

UC San Diego

UC San Diego Previously Published Works

Title

Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals

Permalink

<https://escholarship.org/uc/item/69w3r1gm>

Journal

ACM Transactions on Computing for Healthcare, 7(3)

ISSN

2691-1957

Authors

Luo, Yunfei

Chen, Yuliang

Salekin, Asif

et al.

Publication Date

2026-07-30

DOI

10.1145/3803808

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed

Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals

YUNFEI LUO and YULIANG CHEN, Halicioglu Data Science Institute, University of California San Diego, La Jolla, California, USA

ASIF SALEKIN, Arizona State University, Tempe, Arizona, USA

TAUHIDUR RAHMAN, University of California San Diego, La Jolla, California, USA

Time-series foundation models excel at tasks like forecasting across diverse data types by leveraging informative waveform representations. Wearable sensing data, however, pose unique challenges due to their variability in patterns and frequency bands, especially for healthcare-related outcomes. The main obstacle lies in crafting generalizable representations that adapt efficiently across heterogeneous sensing configurations and applications. To address this, we propose NormWear, the first paradigm multi-modal and ubiquitous framework of foundation model designed to extract generalized and informative representations from wearable sensing data. Specifically, we design a channel-aware attention mechanism with a shared special liaison [CLS] token to detect signal patterns in both intra-sensor and inter-sensors. This helps the model to extract more meaningful information considering both time series themselves and the relationships between input sensors. This helps the model to be widely compatible with various sensors settings. NormWear is pretrained on a diverse set of physiological signals, including PPG, ECG, EEG, GSR, and IMU, from various public datasets. Our model shows exceptional generalizability across 11 public wearable sensing datasets, spanning 18 applications in mental health, body state inference, vital sign estimation, and disease risk evaluation. It consistently outperforms competitive baselines under zero-shot, partial-shot, and full-shot settings, indicating broad applicability in real-world health applications.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**; **Temporal reasoning**; • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; • **Applied computing** → **Health informatics**;

Additional Key Words and Phrases: Machine Learning, Deep Learning, Digital Health, Time Series Modeling, Signal Processing

ACM Reference format:

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. 2026. Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals. *ACM Trans. Comput. Healthcare* 7, 3, Article 43 (May 2026), 43 pages.
<https://doi.org/10.1145/3803808>

Yunfei Luo and Yuliang Chen contributed equally to this research.

TR and YL were supported by NSF Smart and Connected Health (2320678), Google Research Scholar Award (Gift), and Qualcomm Award (Gift).

Authors' Contact Information: Yunfei Luo, Halicioglu Data Science Institute, University of California San Diego, La Jolla, California, USA; e-mail: yul268@ucsd.edu; Yuliang Chen, Halicioglu Data Science Institute, University of California San Diego, La Jolla, California, USA; e-mail: yuc204@ucsd.edu; Asif Salekin, Arizona State University, Tempe, Arizona, USA; e-mail: Asif.Salekin@asu.edu; Tauhidur Rahman (corresponding author), University of California San Diego, La Jolla, California, USA; e-mail: trahman@ucsd.edu.



This work is licensed under Creative Commons Attribution-NonCommercial-NoDerivatives International 4.0.

© 2026 Copyright held by the owner/author(s).

ACM 2637-8051/2026/5-ART43

<https://doi.org/10.1145/3803808>

1 Introduction

Mobile and wearable sensors have been shown to be valuable for the field of healthcare by passively and continuously tracking physiological signals such as **photoplethysmography (PPG)** for pulse, **electrocardiography (ECG)** for heart activity, **galvanic skin response (GSR)**, and **electroencephalography (EEG)** for brain activity. These time series signals are beneficial for early diagnosis, personalized health insights, and remote patient monitoring [93].

Recently, several foundation models have emerged for time series modeling, including Ansari et al. [7], Abbaspourazad et al. [1], Woo et al. [84], Xu et al. [86], and Foumani et al. [26]. Another common approach for signal modeling involves converting raw signal series into 2D images or spectrograms, using fixed-size sliding windows, followed by the use of visual encoders like **Vision Transformers (ViT)** to extract representations for making inferences [21, 23, 41, 71, 80, 82]. These works have significantly advanced the field and provided valuable insights, yet two main issues still exist which need further exploration to fully understand their potential in wearable scenarios. First, contrastive learning-based foundation models [1] rely on a predefined set of input signal types, making them unsuitable when transferring to scenarios with different types and numbers of sensors. Second, while both time series foundation models [7, 84, 97] and spectral-based approaches [71, 82] attempt to address this issue by training a generic encoder that can handle type-agnostic series, they remain limited to processing only univariate series. Because of this constraint, these previous works fail to account for the heterogeneity of multivariate input data; specifically, they do not capture the complex relationships between signals from sensors located on different body parts. These two limitations of recent approaches hinder their generalization and usefulness for wearable health monitoring.

Moreover, Wearable-based multimodal physiological signals present unique challenges that distinguish them from general time series data, such as stock prices or weather patterns. Wearable signal modalities, such as PPG and EEG, vary in characteristics like dimensionality, sampling rate, and resolution, often requiring modality-specific preprocessing. Existing methods tokenize raw signals [7, 97] or convert them into image or spectral representations [47, 77, 85]. While effective for specific tasks, these approaches lack generalizability and fail to provide a consistent preprocessing pipeline across multiple modalities. A consistent framework that accommodates diverse signal requirements is essential for training deep learning-based foundation models and advancing multimodal signal analysis.

In this work, we present NormWear, a normative foundation model, whose overview is presented in Figure 1, aiming to learn effective wearable sensing representations, addressing the above-discussed research gaps. NormWear has been pretrained on more than 2.5 million multivariate wearable sensing segments, comprising total of 14,943 hours of sensor signal series, using publicly available datasets. We evaluated NormWear on 18 public downstream tasks against competitive baselines across zero-shot, few-shot, and full-shot settings. Overall, our contributions with the proposed NormWear healthcare modeling framework can be summarized as follows:

- To our knowledge, we are the first to develop a foundation model specifically designed for wearable sensing data, capable of processing arbitrary configuration of multivariate signals from sources such as the heart, skin, brain, and physical body.
- NORMWEAR comprises novel methodologies built upon the advanced practice in both the fields of signal processing and deep learning, including (a) **continuous wavelet transform (CWT)** based multi-scale representations for modality- and number-agnostic tokenization, (b) channel-aware attention layer that enables the model to process arbitrary multivariate inputs, and (c) a human sensing adapted fusion mechanism that enabled NormWear to achieve zero-shot inference on health related wearable sensing tasks.
- We are also the first to integrate and process a comprehensive wearable signals dataset with varied number of input channels for training self-supervised learning algorithms, with thorough downstream evaluation. These datasets cover key health applications, including mental and physical state inference, vital sign estimation, and disease risk evaluation.

Our proposed NORMWEAR aims to provide a generalized data representation solution for smart health monitoring, benefiting the general public, and serving as a fundamental tool for researchers and professionals to address future healthcare challenges. We made the code and cleaned data to be publicly available to spur reproducible research.

2 Related Work

Foundation models have emerged as a transformative paradigm in machine learning, enabling generalizable and reusable representations across diverse downstream tasks [12].

This notion has attracted significant interest in the wearable sensor domain, particularly for applications in physiological signal analysis. Recent studies have utilized self-supervised learning in wearable devices for tasks such as activity recognition [74, 89, 94]. Additionally, it has been applied to physiological signals such as PPG, ECG, and EEG, spanning various healthcare monitoring tasks [1, 49, 52, 60, 95]. However, these studies often rely on a predefined set of devices, which limits the models' adaptability when clinical application settings change. For instance, when new devices or modalities are introduced, these models, which have not been exposed to such data during training, often require fine-tuning to remain functional. Furthermore, several of these models are not publicly accessible due to the sensitivity of healthcare data, which hinders progress in this area. These challenges underscore the need for an open, pretrained model that can accommodate various device configurations and adapt to evolving clinical requirements.

On the other hand, general time series models [7, 84, 97] have shown significant advancements but are predominantly trained and evaluated in domains such as transportation, energy consumption, and finance, with limited exploration in physiological signals. While physiological signals are inherently multivariate time series, these models have not been trained on such data, leaving their transferability to the physiological sensor domain uncertain. Ignoring the correlations between different sensors may result in suboptimal performance when applied to this domain. Motivated by these limitations, this work focuses on developing a foundation model for sensory time series data capable of accommodating arbitrary combinations of device modalities as multivariate series, while investigating strategies to effectively leverage sensor correlations.

3 Method

3.1 Dataset Construction for Model Pretraining and Downstream Evaluation

We curated a collection of nine publicly available datasets (Table 2) exclusively for model pretraining, resulting in approximately 2,30,962 multivariate time series segments, comprising 1,337 hours of total sensor signal series, across various modalities, including PPG, ECG, EEG, GSR, PCG, and **inertial measurement unit (IMU)** data. To address the dataset size limitation, we then applied heuristic data augmentation (Algorithm 1) to expand the pretrain dataset to 2.5 million segments, comprising 14,943 hours of total sensor signal series. Notably, each sample segment may contain a variable number of input channels depending on the sensor signals provided by the respective datasets. This input configuration aligns seamlessly with our model's design, which is optimized to flexibly handle arbitrary numbers and configurations of sensor signal inputs.

To prevent potential data leakage in downstream tasks, we evaluate our model's transferability using an additional 11 publicly available datasets encompassing 18 modeling tasks as shown in Figure 1(c), which include affective state classification, physical state recognition, biological estimation, and disease risk evaluation. Details about the datasets is presented in Table 1 and Figure 3.

3.2 Tokenization

Tokenization is a fundamental term widely used in natural language processing. In the context of wearable sensing, we leverage this term to represent the stage of signal processing before sending the processed data to the deep learning-based encoder. Spectral methods, which utilize the short-time **fast Fourier transform (FFT)** [14]

Algorithm 1: Time Series Mixup Augmentation

```

1: Input: Time series dataset  $\mathcal{X}$ , number of augmentations  $n$ 
2: Output: Augmented Dataset  $\tilde{\mathcal{X}}$ 
3:  $\tilde{\mathcal{X}} \leftarrow \emptyset$ 
4: for  $i = 1$  to  $n$  do
5:   Sample two time series  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \sim \mathcal{X}$ 
6:   Randomly Sample a chunk size  $\lambda \sim \text{UniformDistribution}(0, l)$ 
7:   Randomly Sample start indices  $s_1, s_2 \sim \text{UniformDistribution}(0, l - \lambda)$ 
8:   Substitute chunk in  $\mathbf{x}^{(1)}$  with chunk in  $\mathbf{x}^{(2)}$ :
       
$$\mathbf{x}_{s_1:s_1+\lambda}^{(1)} \leftarrow \mathbf{x}_{s_2:s_2+\lambda}^{(2)}$$

9:   Append  $\mathbf{x}^{(1)}$  into  $\tilde{\mathcal{X}}$ 
10: end for
11:  $\tilde{\mathcal{X}} \leftarrow \tilde{\mathcal{X}} \cup \mathcal{X}$ 
12: return  $\tilde{\mathcal{X}}$ 

```

Table 1. Downstream Evaluation Data That Are Unseen during Pretraining

Downstream Dataset	Sensor	# Channels	Tasks	#Samp. (#Subj.)
WESAD [70]	IMU, PPG, ECG, GSR	10	Stress detection	11,050 (15)
UCI-HAR [68]	IMU	6	HAR	10,299 (30)
DriverFatigue [51]	EEG	4	Fatigue detection	2,400 (12)
Activity Recognition Total	-	-	-	23,749 (57)
Epilepsy [6]	EEG	1	State recognize	11,500 (500)
GAMEEMO [4]	EEG	4	Valence-arousal	5,600 (28)
EEG Main Tasks Total	-	-	-	17,100 (528)
ECG-Abnormal [13]	ECG	1	Abnormal detection	11,640 (249)
PPG-BP [45]	PPG	1	Risk of diseases	657 (219)
PhysioNet EMG [28]	EMG	1	Muscular diseases	163 (3)
Risk Evaluation Total	-	-	-	12,460 (471)
Noninvasive-BP [24]	PPG	3	BP estimate	125 (26)
PPG-Hgb [2]	PPG	2	Hgb estimate	68 (68)
Fetal-fPCG [11]	PCG	1	Fetal HR estimate	47 (47)
Vital Signs Total	-	-	-	240 (141)
Total All	-	-	-	53,549 (1,197)

The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme.

Table 2. Pretraining Data

Pretrain Dataset	Sensors	#Samp (hours)
Cuff-Less-BP [37]	ECG, PPG	42,934 (72)
PPG-Dalia [67]	ECG, PPGIMU, GSR	42,606 (36)
Auditory-EEG [5]	EEG	13,601 (23)
PhyAAat [8]	EEG	19,550 (33)
MAUS [10]	ECG, PPGGSR	13,068 (22)
Mendeley-YAAD [22]	ECG, GSR	2,964 (5)
Brain-Cognitive [25]	EEG	51,201 (85)
EPHNOGRAM [38]	ECG, PCG	36,611 (61)
BIDMC [61]	ECG, PPG	8,427 (14)
Num Segments (# Segm.)	-	230,962 (385)
# Segm. w/ Augment	-	2,576,418 (4,294)
Num Sensor Signals (# Sign.)	-	802,019 (1,337)
# Sign. w/ Augment	-	8,965,538 (14,943)

with a sliding window to compute spectrograms, are widely regarded as the benchmark approach for tokenization. However, due to the inherent tradeoff between time and frequency resolution, the spectral representation with a fixed window size cannot be generalized. This is because the window size has to be modulated accordingly when the modality varies. To enhance transferability, we propose a well-designed signal processing pipeline that preserves information in both the frequency and time domains across multiple scales. We begin by calculating the first and second derivatives for each single signal series, as suggested by Slapničar et al. [73], followed by computing the CWT on both the raw and derivative series, resulting in three scalograms. Then, we stack the three scalograms to form data in RGB-image-like format. The derivatives capture the rate of signal change at

different moments, while the wavelet transform provides a multi-resolution encoding that preserves information from both the time and frequency domains Torrence and Compo [76]. For the wavelet transform, we use the Mexican Hat wavelet for signal convolution, as recommended by previous studies [15, 29, 32, 55, 56]. We apply scales ranging from 1 to 64, following the guidance of [55, 72], which sufficiently covers most frequency bands of interest for physiological signals. Finally, this RGB-like scalogram is divided into patches, which is treated in the same way as tokens in an ViT [23]. In this way, this tokenization approach can be applied to various types of sensing signals without sensor-specific adjustments or reconfigurations.

3.3 Share-Weighted Encoder

Rather than concatenating tokens from all channels into a single long sequence and processing them with a full attention transformer, we treat each channel of the multivariate signal as an independent input stream. Although all channels share the same transformer backbone, the forward pass is executed separately for each one. This design allows the model to first learn the temporal characteristics of each sensor without interference from others. It not only reduces computational cost but also increases flexibility. Because each channel is processed independently, the model can be pretrained on datasets with varying numbers or types of sensors and later fine-tuned on a target task with a different sensor configuration.

3.4 Channel-Aware Attention with Liaison Special Token

Following the tokenization step, we adopt common reconstruction-based pretraining strategies from **masked auto encoder (MAE)** [31, 34, 92], as shown in Figure 1(B.1), where input tokens are randomly masked and the model is trained to reconstruct the original time series using **mean squared error (MSE)** loss. Inspired by Huang et al. [34], we experiment with four masking strategies, as shown in Figure 2(a), including masking on (1) temporal and scale, (2) scale only, (3) temporal only, and (4) unstructured axes. We observe that the temporal and scalar masking yields the best performance for the downstream tasks. For the model architecture, we construct the backbone of our proposed framework with a convolutional patching layer followed by 12 standard Transformer blocks [79]. For the same reason, NORMWEAR uses a lightweight decoder consisting of two Transformer blocks, combined with a linear projection layer and a convolution layer to reconstruct the raw physiological signals both temporally and spatially. We also prepend a [CLS] token to each signal channel, following standard practice in transformer models, for learning a global representation of the input sequence for that channel.

Another important point to consider is that although empirical studies [1, 57] show that channel-independent structures effectively capture local patterns, they fail to account for relationships across channels. To address this, we use the [CLS] token from each signal channel as a liaison token, allowing them to exchange information through the channel-aware fusion layer after every other encoder block. We explore several fusion approaches and different design of liaison token as shown in Figure 2(b), with each method described below.

Specifically, the [CLS] token serves as an abstract global representation for each signal modality. Here, we propose a hybrid fusion approach. We stack the [CLS] tokens from all signal modalities and perform feature fusion using a self-attention mechanism. The fused [CLS] token is then reattached to its original channel, enabling the newly learned information to be propagated to each patch token in subsequent transformer encoder layers.

