Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals

Anonymous Authors[1]

Abstract

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, a foundation model designed to extract generalized and informative representations from wearable sensing data. NORMWEAR is pretrained on a diverse set of physiological signals, including PPG, ECG, EEG, GSR, and IMU, from various public datasets. For evaluation, we benchmark its performance across 11 public wearable sensing datasets, spanning 18 applications in mental health, body state inference, vital sign estimation, and disease risk evaluation, demonstrating superior performance compared to competitive baselines. Additionally, using a novel representation-alignment-match method, we align physiological signal embeddings with text embeddings, enabling zero-shot inference for unseen wearable signal-based health applications.

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 (Zhang et al., 2024a).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Recently, several foundation models have emerged for time series modeling, including Ansari et al. (2024); Abbaspourazad et al. (2023); Woo et al. (2024); Foumani et al. (2024). 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 (Semenoglou et al., 2023; Wimmer & Rekabsaz, 2023; Vishnupriya & Meenakshi, 2018; Chun et al., 2016; Krishnan et al., 2020; Dosovitskiy et al., 2020). These works have significantly advanced the field and provided valuable insights, yet two main issues still exists which need further exploration to fully understand their potential in wearable scenarios. First, contrastive learning-based foundation models (Abbaspourazad et al., 2023) 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 (Ansari et al., 2024; Zhang et al., 2022; Woo et al., 2024) and spectral-based approaches (Semenoglou et al., 2023; Wimmer & Rekabsaz, 2023) 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 (Ansari et al., 2024; Zhang et al., 2022) or convert them into image or spectral representations (Wu et al., 2023; Mathew et al., 2024; Vaid et al., 2023). 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

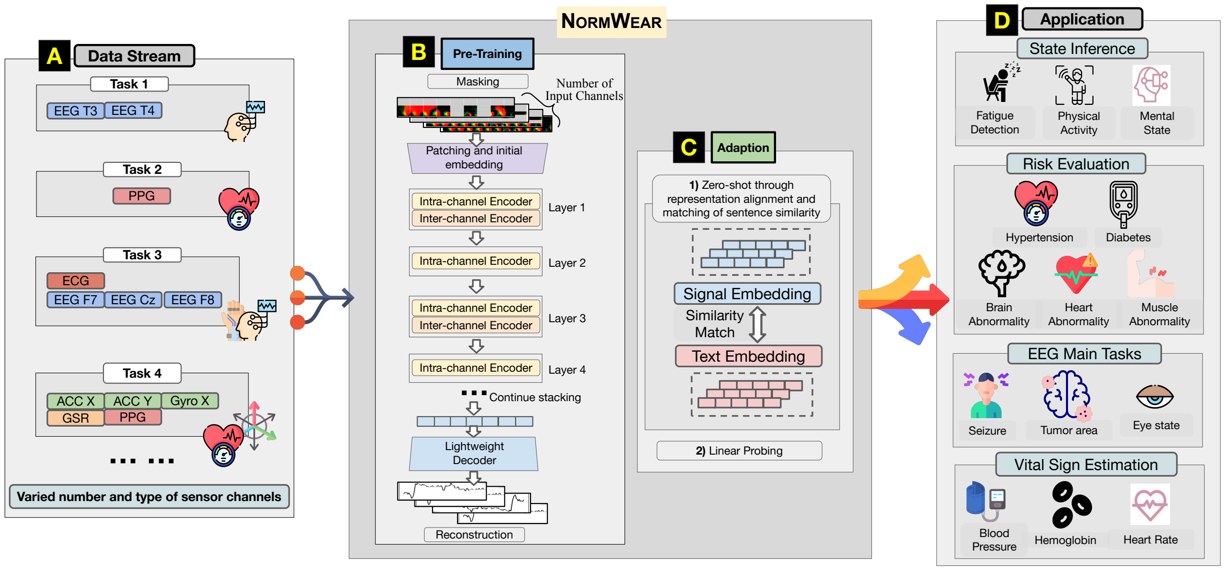


Figure 1. The role of our framework. Several icons from Freepik (n.d.); Zhang et al. (2024a).)

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, 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 publicibly avaliable datasets. We evaluated NormWear on 18 public downstream tasks against competitive baselines under both linear probing and zero-shot inference. 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 makes NormWear the first to achieve zero-shot inference on 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. We make the preprocessed data, codebase, and model weights publicly available.

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.

2. Related Work

Foundation models have emerged as a transformative paradigm in machine learning, enabling generalizable and reusable representations across diverse downstream tasks (Bommasani et al., 2022). In the time series domain, recent works (Ansari et al., 2024; Foumani et al., 2024; Abbaspourazad et al., 2023; Narayanswamy et al., 2024) have demonstrated success in tasks such as forecasting, classification, and anomaly detection. However, their generalizability to health-related wearable signals remains limited due to the lack of in-depth evaluation, reliance on specific sensor types and univariate data, as well as the inability to handle the heterogeneity of multivariate wearable signals. In contrast, NormWear builds upon these principles by introducing a modeling framework that is agnostic to the sensor modality and number of input channels, as stated in section 1, and is presented in details in section 3. NormWear has been evaluated on 18 digital healthcare tasks and demonstrate peak performance against solid time series modeling baselines, including common statistical approach, SoTA model in time series with self-supervised learning (Zhang et al., 2022), SoTA spectrum based modeling approach (Wu et al., 2023), and SoTA time series forecasting model (Ansari et al., 2024). Our work not only generalizes to arbitrary sensor configurations but also ensures compatibility across multivariate data, addressing key limitations of earlier approaches.

Table 1. Downstream evaluation data that are unseen during pretraining.

Downstream Dataset	Sensor	Tasks	#Samp. (#Subj.)
WESAD (Schmidt et al., 2018)	IMU, PPG, ECG, GSR	Stress Detection	11050(15)
UCI-HAR (Reyes-Ortiz et al., 2012)	IMU	HAR	10299(30)
DriverFatigue (Min et al., 2017)	EEG	Fatigue Detection	2400(12)
Activity Recognition Total	-	-	**23749(57)**
Epilepsy (Andrzejak et al., 2023)	EEG	State Recognize	11500(500)
GAMEEMO (Alakus et al., 2020)	EEG	Valence-Arousal	5600(28)
EEG Main Tasks Total	-	-	**17100(528)**
ECG-Abnormal (Bousseljot et al., 2009)	ECG	Abnormal Detection	11640(249)
PPG-BP (Liang et al., 2018)	PPG	Risk of Diseases	657(219)
PhysioNet EMG (Goldberger et al., 2000)	EMG	Muscular Diseases	163(3)
Risk Evaluation Total	-	-	**12460(471)**
Noninvasive-BP (Esmaili et al., 2017)	PPG	BP Estimate	125(26)
PPG-Hgb (Esmaili et al., 2017)	PPG	Hgb Estimate	68(68)
Fetal-fPCG (Bhaskaran et al., 2022)	PCG	Fetal HR Estimate	47(47)
Vital Signs Total	-	-	**240(141)**
Total All	-	-	**53549(1197)**

Table 2. Baselines and pretraining data.

Baseline Methods	Modeling Strategies
TF-C (Zhang et al., 2022)	SoTA in TS SSL; modeling time and frequency domain information at same time.
CLAP (Wu et al., 2023)	SoTA in audio modeling; process signal as spectrogram
Chronos (Ansari et al., 2024)	SoTA in TS forecasting, leverage LLM for modeling
Statistical approach	Reserve full interpretability

Pretrain Dataset	Sensors	#Samp (hours).
Cuff-Less-BP (Kachuee et al., 2016)	ECG, PPG	42934(72)
PPG-Dalia (Reiss Attila, 2019)	ECG, PPG IMU, GSR	42606(71)
Auditory-EEG (Alzahab et al., 2022)	EEG	13601(23)
PhyAAt (Bajaj et al., 2020)	EEG	19550(33)
MAUS (Beh et al., 2021)	ECG, PPG GSR	13068(22)
Mendeley-YAAD (Dar et al., 2022)	ECG, GSR	2964(5)
Brain-Cognitive (Dar et al., 2022)	EEG	51201(85)
EPHNOGRAM (Dar et al., 2022)	ECG, PCG	36611(61)
BIDMC (Dar et al., 2022)	ECG, PPG	8427(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)**

3. Method

3.1. Dataset construction for model pretraining and downstream evaluation

We curated a collection of 9 publicly available datasets (Table 2) exclusively for model pretraining, resulting in approximately 230,962 multivariate time series segments, comprising 4,294 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 herustic 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, which include affective state classification, physical state recognition, biological estimation, and disease risk evaluation. Details about the datasets is presented in Table 1.

