# Wearable data from students, teachers or subjects with alcohol use disorder help detect acute mood episodes via self-supervised learning

Filippo Corponi, Bryan Li, Gerard Anmella, Clàudia Valenzuela-Pascual, Ariadna Mas, Isaella Pacchiarotti, Marc Valentí, Iria Grande, Antonio Benabarre, Marina Garriga, Eduard Vieta, Allan Young, Stephen Lawrie, Heather Whalley, Diego Hidalgo-Mazzei, Antonio Vergari

Submitted to: JMIR mHealth and uHealth
on: December 02, 2023

# *Table of Contents*

# Wearable data from students, teachers or subjects with alcohol use disorder help detect acute mood episodes via self-supervised learning

Filippo Corponi[1] MD, MSc; Bryan Li[1] MSc; Gerard Anmella[2] MD, PhD; Clàudia Valenzuela-Pascual[2] MSc; Ariadna Mas[2] MSc; Isaella Pacchiarotti[2] MD, PhD; Marc Valentí[2] MD, PhD; Iria Grande[2] MD, PhD; Antonio Benabarre[2] MD, PhD; Marina Garriga[2] MD, PhD; Eduard Vieta[2] MD, PhD; Allan Young[3] MD, PhD; Stephen Lawrie[4] MD, PhD; Heather Whalley[4] PhD; Diego Hidalgo-Mazzei[2] PhD; Antonio Vergari[1] PhD

[1]University of Edinburgh Edinburgh GB
[2]University of Barcelona Barcelona ES
[3]King's College London London GB
[4]University of Edinburgh, Department of Psychiatry Edinburgh GB

**Corresponding Author:**
Filippo Corponi MD, MSc
University of Edinburgh
Informatics Forum, 10 Crichton St, Newington, Edinburgh
Edinburgh
GB



## *Abstract*

**Background:** Personal sensing, leveraging data passively and near-continuously collected with wearables from patients in their ecological environment, is a promising paradigm to monitor mood disorders (MDs), a major determinant of worldwide disease burden. However, collecting and annotating wearable data is very resource- intensive. Studies of this kind can thus typically afford to recruit only a couple dozens of patients. This constitutes one of the major obstacles to applying modern supervised machine learning techniques to MDs detection.

**Objective:** In this paper, we overcome this data bottleneck and advance the detection of MDs acute episode vs stable state from wearables data on the back of recent advances in self-supervised learning (SSL). This approach leverages unlabeled data to learn representations during pre-training, subsequently exploited for a supervised task.

**Methods:** We collected open-access datasets recording with an Empatica E4 spanning different, unrelated to MD monitoring, personal sensing tasks – from emotion recognition in Super Mario players to stress detection in undergraduates – and devised a pre-processing pipeline performing on-/off-body detection, sleep-wake detection, segmentation, and (optionally) feature extraction. With 161 E4-recorded subjects, we introduce E4SelfLearning, the largest to date open access collection, and its pre-processing pipeline1. We developed a novel E4-tailored Transformer architecture (E4mer), serving as blueprint for both SSL and fully supervised learning; we assessed whether and under which conditions self-supervised pretraining led to an improvement over two fully supervised baselines, i.e. the fully supervised E4mer and a classical baseline (XGBoost), in detecting acute mood episodes from recording segments taken in 64 (half acute, half stable) patients.

**Results:** SSL confidently outperforms fully-supervised pipelines using either our novel E4mer or XGBoost: 81.23% against 75.35% (E4mer) and 72.02% (XGBoost) correctly classified recording segments. SSL performance is strongly associated with the specific surrogate task employed for pre-training as well as with unlabeled data availability.

**Conclusions:** We showed that SSL, a paradigm where a model is pre-trained on unlabeled data with no need for human annotations prior to deployment on the supervised target task of interest, helps overcome the annotation bottleneck; the choice of the pre-training surrogate task and the size of unlabeled data for pre-training are key determinants of SSL success. We introduced an E4-tailor Transformer architecture (E4mer) that can be used for SSL and share the E4SelfLearning collection, along with its preprocessing pipeline, which can foster and expedite future research into SSL for personal sensing.