We denote the input signals as x , where x_i represent the i th channel. We denote the process of achieving initial embedding, including applying CWT operation, patching, and the first convolution layer, as F_{init} , then we have:

$$F_{init}(x) \in \mathbb{R}^{N \times L \times E}, \quad \text{s.t. } F_{init}(x_i) \in \mathbb{R}^{L \times E}, \quad (1)$$

where N is the number of input sensing channels, L is the patch size, and E is the embedding size. We then denote the shared special token as $z = [CLS] \in \mathbb{R}^E$, and $z_i \in \mathbb{R}^E$ as the deep-copied version of special token being prepended at input channel i . After concatenation with the initial embeddings and pass through the first shared

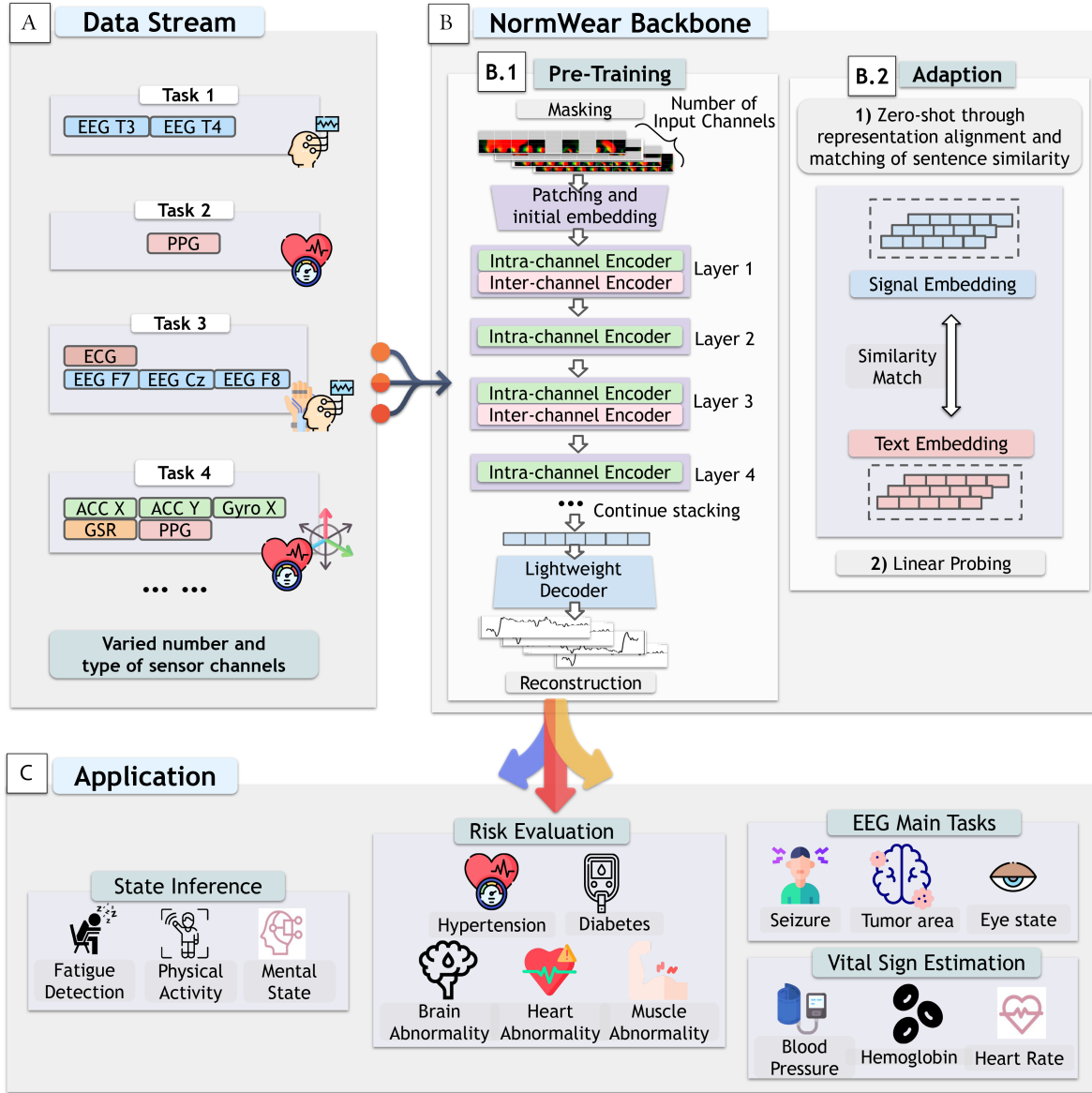


Fig. 1. The role of our framework. In the figure, module A indicate the data source with inter-device and inter-scenario variability. Module B briefly present our proposed NormWear model, a sensor agnostic wearable foundation model that have been pretrained on varied type of sensing signals, with signal-text representation alignment, as shown in module B.2, to confront the zero-shot evaluation. Module C depict different type of health applications that are being leveraged to evaluate the quality of our proposed foundation model. Several icons from Freepik [27] and Zhang et al. [93].

intra-channel encoder (self-attention layer), we will get the channel specific special token representation:

$$z_i = \text{softmax} \left(\frac{Q(z) \cdot K(F_{init}(x_i))^T}{\sqrt{d_k}} \right) \cdot V(z), \quad \text{s.t. } z_i \in \mathbb{R}^E, \quad (2)$$

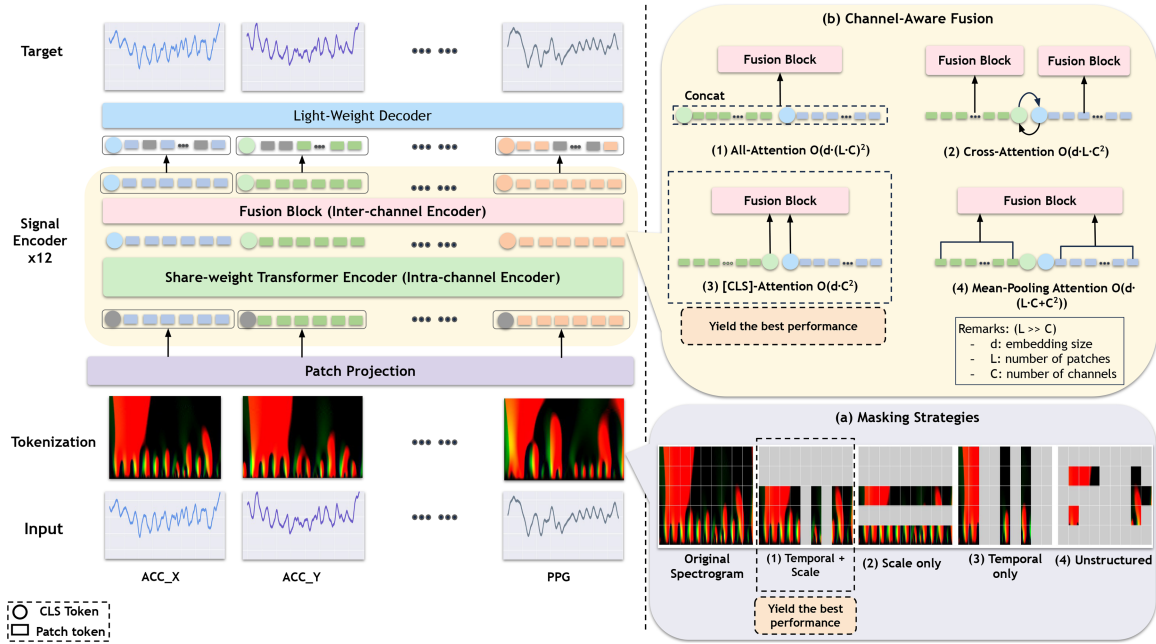


Fig. 2. Overview of the pretraining framework. The main structure of the backbone of NormWear is presented on the left. The model is pretrained following the standard masked autoencoder approach. Incoming sensor signals are transformed via continuous wavelet transform (CWT) into the initial multi-scale embeddings. These embeddings are then fed through alternating intra-channel and inter-channel encoder blocks in the backbone to extract high-level latent representations while exchanging information across channels. (A) Masking strategies applied to scalogram patches; (B) variations of the information fusion mechanism between input channels.

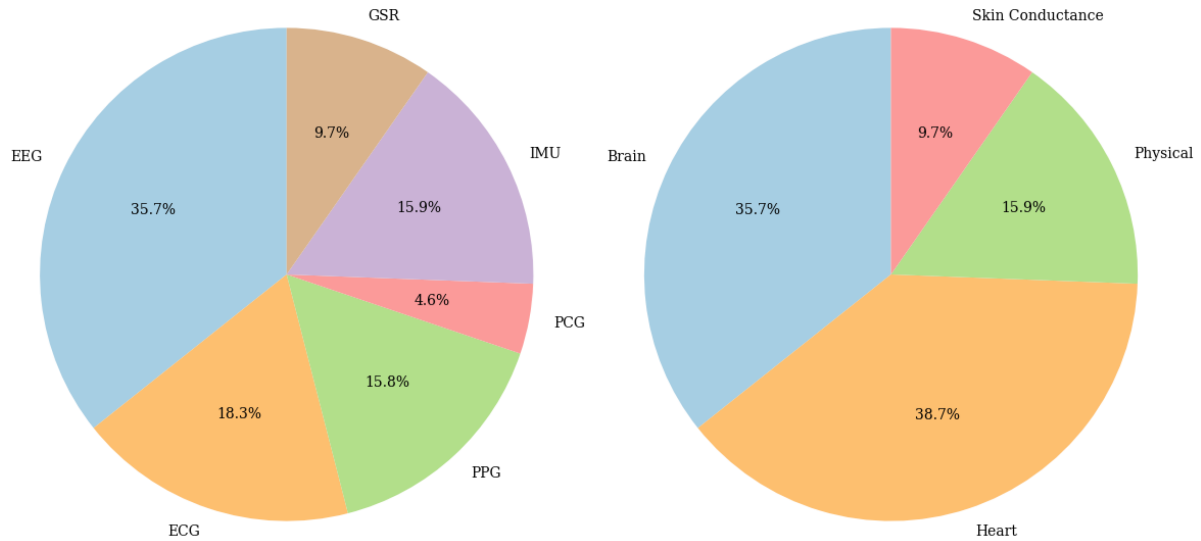


Fig. 3. Distribution of sensor signals used for pretraining. *Left*: Distribution by sensor modality. *Right*: Distribution by type of physiological information.

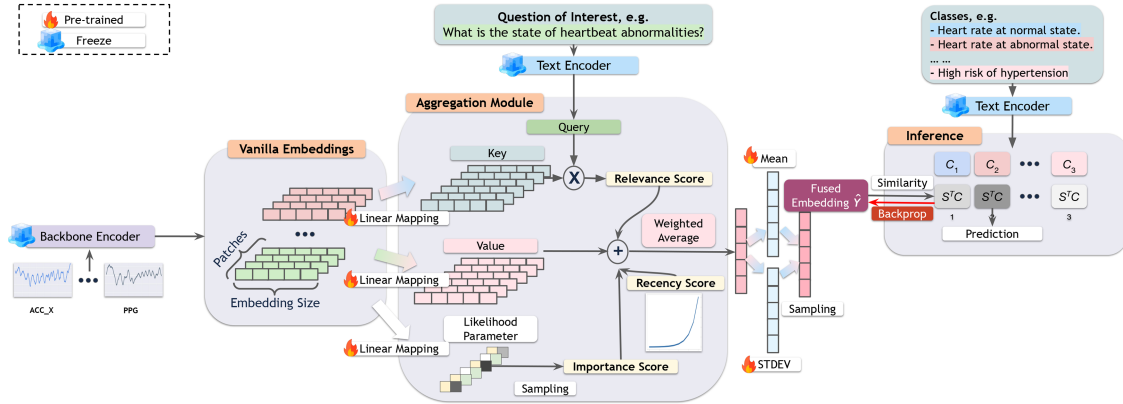


Fig. 4. Memory stream inspired temporal fusion mechanism for representation alignment. This module aggregate the sequence of representations, after the signal data pass through the backbone encoders, into a single fixed length embedding, according to three metrics: (i) relevance, which is conditioned by the semantic embedding of the question associated with the task of interest, (ii) recency, which is determined from the timestamp of each embedding, and (iii) importance, which is controlled by a gating function. The module is pretrained with contrastive learning to map the signal embedding, modulated by the prompted question, to the sentence of ground truth answer.

where d_k is the dimensionality of the key vectors, and Q, K, V are the linear mapping function of query, key, and value that is consistent with self-attention layer [79]. Then these channel specific special token representations will be further modulated through the inter-channel encoder:

$$z_{all} = [z_1, z_2, \dots] \in \mathbb{R}^{N \times E} \quad (3)$$

$$z_{modulate} = \text{softmax} \left(\frac{Q(z_{all}) \cdot K(z_{all})^T}{\sqrt{d_k}} \right) \cdot V(z_{all}), \quad \text{s.t. } z_{modulate} \in \mathbb{R}^{N \times E} \quad (4)$$

where each channel specific special token representation in $z_{modulate}$ will be copied back to each channel and prepend to each of their corresponding $F_{init}(x_i)$, and the result embeddings will be then send to the shared intra-channel encoder in the second encoding block.

Mean-Pooling Fusion. Similar to the [CLS]-Attention Fusion approach, we employ mean-pooling within each channel instead of using the [CLS] token as an abstract global representation.

Our empirical results show that [CLS]-attention fusion achieves the best downstreaming performance for our proposed NORMWEAR model. Details of all the ablation studies are reported in Appendix C. Beyond accuracy, we want to emphasize that the [CLS]-Attention Fusion design is highly flexible. This flexibility arises from the fact that self-attention is length-flexible and permutation-invariant [79]. Consequently, it integrates naturally with our shared-weight encoder, allowing the model to accommodate a variable number of sensor channels presented in any order. We provide additional empirical evidence of NormWear’s permutation invariance in Appendix C, Table C5.

3.5 Sensor-Semantic Representation Alignment

Zero-shot inference is an important aspect to evaluate foundation model. We evaluate our model in this setting by retrieving the closest text-derived label for each unseen task in the shared embedding space. Specifically, to unify information across diverse modalities, we incorporate a representation alignment objective that encourages the embeddings of physiological sensor data to reside in the same latent space as paired textual descriptions.

Once this shared space is established, it naturally supports zero-shot inference by allowing unseen sensor inputs to be interpreted through their proximity to text-derived anchors, without additional task-specific training. Several important works in this direction focusing on domains of vision-language [66], audio-language [85], and motion-language [96]. These works leverage end-to-end training to bind their modality of interest into semantic space. In this work, we extend this methodology to explore NORMWEAR’s ability to generalize across unseen datasets and tasks.

Building on prior work in representation alignment, we notice that in healthcare-related tasks where flexible inference across diverse scenarios is often required, the ground truth labels often have substantial overlap. For instance, depression is inferred from stress levels [42], and running and cycling produce similar IMU signals [43]. Due to these nested relationships, it create potential challenge to representation alignment when using contrastive learning, which requires clearly defined positive and negative pairs. To address this, we first propose a novel way to fuse the signal representations together with improved qualities, then align the representation with vector distance as an auxiliary loss for contrastive learning method. In addition, to reduce computation cost and counteract the issue of catastrophic forgetting [44], we use off-the-shelf frozen encoders for both signal and text modalities.

Human physiological signals are task-specific, dynamic, and often weakly labeled [30, 39, 46, 64]. To address these characteristics, we introduce three complementary scoring mechanisms during feature aggregation: *relevance scores* prioritize patches aligned with the task objective (e.g., IMU for activity recognition), guided by query sentences such as “*What activity is the subject doing?*” *recency scores* emphasize recent segments to better reflect the current physiological or emotional state [18, 20, 62, 65, 69]; and *importance scores* weigh signal segments that contain meaningful or transient patterns often buried in weakly labeled sequences. Together, these scores guide the **Memory Stream inspired Temporal Fusion (MSiTF)** fusion module to generate compact, task-aware representations. This design is inspired by memory-stream retrieval mechanisms [59] and is tailored to the demands of human-centered sensing tasks such as risk assessment, affect detection, and activity recognition.

MSiTF. Our Aggregation or Fusion Module, MSiTF as visually presented in Figure 4, is designed to addresses the above-discussed three challenges through three scores discussed below. Specifically, we denote f as the function that takes the semantic embedding of query sentence q and backbone output $H \in \mathbb{R}^{P \times E}$ as input, where P is the patch size and E is the embedding size, thus having the final fused representation $f(q, H) = \hat{Y} \in \mathbb{R}^E$.

We define the *Relevance* score as the cross attention between the key representations of each sensor timestep and the query sentence embedding, obtained from a pretrained language model (Clinical TinyLlama [53]). This mechanism allows the model to identify distinct but contextually relevant segments in the sensor input. For the *Recency* score, we use an exponential decay function to reflect the intuition that recent time steps are more important than earlier ones. Finally, we consider the importance score IMP in this case to be whether to keep the representation at each timestep or not. In order to achieve this, we assign binary parameters to each timestep, denoted as $\theta_t = p(v_t) \in \mathbb{R}^2$ where $v_t \in \mathbb{R}^E$ is the representation vector at timestep t and p is a trainable linear transformation function which will be optimized during pretraining. We then have the importance score for each patch defined as:

$$W_{imp}(t) = \arg \max_{i \in \{0,1\}} \frac{\exp \left(\left(\log(\theta_{t,i}) + \epsilon \right) / \tau \right)}{\sum_{j \in \{0,1\}} \exp \left(\left(\log(\theta_{t,j}) + \epsilon_j \right) / \tau \right)}, \quad (5)$$

where ϵ is the noise term sampled from Gumbel distribution [35], and τ is the temperature controlling the sharpness of the softmax function. Because $\arg \max$ is not a differentiable function, we will directly take the resulting probability corresponding to index at $j = 1$ to be the *importance* score, with τ being set to a small number to push the result closer to one hot vector from the softmax function. As a result, this logit function will

determine to what extent to activate the gate during forward pass on each patch of the input signals. The final score for each patch is the summation of the three scores as described above. This score will be treated as the weight for aggregating the representations from all the patches to form the fixed length embedded output (vector with size of 768 in our case).

Once the signal embeddings are aggregated, we adopt the reparameterization trick [40]. This design injects stochasticity into the representation, encouraging the model to explore and capture nuanced variations in semantic representations. Finally, we leverage contrastive learning with auxiliary loss on vector distance to train the MSiTF module with a projection layer to text representation on the pretraining datasets. The sentence template formation and training details are presented in Appendix B.5.

3.6 Data Augmentation

Since there are very few publicly available datasets containing multiple devices or modalities, we aim to expand our curated training set to fully leverage the potential of self-supervised learning. Inspired by data augmentation techniques in computer vision and natural language processing [16, 91], we adopt a heuristic approach to augment the dataset. Specifically, we augment each sub-dataset by a factor of 10. For each dataset, we sample two time series, randomly extract a segment from one, and substitute it with a transformed counterpart, as outlined in the pseudocode in Algorithm 1. Intuitively, this can be seen as synthesizing realistic temporal transitions that are rare in short recordings. For example, concatenating a walking segment with a running segment produces a signal that contains a clear transition where motion amplitude and frequency increase over time, resembling a short burst where a person accelerates from walking into running. Although the resulting sequence is synthetic, it preserves local sensor statistics while introducing plausible nonstationary dynamics that encourage the model to learn representations sensitive to temporal change rather than only steady state patterns. As a result, our training set is expanded to 2,586,404 segments, corresponding to 4,294 hours of data.

4 Experiments

NormWear is pretrained exclusively on the data shown in Table 2. In this section, we present a comprehensive evaluation across 11 downstream publicly available datasets, focusing on 18 widely recognized digital healthcare tasks. We evaluate the methods following order of zero-shot capability, partial-shot learning, and full-shot learning.

4.1 Selection of Baselines Covering Representative Modeling Strategies

Modeling multivariate wearable signals with arbitrary input channels and sensor types, such as those capturing activities of heart, brain, and body physical motions, presents unique challenges, as no universally recognized open source baseline or **state-of-the-art (SoTA)** model exists in this domain. To evaluate our approach, we selected diverse and representative baselines (as shown in Table 4).

In the literature, various modeling strategies have been proposed. Firstly, early approaches involved handcrafting statistical features, which was a widely adopted practice in signal processing [50, 68, 87]. We include this simple baseline as sanity check. Secondly, since sensory data can be naturally represented as time series [71, 84], we benchmarked our model against Chronos [7], as well as a self-supervised framework TF-C [97]. Finally, the spectrum-based modeling methods [21, 41, 80] are widely used for signal modeling. Therefore, we incorporate CLAP [85] into baselines that has demonstrates SoTA performance in spectrogram-based modeling. Regarding the comparison with concurrent works proposing foundation models for a specific sensor modality, we leverage PaPaGei [60] for PPG datasets, ECG-FM [48] for ECG datasets, and CBraMod [81] for EEG datasets. These baselines span distinct paradigms, providing a solid foundation to demonstrate the strengths of our model in wearable signal tasks. For uni-modal baselines like Chronos and CLAP, we feed each signal separately into model and

Table 3. Zero-Shot Performance on the Downstream Datasets, with AUC ROC Being Reported

Model	WESAD	UCI-HAR	DriverFatigue	GAMEMO	Epilepsy (eye open state)	Epilepsy (eye relaxation)	Epilepsy (health area)	Epilepsy (tumor area)	Epilepsy (seizure)	PPG-BP (HTN)	PPG-BP (DM)	PPG-BP (CVA)	PPG-BP (CVD)	ECG-Abnormal	PhysioNet EMG	Micro Avg.	Macro Avg.
CLAP—MD	45.3	62.8	58.5	53.1	44.9	45.1	47.6	30.5	84.9	59.4	41.8	46.0	57.4	22.9	55.4	50.4	51.2
CLAP—DP	50.7	52.3	61.1	51.6	54.4	41.9	58.6	46.4	74.3	52.2	41.4	50.6	58.9	42.7	38.3	51.7	52.2
before bind	44.1	48.2	52.1	48.4	54.1	62.6	53.9	52.5	24.6	48.8	49.6	46.3	56.8	54.3	48.2	49.6	49.4
NormWear	55.8	71.2	57.2	51.0	55.7	61.3	67.6	55.8	66.0	57.1	62.5	70.0	59.0	63.1	70.1	61.6	61.5
— w/o IMP	56.2	70.3	55.4	49.8	54.0	56.5	66.9	57.3	52.9	56.5	54.3	61.7	60.7	73.4	65.2	59.4	59.6
— w/o text aug	54.8	65.8	55.2	49.2	31.0	58.4	58.6	32.8	58.1	50.2	52.6	50.8	50.6	47.7	33.6	50.0	51.4
— w/o refine	59.5	72.8	42.7	57.3	50.6	69.0	43.3	50.5	74.8	48.3	38.8	44.6	44.1	72.4	75.7	56.3	56.6

The last two columns show the average across the tasks and across group types, respectively. The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme. The highest performance achieved for each task is highlighted in bold.

concatenate their representations after the forward pass. This ensures that all models have the same field of view, making the comparison fair.