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) (Brigham, 1988) with a sliding window to compute spectrograms, are widely regarded as the benchmark approach for tokenization. However, due to the inherent trade-off 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. (2019), followed by computing the continuous wavelet transform (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 & Compo (1998). For the wavelet transform, we use the Mexican Hat wavelet for signal convolution, as recommended by previous studies (Burke & Nasor, 2004; Hosni & Atef, 2023; Hassani, 2021; Negi et al., 2024; Nedorubova et al., 2021b). We apply scales ranging from 1 to 64, following the guidance of (Sengupta et al., 2022; Nedorubova et al., 2021a), 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

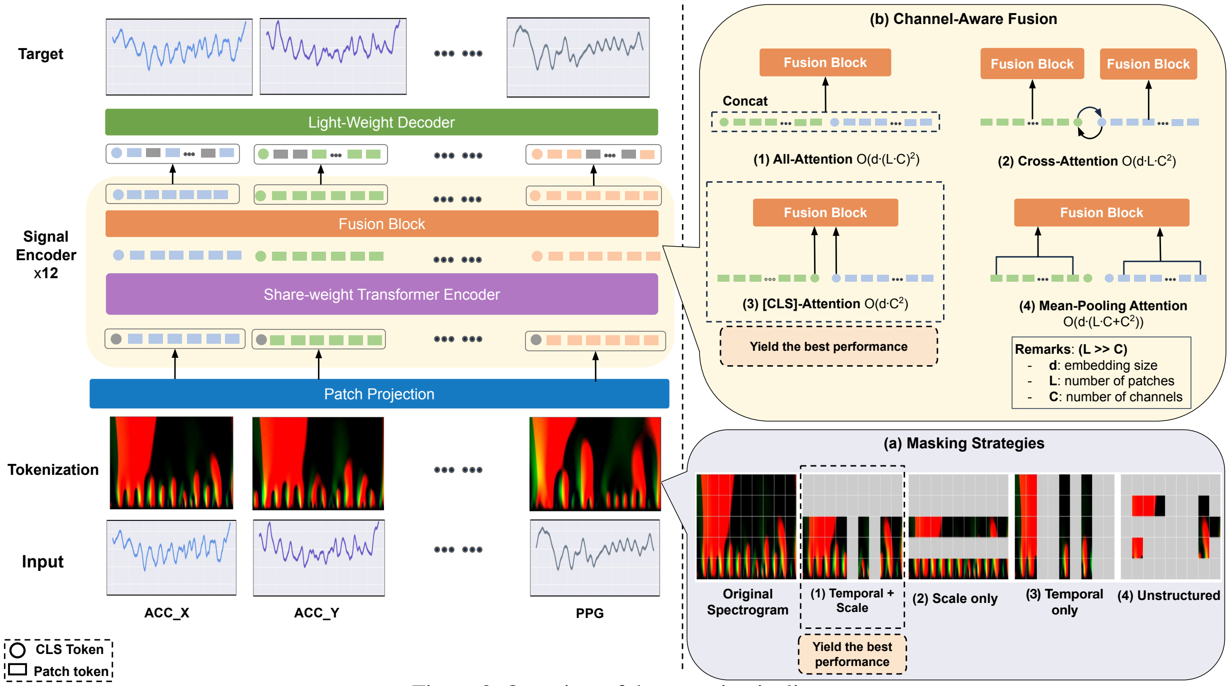


Figure 2. Overview of the pretrain pipeline.

ViT (Dosovitskiy et al., 2020). In this way, this tokenization approach can be applied to various types of sensing signals without sensor-specific adjustments or reconfigurations.

3.3. Model architecture and pretrain strategies

Following the tokenization step, we adopt common reconstruction-based pretraining strategies from Masked Auto Encoder (MAE) (He et al., 2021; Huang et al., 2023; Zhang et al., 2023), which applying masking to input tokens and and optimizing the model using mean squared error (MSE) for reconstructing the raw time series. Inspired by Huang et al. (2023), 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 (Vaswani et al., 2023). For the same reason, NormWear uses a lightweight decoder consisting of 2 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 special token [CLS] at each signal channel, aiming to learn and extract a generic representation for each signal.

Another important point to consider is that although empirical studies (Nie et al., 2023; Abbaspourazad et al., 2023) show that channel-independent structures effectively capture local patterns, they fail to account for relationships across channels. To address this, we introduce a channel-aware attention (fusion) layer after every other encoder block to incorporate cross-channel information. We explore several fusion approaches 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 (Chen et al., 2021), 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, which deviates from our goal of building a general model capable of processing an arbitrary number of channels.

(3) **[CLS]-Attention Fusion** 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

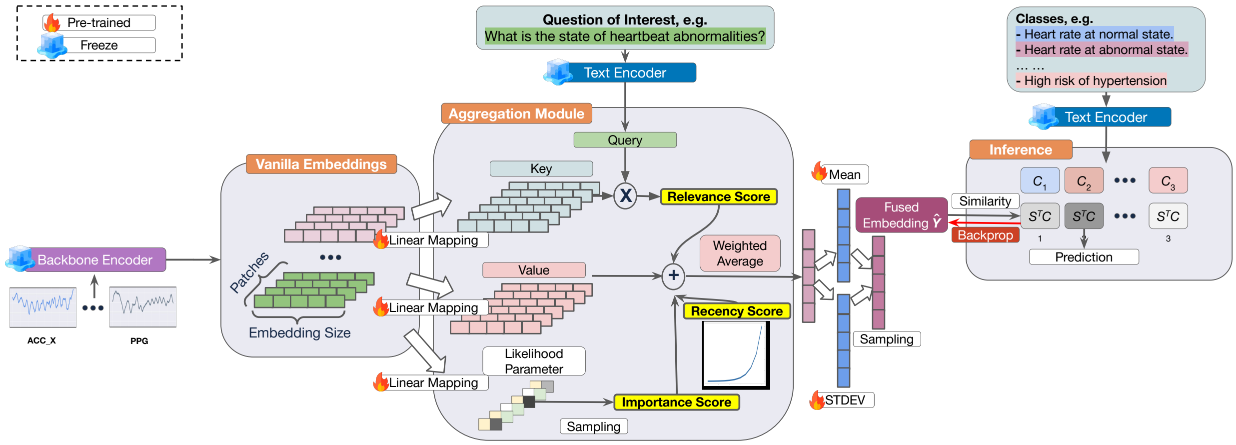


Figure 3. Memory stream inspired temporal fusion mechanism for representation alignment.

token in subsequent transformer encoder layers.

(4) **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 performance for downstream tasks for our proposed NormWear model. Details of all the ablation studies are reported in appendix B.

3.4. Zero-shot inference

Zero-shot learning traditionally refers to generalizing to unseen object categories in classification tasks (Pourpanah et al., 2022). In this work, we extend this concept to explore NormWear's ability to generalize across unseen datasets and tasks. While extensively studied in vision-language (Radford et al., 2021) and audio-language (Wu et al., 2023) domains, zero-shot learning remains underexplored in wearable sensors. Existing frameworks, such as (Zhang et al., 2024b; Wu et al., 2023), align signal time series with text descriptions leveraging end-to-end training. To reduce computation cost and counteract the issue of catastrophic forgetting (Li et al., 2023), we use off-the-shelf frozen encoders for both signal and text modalities. Specifically, we enable zero-shot inference by first training a human-sensing-driven fusion module, followed by an aligner module to map input signals into textual space.

Addressing Three Human Sensing Challenges. Different human-sensing tasks require different sets of signals and patterns. To guide better information extraction from distinct embedding space, our zero-shot inference module incorporates query sentences that guide the target task as heuristic input. For instance, if the task is physical activity detection, the focus should be on IMU, rather than EEG. Such signal modality identification will be guided in our zero-shot inference through query sentences such as: "*What activity is the subject doing?*". To effectively retrieve task-relevant signal information guided by the query sentence, we introduce *relevance score.*

Human physiological and behavioral signals are highly dynamic, meaning they can change rapidly in response to external stimuli or internal states. Focusing on the recent past ensures that the data being analyzed remains contextually relevant and accurately reflects the person's current physiological or emotional state. For instance, the freezing effect on heart rate occurs as a rapid fluctuation in response to an acute external stressor (Roelofs, 2017; Chowdhury et al., 2020; Chaudhury et al., 2021), such as a sudden loud noise, a perceived threat, or an anxiety-inducing situation. Hence, systems designed for physiological signals need to focus more on recent physiological and behavioral signals to ensure effective and meaningful assessments. To this end, we introduce *recency scores*, which assign higher weights to patches closer to the most recent time step in the sequence.

Moreover, due to the challenges of precise labeling, human sensing signals are often weakly labeled (He et al., 2018; Kim et al., 2022; Qian et al., 2021; Ma et al., 2021), mostly consisting of generic state sequences. Only a relatively small fraction of the sequences contain meaningful active segments, anomalies, or transient states, which are crucial for accurate inference of human-centric tasks such as state recognition and risk assessment. To address this characteristic, we introduce an *importance score* that dynamically determines the retention weights across temporal steps, thereby regulating the information flow for subsequent processing.