(JMIR Preprints 02/12/2023:55094)
DOI: https://doi.org/10.2196/preprints.55094

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

# Original Manuscript

# Wearable data from students, teachers or subjects with alcohol use disorder help detect acute mood episodes via self-supervised learning

**Filippo Corponi[1]✉, Bryan M. Li[1], Gerard Anmella[2], Clàudia Valenzuela-Pascual[2], Ariadna Mas[2], Isabella Pacchiarotti[2], Marc Valentí[2], Iria Grande[2], Antonio Benabarre[2], Marina Garriga [2], Eduard Vieta[2], Allan H Young[3], Stephen M. Lawrie[4], Heather C. Whalley[4], Diego Hidalgo-Mazzei[2*], Antonio Vergari[1*]**

[1]School of Informatics, University of Edinburgh

[2]Hospital Clínic de Barcelona, University of Barcelona

[3]Institute of Psychiatry, Psychology and Neuroscience, King's College London

[4]Division of Psychiatry, University of Edinburgh

* Shared supervision

✉ *Corresponding author:* Filippo Corponi, MD, MSc, School of informatics, University of Edinburgh, Informatics Forum, 10 Crichton St, Newington, Edinburgh EH8 9AB, UK. Email: filippo.corponi@ed.ac.uk



## Abstract

**Background** – Personal sensing, leveraging data passively and near-continuously collected with wearables from patients in their ecological environment, is a promising paradigm to monitor mood disorders (MDs), a major determinant of worldwide disease burden. However, collecting and annotating wearable data is very resource-intensive. Studies of this kind can thus typically afford to recruit only a couple dozen patients. This constitutes one of the major obstacles to applying modern supervised machine learning techniques to MDs detection.

**Objective** – In this paper, we overcome this data bottleneck and advance the detection of MDs acute episode vs stable state from wearables data on the back of recent advances in self-supervised learning (SSL). This approach leverages unlabeled data to learn representations during pre-training, subsequently exploited for a supervised task.

**Methods** – We collected open-access datasets recording with an Empatica E4 spanning different, *unrelated to MD monitoring*, personal sensing tasks – from emotion recognition in Super Mario players to stress detection in undergraduates – and devised a pre-processing pipeline performing on-/off-body detection, sleep-wake detection, segmentation, and (optionally) feature extraction. With 161 E4-recorded subjects, we introduce E4SelfLearning, the largest to date open access collection, and its pre-processing pipeline[1]. We developed a novel E4-tailored Transformer architecture (E4mer), serving as

the blueprint for both SSL and fully-supervised learning; we assessed whether and under which conditions self-supervised pretraining led to an improvement over fully-supervised baselines, i.e. the fully-supervised E4mer and pre-deep-learning algorithms, in detecting acute mood episodes from recording segments taken in 64 (half acute, half stable) patients.

**Results** – SSL confidently outperforms fully-supervised pipelines using either our novel E4mer or XGBoost: 81.23% against 75.35% (E4mer) and 72.02% (XGBoost) correctly classified recording segments. SSL performance is strongly associated with the specific surrogate task employed for pre-training as well as with unlabeled data availability.

**Conclusions** – We showed that SSL, a paradigm where a model is pre-trained on unlabeled data with no need for human annotations before deployment on the supervised target task of interest, helps overcome the annotation bottleneck; the choice of the pre-training surrogate task and the size of unlabeled data for pre-training are key determinants of SSL success. We introduced an E4-tailored Transformer architecture (E4mer) that can be used for SSL and share the E4SelfLearning collection, along with its preprocessing pipeline, which can foster and expedite future research into SSL for personal sensing.

**Keywords:** mood disorders; time-series classification; wearables; personal sensing; deep learning; self-supervised learning; transformer

[1]repository to be released upon acceptance for publication

# Introduction

Mood disorders (MDs) are a group of mental health conditions in the Diagnostic and Statistical Manual 5th edition (DSM-5)[1] classification system. They are chronic, recurrent disorders featuring disturbances in emotions, energy, and thought, standing out as a leading cause of worldwide disability[2,3] and suicidality[4]. Timely recognition of mood episodes is critical towards better outcomes[5]. However, this is challenging due to generally limited patient insight[6] compounded with the low availability of specialized care for MDs, with rising demand straining current capacity[7,8].