4.2 Zero-Shot Evaluation

We achieve zero-shot inference by pretraining our proposed novel temporal fusion module on the task of representation alignment. We include the SoTA spectral-based model CLAP [85] as a baseline to provide a more comprehensive comparison of the results. For CLAP, we experimented with both **Manhattan distance (MD)** and **dot product (DP)** as similarity metrics during inference. We observe that there are no statistically significant differences in performance when using MD and DP for label retrieval in CLAP. From Table 3, we could observe that overall, NORMWEAR equipped with MSiTF outperforms the baselines. We compare NORMWEAR with a few ablations by removing importance score (w/o IMP) and removing text augmentation (w/o text aug). We can observe that performance drop after removing each of the above components, verifying their respective importance in improving generalization across various downstream tasks. We present this outcome to demonstrate the zero-shot capability in the wearable signal domain, an aspect not present in recent studies. We also hope this outcome could potentially provide a new perspective that can help drive progress in this direction within the field.

4.3 Partial-Shot and Full-Shot Evaluation

We evaluate the learned representations using linear probing through supervised training on each downstream dataset, and report performance on the corresponding held-out test set. To ensure fair comparison, we use a unified evaluation protocol with identical hyperparameter settings and implementation across all models and the dataset [88]. This design ensures that performance differences are not due to variations in learning rate, regularization, or data augmentation [58]. Specifically, the classification tasks, using logistic regression, are solved by Newton’s method with conjugate gradient, with AUC ROC being reported as main metric. The regression (vital signs) tasks, using ridge regression, are solved by Cholesky’s method with closed form solution, with relative accuracy being reported. For partial-shot evaluation, we leverage 10% of the training data for the linear probing, and detailed performance result is presented in Appendix C, Table C4. The full-shot evaluation results is presented in Table 5. All scores are the higher the better.

Table 4. Comparisons of Studies on Wearable-Sensor Modeling

Framework	Models	Sensor Agnostic	Frequency Modeling	MTS Modeling	Zero Shot Inference
Statistical	Statistical Approach	✓	✗	✗	✗
Sensor Specific	PaPaGei [60]	✗	✗	✓	✗
	ECG-FM [48]	✗	✗	✓	✗
	CBraMod [81]	✗	✓	✓	✗
	CLAP [85]	✗	✓	✗	✓
SSL	Chronos [7]	✓	✗	✗	✗
	TF-C [97]	✓	✓	✗	✗
	NormWear (ours)	✓	✓	✓	✓

MTS, multivariate time series. The highest performance achieved for each task is highlighted in bold.

From Figure 5(A), Table 5, and Appendix E Table E1, we observe that NORMWEAR consistently achieves peak performance across all task groups, including activity recognition, EEG signal analysis, disease risk evaluation, and vital sign estimation. Furthermore, its leading performance remains consistent across various evaluation metrics. Based on the macro-averaged total score across task groups, NORMWEAR delivers a 3.9% improvement over the SoTA time-series self-supervised learning framework [97], a 5.3% improvement over the SoTA spectrum-based modeling method [85], a 6.1% improvement over SoTA time-series forecasting models with LLM backbones [7], and a 5.2% improvement over standard statistical baselines. On larger datasets, NORMWEAR significantly outperforms the statistical baseline by 9.0% and 7.5% for activity recognition and EEG brain activity monitoring tasks, respectively. On smaller datasets, it still achieves peak performance in disease risk evaluation. For vital sign estimation, all methods yield comparable results, suggesting inherent challenges in these regression tasks that warrant further investigation but are beyond the scope of this study.

When comparing with recent modality specific foundation models, NormWear’s main benefit is that it capture cross-modal relationships, making it more versatile for wearable health tasks. While it sacrifices modality-specific optimization for adaptability, this may slightly reduce performance in highly specialized tasks. Single-signal models excel in their domains due to deeper modality-focused training. Instead of maximizing single-modality data, we prioritize signal diversity for better generalization. Benchmarking shows that NormWear, trained on a smaller dataset than EEG-only models, still achieves competitive results, highlighting the effectiveness of our pretraining approach. These findings illustrate NORMWEAR’s capacity to balance consistency and adaptability across a diverse range of tasks and conditions. By excelling across standard benchmarks while addressing the intricacies of varied applications, NORMWEAR exemplifies the philosophy of a foundation model: a reliable generalist capable of performing robustly across both typical and challenging scenarios.

4.4 Inspect Model Output on Regression Tasks

In our evaluation, regression tasks are generally more intuitive, and all methods, including the baseline, achieve reasonable performance. While the proposed model shows statistically significant improvements, the gains are less pronounced than in classification tasks. In this study, regression evaluation serves mainly as a sanity check: it confirms that, while the model pursues better performance on more complex tasks, it maintains basic predictive

Table 5. Detailed Performance on Various Downstream Wearable-Signal-Based Health Related Applications under Full-Shot Linear Probing Evaluation

Downstream Tasks	Statistical	Chronos	CLAP	TF-C	Modality-Specific	NormWear (Ours)
WESAD	66.213	71.489	72.383	69.865	56.656 (max (PPG, ECG))	76.060
UCI-HAR	95.784	91.593	96.420	96.892	-	98.954
DriverFatigue	63.249	76.722	61.889	66.882	80.430 (EEG)	74.292
Activity Recognition Avg.	75.082	79.935	76.897	77.880	-	83.102
Epilepsy (eye open state)	82.489	82.41	85.094	89.153	90.436 (EEG)	92.743
Epilepsy (eye relaxation)	87.457	88.218	89.867	94.416	95.552 (EEG)	94.828
Epilepsy (health area)	86.274	81.08	83.711	85.619	88.065 (EEG)	88.541
Epilepsy (tumor area)	82.816	81.034	83.644	86.348	87.258 (EEG)	87.197
Epilepsy (seizure)	88.272	97.572	97.734	93.998	94.616 (EEG)	97.053
GAMEEMO	51.009	53.747	52.551	56.275	55.420 (EEG)	54.937
EEG Main Tasks Avg.	79.720	80.677	82.100	84.302	85.225	85.883
ECG-Abnormal	97.092	98.585	97.23	98.275	89.898 (ECG)	99.140
PPG-BP (HTN)	59.499	52.425	56.757	65.229	61.839 (PPG)	62.341
PPG-BP (DM)	47.823	51.164	42.455	57.883	55.668 (PPG)	55.893
PPG-BP (CVA)	71.250	50.278	51.667	58.125	73.125 (PPG)	70.625
PPG-BP (CVD)	51.219	58.31	50.91	58.674	49.066 (PPG)	51.773
PhysioNet EMG	99.309	61.6	98.627	78.308	-	99.216
Risk Evaluation Avg.	71.032	62.060	66.274	69.416	-	73.165
Noninvasive-BP	92.310	91.79	91.922	87.481	90.596 (PPG)	92.420
PPG-Hgb	94.219	95.005	94.291	93.408	94.912 (PPG)	94.632
Fetal-fPCG	98.929	99.048	99.195	99.077	-	99.072
Vital Signs Avg.	95.153	95.281	95.136	93.322	-	95.375
Micro Avg.	78.623	76.782	78.130	79.773	-	82.762
Macro Avg.	80.247	79.488	80.103	81.230	-	84.381

The signal name, in column “Modality-Specific,” following each performance score denotes the model specialized for that modality: PPG [60], ECG-FM [48], and EEG [81]. The highest performance achieved for each task is highlighted in bold. The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme.

capabilities comparable to baseline methods. Further analysis shows that the model outputs generally align with the ground truth, exhibiting a strong linear correlation, as illustrated in Figure 6. Some tasks, such as heart rate estimation via Fourier transform or **blood pressure (BP)** estimation from PPG [9], are inherently predictable due to known physiological relationships, which explains why the model can capture these patterns effectively.

4.5 Comparison with Existing Foundation Models for Physiological Signals

Because our proposed framework focuses on signal-level foundation modeling, aiming to learn generalizable representations directly from continuous physiological waveforms (e.g., ECG, EEG, PPG), we therefore added comparisons with several recent signal-specific foundation models designed specifically for PPG [60], EEG [81], and ECG [48] as modality-specific expert models. Detailed quantitative results are presented in Table 5, and a cleaned conclusive result is presented in Table 6. We observe that NormWear achieves competitive and often superior performance across diverse downstream tasks.

While NormWear aims to generalize across diverse sensor modalities, this introduces several design and deployment tradeoffs compared to modality-specific models. During training, generalist models require careful balance across modalities to prevent overfitting to dominant signal types, whereas modality-specific models can focus more narrowly on a single data distribution. This balancing process, however, exposes the model to richer and

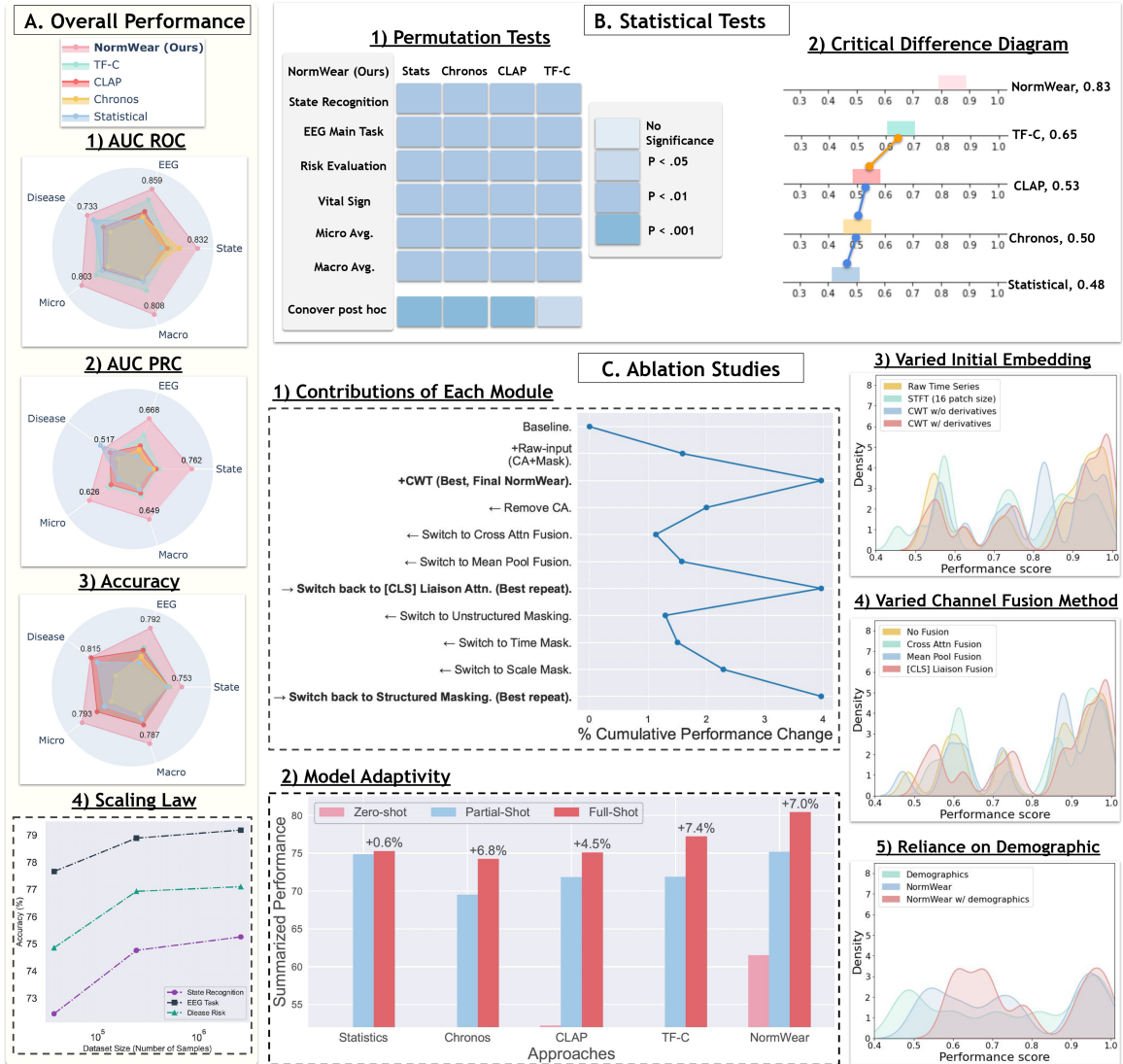


Fig. 5. Overview of performance trend of NORMWEAR. (A) Performance comparison under full-shot linear probing, against competitive baselines in downstream tasks: (1) Disease risk predictions. (2) EEG main tasks (mental and abnormal states prediction). (3) State recognition: physical and mental activities. (4) Macro: Average performance over types of tasks. (5) Micro: Average performance over each task. (B) Statistical texts justifying the leading performance of our proposed model. (C) Summarization of the extensive ablation studies justifying the scaling properties, downstream adaptivity, and the performance contribution of each module within our proposed framework.

more varied signal patterns, potentially enabling broader generalization. In deployment, modality-specific models may be easier to fine-tune when the target modality remains fixed, but become cumbersome when extending to new sensor types. In contrast, NormWear's generalist design supports more scalable deployment across heterogeneous signals without needing a separate model per modality, though fine-tuning still demands attention to cross-modal data balance. These tradeoffs reflect a broader tension between specialization and adaptability in

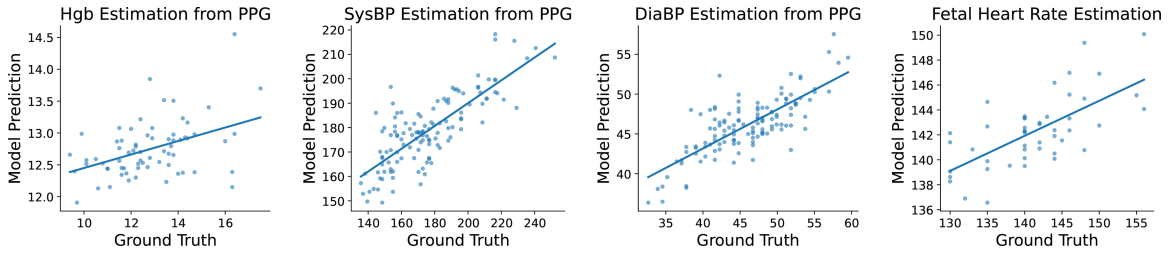


Fig. 6. Visualization of model output on regression tasks.

Table 6. Performance Comparison with Existing Foundation Models

Modality	BioBART [90]	Existing Foundation Model	NormWear
PPG	53.834	59.271	63.338
EEG	77.920	84.540	84.227
ECG	65.702	73.229	87.600
Total Average	68.493	73.397	77.476

The averaged AUROC scores across all the downstream tasks with same modality are presented. The best performance achieved for each task is highlighted in bold.

Table 7. Performance Comparison across Different Input Variants on Downstream Tasks (%)

Task	Raw Time Series	STFT	CWT	CWT w/ Derivatives
State Recognition	80.895	76.555	<u>82.093</u>	83.162
EEG Tasks	<u>85.466</u>	79.565	81.522	85.881
Vital Sign	94.220	94.427	<u>95.064</u>	95.441
Disease Risk Evaluation	69.513	64.153	<u>70.922</u>	73.317
Micro Average	<u>80.845</u>	76.403	80.341	82.833

The highest performance is highlighted in bold, and the second-highest is denoted by underlining.

foundation model design for wearables. Future work could further explore modular or adaptive architectures to mitigate this tension and support both modality-specific refinement and cross-modal generalization.

4.6 Ablation Study: Justification of the CWT-Based Initial Embedding

Considering that there are multiple steps contained in the proposed CWT-based initial signal embedding, to justify the effectiveness of the proposed mechanism, we conduct ablation study to inspect the contribution of each proposed component. From Table 7, we observe that the **short-time Fourier transform (STFT)**, although widely adopted in prior literature, did not bring benefit in performance. In contrast, using the CWT as the initial embedding generally improves performance over directly using raw time-series signals, with the exception of EEG data. Furthermore, when CWT is combined with an additional CWT representation computed from the derivatives of the raw signals, the proposed model consistently achieves the best performance across all downstream tasks and domains. These findings demonstrate that each added component contributes meaningfully to the final performance, thereby justifying the increased model complexity.

Table 8. Performance Comparison under Different Pretraining Data Sizes

Metric	Task	10^4	10^5	10^6
AUC-ROC	State Recognition	80.554	<u>81.842</u>	83.102
	EEG Tasks	84.469	<u>85.689</u>	85.883
	Disease Risk Evaluation	69.357	69.666	<u>69.413</u>
	Micro Average	77.123	<u>77.958</u>	78.156
Accuracy	State Recognition	72.426	<u>74.761</u>	75.254
	EEG Tasks	77.658	<u>78.885</u>	79.182
	Disease Risk Evaluation	74.854	<u>76.930</u>	77.101
	Micro Average	75.450	<u>77.256</u>	77.535

The highest performance is highlighted in bold, and the second-highest is denoted by underlining.

In practice, the initial embedding incurs minimal computational cost. The derivative reduces to a first-order difference with linear complexity, while the CWT is implemented as a convolution with a fixed kernel. Implemented in PyTorch and executed on GPU, the entire process is equivalent to a fixed convolutional layer and is substantially lighter than the successive layers in the proposed model.

4.7 Validation of Model's Compatibility with Scaling Law

We investigate how the amount of pretraining data affects model performance as an important validation of the proposed foundation model. The result is summarized in Figure 5 panel A.4, and quantitative result in Table 8. Our results demonstrate that increasing pretraining data size leads to consistent improvements across multiple downstream tasks and domains, indicating favorable generalization behavior. Notably, the performance gains exhibit a logarithmic-like growth pattern rather than exponential scaling, a phenomenon widely observed in prior large-scale learning studies [54]. Taken together, these results suggest that scaling pretraining data acts as a reliable mechanism for incremental performance improvement, rather than inducing a qualitative change in model behavior. In this work, scaling is therefore viewed as a complementary validation of the proposed foundation model, demonstrating its capacity to effectively leverage additional data and computation when available.