To summarize, the latent signal representations from NormWear have to be aggregated according to **(1)** *how relevant they are to the objective tasks*, **(2)** *how close they are to the most current time step*, and **(3)** *how important they are to the data itself*. Inspired by the philosophy of memory stream retrieval mechanism (Park et al., 2023), we implemented such a fusion mechanism named MSiTF to generate representations optimized for human sensing, shown in Figure 3.

Memory Stream inspired Temporal Fusion (MSiTF). Our Aggregation or Fusion Module, MSiTF, 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 consider the *Relevance* score to be the cross-attention score between the sentence embedding generated by the pretrained language model (Muzammil, 2021) of the query sentence and the key representation of the embedding of each time step, enabling distinct but relevant contextual information identification. For the *Recency* score, to prioritize recent time steps, we use an exponential decay function, where the further the time step to the most recent time step, the lower the score. Finally, we consider the importance score IMP in this case to be whether to keep the representation at each time step or not. In order to achieve this, we assign binary parameters to each time step, denoted as $\theta_t = p(v_t) \in \mathbb{R}^2$ where $v_t \in \mathbb{R}^E$ is the representation vector at time step 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

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

where ϵ is the noise term sampled from Gumbel distribution (Jang et al., 2017), 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 a variational-inspired approach (Kingma & Welling, 2022). This design injects stochasticity into the representation, encouraging the model to explore and capture nuanced variations in semantic representations.

Aligner Module, Objective Function, and Pretraining. Finally, 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 pre-trained language model (Muzammil, 2021). At this stage, Y is leveraged to supervise the fused output $\hat{Y}$ with integrated loss function with penalty on Manhattan distance and cosine similarity, aiming to align the physiological and semantic vectors to have the same magnitude and direction:

$$Loss(Y, \hat{Y}) = \lambda |Y - \hat{Y}| + \left(1 - \frac{Y \cdot \hat{Y}}{\|Y\| \|\hat{Y}\|}\right) \tag{2}$$

where λ is hyper-parameters controlling the weight of loss components. We train the MSiTF on the pretraining datasets stated in Table 2, with both classification and regression tasks. 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) (Achiam et al., 2023) to generate 20 alternative variations for each sentence, from which only one is randomly sampled during pre-training.

During test-time inference on downstream datasets, each ground truth label is converted into a sentence (details in appendix A), 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.

4. Experiments

In this section, we present a comprehensive evaluation across 11 publicly available datasets, focusing on 18 widely-recognized digital healthcare tasks. We first assess the transferability advantage of our proposed model compared to the solid baselines. Additionally, we examine the zero-shot capabilities of NormWear.

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 2).

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 (Yan et al., 2023a; Reyes-Ortiz et al., 2012; Mikelsons et al., 2017). We include this simple baseline as sanity check. Secondly, since sensory data can be naturally represented as time series (Woo et al., 2024; Semenoglou

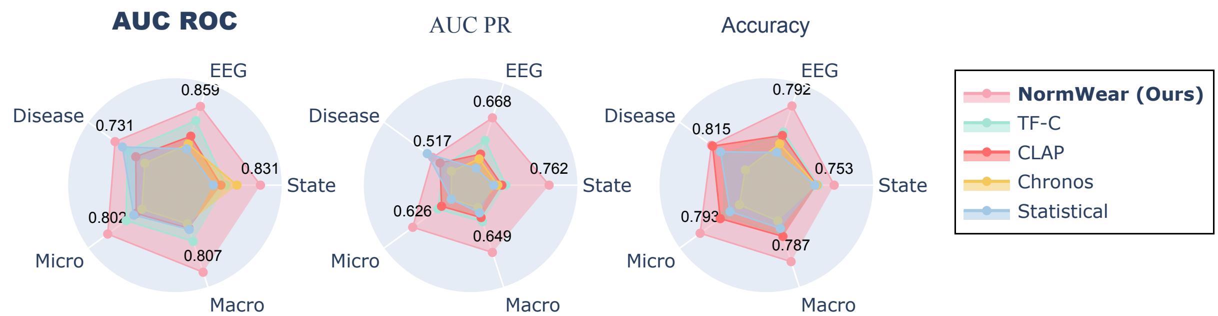


Figure 4. Overview of performance trend of NORMWEAR against competitive baselines in downstream tasks: **(1) Disease** risk predictions. **(2) EEG** main tasks (mental and abnormal states prediction). **(3) State** recognition: physical and mental activities. **(4) Macro**: Average performance over types of tasks. **(5) Micro**: Average performance over each task.

Table 3. Performance on various downstream wearable-signal-based health related applications under linear probing evaluation.

Downstream Tasks	Statistical	Chronos	CLAP	TF-C	NORMWEAR (Ours)
WESAD	66.213	71.489	72.383	69.865	**76.060**
UCI-HAR	95.784	91.593	96.420	96.892	**98.954**
DriverFatigue	63.249	**76.722**	61.889	66.882	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	**92.743**
Epilepsy (eye relaxation)	87.457	88.218	89.867	94.416	94.828
Epilepsy (health area)	86.274	81.08	83.711	85.619	**88.541**
Epilepsy (tumor area)	82.816	81.034	83.644	86.348	87.197
Epilepsy (seizure)	88.272	97.572	**97.734**	93.998	97.053
GAMEEMO	51.009	53.747	52.551	**56.275**	54.937
EEG Main Tasks Avg.	79.720	80.677	82.100	84.302	**85.883**
ECG-Abnormal	97.092	98.585	97.23	98.275	**99.140**
PPG-BP (HTN)	59.499	52.425	56.757	**65.229**	62.341
PPG-BP (DM)	47.823	51.164	42.455	**57.883**	55.893
PPG-BP (CVA)	**71.25**	50.278	51.667	58.125	70.625
PPG-BP (CVD)	51.219	58.31	50.91	**58.674**	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	92.31	91.79	91.922	87.481	**92.420**
PPG-Hgb	94.219	**95.005**	94.291	93.408	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**

et al., 2023), we benchmarked our model against Chronos (Ansari et al., 2024) , as well as a self-supervised framework TF-C (Zhang et al., 2022). Finally, the spectrum-based modeling methods (Vishnupriya & Meenakshi, 2018; Chun et al., 2016; Krishnan et al., 2020) are widely used for signal modeling. Therefore, we incorporate CLAP (Wu et al., 2023) into baselines that has demonstrates SoTA performance in spectrogram-based modeling. These baselines span distinct paradigms, providing a solid foundation to demonstrate the strengths of our model in wearable signal tasks.

4.2. Downstream evaluation, NORMWEAR achieves the peak performance

We perform supervised training to evaluate the representation with linear probing on each downstream dataset. Performance is then assessed in the test set of these datasets. 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. All scores are the higher the better. Such

Table 4. Zero-shot performance on the downstream datasets, with AUC ROC being reported. The last two columns show the average across the tasks and across group types respectively.

Model	WESAD	UCI-HAR	DriverFatigue	GAMEEMO	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
NORMWEAR w/ MSiTF	55.9	**71.4**	54.9	50.2	54.0	56.4	**66.9**	**57.4**	53.7	56.5	53.2	**65.0**	**63.1**	**74.3**	**65.7**	**59.9**	**60.1**
- 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

evaluation framework ensure better reproducibility, numerical stability, and fairness in performance comparison.

From Figure 4, Table 3, and Table 12, 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 state-of-the-art (SoTA) time-series self-supervised learning framework (Zhang et al., 2022), a 5.3% improvement over the SoTA spectrum-based modeling method (Wu et al., 2023), a 6.1% improvement over SoTA time-series forecasting models with LLM backbones (Ansari et al., 2024), 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. 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.

4.3. The first zero-shot enabled foundation model for wearable sensing health applications

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 (Wu et al., 2023) 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 4, 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 that, even without fine-tuning, the model is capable of learning informative representations that can be directly leveraged for downstream tasks. Furthermore, as shown in Section 4.2, even a straightforward adaptation, such as linear probing, can yield notably improved results.

5. Conclusion

In this work, we mainly propose a foundation model for wearable physiological signals. NORMWEAR is 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.

Impact Statement

NORMWEAR addresses a significant need—handling complex, heterogeneous biosignals across a variety of healthcare tasks—and provides promising evidence of strong performance and generalizability. The methodological innovations, particularly the tokenization and fusion mechanisms, are valuable contributions that could inspire future work in multimodal healthcare analytics.

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 is submitted in supplementary material referred to as *NormWear_main.zip*, 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 paper. The GitHub repository containing all the documentation will be published simultaneously with the paper.

References

Abbaspourazad, S., Elachqar, O., Miller, A. C., Emrani, S., Nallasamy, U., and Shapiro, I. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alakus, T. B., Gonen, M., and Turkoglu, I. Database for an emotion recognition system based on eeg signals and various computer games–gameemo. *Biomedical Signal Processing and Control*, 60:101951, 2020.