Personal sensing, involving the use of machine learning (ML) to harness data passively and near-continuously collected with wearable devices from patients in their ecological environment, has been attracting interest as a promising paradigm to address this gap[9]. Indeed, some of the core MD clinical features (e.g. disturbance in mood and energy levels) translate into changes in physiological parameters measurable with wearable devices[10–12]. A major barrier towards the development of clinical decision support systems featuring personal sensing has been the scarcity of labelled data, that is data with annotations by clinicians about the MD state (e.g. diagnosis, disease phase, symptoms severity). Collecting and annotating data for personal sensing in MDs is indeed an expensive and time-consuming enterprise; thus, studies typically use samples running into only few dozens of patients[13–20].

In this work, we take a different perspective and leverage *unlabeled* data collected with the Empatica E4 wristband[21], a popular research-grade device for personal sensing studies[22], as well as recent advancements in self-supervised learning (SSL) techniques that can learn meaningful

representations from such unlabeled data. Specifically, we take advantage of open-access datasets which record physiological data with the E4 across different settings but do not address MDs and therefore do not provide information about the mood state of the subjects involved. While each such dataset has only a limited number of subjects, our aggregated and preprocessed dataset E4SelfLearning can break the labelled data bottleneck for personal sensing in MDs (Figure 1).

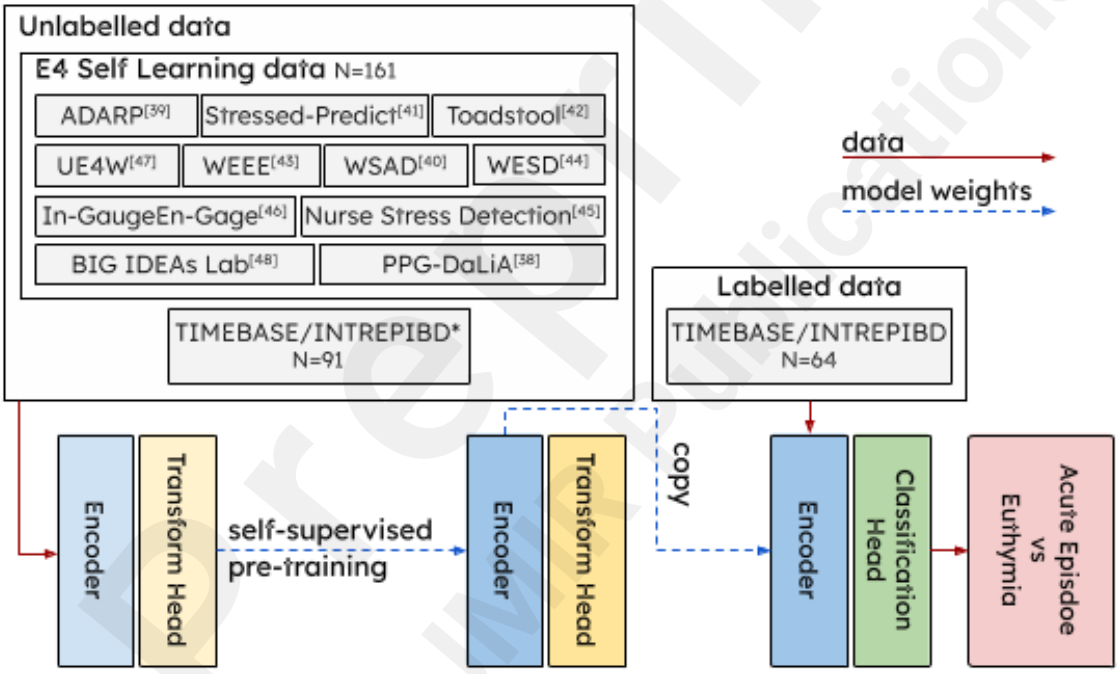



Figure 1: **A total of ~6254 hours (261 days) of unlabeled recordings from 252 subjects while awake were used for self-supervised pre-training**. Unlabeled data comprised a collection of eleven open-access datasets, whose pre-processed version we make publicly available (E4SelfLearning), along with part of the iden**T**ifying d**I**gital bio**M**arkers of illn**E**ss activity in **B**ipol**A**r di**S**ord**E**r/**I**dentifying digital biomarkers of ill**N**ess activity and **T**reatment **RE**s**P**onse **I**n **B**ipolar **D**isorder (TIMEBASE/INTREPIBD) study that was not relevant for the target task

under investigation, i.e. acute episode vs euthymia classification. Unlabeled data was passed through a model consisting of an encoder and a transform head for self-supervised pre-training; the pre-trained encoder block was then retained for the target task while the transform head was replaced with a new, randomly initialized classification head. N: subjects #; * the target task (labelled) training set from the TIMEBASE/INTREPIBD study was also used during self-supervised pre-training. Further details on the datasets used in the present study are available in Supplementary Table 1.

