clifton, "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1914–1923, 2017.

[15] Zhilin Zhang, Zhouyue Pi, and Benyuan Liu, "Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 522–531, 2015.

[16] D. Anguita, Alessandro Ghio, L. Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *The European Symposium on Artificial Neural Networks (ESANN)*, 2013.

[17] Dian-Marie Ross and Edmond Cretu, "Probabilistic modelling of sleep stage and apneaic events in the university college of dublin database (ucddb)," in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2019, pp. 0133–0139.

[18] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan, "Time-series representation learning via temporal and contextual contrasting," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2021.

[19] Minsoo Yeo, Hoonsuk Byun, Jiyeon Lee, Jungick Byun, Hak-Young Rhee, Wonchul Shin, and Heenam Yoon, "Respiratory event detection during sleep using electrocardiogram and respiratory related signals: Using polysomnogram and patch-type wearable device data," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 550–560, 2022.

[20] Weiqi Zhang, Jia Li Jianfeng Zhang, and Fugee Tsung, "A co-training approach for noisy time series learning," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2023.