Alzahab, N. A., Di Iorio, A., Apollonio, L., Alshalak, M., Gravina, A., Antognoli, L., Baldi, M., Scalise, L., and Alchalabi, B. Auditory evoked potential eeg-biometric dataset, 2022.

Andrzejak, R. G., Lehnertz, K., Rieke, C., Mormann, F., David, P., and Elger, C. E. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state [dataset]. *Physical Review E*, 2023. doi: 10.34810/data490. URL https://doi.org/10.34810/data490.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Bajaj, N., Carrión, J. R., and Bellotti, F. Phyaat: Physiology of auditory attention to speech dataset. *arXiv preprint arXiv:2005.11577*, 2020.

Beh, W.-K., Wu, Y.-H., and Wu, A.-Y. A. Maus: A dataset for mental workload assessment on n-back task using wearable sensor, 2021. URL https://dx.doi.org/10.21227/q4td-yd35.

Bhaskaran, A., J, S. K., George, S., and Arora, M. Heart rate estimation and validation algorithm for fetal phonocardiography. *Physiological Measurement*, 43(7):075008, jul 2022. doi: 10.1088/1361-6579/ac7a8c. URL https://dx.doi.org/10.1088/1361-6579/ac7a8c.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Bousseljot, R., Kreiseler, D., and Schnabel, A. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. In *PTB-XL, a large publicly available electrocardiography dataset*, 2009. URL https://api.semanticscholar.org/CorpusID:111121953.

Brigham, E. O. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988.

Burke, M. and Nasor, M. Wavelet based analysis and characterization of the ecg signal. *Journal of Medical Engineering & Technology*, 28(2):47–55, 2004.

Carmona, C. U., Aubet, F.-X., Flunkert, V., and Gasthaus, J. Neural contextual anomaly detection for time series, 2021. URL https://arxiv.org/abs/2107.07702.

Chaudhury, S., Yu, C., Liu, R., Kumar, K., Hornby, S., Duplessis, C., Sklar, J. M., Epstein, J. E., and Reifman, J. Wearables detect malaria early in a controlled human-infection study. *IEEE Transactions on Biomedical Engineering*, 69(6):2119–2129, 2021.

Chen, C.-F., Fan, Q., and Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

Chowdhury, M. H., Shuzan, M. N. I., Chowdhury, M. E., Mahbub, Z. B., Uddin, M. M., Khandakar, A., and Reaz, M. B. I. Estimating blood pressure from the photoplethysmogram signal and demographic features using machine learning techniques. *Sensors*, 20(11):3127, 2020.

Chun, S. Y., Kang, J.-H., Kim, H., Lee, C., Oakley, I., and Kim, S.-P. Ecg based user authentication for wearable devices using short time fourier transform. In *2016 39th international conference on telecommunications and signal processing (tsp)*, pp. 656–659. IEEE, 2016.

Dar, M. N., Rahim, A., Akram, M. U., Khawaja, S. G., and Rahim, A. Yaad: young adult's affective data using wearable ecg and gsr sensors. In *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pp. 1–7. IEEE, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Esmaili, A., Kachuee, M., and Shabany, M. Nonlinear cuffless blood pressure estimation of healthy subjects using pulse transit time and arrival time. *IEEE Transactions on Instrumentation and Measurement*, 66(12):3299–3308, 2017.

Foumani, N. M., Tan, C. W., Webb, G. I., and Salehi, M. Improving position encoding of transformers for multivariate time series classification. *Data Mining and Knowledge Discovery*, 38(1):22–48, 2024.

Freepik. Hypertension; blood pressure gauge; motion sensor; student sleeping in class; diabetes; blood cells; edge computing; galvanic skin response; motion sensor; accelerometer sensor; eeg, n.d. URL prefix: https://www.flaticon.com/free-icon/ , 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.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Hassani, T. Federated emotion recognition with physiological signals-gsr, 2021.

He, J., Zhang, Q., Wang, L., and Pei, L. Weakly supervised human activity recognition from wearable sensors by recurrent attention learning. *IEEE Sensors Journal*, 19 (6):2287–2297, 2018.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Hosni, A. and Atef, M. Remote real-time heart rate monitoring with recursive motion artifact removal using ppg signals from a smartphone camera. *Multimedia Tools and Applications*, 82(13):20571–20588, 2023.

Hu, K., Ivanov, P. C., Chen, Z., Carpena, P., and Stanley, H. E. Effect of trends on detrended fluctuation analysis. *Physical Review E*, 64(1):011114, 2001.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., and Feichtenhofer, C. Masked autoencoders that listen, 2023. URL https://arxiv.org/abs/2207.06405.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax, 2017.

Jolliffe, I. T. and Cadima, J. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Kachuee, M., Kiani, M. M., Mohammadzade, H., and Shabany, M. Cuffless blood pressure estimation algorithms for continuous health-care monitoring. *IEEE Transactions on Biomedical Engineering*, 64(4):859–869, 2016.

Kim, D., Lee, J., Cho, M., and Kwak, S. Detector-free weakly supervised group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20083–20093, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Krishnan, P., Yaacob, S., Krishnan, A. P., Rizon, M., and Ang, C. K. Eeg based drowsiness detection using relative band power and short-time fourier transform. *J. Robotics Netw. Artif. Life*, 7(3):147–151, 2020.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Liang, Y., Chen, Z., Liu, G., and Elgendi, M. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific data*, 5(1):1–7, 2018. doi: 10.6084/m9.figshare.5459299.v5.

Ma, H., Zhang, Z., Li, W., and Lu, S. Unsupervised human activity representation learning with multi-task deep clustering. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–25, 2021.

Mathew, G., Barbosa, D., Prince, J., and Venkatraman, S. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. *npj Cardiovascular Health*, 1(1):25, Oct 2024. ISSN 2948-2836. doi: 10.1038/s44325-024-00027-5. URL https://doi.org/10.1038/s44325-024-00027-5.

Mikelsons, G., Smith, M., Mehrotra, A., and Musolesi, M. Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities. In *ML4H Workshop at 31st Conference on Neural Information Processing Systems (NIPS)*, 2017. URL https://arxiv.org/abs/1711.06350.

Min, J., Wang, P., and Hu, J. The original EEG data for driver fatigue detection. *figshare.Dataset.*, 7 2017. doi: 10.6084/m9.figshare.5202739.v1.

Muzammil, M. Finetuning endevsols/tinyllama-2.5t-clinical model on clinical dataset., 2021. URL https://huggingface.co/muzammil-eds/tinyllama-2.5T-Clinical-v2.

Narayanswamy, G., Liu, X., Ayush, K., Yang, Y., Xu, X., Liao, S., Garrison, J., Tailor, S., Sunshine, J., Liu, Y., Althoff, T., Narayanan, S., Kohli, P., Zhan, J., Malhotra, M., Patel, S., Abdel-Ghaffar, S., and McDuff, D. Scaling wearable foundation models, 2024. URL https://arxiv.org/abs/2410.13638.

Nedorubova, A., Kadyrova, A., and Khlyupin, A. Human activity recognition using continuous wavelet transform and convolutional neural networks. *arXiv preprint arXiv:2106.12666*, 2021a.

Nedorubova, A., Kadyrova, A., and Khlyupin, A. Human activity recognition using continuous wavelet transform and convolutional neural networks. *arXiv preprint arXiv:2106.12666*, 2021b.

Negi, P. C., Giri, H., Sharma, S., Sharma, N., et al. A comparative study of scalograms for human activity classification. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pp. 1–5. IEEE, 2024.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior, 2023.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z., and Wu, Q. J. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4051–4070, 2022.

Qian, B. and Rasheed, K. Hurst exponent and financial market predictability. In *IASTED conference on Financial Engineering and Applications*, pp. 203–209. Proceedings of the IASTED International Conference Cambridge, MA, 2004.

Qian, H., Pan, S. J., and Miao, C. Weakly-supervised sensor-based activity segmentation and recognition via learning from distributions. *Artificial Intelligence*, 292:103429, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Reiss Attila, Indlekofer Ina, S. P. PPG-DaLiA. UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C53890.

Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., and Parra, X. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C54S4K.

Roelofs, K. Freeze for action: neurobiological mechanisms in animal and human freezing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372 (1718):20160206, 2017.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 400–408, 2018.

Semenoglou, A.-A., Spiliotis, E., and Assimakopoulos, V. Image-based time series forecasting: A deep convolutional neural network approach. *Neural Networks*, 157:39–53, 2023. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2022.10.006. URL https://www.sciencedirect.com/science/article/pii/S0893608022003902.

Sengupta, R., Polian, I., and Hayes, J. P. Wavelet transform assisted neural networks for human activity recognition. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1254–1258. IEEE, 2022.

Slapničar, G., Mlakar, N., and Luštrek, M. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors*, 19(15):3420, 2019.

Thompson, J. M. T., Stewart, H. B., and Turner, R. Nonlinear dynamics and chaos. *Computers in Physics*, 4(5): 562–563, 1990.

Torrence, C. and Compo, G. P. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998.