Fully-supervised systems require vast amounts of data to train, thus limiting their application in fields, such as healthcare, where amassing large, high-quality datasets is demanding in terms of time and human resources[23]. While previous studies in personal sensing for MDs investigated different tasks, including acute episode detection[13–16], regression of a psychometric scale total score[17–19], and more recently multi-task inference of all items in two commonly used psychometric scales[24], they all developed their models in a fully-supervised fashion, i.e. they were trained on samples for which ground-truth labels were available. As a result, considering that obtaining clinical annotations from patients, especially when on an acute MD episode, is a challenging and expensive enterprise, the sample size is generally modest (e.g. 52 in Côté-Allard et al.[15], 45 in[13], or 31 Pedrelli et al.[18]).

SSL, on the other hand, is a framework where the model creates proxy supervisory signals within the data itself, therefore alleviating the annotation bottleneck and allowing us to repurpose existing unlabeled

datasets[25]. Specifically, SSL derives supervisory signal from the data itself thanks to pretext tasks, that is new supervised challenges, for example imputing occluded parts of the input data. Through such preparatory pretext tasks, not requiring expert annotation, the model learns useful representations, partial solutions to the downstream target task of interest, for which only a comparatively small amount of annotated data is available[26]. On the back of the great success of SSL in Computer Vision (CV)[27] and Natural Language Processing (NLP)[28], and with encouraging findings in other healthcare applications[29], we extend pioneering SSL works on multi-variate time-series[30–32] to personal sensing in MDs.

In this work, we make the following contributions:

- We gathered eleven open-access datasets recording physiological data with an Empatica E4 device and developed a pipeline for pre-processing such data that does on-/off-body detection, sleep-wake detection, segmentation, and (optionally) feature-extraction. We make the pre-processing pipeline and the pre-processed data publicly available. This collection (E4SelfLearning), with 161 subjects, is the biggest open access to date. We believe that this effort can stimulate future research into SSL with multi-variate time-series sensory data by removing two barriers, pre-processing and data availability.
- We propose a novel Transformer[33] architecture (E4mer, Figure 2) and show that SSL is a viable paradigm, outperforming both the fully-supervised E4mer and classical machine learning (CML) models using handcrafted features in distinguishing MD acute episode from clinical stability (euthymia in psychiatric

parlance), i.e. a time-series (binary) classification task.

- We investigate what makes SSL successful. Specifically, we compare two main pretext task designs (i.e. masked prediction and transformation prediction)[26] and, for the best-performing routine, we study its sensitivity to the unlabeled data availability in ablation analyses. We inspect learned embeddings and show that they capture meaningful semantics about the underlying context, i.e. sleep-wake status, and symptom severity.