4.8 Quantify the Intrinsic Behaviors: Nonlinear Dynamics Analysis on the Layer-Wise Waveforms

Understanding the representations extracted by intermediate layers is crucial to interpreting our model's behavior. While one could directly visualize the millions of parameters and hundreds of waveform features per layer such as the example we provided in Appendix I.2, such raw visualization becomes inefficient and difficult to generalize. Instead, we adopt nonlinear dynamics analysis as a compact yet powerful way to capture the overall temporal patterns of these internal representations. Specifically, we draw from chaos theory to measure the dynamics of the layer-wise signals using metrics such as the Lyapunov exponent [83] (sensitivity to initial conditions), Hurst exponent [63] (long-term self-correlation), and persistence entropy [87] (unpredictability in system states). We obtain the following key observations from the analysis result presented in Figure 7(B):

Observation 1. Deeper Layers Capture Higher-Order Complexity. For signals such as GSR, EEG, and ACC, deeper layers show lower long term self-correlation (DFA [33]) and higher unpredictability (persistence entropy), indicating a transition to representations that are less periodic and more chaotic. The decrease in the Lyapunov exponent across layers suggests reduced variation in extracted features, aligning with the idea that deeper layers capture more abstract, long-term patterns with broader receptive fields.

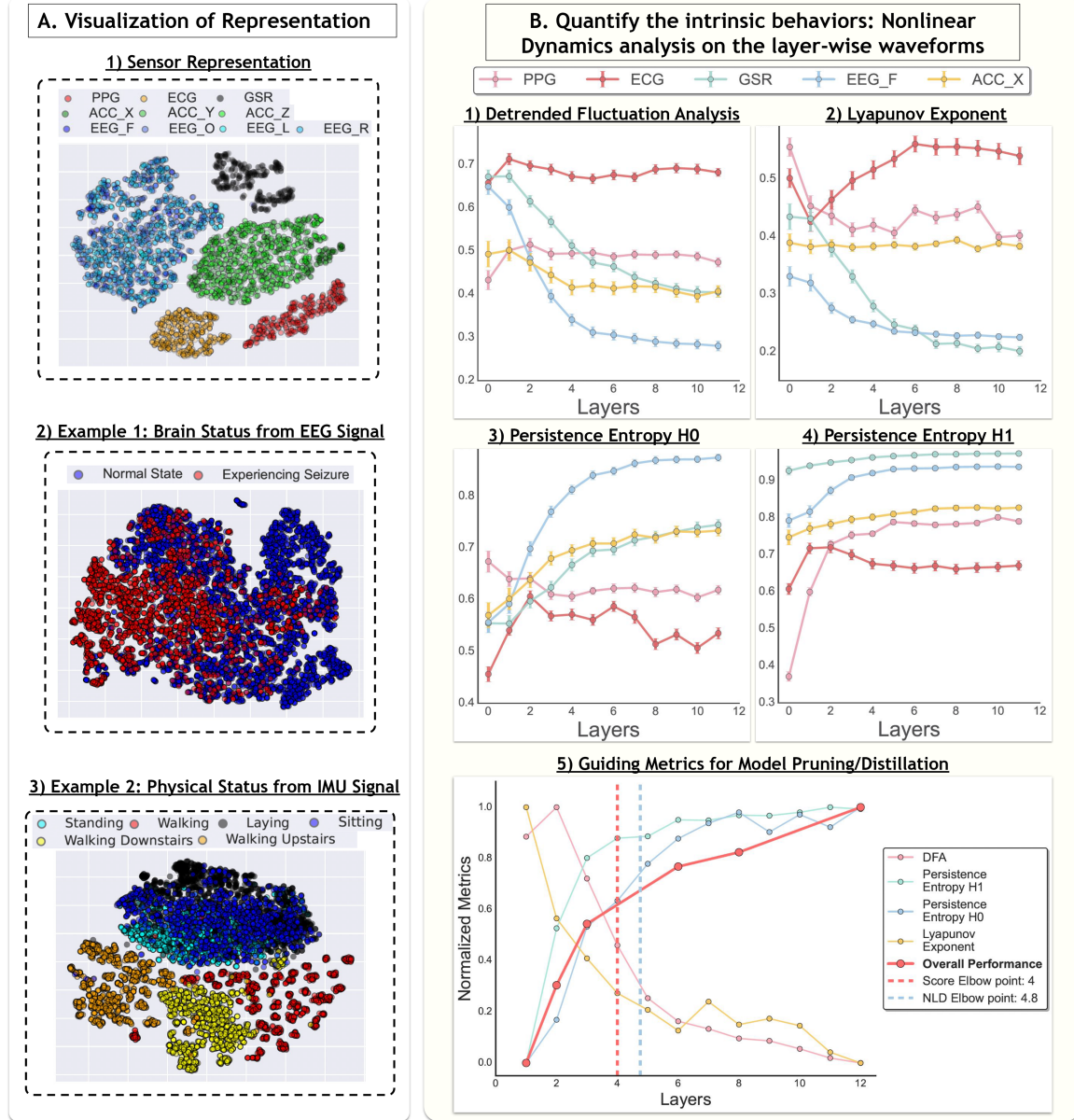


Fig. 7. Quantitative analysis of the representation (A) T-SNE plot of the representation learned by the model. (B) Quantify the representation learned by the model through Nonlinear Dynamics analysis on the intermediate output.

Observation 2. Reflecting Modalities with Simpler Dynamics. In contrast, PPG and ECG signals, dominated by regular heart activity, exhibit more stable patterns across layers. This aligns with their simpler waveform structures and less complex dynamics compared to signals related to neural and physical activities.

Observation 3. Quantitative Metrics for Guiding Model Pruning or Distillation. We observe that the nonlinear dynamical metrics naturally serves as a guidance toward model post-training optimization. The bottom sub-figure

in Figure 7(B) present the observation that such nonlinear dynamical analysis also reveals the number of layers that carry sufficient information, which is important for optimization strategies such as layer pruning or model distillation. In this figure, we firstly put all the nonlinear dynamical metrics together after min–max normalization, along with the performance from ablation studies on the 18 downstream tasks with output representations from different layers. We then calculate the elbow point for each of the curve. Secondly, the average elbow point of the four nonlinear dynamical metrics is presented. Regarding the curve for the performance, the average elbow point of the performance across state recognition, EEG main tasks, and disease risks is presented, and the macro average score is plotted in the figure for cleanliness. Finally, considering the design of our architecture with inter-channel encoder placed at every other encoding blocks starting at the first block, if we are about to pruning or distillation layers, we should always keep even number of layers since the information need to be fused back at layers with even indexes after the [CLS] liaison fusion happens at layers with odd indexes. Therefore, the nearest layer with even index to the elbow point provided by the average elbow points across nonlinear dynamical metrics is 4, which is consistent with the average elbow point across performances as the empirical evidence from our ablation studies.

These visualizations reveal that the model progressively transforms raw sensory data into representations aligned with the complexity of each signal. For GSR and EEG, deeper layers exhibit increased unpredictability and reduced periodicity, highlighting the extraction of nuanced, higher-order patterns critical for human sensing. In contrast, the stability of representations for PPG and ECG reflects their simpler dynamics, demonstrating the model's adaptability to varying signal characteristics. This analysis confirms that the intermediate representations are purposefully optimized to capture the temporal and structural nuances of each modality, enhancing our understanding of model's intrinsic behavior in a quantitative way. Finally, this analysis serves as a principled quantitative tool for guiding post-training optimization, which usually involve layer pruning and model distillation.

5 Discussion and Conclusion

Engineering Breakthrough. This study introduces NORMWEAR, a foundation model for wearable sensing that establishes a new paradigm for universal health monitoring. Unlike conventional approaches tied to fixed sensor configurations, NORMWEAR employs a channel-aware mechanism to accommodate heterogeneous inputs. This enables robust generalization across devices, sensor types, and deployment settings, creating a form of universal wearable AI that is independent of hardware uniformity. By providing a unified intelligence layer for wearable systems, NORMWEAR lowers integration barriers and paves the way for scalable, inclusive digital health solutions.

Generalizability. NORMWEAR aims to serve as a practical tool that could serve as a starting point for researchers and clinicians when tackling a problem with wearable based signal data. Our proposed model could extract informative representations from raw signal series, which can be leveraged for further machine learning modeling, clustering, embedding vector-based information retrieval, and deployment of real-time health states monitoring with minimal tuning. We've justified the utilizability and generalization of NORMWEAR through an extensive evaluation of various ubiquitous health applications. As for future works, it is important to leverage our framework on larger scale clinical applications and explore the applicability of embedding vectors as state representations for intervention modeling problems that comprise the decision-making process.

Insights into Model Behavior and Future Optimization. Our analysis of intermediate representations further highlights how NORMWEAR adapts to the intrinsic complexity of different physiological signals. For example, deeper layers produced increasingly irregular and less periodic representations for GSR and EEG, reflecting their higher dynamical complexity, while maintaining stable structures for PPG and ECG with simpler dynamics. This adaptability demonstrates that the model is not only performing predictive tasks but is also internally optimized to capture modality-specific temporal and structural features. Such quantitative characterization provides a principled lens for interpreting the model's intrinsic behavior and, importantly, offers guidance for post-training optimization strategies such as layer pruning or model distillation.

Limitation and Future Work. We acknowledge several limitations to be addressed in future work. (1) The representation alignment component is currently trained on a limited set of healthcare-related objectives, and expanding the pretraining corpus with more diverse semantic labels may improve generalization. (2) While our design supports classification tasks well, adapting the framework for regression remains an open challenge, and future work may explore alternative formulations beyond label discretization. (3) NormWear currently focuses on physiological signals with relatively narrow frequency bands; extending its applicability to higher-frequency modalities such as audio or lower-resolution clinical summaries is a promising direction.

Regarding the second limitation specifically, although our framework achieves reasonable performance on regression tasks in supervised settings, zero-shot regression remains inherently challenging, particularly for a general-purpose model. Our additional analysis shows that the model’s predictions still align reasonably well with the ground truth, demonstrating that it maintains basic predictive capability. A more thorough exploration of zero-shot regression, potentially involving alternative formulations beyond label discretization, is an interesting direction for future work.

Broad Impact. NORMWEAR is the first foundation model tailored for multivariate physiological signals that supports a wide range of wearable health tasks across sensor modalities, device types, and clinical applications. Through a unified CWT-based tokenization pipeline and a channel-aware fusion mechanism, it enables robust, modality-agnostic representation learning. Our extensive evaluation across zero-shot, partial-shot, and full-shot settings demonstrates NormWear’s strong generalizability and practical relevance. We believe NormWear provides a valuable resource for advancing foundation modeling in digital health and promoting more unified benchmarks in the community.

Ethics Statement

This study contains applications in the field of healthcare. We ensured that all the data being used during pretraining and evaluations were made publicly available by the original authors, and all these works were cited properly.

Reproducibility Statement

The full code base, comprising all the scripts for exploratory data analysis and preprocessing, model construction, pretraining, downstream evaluation, result analysis, and all the visualizations that are described in this article, is published at <https://github.com/Mobile-Sensing-and-UbiComp-Laboratory/NormWear>.

Acknowledgments

We thank the Supercomputing Center and Halicioğlu Data Science Institute at University of California San Diego for providing the working space and the computational resources necessary to conduct all experiments.

References

- [1] Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. 2024. Large-scale training of foundation models for wearable biosignals. arXiv:2312.05409. Retrieved from <https://arxiv.org/abs/2312.05409>
- [2] Tomy Abuzairi, Ester Vinia, Muhammad Arkana Yudhistira, Mia Rizkinia, and Winda Eriska. 2024. A dataset of hemoglobin blood value and photoplethysmography signal for machine learning-based non-invasive hemoglobin measurement. *Data in Brief* 52 (2024), 109823. DOI: <https://doi.org/10.1016/j.dib.2023.109823>
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv:2303.08774. Retrieved from <https://arxiv.org/abs/2303.08774>
- [4] Talha Burak Alakus, Murat Gonen, and Ibrahim Turkoglu. 2020. Database for an emotion recognition system based on EEG signals and various computer games—GAMEEMO. *Biomedical Signal Processing and Control* 60 (2020), 101951.
- [5] Nibras Abo Alzahab, Angelo Di Iorio, Luca Apollonio, Muaaz Alshalak, Alessandro Gravina, Luca Antognoli, Marco Baldi, Lorenzo Scalise, and Bilal Alchalabi. 2021. Auditory evoked potential EEG-biometric dataset. (version 1.0.0). *PhysioNet* (2021). RRID:SCR_007345. DOI: <https://doi.org/10.13026/ps31-fc50>

- [6] Ralph Gregor Andrzejak, Klaus Lehnertz, Christoph Rieke, Florian Mormann, Peter David, and Christian E. Elger. 2023. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E* 64 (2023), 061907. DOI: <https://doi.org/10.34810/data490>
- [7] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the language of time series. arXiv:2403.07815. Retrieved from <https://arxiv.org/abs/2403.07815>
- [8] Nikesh Bajaj, Jesús Requena Carrión, and Francesco Bellotti. 2020. PhyAAt: Physiology of auditory attention to speech dataset. arXiv:2005.11577. Retrieved from <https://arxiv.org/abs/2005.11577>
- [9] Gholamreza Bazmandegan, Mitra Abbasifard, Hamid Ostadebrahimi, Mohammadreza Gholamrezapour, and Zahra Kamiab. 2023. Positive association between hemoglobin concentration and blood pressure in adults: A cross-sectional study based on Rafsanjan cohort study. *International Journal of Hypertension* 2023, 1 (2023), 6283711.
- [10] Win-Ken Beh, Yi-Hsuan Wu, and An-Yeu (Andy) Wu. 2021. MAUS: A dataset for mental workload assessment on N-back task using wearable sensor. *DataPort*. DOI: <https://doi.org/10.21227/q4td-yd35>
- [11] Amrutha Bhaskaran, Sidhesh Kumar, J. Shirley George, and Manish Arora. 2022. Heart rate estimation and validation algorithm for fetal phonocardiography. *Physiological Measurement* 43, 7 (2022), 075008. DOI: <https://doi.org/10.1088/1361-6579/ac7a8c>
- [12] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2022. On the opportunities and risks of foundation models. arXiv:2108.07258. Retrieved from <https://arxiv.org/abs/2108.07258>
- [13] Ralf Boussetjolot, Dieter Kreiseler, and Allard Schnabel. 2009. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Retrieved from <https://api.semanticscholar.org/CorpusID:111121953>
- [14] E. Oran Brigham. 1988. *The Fast Fourier Transform and Its Applications*. Prentice-Hall, Inc., Hoboken, New Jersey.
- [15] M. J. Burke and M. Nasor. 2004. Wavelet based analysis and characterization of the ECG signal. *Journal of Medical Engineering & Technology* 28, 2 (2004), 47–55.
- [16] Chris U. Carmona, François-Xavier Aubet, Valentin Flunkert, and Jan Gasthaus. 2021. Neural contextual anomaly detection for time series. arXiv:2107.07702. Retrieved from <https://arxiv.org/abs/2107.07702>
- [17] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2021. Unsupervised learning of visual features by contrasting cluster assignments. arXiv:2006.09882. Retrieved from <https://arxiv.org/abs/2006.09882>
- [18] Sidhartha Chaudhury, Chenggang Yu, Ruifeng Liu, Kamal Kumar, Samantha Hornby, Christopher Duplessis, Joel M. Sklar, Judith E. Epstein, and Jaques Reifman. 2021. Wearables detect malaria early in a controlled human-infection study. *IEEE Transactions on Biomedical Engineering* 69, 6 (2021), 2119–2129.
- [19] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-attention multi-scale vision transformer for image classification. arXiv:2103.14899. Retrieved from <https://arxiv.org/abs/2103.14899>
- [20] Moajjem Hossain Chowdhury, Md Nazmul Islam Shuzan, Muhammad E. H. Chowdhury, Zaid B. Mahbub, M. Monir Uddin, Amith Khandakar, and Mamun Bin Ibne Reaz, Amith Khandakar, and Mamun Bin Ibne Reaz. 2020. Estimating blood pressure from the photoplethysmogram signal and demographic features using machine learning techniques. *Sensors* 20, 11 (2020), 3127.
- [21] Se Young Chun, Jae-Hwan Kang, Hanvit Kim, Chung-ho Lee, Ian Oakley, and Sung-Phil Kim. 2016. ECG based user authentication for wearable devices using short time Fourier transform. In *Proceedings of the IEEE 2016 39th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, New York, NY, 656–659. DOI: <https://doi.org/10.1109/TSP.2016.7760964>
- [22] Muhammad Najam Dar, Amna Rahim, Muhammad Usman Akram, Sajid Gul Khawaja, and Aqsa Rahim. 2022. YAAD: Young adult's affective data using wearable ECG and GSR sensors. In *Proceedings of the IEEE 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*. IEEE, New York, NY, 1–7.
- [23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929. Retrieved from <https://arxiv.org/abs/2010.11929>
- [24] Amirhossein Esmaili, Mohammad Kachuee, and Mahdi Shabany. 2017. Nonlinear cuffless blood pressure estimation of healthy subjects using pulse transit time and arrival time. *IEEE Transactions on Instrumentation and Measurement* 66, 12 (2017), 3299–3308.
- [25] Hamid Fekri Azgomi, Luciano R. F. Branco, Md Rafiul Amin, Saman Khazaei, and Rose T. Faghieh. 2023. Regulation of brain cognitive states through auditory, gustatory, and olfactory stimulation with wearable monitoring. *Scientific Reports* 13 (2023), 12399. DOI: <https://doi.org/10.1038/s41598-023-37829-z>
- [26] Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I. Webb, and Mahsa Salehi. 2024. Improving position encoding of transformers for multivariate time series classification. *Data Mining and Knowledge Discovery* 38, 1 (2024), 22–48.
- [27] Freepik. n.d. Hypertension Blood Pressure Gauge; Motion Sensor; Student Sleeping in Class; Diabetes; Blood Cells; Edge Computing; Galvanic Skin Response; Motion Sensor; Accelerometer SensorEeg (IDs: hypertension_4939229; blood-pressure-gauge_3184052; motion-sensor_2818201; student-sleeping-in-class_43739; diabetes_2750352; blood-cells_3400003; edge-computing_11068838; galvanic-skin-response_11228469; motion-sensor_17881894; accelerometer-sensor_11330476; eeg_9851782). Retrieved from <https://www.flaticon.com/free-icon/>