Vaid, A., Jiang, J., Sawant, A., Lerakis, S., Argulian, E., Ahuja, Y., Lampert, J., Charney, A., Greenspan, H., Narula, J., Glicksberg, B., and Nadkarni, G. N. A foundational vision transformer improves diagnostic performance for electrocardiograms. *npj Digital Medicine*, 6(1):108, Jun 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00840-9. URL https://doi.org/10.1038/s41746-023-00840-9.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Vishnupriya, S. and Meenakshi, K. Automatic music genre classification using convolution neural network. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–4, 2018. doi: 10.1109/ICCCI.2018.8441340. URL https://ieeexplore.ieee.org/document/8441340.

Wimmer, C. and Rekabsaz, N. Leveraging vision-language models for granular market change prediction, 2023. URL https://arxiv.org/abs/2301.10166.

Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. Determining lyapunov exponents from a time series. *Physica D: nonlinear phenomena*, 16(3):285–317, 1985.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yan, Y., Huang, Y.-C., Zhao, J., Liu, Y.-S., Ma, L., Yang, J., Yan, X.-D., Xiong, J., and Wang, L. Topological nonlinear analysis of dynamical systems in wearable sensor-based human physical activity inference. *IEEE Transactions on Human-Machine Systems*, 53(4):792–801, 2023a. doi: 10.1109/THMS.2023.3275774.

Yan, Y., Huang, Y.-C., Zhao, J., Liu, Y.-S., Ma, L., Yang, J., Yan, X.-D., Xiong, J., and Wang, L. Topological nonlinear analysis of dynamical systems in wearable sensor-based human physical activity inference. *IEEE Transactions on Human-Machine Systems*, 53(4):792–801, 2023b. doi: 10.1109/THMS.2023.3275774.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization, 2017.

Zhang, W., Yang, L., Geng, S., and Hong, S. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

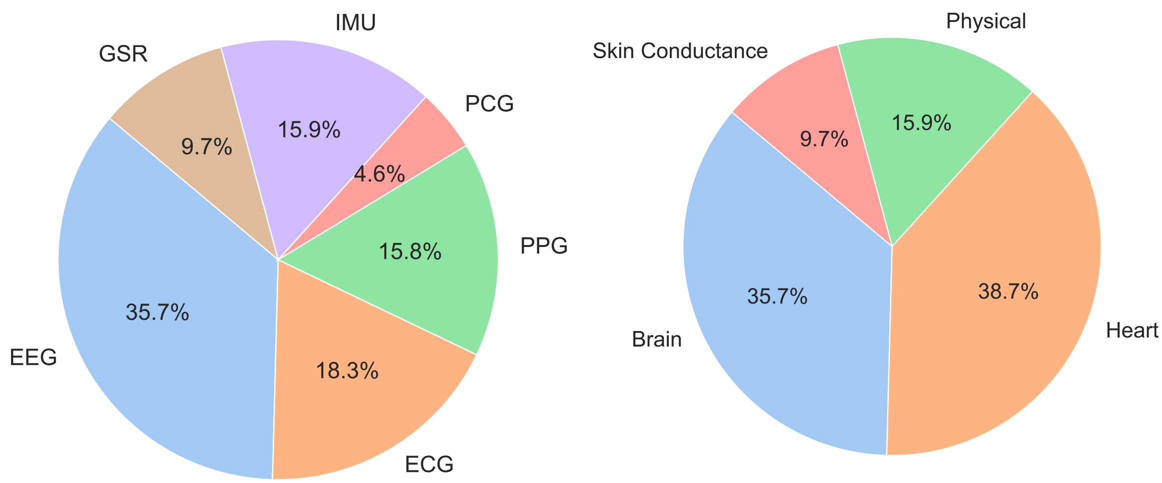
Zhang, X., Zhao, Z., Tsiligkaridis, T., and Zitnik, M. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022.

Zhang, X., Chowdhury, R. R., Gupta, R. K., and Shang, J. Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*, 2024a.

Zhang, X., Teng, D., Chowdhury, R. R., Li, S., Hong, D., Gupta, R. K., and Shang, J. Unimts: Unified pre-training for motion time series, 2024b. URL https://arxiv.org/abs/2410.19818.

A. Implementation Detail

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 4294 hours. The distribution of the pretraining dataset, as shown in Figure 5, 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.



(a) Distribution of sensor signals for pretrain. **(b)** Distribution of type of information for pretrain.

Figure 5. Pretrain data distribution.

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 (Reyes-Ortiz et al., 2012; Liang et al., 2018; Bousseljot et al., 2009). 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.

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 (Zhang et al., 2017; Carmona et al., 2021), 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. As a result, our training set is expanded to 2,586,404 segments, corresponding to 4,294 hours of data.

Pretraining Framework. Normwear is derived from the Masked Autoencoder (MAE) (He et al., 2021). The detailed hyper-parameter choice is descibe in 5. We use a Conv2D layer with a kernel size of (9, 5) and a stride of (9, 5), ensuring no

Algorithm 1 Time Series Mixup Augmentation

Input: Time series dataset $\mathcal{X}$, number of augmentations n
Output: Augmented Dataset $\tilde{\mathcal{X}}$

1: **for** $i = 1$ to n **do**
2: Sample two time series $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \sim \mathcal{X}$
3: Sample a chunk size $\lambda \sim \mathcal{U}(0, l)$
4: Sample start indices $s_1, s_2 \sim \mathcal{U}(0, l - \lambda)$
5: Swap chunk from $\mathbf{x}^{(2)}$ into $\mathbf{x}^{(1)}$:

$$\mathbf{x}^{(1)}_{s_1:s_1+\lambda} \leftarrow \mathbf{x}^{(2)}_{s_2:s_2+\lambda}$$

6: Append $\mathbf{x}^{(1)}$ into $\tilde{\mathcal{X}}$
7: **end for**
8: **return** $\tilde{\mathcal{X}}$

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, 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 six 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 Mean Squared Error (MSE) loss on all data points. Our models are pre-trained 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 Fig. 6)

MSiTF. For pretraining the representation alignment module, we have the training hyper-parameters in Table 6.

Sentence template example for signal-sext alignment.

- Emotion Task:
  - 'The emotion detected is {}.',
  - 'This subject is feeling {}.',
  - 'The emotional state is {}.',
  - 'The identified emotion is {}.'
- Activity Task:
  - 'This subject is currently {}.',
  - 'The subject is engaged in {}.',
  - 'Current activity is {}.',
  - 'Subjectś activity is {}.'

where {} is the placeholder for the corresponding label of each sample in pretraining datasets.

Statistical Feature list:

Features in *time domain*: mean, std, max, min, skew, kurtosis, 25% quantile, median, 75% quantile.

Features in *frequency domain*: centroid, spread, mean frequency, peak frequency, 25% quantile frequency, median frequency, 75% quantile frequency.