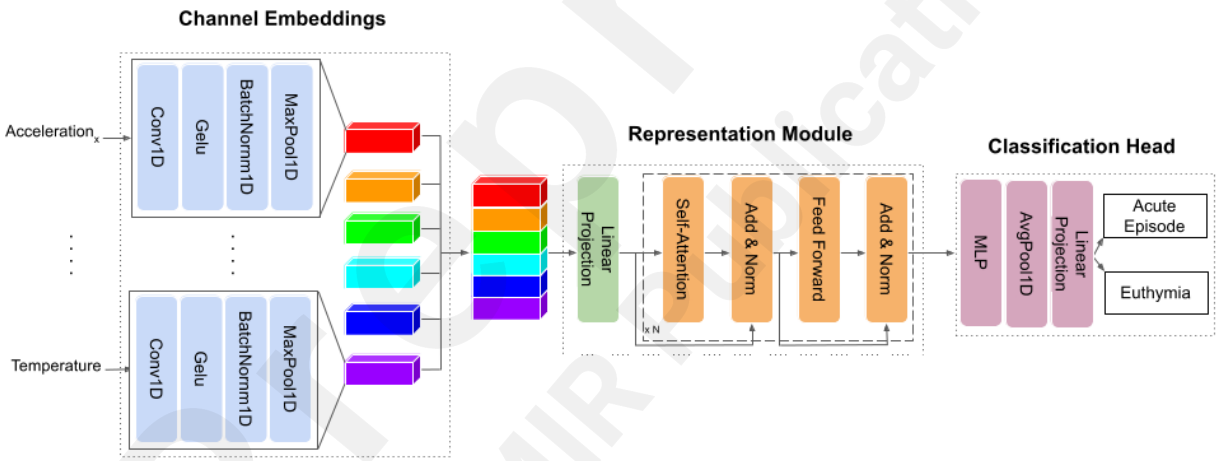



Figure 2: **E4mer is a Transformer model tailored to the Empatica E4 input data**. The E4mer is constituted of three sequential modules: 1) Channel embeddings set in parallel, one for each Empatica E4 raw input channel (i.e. acceleration$_x$, acceleration$_y$, acceleration$_z$, blood volume pressure, electrodermal activity, temperature), extracting features and mapping channels to tensors of dimensionality ($B$=batch size, $N$ = time steps, $F$ = filters #) so that they can be conveniently concatenated along the dimension $F$; 2) Representation Module learning contextual representations of the input time steps within the input segment thanks to the multi-head self-attention

mechanism; 3) Classification Head outputting probabilities for the two target classes, i.e. acute episode and euthymia. Self-supervised learning models employed in our experiments feature the same E4mer architecture described above, where, however, the Classification Head is replaced with a Transform Head projecting onto a label space compatible with the pretext task at hand.

# Methods

## Study sample

**TIMEBASE/INTREPIBD Cohort -** Our target task is distinguishing MD acute episodes from euthymia using wearable data. We start from a dataset for which we have labelled samples, the iden**T**ifying d**I**gital bio**M**arkers of illn**E**ss activity in **B**ipol**A**r di**S**ord**E**r/**I**dentifying digital biomarkers of ill**N**ess activity and **T**reatment **RE**s**P**onse **I**n **B**ipolar **D**isorder (TIMEBASE/INTREPIBD) cohort[34]. A detailed description of the data collection campaign is given in Anmella et al. [34]. In brief, this is a prospective, exploratory, observational study conducted at Hospital Clìnic, Barcelona, Spain. Patients with a DSM-5 diagnosis of either Major Depressive Disorder (MDD) or Bipolar Disorder (BD) were enrolled either on acute affective episodes (defined according to Structured Clinical Interview for DSM-5 Disorders criteria) or in euthymia (score ≤ 7 on the Hamilton Depression Rating Scale-17[35] and Young Mania Rating Scale[36] for at least 8 weeks[37] as confirmed with weekly ambulatory assessments). The former group had post-acute-phase follow-ups, which were however excluded from all analyses herewith presented. At the moment of

conducting this study, a total of 64 patients were available for the target task, half on an acute affective episode, and half in euthymia. Additionally, an extra 91 subjects (including healthy controls, subjects with schizophrenia, and subjects with a substance abuse disorder), whose status was not relevant to the target task, were available from the TIMEBASE/INTREPIBD cohort for self-supervised pretraining.

Patients were interviewed by a psychiatrist collecting clinical demographics (Table 1 & Supplementary Table (ST) 2) and were required to wear on their non-dominant wrist an E4 device until battery ran out ( ~ 48 hours). The E4 records (sampling rate between parentheses) 3D acceleration (32Hz), blood volume pressure (BVP, 64Hz), electrodermal activity (EDA, 4Hz), heart rate (HR, 1Hz), inter-beat intervals (IBI, i.e. the time between two consecutive heart ventricular contractions) and skin temperature (TEMP, 1Hz).