- [28] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* 101, 23 (2000), e215–e220. DOI: <https://doi.org/10.1161/01.CIR.101.23.e215>
- [29] Tara Hassani. 2021. Federated Emotion Recognition with Physiological Signals-GSR. (Dissertation). Retrieved from <https://urn.kb.se/resolve?urn=urn:nbn:se:bth-22287>
- [30] Jun He, Qian Zhang, Liqun Wang, and Ling Pei. 2018. Weakly supervised human activity recognition from wearable sensors by recurrent attention learning. *IEEE Sensors Journal* 19, 6 (2018), 2287–2297.
- [31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked autoencoders are scalable vision learners. arXiv:2111.06377. Retrieved from <https://arxiv.org/abs/2111.06377>
- [32] Asmaa Hosni and Mohamed Atef. 2023. Remote real-time heart rate monitoring with recursive motion artifact removal using PPG signals from a smartphone camera. *Multimedia Tools and Applications* 82, 13 (2023), 20571–20588.
- [33] K. Hu, P. C. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley. 2001. Effect of trends on detrended fluctuation analysis. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 64, 1 (2001), 011114.
- [34] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2023. Masked autoencoders that listen. arXiv:2207.06405. Retrieved from <https://arxiv.org/abs/2207.06405>
- [35] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with Gumbel-Softmax. arXiv:1611.01144. Retrieved from <https://arxiv.org/abs/1611.01144>
- [36] Ian T. Jolliffe and Jorge Cadima. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2065 (2016), 20150202.
- [37] Mohammad Kachuee, Mohammad Mahdi Kiani, Hoda Mohammadzade, and Mahdi Shabany. 2016. Cuffless blood pressure estimation algorithms for continuous health-care monitoring. *IEEE Transactions on Bio-Medical Engineering* 64, 4 (2016), 859–869.
- [38] A. Kazemnejad, P. Gordany, and R. Sameni. 2021. EPHNOGRAM: A Simultaneous Electrocardiogram and Phonocardiogram Database. (version 1.0.0). *PhysioNet*. DOI: <https://doi.org/10.13026/tjtq-5911>
- [39] Dongkeun Kim, Jinsung Lee, Minsu Cho, and Suha Kwak. 2022. Detector-free weakly supervised group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, New York, NY, 20083–20093.
- [40] Diederik P. Kingma and Max Welling. 2022. Auto-encoding variational bayes. arXiv:1312.6114. Retrieved from <https://arxiv.org/abs/1312.6114>
- [41] Pranesh Krishnan, Sazali Yaacob, Annapoorni Pranesh Krishnan, Mohamed Rizon, and Chun Kit Ang. 2020. EEG based drowsiness detection using relative band power and short-time fourier transform. *Journal of Robotics, Networking and Artificial Life* 7, 3 (2020), 147–151.
- [42] Joelle LeMoult. 2020. From stress to depression: Bringing together cognitive and biological science. *Current Directions in Psychological Science* 29, 6 (2020), 592–598.
- [43] Haoyu Li, Stéphane Derrode, and Wojciech Pieczynski. 2019. An adaptive and on-line IMU-based locomotion activity classification method using a triplet Markov model. *Neurocomputing* 362 (2019), 94–105.
- [44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.), PMLR, Cambridge, MA, 19730–19742. Retrieved from <https://proceedings.mlr.press/v202/li23q.html>
- [45] Yongbo Liang, Zhencheng Chen, Guiyong Liu, and Mohamed Elgendi. 2018. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China. *Scientific Data* 5, 1 (2018), 1–7. DOI: <https://doi.org/10.6084/m9.figshare.5459299.v5>
- [46] Haojie Ma, Zhijie Zhang, Wenzhong Li, and Sanglu Lu. 2021. Unsupervised human activity representation learning with multi-task deep clustering. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.
- [47] George Mathew, Daniel Barbosa, John Prince, and Subramaniam Venkatraman. 2024. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. *NPJ Cardiovascular Health* 1, 1 (2024), 25. DOI: <https://doi.org/10.1038/s44325-024-00027-5>
- [48] Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. 2024. ECG-FM: An open electrocardiogram foundation model. arXiv:240805178. Retrieved from <https://arxiv.org/abs/2408.05178>
- [49] Temesgen Mehari and Nils Strodthoff. 2022. Self-supervised representation learning from 12-lead ECG data. *Computers in Biology and Medicine* 141, (2022), 105114. DOI: <https://doi.org/10.1016/j.combiomed.2021.105114>
- [50] Gatis Mikelsons, Matthew Smith, Abhinav Mehrotra, and Mirco Musolesi. 2017. Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities. arXiv:1711.06350. Retrieved from <https://arxiv.org/abs/1711.06350>
- [51] Jianliang Min, Ping Wang, and Jianfeng Hu. 2017. *The Original EEG Data for Driver Fatigue Detection*. DOI: <https://doi.org/10.6084/m9.figshare.5202739.v1>
- [52] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. 2020. Contrastive representation learning for electroencephalogram classification. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, vol. 136. Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland (Eds.), PMLR, 238–253. Retrieved from <https://proceedings.mlr.press/v136/mohsenvand20a.html>

- [53] Muhammad Muzammil. 2021. Finetuning EnDevSols/Tinyllama-2.5T-Clinical Model on Clinical Dataset. Retrieved from <https://huggingface.co/muzammil-eds/tinyllama-2.5T-Clinical-v2>
- [54] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. 2024. Scaling wearable foundation models. arXiv:2410.13638. Retrieved from <https://arxiv.org/abs/2410.13638>
- [55] Anna Nedorubova, Alena Kadyrova, and Aleksey Khlyupin. 2021. Human activity recognition using continuous wavelet transform and convolutional neural networks. arXiv:2106.12666. Retrieved from <https://arxiv.org/abs/2106.12666>
- [56] Pranshu Cbs Negi, Himanshu Giri, Shiru Sharma, Neeraj Sharma, et al. 2024. A comparative study of scalograms for human activity classification. In *Proceedings of the 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*. IEEE, New York, NY, 1–5.
- [57] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A time series is worth 64 words: Long-term forecasting with transformers. arXiv:2211.14730. Retrieved from <https://arxiv.org/abs/2211.14730>
- [58] Avital Oliver, Augustus Odena, Colin A. Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., Red Hook, NY. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2018/file/c1fea270c48e8079d8ddf7d06d26ab52-Paper.pdf
- [59] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. arXiv:2304.03442. Retrieved from <https://arxiv.org/abs/2304.03442>
- [60] Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. 2024. PaPaGei: Open foundation models for optical physiological signals. arXiv:2410.20542. Retrieved from <https://arxiv.org/abs/2410.20542>
- [61] Marco A. F. Pimentel, Alistair E. W. Johnson, Peter H. Charlton, Drew Birrenkott, Peter J. Watkinson, Lionel Tarassenko, and David A. Clifton. 2017. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Bio-Medical Engineering* 64, 8 (2017), 1914–1923. DOI: <https://doi.org/10.1109/TBME.2016.2613124>
- [62] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *Proceedings of the International Conference on Machine Learning*. PMLR, 28043–28078.
- [63] Bo Qian and Khaled Rasheed. 2004. Hurst exponent and financial market predictability. In *Proceedings of the IASTED Conference on Financial Engineering and Applications*, 203–209.
- [64] Hangwei Qian, Sinno Jialin Pan, and Chunyan Miao. 2021. Weakly-supervised sensor-based activity segmentation and recognition via learning from distributions. *Artificial Intelligence* 292 (2021), 103429.
- [65] Zhen Qin, Xiaodong Han, Weixuan Sun, Bowen He, Dong Li, Dongxu Li, Yuchao Dai, Lingpeng Kong, and Yiran Zhong. 2023. Toeplitz neural network for sequence modeling. arXiv:2305.04749. Retrieved from <https://arxiv.org/abs/2305.04749>
- [66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. arXiv:2103.00020. Retrieved from <https://arxiv.org/abs/2103.00020>
- [67] Schmidt Philip, Reiss Attila, and Indlekofer Ina. 2019. PPG-DaLiA. *UCI Machine Learning Repository*. DOI: <https://doi.org/10.24432/C53890>
- [68] Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. 2012. *Human Activity Recognition Using Smartphones*. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C54S4K>
- [69] Karin Roelofs. 2017. Freeze for action: Neurobiological mechanisms in animal and human freezing. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372, 1718 (2017), 20160206.
- [70] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*. ACM, New York, NY, 400–408. DOI: <https://doi.org/10.1145/3242969.3242985>
- [71] Artemios-Anargyros Semenoglou, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2023. Image-based time series forecasting: A deep convolutional neural network approach. *Neural Networks* 157 (2023), 39–53. DOI: <https://doi.org/10.1016/j.neunet.2022.10.006>
- [72] Roshwin Sengupta, Ilia Polian, and John P. Hayes. 2022. Wavelet transform assisted neural networks for human activity recognition. In *Proceedings of the 2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, New York, NY, 1254–1258.
- [73] Gašper Slapničar, Nejc Mlakar, and Mitja Luštrek. 2019. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors* 19, 15 (2019), 3420.
- [74] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J. Wareham, and Cecilia Mascolo. 2021. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning (ACM CHIL '21)*. ACM, New York, NY, 69–78. DOI: <https://doi.org/10.1145/3450439.3451863>
- [75] J. M. T. Thompson, H. B. Stewart, and Rick Turner. 1990. Nonlinear dynamics and chaos. *Computers in Physics* 4, 5 (1990), 562–563.
- [76] Christopher Torrence and Gilbert P. Compo. 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* 79, 1 (1998), 61–78.

- [77] Akhil Vaid, Joy Jiang, Ashwin Sawant, Stamatios Lerakis, Edgar Argulian, Yuri Ahuja, Joshua Lampert, Alexander Charney, Hayit Greenspan, Jagat Narula, et al. 2023. A foundational vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digital Medicine* 6, 1 (2023), 108. DOI: <https://doi.org/10.1038/s41746-023-00840-9>
- [78] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 11.
- [79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. arXiv:1706.03762. Retrieved from <https://arxiv.org/abs/1706.03762>
- [80] S. Vishnupriya and K. Meenakshi. 2018. Automatic music genre classification using convolution neural network. In *Proceedings of the 2018 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, New York, NY, 1–4. DOI: <https://doi.org/10.1109/ICCCI.2018.8441340>
- [81] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. 2025. CBraMod: A criss-cross brain foundation model for EEG decoding. arXiv:2412.07236. Retrieved from <https://arxiv.org/abs/2412.07236>
- [82] Christopher Wimmer and Navid Rekabsaz. 2023. Leveraging vision-language models for granular market change prediction. arXiv:2301.10166. Retrieved from <https://arxiv.org/abs/2301.10166>
- [83] Alan Wolf, Jack B. Swift, Harry L. Swinney, and John A. Vastano. 1985. Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena* 16, 3 (1985), 285–317.
- [84] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. arXiv:2402.02592. Retrieved from <https://arxiv.org/abs/2402.02592>
- [85] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '23)*. IEEE, New York, NY, USA, 1–5.
- [86] Maxwell A. Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrímsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik J. Raghuram, James M. Rehg, and Shirley Ren. 2025. RelCon: Relative contrastive learning for a motion foundation model for wearable data. arXiv:2411.18822. Retrieved from <https://arxiv.org/abs/2411.18822>
- [87] Yan Yan, Yi-Chun Huang, Jinjin Zhao, Yu-Shi Liu, Liang Ma, Jing Yang, Xu-Dong Yan, Jing Xiong, and Lei Wang. 2023. Topological nonlinear analysis of dynamical systems in wearable sensor-based human physical activity inference. *IEEE Transactions on Human-Machine Systems* 53, 4 (2023), 792–801. DOI: <https://doi.org/10.1109/THMS.2023.3275774>
- [88] Hang Yuan, Shing Chan, Andrew P. Creagh, Catherine Tong, Aidan Acquah, David A. Clifton, and Aiden Doherty. 2024. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ Digital Medicine* 7, 1 (2024), 91. DOI: <https://doi.org/10.1038/s41746-024-01062-3>
- [89] Hang Yuan, Tatiana Plekhanova, Rosemary Walmsley, Amy C. Reynolds, Kathleen J. Maddison, Maja Bucan, Philip Gehrman, Alex Rowlands, David W. Ray, Derrick Bennett, et al. 2024. Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. *NPJ Digital Medicine* 7, 1 (2024), 86. DOI: <https://doi.org/10.1038/s41746-024-01065-0>
- [90] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (Eds.). Association for Computational Linguistics, 97–109. DOI: <https://doi.org/10.18653/v1/2022.bionlp-1.9>
- [91] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. arXiv:1710.09412. Retrieved from <https://arxiv.org/abs/1710.09412>
- [92] Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. 2024. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems* 35, 11 (2024), 16129–16138. DOI: <https://doi.org/10.1109/TNNLS.2023.3292066>
- [93] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. 2024. Large language models for time series: A survey. arXiv:2402.01801. Retrieved from <https://arxiv.org/abs/2402.01801>
- [94] Xiyuan Zhang, Ranak Roy Chowdhury, Jiayun Zhang, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. 2023. Unleashing the power of shared label structures for human activity recognition. arXiv:2301.03462. Retrieved from <https://arxiv.org/abs/2301.03462>
- [95] Xiao Zhang, Yongqiang Lyu, Tong Qu, Pengfei Qiu, Xiaomin Luo, Jingyu Zhang, Shunjie Fan, and Yuanchun Shi. 2019. Photoplethysmogram-based cognitive load assessment using Multi-Feature fusion model. *ACM Transactions on Applied Perception* 16, 4, Article 19 (2019), 1–17. DOI: <https://doi.org/10.1145/3340962>
- [96] Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, Shuheng Li, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. 2024. UniMTS: Unified pre-training for motion time series. arXiv:2410.19818. Retrieved from <https://arxiv.org/abs/2410.19818>
- [97] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Advances in Neural Information Processing Systems*, Vol. 35, 3988–4003.

Appendices

A Datasets

Few openly accessible multi-channel or multi-device datasets for physiological signals exist, limiting advancements in this field. To address this gap, we curated a dataset comprising approximately 385 hours of recordings. Using the augmentation algorithm described below, we expanded this dataset to 4,294 hours. The distribution of the pretraining dataset, as shown in Figure 3, reflects the inherent diversity of the original recordings, ensuring balanced representation across channels and devices. This curated and augmented dataset provides a critical resource for developing robust models, facilitating progress in multi-channel physiological signal research.

Table A1 overviews used dataset in our experiment along with the modality and task type. We will give further details for each dataset below:

WESAD [70] is a publicly available multimodal dataset used for wearable stress and affect detection, formulated as a classification task with labels: neutral, stress, and amusement. The dataset includes physiological and motion data collected from 15 subjects during a lab study, using a chest-worn RespiBAN device and a wrist-worn Empatica E4 device. From the chest device, we use ECG, GSR, and triaxial acceleration (ACC-X, ACC-Y, ACC-Z), all sampled at 700 Hz. From the wrist device, we use PPG, GSR (4 Hz), and triaxial acceleration (ACC-X, ACC-Y, ACC-Z, 32 Hz). The selected channels span multiple physiological and motion modalities from both chest and wrist sensors. Each data segment is labeled with one of the three affective states, serving as the target output for classification tasks.

UCI-HAR [68] dataset is publicly available and is used for classifying human activities based on sensor data. It comprises data from 30 volunteers, aged 19–48, each performing six activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying. During these activities, participants carried a waist-mounted smartphone equipped with embedded accelerometer and gyroscope sensors. The input channels consist of triaxial linear acceleration and triaxial angular velocity, totaling six channels. Each data segment is labeled with one of the six activities, serving as the target output for classification tasks. The sensors recorded data at a constant rate of 50 Hz.

Table A1. NormWear Pretraining Hyper-Parameters

Hyper-parameter	Value
# Cross-Patches Transformer Encoder	12
# Cross-Channels Transformer Encoder	6
# Transformer Decoder	2
# Attention Heads	12
Encoder Latent Size	768
Decoder Latent Size	512
Feedforward Latent Size	3,072
Normalization	LayerNorm
Patch size (time axis)	9
Patch size (scale axis)	5
Optimizer	AdamW
Loss Scalar	NativeScaler
Base Learning Rate (blr)	1e-3
Epochs	140
Batch size	192

Table B1. MSiTF Hyper-Parameter

Hyper-parameter	Value
Learning rate (lr)	1e-3
Epochs	40
Batch size	32
L2 regularization	5e-6
lr decay rate	0.997
λ	0.5
τ	0.5

Driver Fatigue EEG Dataset [51] is a publicly available dataset used for detecting driver fatigue based on EEG signals. EEG data were collected using a 40-channel Neuroscan amplifier. The recordings include EEG data corresponding to two states: alert and fatigued. Each data segment is labeled with one of these states, serving as the target output for classification tasks.

Epileptic Seizure Recognition [6] dataset is publicly available and is used for classifying neurological and physiological states based on EEG signals. It comprises data from 500 subjects, each recorded for 23.6 seconds using a single EEG channel at a sampling rate of 178 Hz. Each sample is labeled with one of five brain states, allowing for the construction of multiple binary classification tasks that target different aspects of neurological assessment. Specifically, we formulated five tasks:

- *Eye Relaxation*: Detects eye fatigue by distinguishing between relaxed and alert states based on EEG changes related to eye closure.
- *Health Area*: Classifies brain regions as healthy or affected by neurological abnormalities.
- *Tumor Area*: Detects EEG patterns indicative of tumor presence in specific brain regions.
- *Seizure*: Identifies seizure activity from non-seizure states.
- *Eyes Open vs. Closed*: Differentiates EEG signals associated with visual input states.

GAMEEMO [4] is a publicly available dataset used for emotion recognition based on EEG signals. It comprises data from 28 subjects, each playing four emotion-inducing computer games (boring, calm, horror, and funny) for five minutes per game, totaling 20 minutes of EEG data per subject. EEG signals were recorded using the EMOTIV EPOC+ headset, which includes 14 channels (AF3, AF4, F3, F4, F7, F8, FC5, FC6, O1, O2, P7, P8, T7, and T8) positioned according to the 10–20 system. The signals were sampled at 128 Hz. After each gameplay session, subjects rated their emotional response using the Self-Assessment Manikin form, providing continuous scores for arousal and valence. These scores were quantized into binary values using subject-specific median thresholds: arousal and valence ratings above the median were labeled as high, and those below or equal to the median as low. Combining the binarized arousal and valence ratings yields four discrete emotional classes: low arousal and low valence, low arousal and high valence, high arousal and low valence, and high arousal and high valence. Each data segment is labeled with one of these four classes, serving as the target output for four-class emotion classification tasks.

ECG Heartbeat Categorization [13] is a publicly available dataset used for classifying heartbeat signals based on ECG recordings. It comprises two collections of heartbeat signals derived from PhysioNet’s MIT-BIH Arrhythmia Dataset and the PTB Diagnostic ECG Database. The first collection includes 1,09,446 samples categorized into five classes: normal (N), supraventricular ectopic (S), ventricular ectopic (V), fusion (F), and unknown (Q), with ECG signals sampled at 125 Hz. The second collection consists of 14,552 samples categorized into two classes: normal and abnormal, also sampled at 125 Hz. For our analysis, we restructured the dataset into a binary classification framework by consolidating the original categories into two classes: normal and abnormal heartbeats.

PPG-China [45] is a publicly available dataset used for classifying cardiovascular and metabolic conditions based on PPG signals. It comprises 657 data records from 219 subjects, aged 20–89 years, including individuals with conditions such as hypertension and diabetes. PPG signals were recorded using a single channel at a sampling rate of 125 Hz. Each subject’s data includes PPG waveforms and corresponding clinical information, facilitating the construction of multiple classification tasks focused on cardiovascular and systemic health monitoring. Specifically, we formulated four tasks:

- *PPG-HTN*: Identifies stages of hypotension severity by classifying PPG signals into four levels.
- *PPG-DM*: Detects diabetes by distinguishing between diabetic and non-diabetic individuals.
- *PPG-CVA*: Identifies the presence or absence of cerebrovascular accidents (strokes) based on PPG patterns.
- *PPG-CVD*: Assesses cardiovascular disease by classifying PPG signals into three cardiovascular health categories.

PhysioNetEMG [28] is a publicly available dataset used for classifying neuromuscular conditions based on **electromyography (EMG)** signals. It comprises single-channel EMG recordings from the tibialis anterior muscle of three subjects: one healthy, one with neuropathy, and one with myopathy. The EMG signals were recorded at a sampling rate of 4,000 Hz. Each recording was segmented into time series samples using a fixed-length window of 6 second. Each segment is labeled according to the subject's condition—healthy, neuropathy, or myopathy—serving as the target output for classification tasks.

Non-Invasive BP Estimation [24] is a publicly available dataset used for cuff-less BP estimation. It comprises data from 26 subjects, each with recorded ECG and PPG signals, sampled at 1,000 Hz. Reference BP measurements were taken during signal acquisition. Each subject's data also includes demographic information such as age, weight, and height. The dataset is structured to facilitate regression tasks aimed at predicting systolic and diastolic BP values.