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,

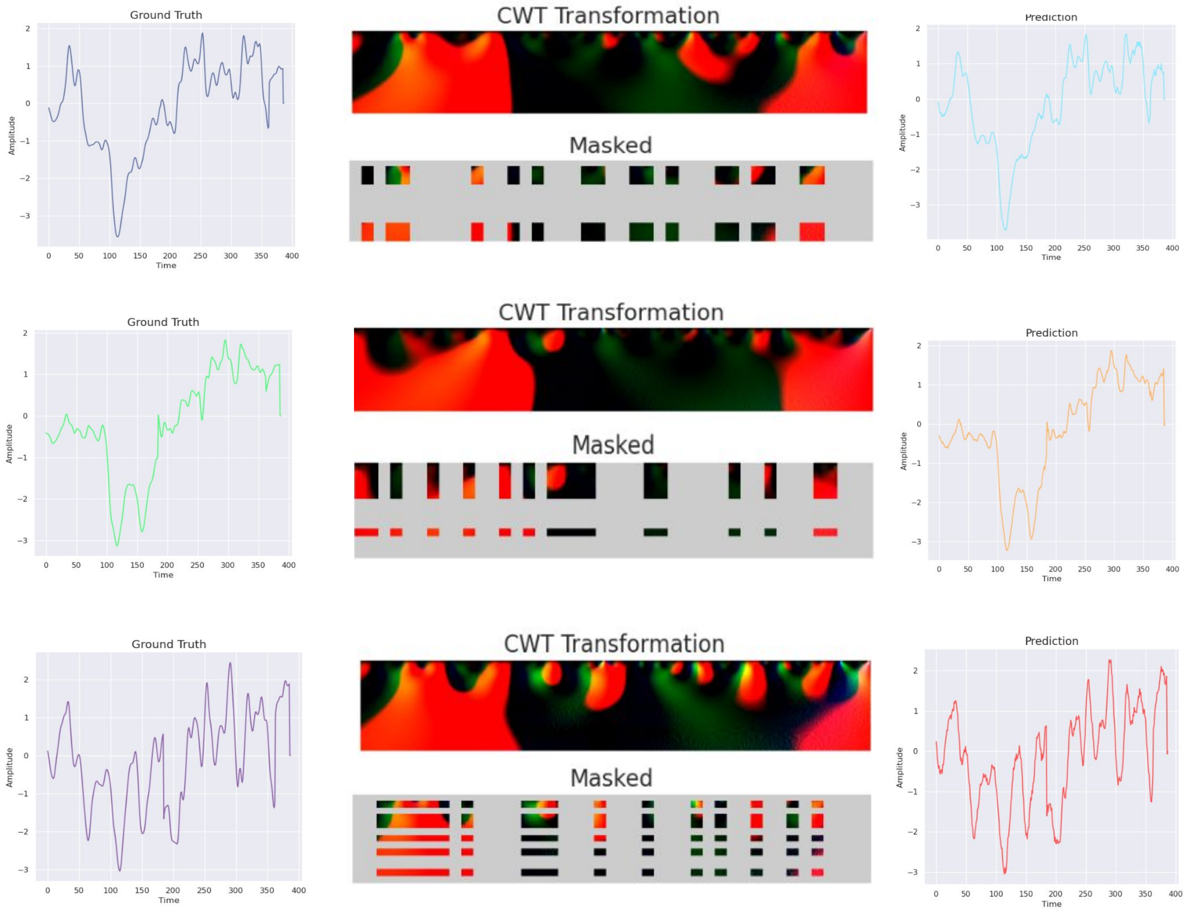


Figure 6. Visualization of original time series (left), CWT transformation image with structured masking (middle), and reconstructed time series (right).

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.

B. 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 7 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 prohibitible. Among all the compared strategies, the [CLS] token fusion generally achieves the best accuracy with a minor increase in parameters.

Masking Strategies in Pre-training. We ablated our masking strategy introduced in Section 3.3. 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 8).

Input Representations. Table9 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

Table 5. 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	3072
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 6. MSiFT 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

Table 7. 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**

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.

Table 8. 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**

Table 9. 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**

Table 10. Performance on various downstream wearable-signal-based health related applications under linear probing evaluation using 5 fold cross validation stratified by subject ID (if provided by the data source). In this table, The classification tasks are solved by Newton's method with conjugate gradient, and the AUC ROC are reported. The regression (noninvasive BP estimate) tasks are solved by Cholesky's method with closed form solution for ridge regression, and the relative accuracy (1 minus relative error) are reported. All the scores are the higher the better.

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	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**

From Table 11, we observe that demographic information and represenations extracted from wearable signals have their own strength on different tasks, and most of the time, when we concatenate them together, the overall performance will be better. The performance drop in some cases after concatenation, which indicate that there might be some confounding relationship between these two representations, hence further indicated that the information lies in demographic and the wearable represenation from NormWear are focused on different aspects. Same observation are observed with arbitrary model checkpoints during pretraining (denoted as Medium and Large marker representing different stage of training when we do the study on increasing the pretrain size.)

Table 11. Checking the reliance on demographic information.

Downstream Tasks	Simple Baseline Mode and Mean	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	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

Table 12. 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**

C. Statistical significance on the model comparison

We conduct statistical analysis to check the significance of the difference between models' performance. We first run the downstream evaluation 100 times for each model on all the tasks, without fixing random seed. We observed that the outcomes stay consistent due to the stability of the optimization process.

We then conduct a permutation test, across the results from these 100 runs, to assess whether our method significantly outperforms the baselines. We declare the alternative hypothesis as whether the score (AUC ROC) of NormWear is greater than the baselines in comparison. The reported P value represents the probability of observing a test statistic as extreme as, or more extreme than, the observed difference under the null hypothesis, assuming that the AUC ROC score of NormWear is not greater than the baseline. The results indicate that in nearly all cases, the statistical significance (p-value) is less than 0.01, providing statistical significance evidence of the robustness and superiority of our approach. In Table 7, we include the results from conducting the statistical test across different task groups (the groups were highlighted with different colors in the tables in main sections) and the total average scores.

We also present the critical difference diagram (CD) to visually compare the performances of multiple models across datasets, highlighting whether their performance differences are statistically significant. In order to achieve CD diagram, we first conduct Friedman Chi square test on the scores achieved by the models across all the downstream tasks, and observe P value of P ¡ .001, making sure all the models' performance are coming from different distribution. Then we conduct Conover post hoc test to check the pair-wise model performance difference, where the P values corresponding to NormWear vs. baselines are presented in the last row of Table 7. Finally, we create CD diagram based on these results, and result in the diagram shown in Figure 8. Our proposed model, NormWear is far apart from the bar, indicating its statistical significance against competitive baselines.

Ours/Baselines	Stats	Chronos	CLAP	TFC
NormWear - activity	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - eeg	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - risk	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - vital	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - micro avg.	$P < .01$	$P < .01$	$P < .01$	$P < .01$
NormWear - macro avg.	$P < .01$	$P < .01$	$P < .01$	$P < .01$
Conover post hoc	$P < .001$	$P < .001$	$P < .001$	$P < .05$

Figure 7. Permutation test on models' performance.

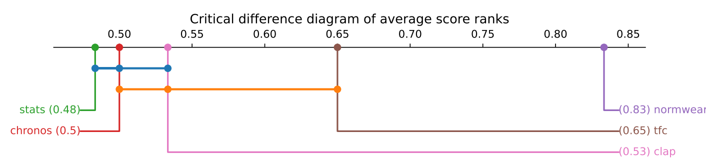


Figure 8. Critical Difference Diagram

D. 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 9, 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.

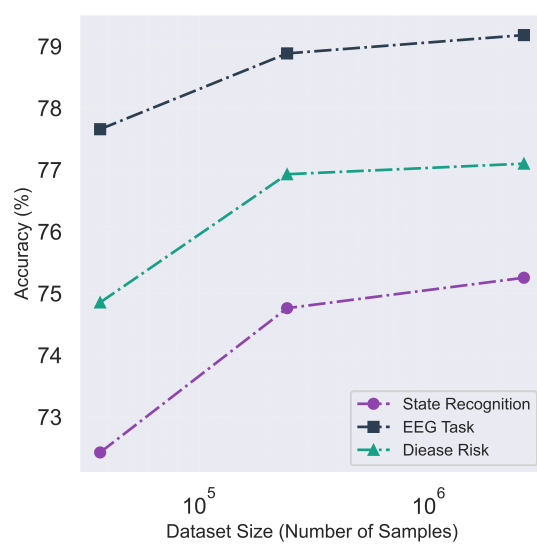


Figure 9. **Impact of scaling the pretraining dataset on downstream tasks.** The y-axis represents the average accuracy across tasks, while the x-axis denotes the size of the pretraining dataset in terms of the number of samples.

E. Deployment of NORMWEAR: testing on the edge

As shown in the table 13, the GPU setup on an NVIDIA RTX 3090 significantly outperforms other configurations in inference speed, achieving an inference time of only 0.18 seconds while maintaining low RAM usage (8.04 MB) and moderate VRAM requirements (732.82 MB). In contrast, the CPU setup on MacOS M1 requires 4.21 seconds, reflecting a considerably slower performance despite similar RAM usage (9.12 MB) and no VRAM consumption. On edge devices, such as the Jetson Nano 4GB, the CPU-based setup exhibits the slowest inference time of 40.69 seconds, while the GPU variant improves this to 34.87 seconds with a VRAM requirement of 504.46 MB. Storage requirements remain constant across all configurations at 1.63 GB.

Table 13. Computation resources consumed across various devices, on 6 channels data for 6 seconds.

Dataset/Task	Infer time	RAM	VRAM	Storage
CPU (MacOS, M1)	4.21 s	9.12 MB	-	1.63 GB
GPU - Debian GNU/Linux - NVIDIA-RTX-3090	0.18 s	8.04 MB	732.82 MB	1.63 GB
Edge (Jetson Nano 4GB, CPU)	40.69 s	9.12 MB	-	1.63 GB
Edge (Jetson Nano 4GB, GPU)	34.87 s	8.17 MB	504.46 MB	1.63 GB

F. Feature Visualization

F.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 (Van der Maaten & Hinton, 2008), with PCA (Jolliffe & Cadima, 2016) 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.

F.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 time step 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 (Thompson et al., 1990) to quantify the overall patterns of these extracted features, which are discussed in detail in section F.3.

F.3. 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. To quantify the meaningfulness of these representations, we conducted a nonlinear dynamics analysis inspired by chaos theory. This method analyzes the features' intrinsic behaviors through metrics like the Lyapunov exponent (Wolf et al., 1985) (sensitivity to initial conditions), Hurst exponent (Qian & Rasheed, 2004) (self-correlation/seasonality), and persistence entropy (Yan et al., 2023b) (unpredictability in system states). We obtain the following key observations:

1. Deeper Layers Capture Higher-Order Complexity.

- For signals such as GSR, EEG, and ACC, deeper layers show lower self-correlation (DFA (Hu et al., 2001)) 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.