**E4SelfLearning** - For self-supervised pre-training, we gathered eleven open-access datasets recording with an E4 [38,38–40,40–48]. While they all used the same hardware, software, and firmware, such datasets could differ substantially for population, recording setting, and task: from students taking exams[44] or attending classes[46], to nurses carrying out their duty[45] and subjects performing different physical activities[43] or playing Super Mario[42]. Subjects that are not part of the target classes from the TIMEBASE/INTREPIBD study are also included in the unlabeled data for SSL.

| | Age | Females | Diagnosis | HDRS | YMRS |
|---|---|---|---|---|---|
| | Mean (std) | N (Percentage) | | Mean (std) | Mean (std) |
| Euthymia N=32 | 47.22 (16.06) | 14 (43.75%) | BD (N=26) | 2.93 (1.73) | 1.3 (1.61) |
| | | | MDD (N=6) | 3.14 (1.95) | 0.29 (0.76) |
| Acute Episode N=32 | 50.56 (13.05) | 15 (46.88%) | MDE-BD (N=9) | 20.22 (6.34) | 2.56 (3.94) |
| | | | MDE-MDD (N=7) | 25.14 (4.78) | 1.86 (2.41) |
| | | | ME (N=14) | 5.67 (4.37) | 20.13 (6.28) |
| | | | MX (N=2) | 16 (4.24) | 13.5 (4.95) |

Table 1: **Clinical-demographic features of target task (acute episode vs**

**euthymia classification) population.** Mood episodes clinically lie on a spectrum, with depression on one end and mania on the other; mixed episodes, featuring symptoms from both polarities, are a bridge between the two spectrum extremes. In this study we considered acute episodes of any polarity and, similarly, we considered euthymia as a unique class, whether in the context of a bipolar or major depression diagnosis. Medication classes administered in the cohort are given in Supplementary Table 2; Bonferroni-corrected chi-squared tests found no significant association between treatment status (being on a given drug class or not) and target class (acute episode vs euthymia). BD: Bipolar Disorder; HDRS: Hamilton Depression Rating Scale-17; MDD (Major Depressive Disorder); MDE-BD (Major Depressive Episode in Bipolar Disorder); MDE-MDD (Major Depressive Episode in Major Depressive Disorder); ME (Manic Episode); MX (Mixed Episode); YMRS: Young Mania Rating Scale.

## Data Pre-processing

Our pre-processing encompassed the following sequential stages: on-/off-body detection, sleep/wake detection, segmentation, and (when preparing data for CML models) features extraction.

During free-living wear, subjects might remove their device or contact with the wrist might be suboptimal. As a result, off-body periods can be erroneously mistaken for periods of sleep or sedentary behaviour, due to the shared feature of an absence of movement. Signal discontinuity in

biopotentials, such as EDA, due to a lack of skin contact can be reliably leveraged to detect non-wear periods. Similarly to[49,50] we considered measurements smaller than 0.05 $\mu$S as indicative of off-body status. Furthermore, as we noticed occurrences of values greater than the EDA sensor range (i.e. 100 $\mu$S[51]), as well as instances of TEMP values outside the physiological range (30-40 $C^{\circ}$), we set both to off-body.

As physiological data vary wildly across sleep and wake, we used sleep-wake detection as a form of data-cleaning to reduce the variance in the signal and considered wake time only in our analyses, especially as most publicly available datasets recorded in wake conditions. We opted for the algorithm by Van Hees et al.[52] (*Van Hees*) which was reported as the best performing in a recent benchmark study on sleep-wake detection (average F1-score 79.1)[53]. Like most non-proprietary algorithms, *Van Hees* uses triaxial acceleration and, specifically, relies on a simple heuristic defining sleep with the absence of change in arm angle greater than 5 degrees for 5 minutes or more. To accommodate this rule, wherever on-body sampling cycles did not constitute unbroken sequences of at least 5 minutes duration, all the measurements in that period were considered as off-body and discarded from further analysis.