PPG-HGB [2] is a publicly available dataset used for non-invasive hemoglobin (Hb) measurement based on PPG signals. It comprises data from 68 subjects, aged 18–65 years, with a gender distribution of 56% female and 44% male. PPG signals were recorded using the MAX30102 sensor, which emits red and infrared light. The sensor's analog-to-digital converter output data rate can be programmed from 50 **samples per second (sps)** to 3,200 sps. Each subject contributed 12 sets of PPG signals, totaling 816 data records. We formulate regression tasks aimed at predicting Hb concentration levels.

Fetal-fPCG [11] is a publicly available dataset designed for estimating **fetal heart rate (FHR)** using **fetal phonocardiography (fPCG)** signals. It includes recordings from 60 pregnant women, aged 18–37 years, with gestational ages between 31 and 40 weeks. The recordings were collected at St. John's Hospital, Bangalore using an electronic stethoscope (SS30LA) connected to a Biopac MP36 data acquisition system. The stethoscope was placed on the lower abdomen of each subject to capture the fPCG signal, which was sampled at 2,000 Hz. The dataset supports regression tasks, where the goal is to predict continuous FHR values directly from the fPCG waveforms.

B Implementation Detail

B.1 Data Preprocess

For the data preparation, we set the uniform sampling rate and interval length to 65 HZ and 6 seconds, respectively. In our case, 65 Hz covers most of the frequency bands of interest such as heart activity, physical motions, and neuron activity up to the beginning of Gamma power (above 30 Hz). And a great amount of samples are less than 6 seconds such as [13, 45, 68]. We conduct basic pre-processing for each signal with identical setting: (1) de-trended by subtract the result of a linear least-squares fit to series data from the raw time series, and (2) Gaussian smoothed with standard deviation of 1.3 (0.02 seconds), ensuring a highly consistent dataset for training.

Since the Transformer's computational requirements scale quadratically with input length, to release the full potential of our self-supervised algorithm, we segment our multivariate time series into intervals with a uniform length and pad shorter samples with zeros. This approach not only enables parallel processing of samples in large minibatches but also addresses variation in the length of individual samples.

For the downstream task, we split the data into train and test sets for linear probing evaluation with portion of 80% and 20% correspondingly. The split is stratified on the anonymized subject ID if this information is provided by the dataset.

B.2 Pretraining Framework

Normwear is derived from the MAE [31]. The detailed hyper-parameter choice is described in Table B1. We use a Conv2D layer with a kernel size of (9, 5) and a stride of (9, 5), ensuring no overlapping patches. This layer takes input with 3 channels and projects it to 768 channels, matching the hidden size of our encoders. In Normwear, we apply structured masking independently to each variate along both the frequency and time axes,

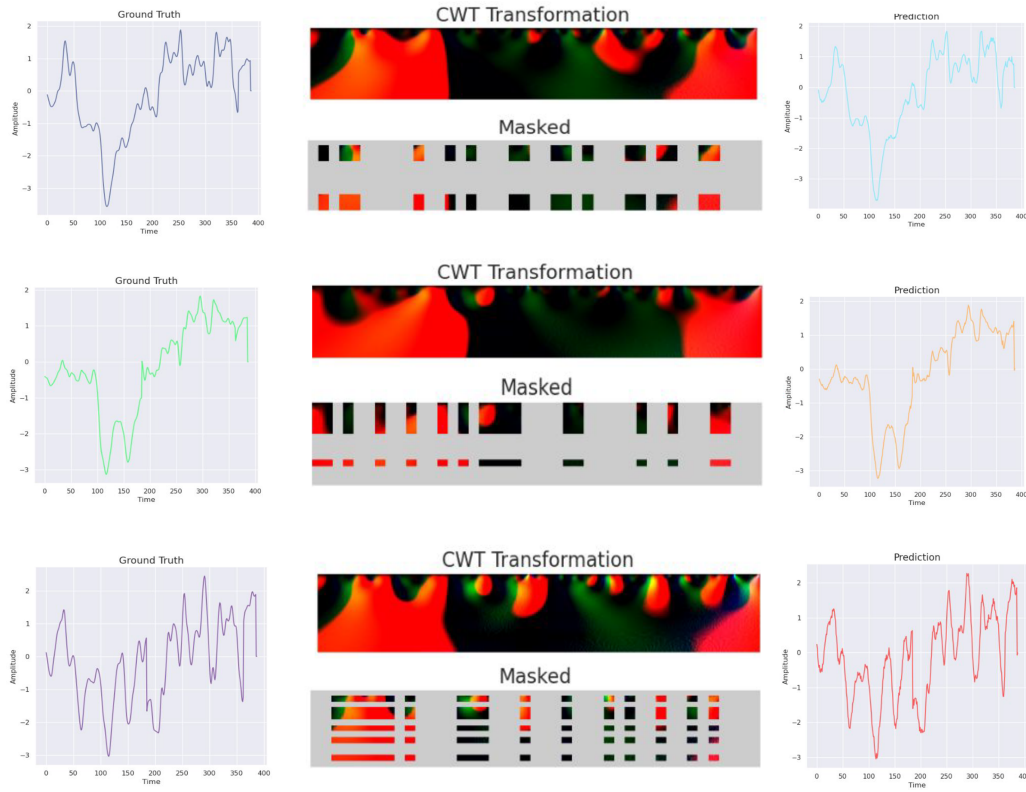


Fig. B1. Visualization of original time series (left), CWT transformation image with structured masking (middle), and reconstructed time series (right).

with respective masking ratios of 0.6 and 0.5. This results in an expected overall masking ratio of 0.8 for each variate. Only the unmasked tokens are passed to the encoder, reducing computational complexity. To enhance representation learning, we introduce 6 additional transformer blocks as fusion layers, interleaved with the original 12 encoder blocks, creating a total of 18 blocks. Each transformer block has a hidden dimension of 768 and uses LayerNorm as in the original MAE. The latent embeddings obtained from the encoder are projected from 768 to 512 dimensions. Learnable masked tokens are reinserted at their original positions, and positional embeddings are added to guide the decoder in reconstructing the input series. The lightweight decoder consists of two transformer blocks with a hidden dimension of 512, followed by two Conv1D layers. The first Conv1D layer maps from the flattened multivariate signal embedding to an intermediate dimension, and the second Conv1D layer maps from this intermediate dimension back to the original multivariate signal space. A GELU activation function is used between these layers, with BatchNorm applied to the input. The decoder reconstructs the original input series, and the model is trained using MSE loss on all data points. Our models are pretrained for 45,000 steps with a batch size of 256, using the AdamW optimizer with a learning rate of 10^{-4} . We did not perform on-the-fly data augmentation, as suggested in the MAE framework, due to the high masking ratio. (An end-to-end example of the input and output of this pretraining pipeline is illustrated in Figure B1).

All the models are pretrained on 4 NVIDIA RTX 3090 **graphical computing unit (GPU)**, with 24 GB of GPU memory on each card.

B.3 MSiTF

For pretraining the representation alignment module, we have the training hyper-parameters in Table B1.

B.4 Detailed Patch Fusion Operation

Regarding the detailed forward operation, we aggregate the signal embeddings using a weighted average operation with weights determined from three aspects: relevance, recency, and importance, as specified in the following.

Relevance. With the semantic embedding of query sentence q and backbone output $H \in \mathbb{R}^{L \times E}$, we have relevance being calculated as:

$$W_{rel}(t) = \cos_sim(q, k(H)) \in \mathbb{R}^L, \quad (B1)$$

where $k(\cdot) \in \mathbb{R}^{L \times E}$ are the key linear mapping functions. “cos_sim” is the function to compute cosine similarity between the query vector and each key vectors which will result in a vectors of length L containing indicating relevance scores.

Recency. We also introduce recency score which is designed based on the assumption that the more recent the information, the more focus should be allocated. We define the recency score W_{rec} of vector at each timestamp t as an exponential decay function:

$$W_{rec}(t) = p^{T-t}, t \in (0, T), p < 1, \quad (B2)$$

where T is the total sequence length and p is the decay factor. When $p = 1$, the result score is equivalent to not considering the recency factor.

Importance. Finally, we consider the importance score in this case to be whether to keep the representation at each timestep or not. In order to achieve this, we assign binary parameters to each timestep, denoted as $\theta_t = p(v_t) \in \mathbb{R}^2$ where $v_t \in \mathbb{R}^E$ is the representation vector at timestep t and p is a linear transformation function. We then have the importance score defined as:

$$W_{imp}(t) = \arg \max_{\{0,1\}} \frac{\exp\left(\left(\log(\theta_t) + \epsilon\right)/\tau\right)}{\sum_{j \in \{0,1\}} \exp\left(\left(\log(\theta_{t,j}) + \epsilon_j\right)/\tau\right)}, \quad (B3)$$

where ϵ is the noise term sampled from Gumbel distribution [35], and τ is the temperature controlling the sharpness of the softmax function. Because $\arg \max$ is not a differentiable function, we will directly take the result probability corresponding to index at $j = 1$ with τ being set to a small number to push the result closer to a one hot vector from softmax function.

Integrated Score. The integrated weight W is finally defined as:

$$W(t) = \frac{W_{rec}(t) + W_{rel}(t) + W_{imp}(t)}{\sum_{i=1}^T W_{rec}(i) + W_{rel}(i) + W_{imp}(i)}, \quad (B4)$$

which will result in final output $\hat{Y} = \sum_{t=1}^T H_t \cdot W(t)$. Intuitively, the latent representations from all timesteps are fused before the final output layer with weighted scores computed according to how relevant they are to the objective tasks, how important they are to the data itself, and how close they are to the most current timestep. The output layer is instructed to select the most informative representations to optimize objective task of representation alignment.

B.5 Aligner Module, Objective Function, and Pretraining

The Aligner Module matches two vectors: the fused representation $f(q, H) = \hat{Y} \in \mathbb{R}^E$ with the semantic embedding (Y) of ground truth sentence, which is obtained from prompting the ground truth label using a template, for

example, “The subject is presently {activity_label}.” In the same manner as the query embedding, the ground truth sentence is encoded using the same pretrained language model [53]. At this stage, Y is leveraged to supervise the fused output $\hat{Y}$. The vanilla contrastive learning loss formula following [96] is:

$$Loss_{ctl}(Y, \hat{Y}) = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\hat{Y}_i^T Y_i)^{\frac{1}{\gamma}}}{\sum_{k=1}^N \exp(\hat{Y}_i^T Y_k)^{\frac{1}{\gamma}}}, \quad (B5)$$

where N is the batch size and γ is the learnable temperature parameter. We denote this loss function as contrastive loss with batch normalizer. We also leverage a refine process after contrastive learning using similarity loss with per sample normalizer, which is essentially cosine similarity loss, with vector distance as supplemental penalty:

$$Loss_{refine}(\hat{Y}, Y) = \frac{1}{N} \sum_{i=1}^N \left(\left(1 - \frac{\hat{Y}_i^T Y_i}{\|\hat{Y}_i\| \|Y_i\|} \right) + \lambda |\hat{Y}_i - Y_i| \right), \quad (B6)$$

where λ is hyper-parameters controlling the weight of the supplemental loss components.

B.6 Sentence Template Example for Signal-Sext Alignment

To enhance the expressiveness and diversity of supervision signals for our MSiTF alignment module, we convert categorical labels into natural language descriptions using varied prompt templates. We apply this strategy to several pretraining tasks. We present example sentence templates below for emotion recognition and activity recognition to demonstrate the general idea of how we derive text modality from the raw label:

For the *emotion recognition* task, we use:

[hidealllines=false,backgroundcolor=gray!30]

- “The emotion detected is {}.”
- “This subject is feeling {}.”
- “The emotional state is {}.”
- “The identified emotion is {}.”

For the *activity recognition* task, we use:

[hidealllines=false,backgroundcolor=gray!30]

- “This subject is currently {}.”
- “The subject is engaged in {}.”
- “Current activity is {}.”
- “Subject’s activity is {}.”

By exposing the model to multiple phrasings for the same label, this design helps it learn modality-invariant representations that are more robust to linguistic variation and better aligned across modalities. Specifically, to increase the diversity of semantic representations of query and ground truth sentences in the pretraining signal corpus, we utilize large language models (GPT-3.5) [3] to generate 20 alternative variations for each sentence, from which only one is randomly sampled during pretraining. During test-time inference on downstream datasets, each ground truth label is converted into a sentence (details in Appendix B.6), which is transformed into a semantic embedding using a frozen text encoder. The sentence with the closest distance with the embedding from our foundation model is used as the final inferential result.

B.7 Statistical Feature List

Our statistical baseline includes features extracted from both the time and frequency domains. In the time domain, we compute the mean, standard deviation, maximum, minimum, skewness, kurtosis, 25% quantile, median, and

Table C1. Performance Comparison of Various Fusion Schemes

Downstream Tasks	No Fusion	Cross-Attention Fusion	Mean Pooling Fusion	[CLS] Token Fusion
WESAD	72.209	74.165	71.99	75.390
UCI-HAR	97.793	96.908	97.566	98.928
DriverFatigue	73.252	60.308	72.552	75.167
Activity Recognition Avg.	81.085	77.127	80.703	83.162
Epilepsy (eye open state)	90.966	84.075	89.817	92.203
Epilepsy (eye relaxation)	94.399	93.589	93.912	94.908
Epilepsy (health area)	87.866	86.899	87.248	88.117
Epilepsy (tumor area)	86.599	86.861	87.152	86.888
Epilepsy (seizure)	97.477	96.351	96.719	96.638
GAMEEMO	57.695	56.724	58.079	56.532
EEG Main Tasks Avg.	85.834	84.083	85.488	85.881
ECG-Abnormal	99.429	99.441	99.268	99.041
PPG-BP (HTN)	61.850	60.983	63.577	60.344
PPG-BP (DM)	58.333	62.800	62.200	59.459
PPG-BP (CVA)	61.319	61.458	59.236	70.278
PPG-BP (CVD)	48.417	53.585	46.961	52.596
PhysioNet EMG	93.715	95.49	86.749	98.184
Risk Evaluation Avg.	70.511	72.293	69.665	73.317
Noninvasive-BP	88.356	92.759	88.719	92.470
PPG-Hgb	95.031	93.413	95.086	94.766
Fetal-fPCG	98.582	99.145	98.771	99.088
Vital Signs Avg.	93.990	95.106	94.192	95.441
Micro Avg.	81.294	80.831	80.867	82.833
Macro Avg.	82.855	82.152	82.512	84.450

The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme. The best performance achieved for each task is highlighted in bold.

75% quantile. In the frequency domain, we extract the spectral centroid, spectral spread, mean frequency, peak frequency, as well as the 25%, 50% (median), and 75% quantile frequencies.

B.8 Radar Plot or Performance Trend

To enhance the visual contrast between model performances across tasks, we applied the Softmax function to the raw performance scores. This transformation rescales the scores to a range between 0 and 1, accentuating relative differences between models. While the Softmax transformation emphasizes the relative improvement of our model over others, we note that the absolute scores may differ from those originally reported.

C Ablation Study

Due to computational constraints, we will conduct the ablation study on our smaller dataset (37k samples) to train and evaluate the model, establishing a proof of concept and demonstrating the effectiveness of our approach in a controlled setting.

Fusion Schemes. Table C1 shows the performance of different fusion schemes, including (1) no fusion, (2) cross-attention fusion, (3) [CLS]-attention fusion, and (4) mean-pooling fusion. We excluded all-attention fusion in our ablation study because it is computationally prohibitive. Among all the compared strategies, the [CLS] token fusion generally achieves the best accuracy with a minor increase in parameters.

Table C2. Performance Comparison of Different Masking Strategies

Downstream Tasks	Unstructured Mask ($P = 0.8$)	Time Mask ($P_t = 0.8, P_f = 0.0$)	Scale Mask ($P_t = 0.0, P_f = 0.8$)	Structured Mask ($P_t = 0.6, P_f = 0.5$)
WESAD	71.46	71.952	72.201	75.390
UCI-HAR	97.097	98.438	98.106	98.928
DriverFatigue	72.719	73.424	78.354	75.167
Activity Recognition Avg.	80.425	81.271	82.887	83.162
Epilepsy (eye open state)	89.521	91.895	89.407	92.203
Epilepsy (eye relaxation)	93.471	94.808	93.786	94.908
Epilepsy (health area)	86.812	88.510	87.317	88.117
Epilepsy (tumor area)	86.524	88.254	85.502	86.888
Epilepsy (seizure)	96.59	97.791	95.29	96.638
GAMEEMO	58.043	56.770	55.771	56.532
EEG Main Tasks Avg.	85.160	86.338	84.512	85.881
ECG-Abnormal	99.085	99.316	98.296	99.041
PPG-BP (HTN)	58.880	55.333	59.230	60.344
PPG-BP (DM)	61.074	48.386	58.896	59.459
PPG-BP (CVA)	56.389	58.472	64.167	70.278
PPG-BP (CVD)	52.572	46.557	55.666	52.596
PhysioNet EMG	85.160	95.490	83.922	98.184
Risk Evaluation Avg.	68.860	67.259	70.030	73.317
Noninvasive-BP	90.124	90.650	91.152	92.470
PPG-Hgb	95.314	95.055	94.713	94.766
Fetal-fPCG	98.630	99.121	98.926	99.088
Vital Signs Avg.	94.689	94.942	94.930	95.441
Micro Avg.	80.526	80.568	81.150	82.833
Macro Avg.	82.284	82.453	83.090	84.450

The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme. The best performance achieved for each task is highlighted in bold.

Masking Strategies in Pretraining. We ablated our masking strategy introduced in Section 3.4. Using a consistent mask ratio of 0.8 in all strategies, we found that applying masking along the scale and time axes produced the best performance (details in Table C2).

Input Representations. Table C3 compares the performance of two input representations: (1) CWT scalogram and (2) raw time series. The CWT scalogram converts the time series into a time-frequency representation, while the raw time series retains the original sensor data. Among the two representations, the model trained on CWT scalograms demonstrates better performance, suggesting that the time-frequency features enhance model accuracy.

Semi-Supervised Learning (Partial-shot). To evaluate the generalizability and quality of learned representations, we conducted a semi-supervised learning evaluation following the protocol established by prior self-supervised methods [17]. Specifically, we assessed performance on the NORMWEAR dataset using frozen features and a limited labeled subset (10%). We deliberately excluded the commonly used 1% label evaluation due to the inherently small sample size of our downstream medical dataset. A 1% labeling scenario would provide fewer than 10 labeled instances, rendering the results statistically unreliable and scientifically unjustified. Instead, we sampled 10% of the training data while preserving the original label distribution, and then trained a linear classifier atop the frozen NORMWEAR features for classification tasks and regression tasks. The results, summarized in Table C4, demonstrate the effectiveness of our method under realistic semi-supervised constraints.