2. 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.

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, supporting the conclusion that the model learns meaningful features tailored to human sensing tasks.

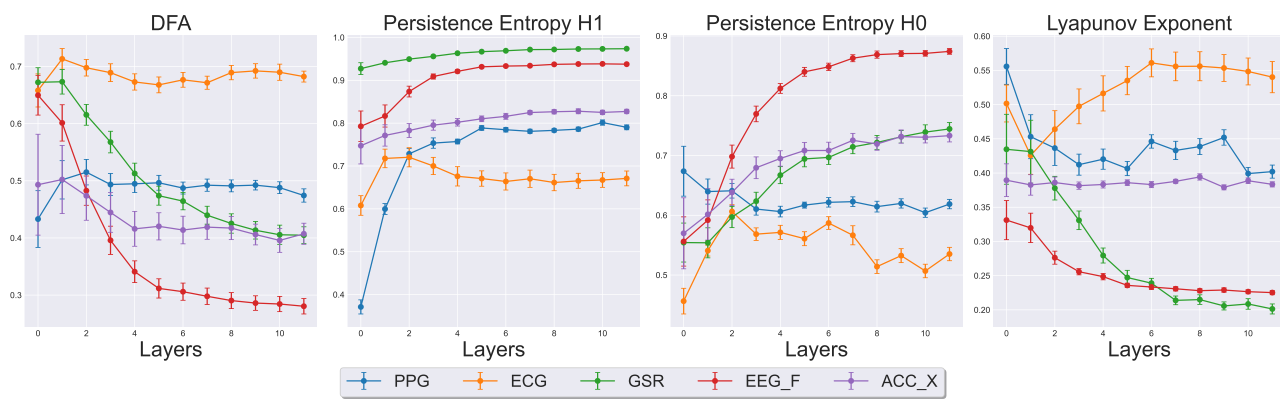
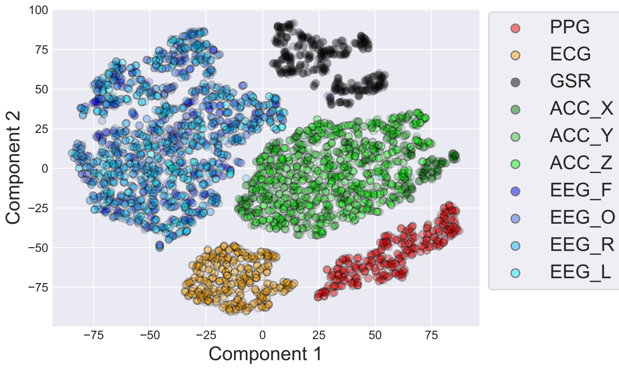


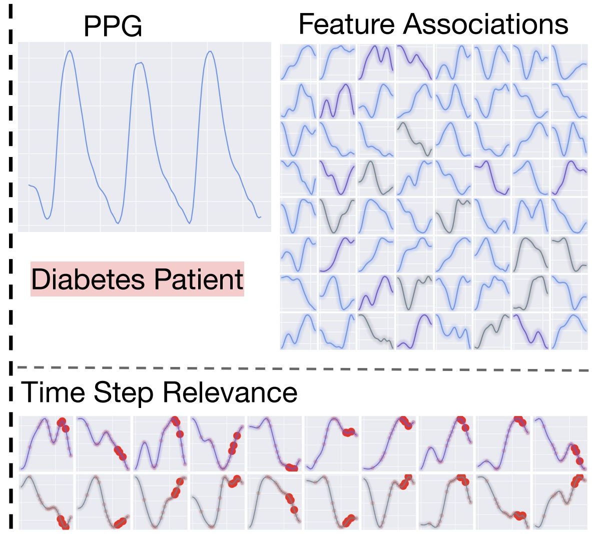
Figure 10. Nonlinear dynamic analysis on the waveforms extract at different layers of our model.

F.4. T-SNE plot among classes

In this section, 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 12, 13, 14, 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.



(a) T-SNE plot of the embeddings of the [CLS] special tokens, including signals of PPG, ECG, GSR, tri-axis accelerometer, and EEG at lobe of front, occiput, right, and left



(b) Visualization of features extracted by the intermediate layer

Figure 11. Feature visualization.

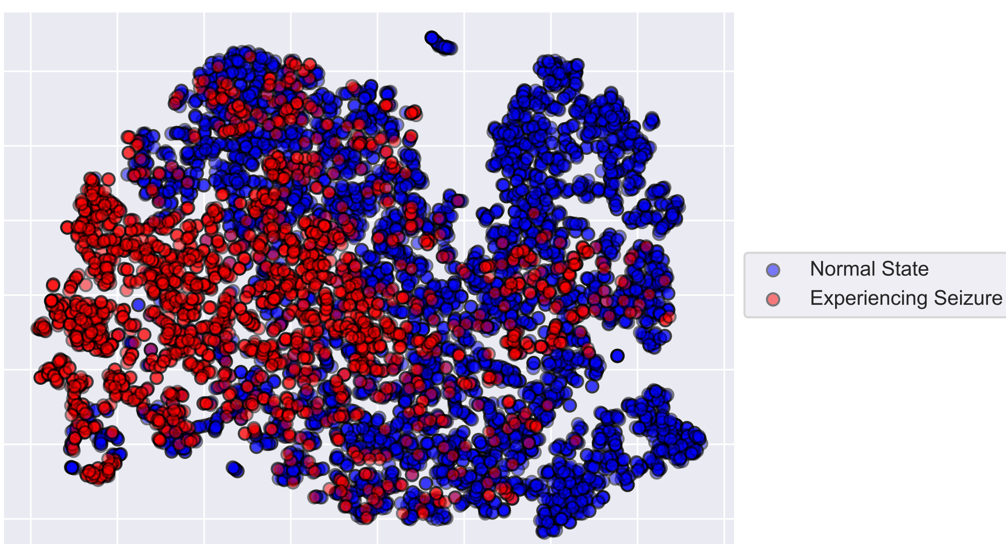


Figure 12. Visualization of embedding on EEG signals.

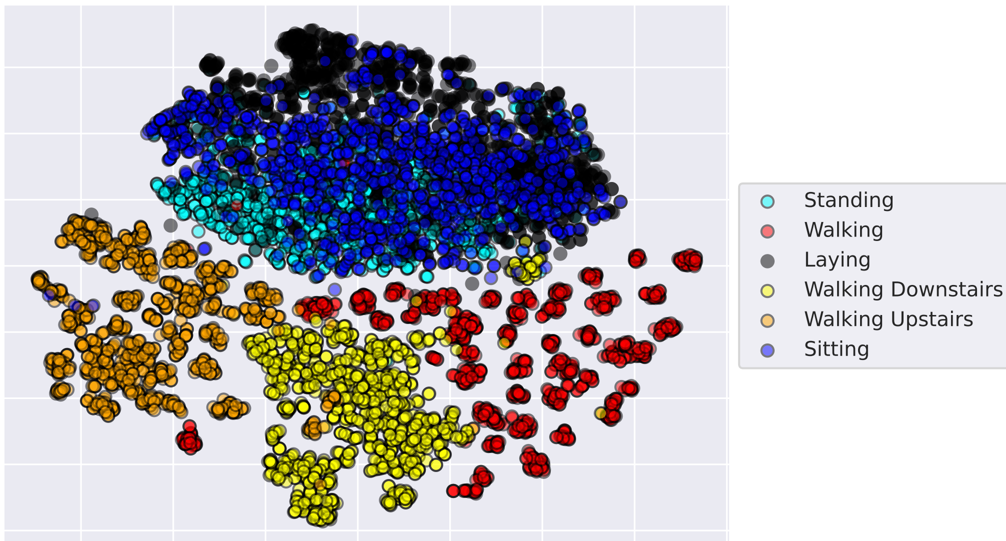


Figure 13. Visualization of embedding on signals from IMU sensors.

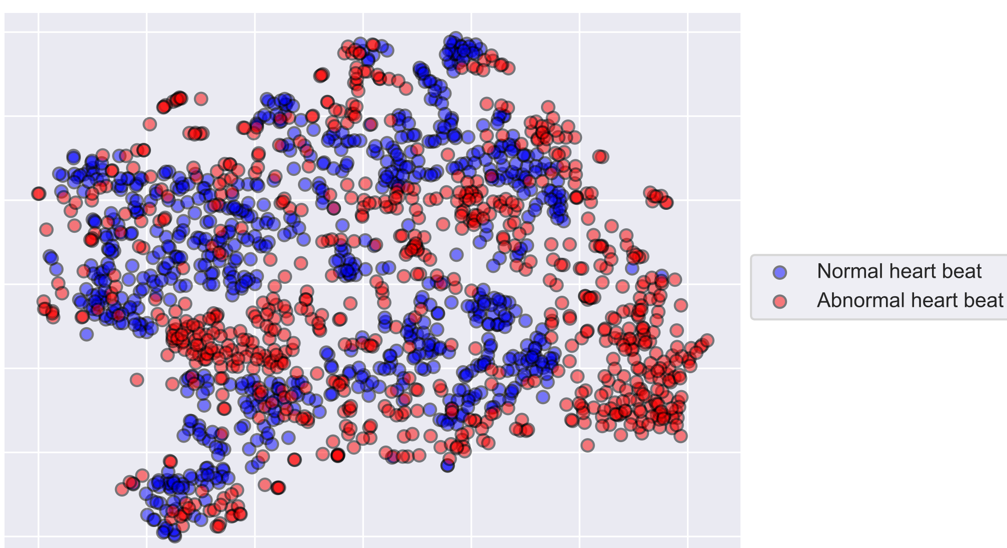


Figure 14. Visualization of embedding of ECG.