Table C3. Performance Comparison Between CWT Scalogram and Raw Time Series as Inputs

Downstream Tasks	Raw Series Input	CWT Scalogram Input
WESAD	70.862	75.390
UCI-HAR	97.969	98.928
DriverFatigue	73.854	75.167
Activity Recognition Avg.	80.895	83.162
Epilepsy (eye open state)	91.978	92.203
Epilepsy (eye relaxation)	94.781	94.908
Epilepsy (health area)	88.045	88.117
Epilepsy (tumor area)	85.619	86.888
Epilepsy (seizure)	97.722	96.638
GAMEEMO	54.651	56.532
EEG Main Tasks Avg.	85.466	85.881
ECG-Abnormal	97.701	99.041
PPG-BP (HTN)	52.614	60.344
PPG-BP (DM)	62.012	59.459
PPG-BP (CVA)	56.181	70.278
PPG-BP (CVD)	54.812	52.596
PhysioNet EMG	93.756	98.184
Risk Evaluation Avg.	69.513	73.317
Noninvasive-BP	89.850	92.470
PPG-Hgb	93.832	94.766
Fetal-fPCG	98.977	99.088
Vital Signs Avg.	94.220	95.441
Micro Avg.	80.845	82.833
Macro Avg.	82.523	84.450

The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme. The best performance achieved for each task is highlighted in bold.

Permutation-Invariant Input Channel Analysis. In many multimodal or multichannel sensing tasks, the input channel order is typically fixed and determined by hardware or preprocessing pipelines, limiting flexibility during deployment. This constraint raises the question of whether *Normwear* relies on a specific channel ordering to perform well. To examine this, we conducted an experiment on datasets with multiple input channels by circularly shifting the channel order by one position and evaluating the resulting model performance. As shown in Table C6, the model performance remains consistent across different permutations. These results suggest that our model does not rely on a fixed input channel configuration and is robust to variations in channel ordering, making it more applicable in practical scenarios where such inconsistencies may occur.

k-Fold Analysis. To evaluate whether Normwear maintains consistent performance on datasets with limited subject diversity, we conducted fivefold cross-validation stratified by subject ID. We applied this protocol to all downstream tasks containing 30 or fewer subjects to ensure a robust assessment. As shown in Table C6, our model consistently outperformed the baselines across all tasks, demonstrating the robustness of our evaluation metric.

Demographic Analysis. Several previous works [1, 54] have used learned representations to infer demographic labels. These results suggest that wearable signals do contain demographic information. In Table C7, we wanted to investigate that NormWear does not extract only demographic information (e.g., age, sex, height, etc. depending on what is available within each dataset), hence indicating that the representation that our proposed model extracted and the demographic could be used as complementary features to each other during downstream modeling. From Table C7, we observe that demographic information and wearable signal representations each excel at different tasks. In most cases, concatenating them improves overall performance. However, the occasional performance

Table C4. Semi-Supervised Learning on Downstream Tasks

Downstream Tasks	Statistical	Chronos	CLAP	TF-C	Modality-Specific	NORMWEAR (Ours)
WESAD	64.869	64.908	68.626	62.218	59.371	70.25
UCI-HAR	94.124	73.124	92.794	92.334	-	98.355
DriverFatigue	63.237	72.454	50.193	54.613	69.004	55.094
Activity Recognition Avg.	74.077	70.162	70.538	69.722	-	74.566
Epilepsy (eye open state)	82.186	80.082	84.103	88.02	89.152	85.456
Epilepsy (eye relaxation)	87.480	81.820	88.716	93.670	95.191	92.369
Epilepsy (health area)	86.096	77.682	82.651	84.940	87.377	85.471
Epilepsy (tumor area)	82.153	78.364	82.579	85.450	86.962	83.033
Epilepsy (seizure)	88.179	96.786	97.386	92.900	94.063	92.345
GAMEMO	54.527	50.176	51.952	49.714	52.046	52.633
EEG Main Tasks Avg.	80.104	77.485	81.231	82.449	84.132	81.885
ECG-Abnormal	96.420	97.613	95.432	94.769	79.918	93.921
PPG-BP (HTN)	52.491	49.407	48.397	53.800	57.544	53.967
PPG-BP (DM)	41.254	48.574	38.664	45.383	56.532	57.545
PPG-BP (CVA)	83.056	51.944	48.125	51.667	64.792	66.597
PPG-BP (CVD)	55.753	47.547	59.505	55.651	47.586	54.614
PhysioNet EMG	92.993	70.248	92.415	79.412	-	87.503
Risk Evaluation Avg.	70.328	60.888	63.756	63.447	-	69.025
Noninvasive-BP	90.589	93.783	91.614	92.707	92.671	90.694
PPG-Hgb	95.068	94.999	94.712	94.981	94.916	94.633
Fetal-fPCG	99.020	99.153	98.889	98.902	-	98.813
Vital Signs Avg.	94.892	95.978	95.072	95.530	-	94.713
Micro Avg.	78.305	73.815	75.931	76.174	-	78.516
Macro Avg.	79.850	76.129	77.649	77.787	-	80.047

The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme. The best performance achieved for each task is highlighted in bold.

drop after concatenation suggests a confounding relationship between the two, implying that demographic data and NormWear’s wearable representations capture different aspects.

D Statistical Significance on the Model Comparison

We performed a statistical analysis to test the significance of the differences in model performance. First, we ran the downstream evaluations 100 times for each model on every task without fixing the random seed. The results remained consistent due to the stable optimization process.

Next, we applied a permutation test on the results from these 100 runs to determine whether NormWear’s AUC ROC score is greater than that of the baselines. The reported p-value represents the probability of observing a test statistic as extreme or more extreme than the observed difference under the null hypothesis, which assumes that NormWear’s score is not higher than the baseline. In nearly all cases, the p-value is less than 0.01, confirming the statistical significance and indicating the robustness and superiority of our approach. Figure 5(B.1) presents the statistical test results across different task groups (as indicated by the color coding in the main tables) along with the overall average scores.

We also include a **critical difference (CD)** diagram to visually compare the performance of multiple models across datasets and highlight statistically significant differences. To generate the CD diagram, we first conducted a Friedman Chi-square test on the models’ scores across all downstream tasks, which yielded a p-value of $p < 0.001$,

Table C5. Performance of NormWear with Original Input Channel Order Compared to Random Shuffling across Tasks

Task	Original Order	Random Shuffle
WESAD (IMU, PPG, ECG, GSR)	0.761	0.763
UCI-HAR (IMU)	0.989	0.975
Drive Fatigue (EEG)	0.743	0.721
GAMEEMO (EEG)	0.549	0.530
Noninvasive-BP (PCG, PPG, ECG)	0.924	0.914
PPG-HGB (Red, IR)	0.946	0.948

The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme.

Table C6. Performance on Downstream Health-Related Tasks under Linear Probing Using 5-Fold Subject-Stratified Cross-Validation

Downstream Tasks	Statistical	Chronos	CLAP	TF-C	NormWear-L (Ours)
WESAD	79.992 ± 0.707	83.332 ± 0.841	87.824 ± 0.463	82.701 ± 0.536	89.585 ± 0.683
UCI-HAR	95.602 ± 0.148	91.956 ± 0.256	96.864 ± 0.175	97.382 ± 0.138	98.179 ± 0.06
DriverFatigue	69.614 ± 1.138	72.48 ± 2.848	66.251 ± 0.471	65.026 ± 1.198	68.971 ± 1.32
GAMEEMO	64.281 ± 1.292	56.694 ± 0.878	64.119 ± 0.543	62.925 ± 0.999	67.863 ± 0.72
Noninvasive-BP	92.83 ± 0.386	92.223 ± 0.356	92.612 ± 0.272	88.707 ± 0.622	93.381 ± 0.516
Avg.	80.464 ± 0.734	79.337 ± 1.036	81.534 ± 0.385	79.348 ± 0.699	83.596 ± 0.660

Classification reports AUC ROC; regression reports relative accuracy. All metrics are higher-is-better. The best performance achieved for each task is highlighted in bold.

Table C7. Checking Reliance on Demographic Information

Downstream Tasks	Empirical Distribution	Demographic	NormWear-Medium	Demographic + NormWear-Medium	NormWear-Large	Demographic + NormWear-Large
WESAD	50.000	49.907	74.227	69.06	76.06	68.755
Noninvasive-BP	92.988	92.954	91.427	90.84	92.42	92.528
PPG-Hgb	94.816	95.634	94.911	95.835	94.632	96.384
Fetal-fPCG	99.033	99.039	98.997	99.001	99.072	99.097
Vital Signs Avg.	95.612	95.876	95.112	95.225	95.375	96.003
PPG-BP (HTN)	50.000	59.899	62.746	64.482	62.341	61.291
PPG-BP (DM)	50.000	47.297	62.613	47.86	55.893	60.135
PPG-BP (CVA)	50.000	81.875	67.639	83.681	70.625	77.847
PPG-BP (CVD)	50.000	71.011	51.504	70.37	51.773	67.466
Risk Evaluation Avg.	50.000	65.021	61.126	66.598	60.158	66.685
Micro Avg.	67.105	74.702	75.508	77.641	75.352	77.938
Macro Avg.	65.204	70.268	76.821	76.961	77.198	77.148

Simple baseline: For regression tasks (yellow), the mean prediction is used; for classification tasks (blue and red), the mode prediction is used. NormWear-Medium and NormWear-Large refer to NormWear's pretrained checkpoints trained on 2.58 million and 8.97 million signal segments, respectively. The best performance achieved for each task is highlighted in bold.

confirming that the models' performances come from different distributions. We then applied the Conover *post hoc* test to examine pairwise differences between model performances; the p-values for NormWear compared with the baselines are shown in the last row of Figure 5(B.1). Finally, based on these results, we generated the CD diagram displayed in Figure 5(B.2). In this diagram, our proposed model, NormWear, is well separated from the others, indicating its statistical superiority over the competitive baselines.

E Supplementary Metrics

Normwear's performance is summarized in Figure 5 and detailed in Table E1. Normwear consistently exceeds the baseline models by a wide margin, demonstrating a clear advantage.

F Scaling up the Pretraining Data Size

In addition to demonstrating that NormWear outperforms all strong baselines, we further investigate the effect of varying pretraining data size on the model's downstream performance to examine whether the scaling law applies to our proposed methodology. As shown in Figure 5(A.4), the overall performance (measured by accuracy) significantly improves as the pretraining data size increases from approximately 37k (62 hours) to nearly 2.5M (4000 hours) samples of wearable signal data. This observation indicates that our model adheres to the scaling law, highlighting its potential scalability and suitability for future large-scale applications.

Table E1. Details of Incidental Performance Metrics

Task Group	Methods	AUC ROC	AUC PR	Accuracy	Precision	Recall	F1 Score
Activity Recognition	Statistical	75.082	63.996	65.298	61.450	61.56	61.034
	Chronos	79.935	65.622	66.175	62.044	61.512	60.522
	CLAP	76.897	67.026	66.349	62.790	62.826	62.435
	TF-C	77.880	68.228	67.175	64.967	64.798	64.783
	NormWear (Ours)	83.102	76.232	75.254	72.606	72.177	72.053
EEG Main Tasks	Statistical	79.720	50.172	73.921	63.567	57.529	57.948
	Chronos	80.677	55.507	75.285	72.442	52.520	47.671
	CLAP	82.100	57.518	76.391	68.506	61.961	62.650
	TF-C	84.302	61.864	76.825	71.702	65.517	67.889
	NormWear (Ours)	85.883	66.841	79.182	72.485	69.158	69.698
Disease Risk Evaluation	Statistical	71.032	53.783	79.688	52.718	53.235	50.807
	Chronos	62.060	40.673	71.910	45.512	43.739	40.569
	CLAP	66.274	48.232	81.327	53.028	54.721	52.804
	TF-C	69.416	46.312	78.929	52.123	52.352	51.349
	NormWear (Ours)	73.165	51.666	81.530	54.133	56.314	54.428
Micro Average	Statistical	75.317	51.596	74.503	58.804	56.618	55.709
	Chronos	73.082	51.596	72.113	59.590	50.806	47.401
	CLAP	74.729	55.705	76.357	61.171	59.238	58.669
	TF-C	77.063	56.916	75.737	62.523	60.107	60.652
	NormWear (Ours)	80.240	62.649	79.336	65.168	64.624	64.061
Macro Average	Statistical	75.278	55.983	72.969	59.245	57.441	56.596
	Chronos	74.224	53.934	71.123	59.999	52.590	49.587
	CLAP	75.091	57.592	74.689	61.441	59.836	59.296
	TF-C	77.199	58.801	74.310	62.931	60.889	61.340
	NormWear (Ours)	80.717	64.913	78.656	66.408	65.883	65.393

The shading colors are used to facilitate the differentiation of various types of downstream tasks. The result tables in the later sections use consistent color coding to indicate task categories, in accordance with this shading scheme. The best performance achieved for each task is highlighted in bold.

G Channel Fusion Complexity Analysis

When conducting multi-channel modeling, for example, when the input comprises an arbitrary number of signals, a fusion operation needs to be conducted across all channels in order to let the model extract correlation information.

We explore several fusion approaches and different design of liaison token as shown in Figure 2(b), with each method described below:

- (1) *All-Attention Fusion*: This approach involves concatenating all tokens from each modality without considering their individual properties and fusing the information through a self-attention module. However, this method requires quadratic computation time, as every token passes through the self-attention module, making it impractical for real-world applications.
- (2) *Cross-Attention Fusion*: In addition to the cross-attention mechanism used in Cross-ViT [19], we introduce a slight modification to fit in our problem setting. We propose a symmetric fusion method, using the [CLS] token from each modality as an intermediary to exchange information between the patch tokens of another modality, then projecting the information back to its original modality in the subsequent Transformer layer. While this strategy is efficient, it restricts the model to handling only two time series signals or modalities,

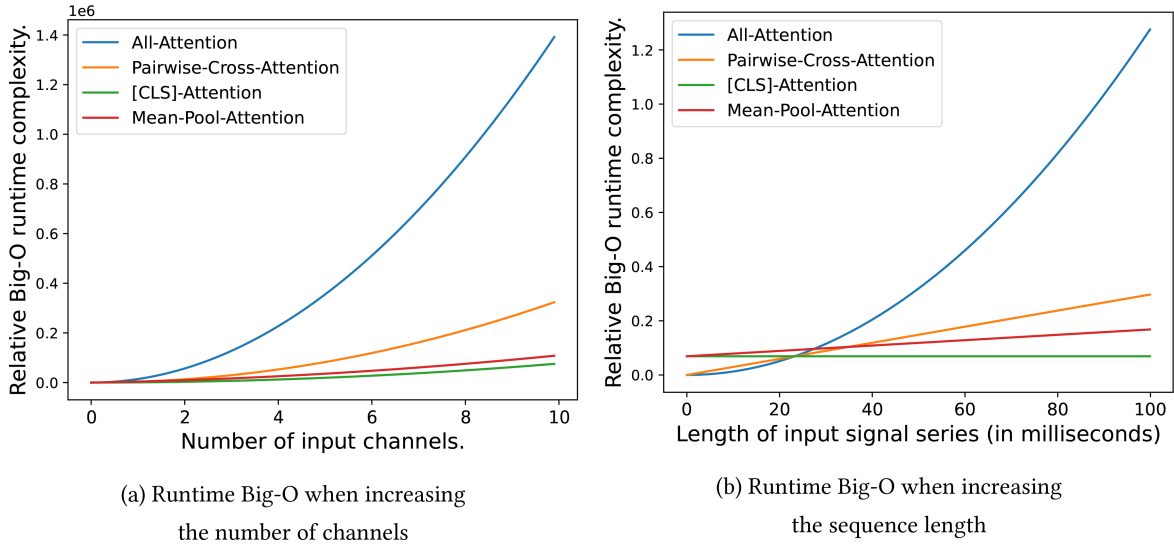


Fig. G1. Visualization of runtime complexity when scaling up the number of channels or the sequence length.

which deviates from our goal of building a general model capable of processing an arbitrary number of channels.

Because we will deploy the model on an edge device like Jetson Nano, other than empirical evidence of the performance, we also have to consider the computation complexity of different approaches. A brief visualization of the runtime complexity of different approaches is presented in Figure G1. The detailed derivation is presented in the following subsections.

G.1 All-Attention

For the approach of conducting self-attention by concatenating all the patches, we arrive the Big-O complexity expression as follows:

- We denote C as the number of input channels, d as the embedding size, L as the number of patches convolved from the time series in each channel (proportional to sequence length), and $x \in \mathbb{R}^{C \times L \times d}$ as the input data before feeding into the fusion block. We have a total of $L \cdot C$ patches.
- When calculating the attention scores, DPs are computed for each pair of the patches, which results in the following calculation process:
where “1), 2), 3)” represents the operations conducted at the first, second, and third rounds of entering the entire nested loops. The complexity for the first round of operation results in a complexity of:

$$\sum_{i=1}^C \sum_{j=1}^L \sum_{k=1}^C \sum_{l=1}^L 2d = \sum_{i=1}^C \sum_{j=1}^L \sum_{k=1}^C L \cdot 2d = \sum_{i=1}^C \sum_{j=1}^L C \cdot L \cdot 2d = O(d \cdot (L \cdot C)^2), \quad (G1)$$

where in the case of multi-head attention, the DP still has the complexity of $O(2d)$, and because the number of heads is a constant, the final complexity is equivalent to the result in Equation (G1).

Algorithm 2: All-Attention Complexity

```

for  $i \in [1, 2, \dots, C]$  do
  for  $j \in [1, 2, \dots, L]$  do
     $N \leftarrow \exp(\text{attn}(x_{i,j})) \implies O(L \cdot C)$ 
    for  $k \in [1, 2, \dots, C]$  do
      for  $l \in [1, 2, \dots, L]$  do
        1) Calculate dot product:  $\text{attn}(x_{i,j}, x_{k,l}) = x_{i,j}^T x_{k,l} \implies O(2d)$ 
        2) Softmax over attention scores:  $\frac{\exp(\text{attn}(x_{i,j}, x_{k,l}))}{N} \implies O(1)$ 
        3) Weighted average:  $x_{i,j} + \text{attn}(x_{i,j}, x_{k,l}) \cdot x_{k,l} \implies O(2d)$ 
      end for
    end for
  end for
end for

```

— Similarly, the softmax operation will result in a complexity of $O((L \cdot C)^2)$, and the final weighted average operation will also have a complexity of $O(d \cdot (L \cdot C)^2)$, which results in total complexity of:

$$O(d \cdot (L \cdot C)^2) + O((L \cdot C)^2) + O(d \cdot (L \cdot C)^2) = O(d \cdot (L \cdot C)^2). \quad (\text{G2})$$

G.2 Cross-Attention

For the pairwise cross-attention approach following guidance of Chen et al. [19], we have the operation defined as with the same notion in the previous subsection. The total complexity is:

$$O(C^2 \cdot L \cdot 2d) + O(C^2 \cdot L) + O(C^2 \cdot L \cdot 2d) = O(d \cdot L \cdot C^2). \quad (\text{G3})$$

Algorithm 3: Cross-Attention Complexity

```

for  $i$  in  $[1, 2, \dots, C - 1]$  do
  for  $j$  in  $[1, 2, \dots, C]$  do
    2)  $N = \exp(\text{attn}(x_{i,1})) \implies O(L)$ 
    for  $k$  in  $[2, 3, \dots, L]$  do
      1) Calculate  $\text{attn}(x_{i,1}, x_{j,k}) \implies O(2d)$ 
      2) Softmax over all-attention scores,  $\frac{\exp(\text{attn}(x_{i,1}, x_{j,k}))}{N} \implies O(1)$ 
      3) Weighted average:  $x_{i,1} + x_{j,k} \implies O(2d)$ 
    end for
  end for
end for

```

G.3 [CLS]-Attention

This is the approach that we adopted for the final version of our proposed foundation model. Only the embedding corresponding to the [CLS] token of each channel is involved during the self-attention operation. Therefore, the complexity is:

$$O(d \cdot C^2). \quad (\text{G4})$$

G.4 Mean-Pool Attention

For fusion with mean-pool attention, we first calculate the mean representation for each channel, resulting in a complexity of $O(C \cdot L \cdot d)$. And self-attention with Tese mean representations has the same complexity as [CLS]-attention, which is $O(d \cdot C^2)$. Thus, the total complexity is:

$$O(C \cdot L \cdot d) + O(d \cdot C^2) = O(d \cdot (L \cdot C + C^2)). \quad (G5)$$

H MSiTF Complexity Analysis

Algorithm 4: MSiTF Runtime Complexity

key embedding $E_k = k(S) \in \mathbb{R}^{p \times d}, \implies O(d^2)$
 value embedding $E_v = v(S) \in \mathbb{R}^{p \times d}, \implies O(d^2)$
 Relevance score $Rel = E_k^T Q \in \mathbb{R}^p, \implies O(pd)$
 likelihood parameter $E_l = l(S) \in \mathbb{R}^{p \times 2}, \implies O(d^2)$
 Importance score sampling $W_{imp} \in \mathbb{R}^p$ (equation 5) $\implies O(p)$
 Fused embedding $E_{final} = E_v^T (\alpha W_{imp} + \beta W_{rel} + \kappa W_{rec}) \in \mathbb{R}^d, \implies O(pd)$
 Inference final score $c = \arg \max_{i \in |C|} C_i^T E_{final}, \implies O(cd)$

where d being the latent size, p being the number of total patches, c being the number of available ground truth choice, k and v being the key and value linear mapping, $S \in \mathbb{R}^{p \times d}$ as the signal embeddings, Q as the query sentence embedding, and C as the list of available answer choice sentences. The total runtime complexity is $O(d^2 + pd + p + cd)$. Since d is constant, we have runtime complexity of $O(p + c)$.