G. Reconstruction Example

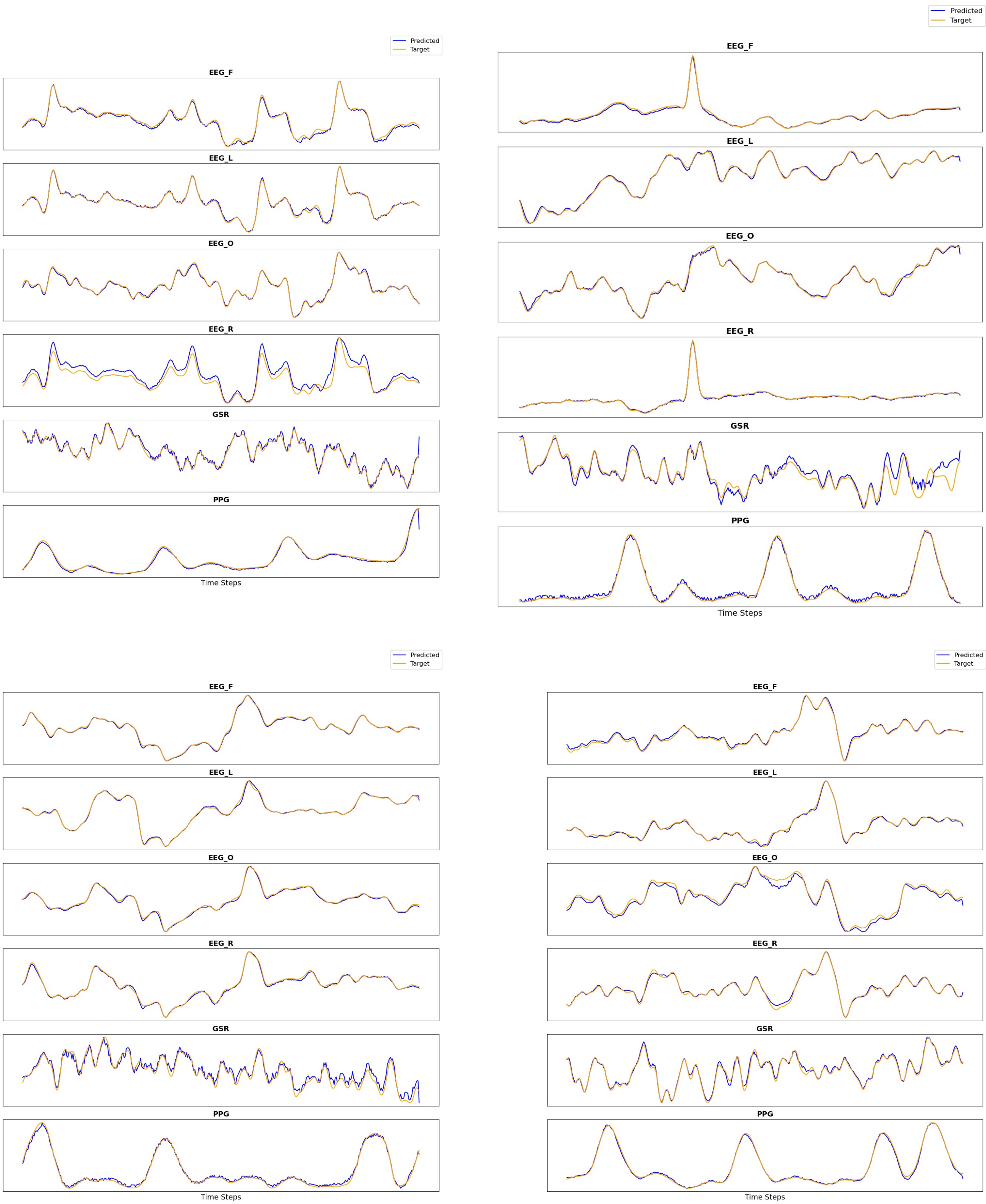


Figure 15. **Uncurated random samples** on Phyatt scalogram, using a NormWear trained in our training set. The masking ratio is 80%.

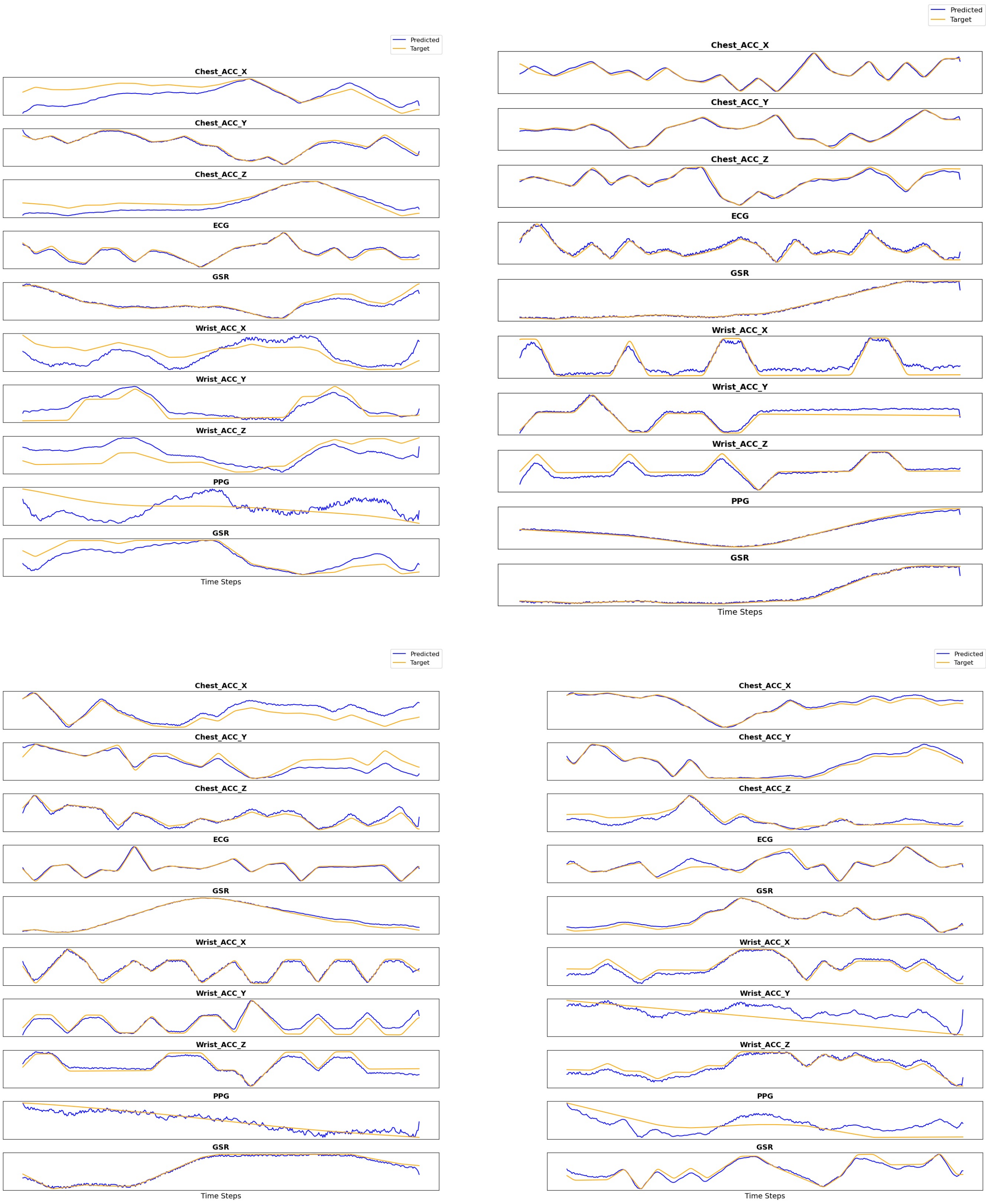


Figure 16. **Uncurated random samples** on WESAD scalogram, using a NormWear trained in our training set. The masking ratio is 80%. Note that the IMU data are not in the training set and, in general, NormWear is able to reconstruct this with high accuracy.