Regarding memory complexity of MSiTF, with m being the size of text encoder, w being the size of normwear, we have (i) Signal representations: $O(pd)$; (ii) Text representations: $O(cd)$; (iii) Total: $O(m + w + d(p + c))$. Since m , w , and d are all constants, we have memory complexity of $O(p + c)$.

I Feature Visualization

Feature visualization serves as a tool to interpret and analyze the internal representations learned by the model. By examining activation patterns or embedding structures at various layers, we aim to understand how the model encodes input signals and whether these representations align with relevant semantic or structural information. This analysis provides insight into the effectiveness of the learned features and can inform architectural or training modifications to improve performance and generalization.

I.1 The Model is Agnostic to the Input Signals

This section investigates whether, without requiring the signal modality type information as input, NORMWEAR can effectively distinguish between different signal sources. We randomly sampled 500 samples for each sensor type and fed them into our pretrained model. We use t-SNE [78], with PCA [36] initialization to visualize the learned representations corresponding to the [CLS] special token at the last layer. The PCA preserves the global structure, while t-SNE emphasizes local relationships in the data. From Figure 11(a), we observe that representations from sensors located at the same body position are clustered closely together, while representations from different body locations are clearly separated. This suggests that our model is signal-agnostic, as it can recognize the signal type differences, map their representations appropriately in the embedding space, and guide feature extraction within each Transformer block.

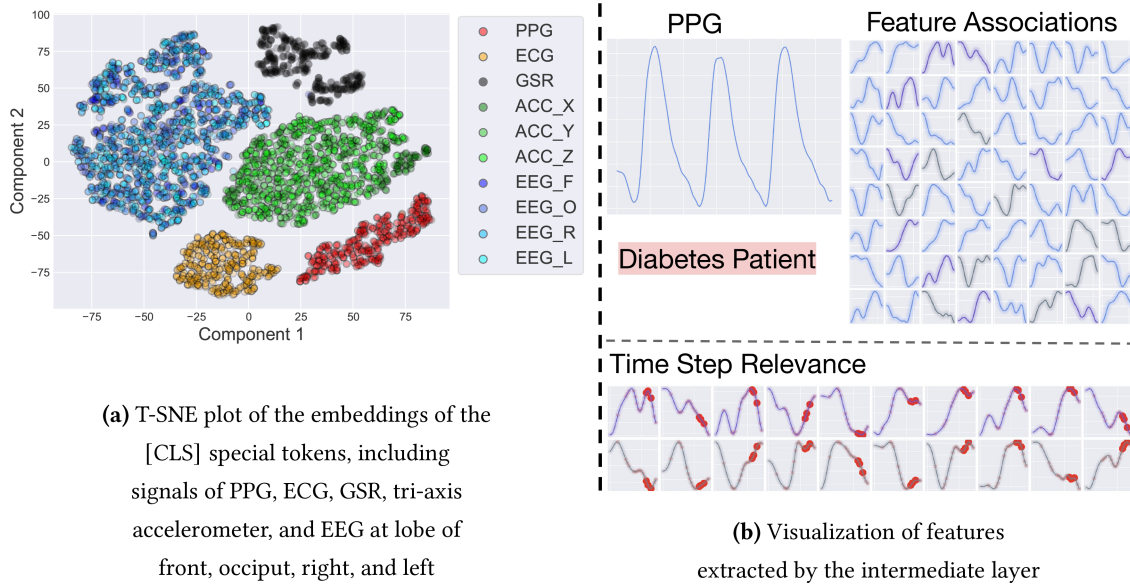


Fig. 11. Feature visualization.

1.2 Waveform Visualization

Figure 11(b) under “Feature Associations” shows the features extracted by our model. Each patch corresponds to a representation with a vector size of $\mathbb{R}^{768}$. When ordered by time sequence, these representations form 768 waveforms per layer, representing the model’s extracted features. The figure displays 64 randomly sampled waveforms from a selected layer. The features highlighted in purple and gray indicate the top 10 patterns positively and negatively associated with the target task (diabetes classification, in this example), with associations determined by linear regression parameters during linear probing. Additionally, our relevance-based fusion mechanism identifies the contribution of each timestep during inference, highlighted by red dots in the “Time Step Relevance” section of Figure 11(b).

Such a visualization pipeline can assist researchers and clinicians by offering insights into how the model reaches its final predictions. Given the millions of parameters and hundreds of waveform features per layer, visualizing these features individually is inefficient for understanding the overall behavior of the proposed foundation model. As a result, we use several techniques in nonlinear dynamic analysis [75] to quantify the overall patterns of these extracted features, which are discussed in detail in Section 4.8.

1.3 T-SNE Plot among Classes

We present T-SNE plots of NormWear’s embeddings across different classes to provide insights into their structure and assess their suitability for sample similarity-based information retrieval. It is important to note that these plots are exploratory in nature and do not serve as a claim of the embeddings’ superiority. As shown in Figures 7(A), clear class separations can be observed in certain scenarios. For example, EEG samples from seizure subjects and normal subjects are distinctly separated, and physical activity types are well-clustered. For ECG data, abnormal heartbeats tend to form cohesive clusters. However, it is essential to recognize that these T-SNE plots reduce the latent representations into a 2D space, which may not fully capture the inherent properties of the embeddings in their original high-dimensional form.

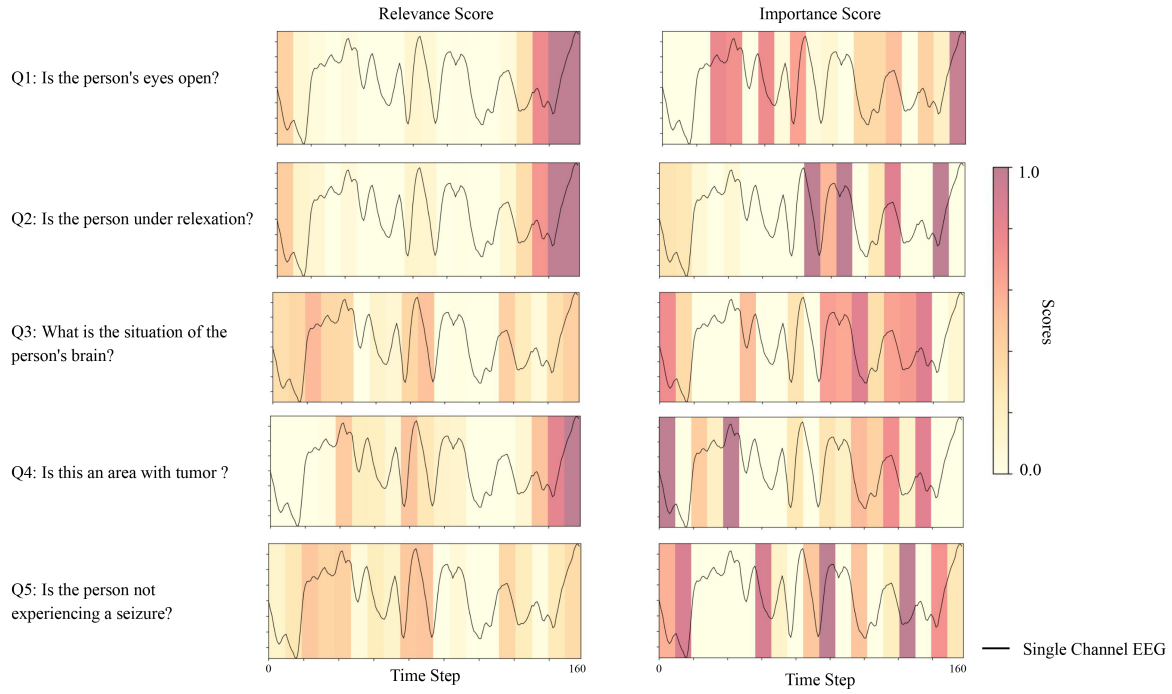


Fig. 12. Visualization of relevance scores (left) and importance scores (right) for a single channel EEG sample from the Epilepsy dataset under five task-specific questions. The background color follows a yellow-to-red scale, where darker regions indicate higher scores.

1.4 Supplementary Qualitative Analysis of MSiTF

To understand how each of our proposed gating modules in MSiTF—relevance, recency, and importance—select useful features for different tasks, we visualize the scores assigned to each time window. As shown qualitatively in Figure 12, the heatmaps reveal that both relevance and importance scores are sensitive to task differences. For example, in the eye closure detection task, the model focuses on the last few patches, whereas in the seizure detection task, it emphasizes patches with large fluctuations. A similar pattern is observed for the importance score, where patches are weighted differently across tasks. This suggests that our gating mechanism can adaptively select relevant features based on the task. We include a figure of the recency score (Figure 13) for completeness. Since the recency score is derived from a fixed decay function and is not learned, it remains the same across tasks.

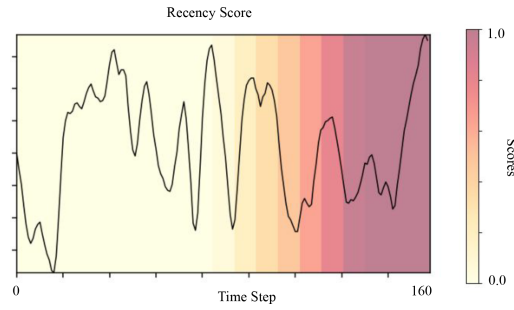


Fig. 13. Recency score generated by a decay function. The sample is selected from the Epilepsy dataset.

To improve visualization, we aggregated token scores using a window size of 9, which matches our tokenization patch size. We then applied Z-score normalization to ensure comparability across tasks. The sample was selected from the Epilepsy dataset due to its multiple and diverse task types.

J Reconstruction Example

Examples of the reconstruction output is presented in Figures [J1](#) and [j2](#).

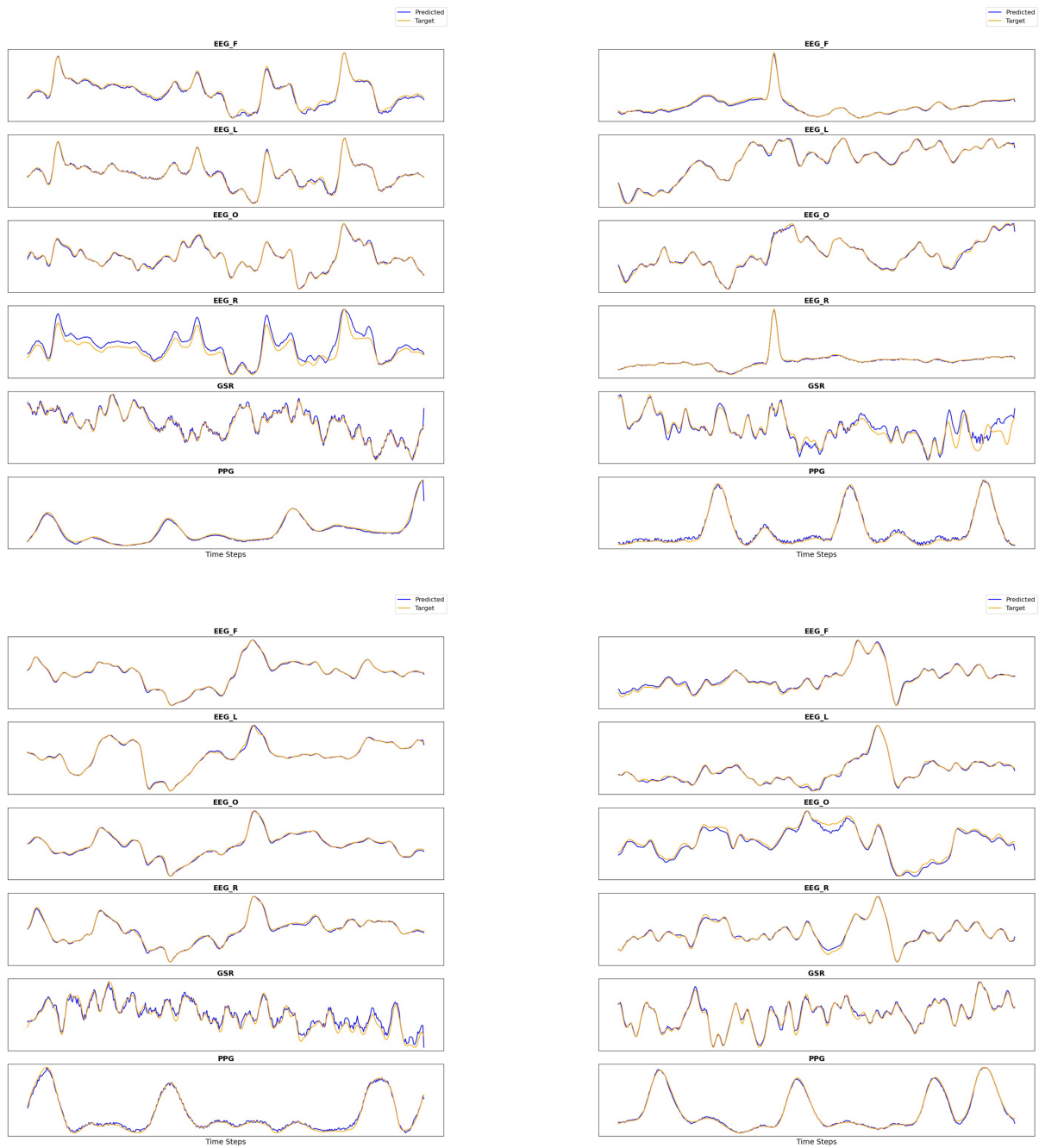


Fig. J1. Uncurated random samples on Phyll scalogram, using a NORMWEAR trained in our training set. The masking ratio is 80%.

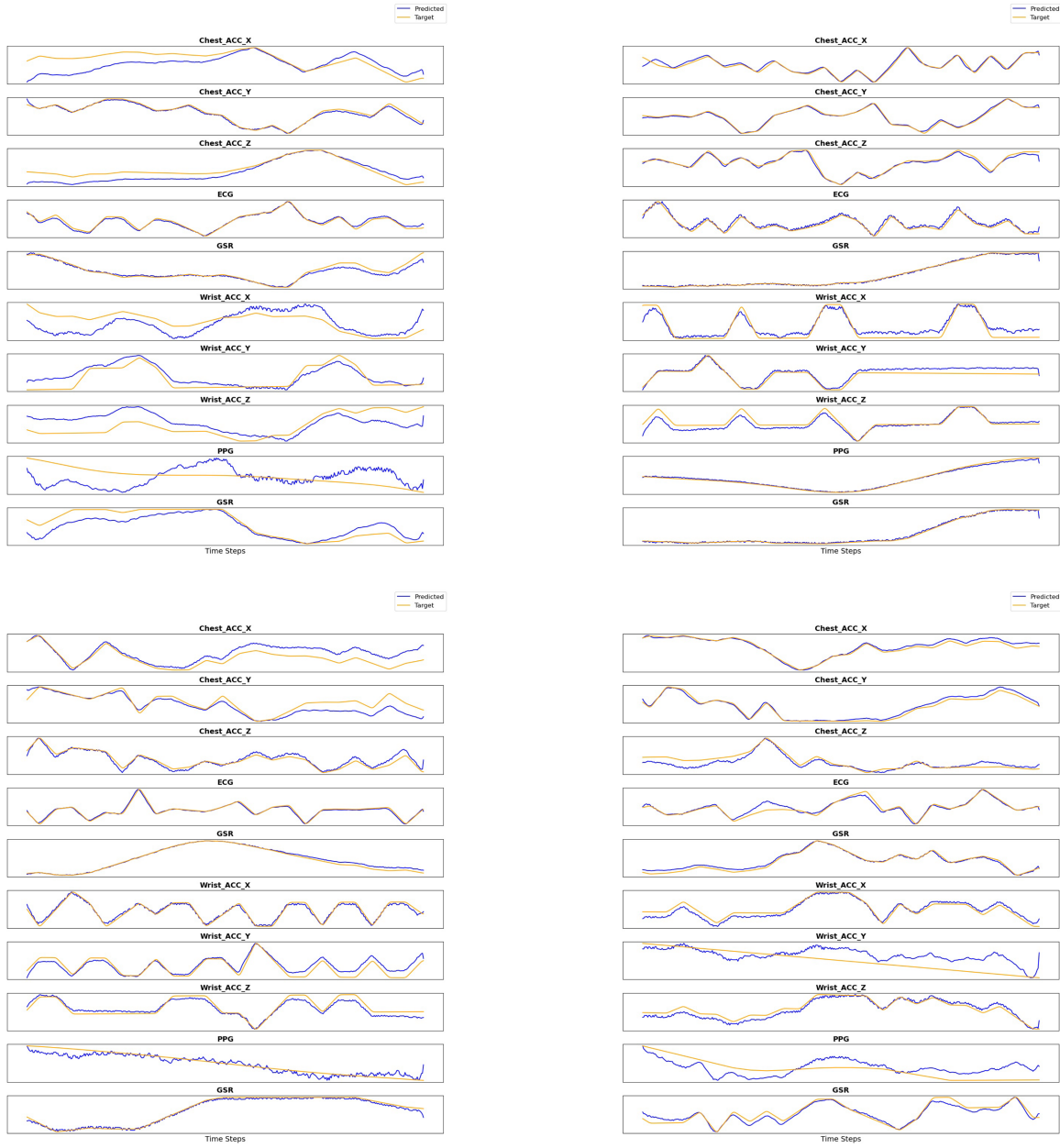


Fig. J2. Uncurated random samples on WESAD scalogram, using a NormWear trained in our training set. The masking ratio is 80%. Note that the IMU data are not in the training set and, in general, NORMWEAR is able to reconstruct this with high accuracy.

Received 12 November 2025; revised 4 February 2026; accepted 10 March 2026