Wake time from each recording was then segmented with a sliding window, whose segment length ($\omega$) and step-size ($\Delta\omega$) (in seconds) we set to 512 and 128, respectively. This approach, also referred to as window slicing[54], is a common form of data augmentation in time-series classification as multiple segments are produced from a single recording, each one marked with the same label, and is common in personal sensing for MDs. Previous relevant works[15,18,55] defined $\omega$ ($\Delta\omega$) based on clinical intuition and convenience concerning the available data. Another work[24]

investigating Hamilton Depression Rating Scale-17 and Young Mania Rating Scale items regression found the optimal $\omega$ through tuning, a very computationally expensive approach in our setting; however, it showed that $\omega$ was not among the most important hyperparameters for the task at hand. Here we opted for 512 (~8.5 minutes, conveniently a power of 2 for computational efficiency in binary computers), similar to the 5-minute intervals used in[55] for training neural autoencoder architectures on anomaly detection by reconstruction error estimation. Our choice was a trade-off between clinical insight and technical constraints. Clinical intuition suggests that too small a value of $\omega$ may be ill-suited to capture enough information towards acute episode vs euthymia discrimination On the other hand, unlabelled datasets used for self-supervised pretraining recorded relatively short sessions (e.g. 1 hour in [41]). As both CML and deep-learning models are trained on individual segments and too long a segment length equates to fewer training data points, a 512-second-long segment allowed us to have enough data for developing ML models[55].

Recording segments constituted our basic unit of analysis and, for the target task, segments from the same recording all shared the same ground truth label (i.e. either acute episode or euthymia). When fed to deep-learning models, segments were channel-wise standardized by subtracting the mean and dividing by the standard deviation. Such statistics were learned from the target task training set or, in the case of SSL, its aggregation with unlabeled data. Acceleration, BVP, EDA, and TEMP were considered in deep-learning models while HR and IBI, as features derived from BVP through a proprietary algorithm, were excluded from the deep-learning experiments herewith shown (see Supplementary Material). On the other hand, when using CML, handcrafted features were

extracted from segments using *FLIRT*[56], a popular open-access feature extraction toolkit for Empatica E4. Note that a single row of features per segment was extracted, in other words, the window size parameter in *FLIRT* was set equal to $\omega$. We used all features available through this package, derived with the functions *flirt.acc.get_acc_features* (e.g. acceleration entropy), *flirt.eda.get_eda_features* (e.g. tonic and phasic EDA components), and *flirt.hrv.get_hrv_features* (e.g. heart rate and heart rate variability measures). As FLIRT does provide built-in functions for TEMP, we also extracted the segment average and standard deviation for this channel. Any missing value was handled with mean imputation. The percentage rate of missing had a range [0, 37.31] with a mean of 10.44.

## Data splits & Metrics

In SSL experiments, we split unlabeled data with a ratio of 85/15 into train and validation set, partitioning recordings across the two sets. As for the target task, we investigated a *time-split* scenario therefore splitting each recording into train/val/test again with a ratio of 70/15/15 along recording time, thus testing generalization across future time points. We made sure that segments with overlapping motifs at the border between target task splits (resulting from using a sliding window with $\Delta\omega < \omega$) were confined to one split only, thus ultimately producing segments# 18896/3904/4128 for train/val/test. The target task validation set doubled as a test set for estimating generalization performance on the SSL pretext task. *Time-split* scenario is common in personal sensing for MDs (e.g. [18,24]) and indeed, despite efforts towards learning subject-invariant representations[57,58], cross-subjects generalization remains an unsolved challenge so personal sensing systems typically require access to each subject's physiological

data distribution at training time[59].

The target task is time-series binary classification. As expected in free-living wear, total wear time as well as off-body and wake time vary across subjects (and, as a result, so does the number of segments). Two-tailed t-tests were used to verify significant mean differences in off-body and wake time across individuals from the two target classes (acute episode and euthymia) but yielded a Bonferroni-corrected $p > 0.05$ ($p=0.5580$ for off-body and for $p=0.8163$ for wake time). An equal number of segments from each class was extracted for the target task. To that end, we found the pairing of euthymia and acute episode recordings that minimized the pairwise difference between the number of segments available per participant; then, within each pair, the first *n* segments were retained, where *n* is the segments number of the shortest recording in the pair. We optimized models on the target task for segment-level accuracy ( $\mathrm{ACC}_{\mathrm{segment}}$ ). Secondarily, in order to provide a subject-level perspective, we reported the subject accuracy:

$$\mathrm{ACC}_{\mathrm{subject}} = \frac{1}{S}\sum_{s=1}^{S} \square 1(\hat{y}_s = y_s)$$

where $y_s$ is the ground truth mood state of the $s^{th}$ subject, which is constant across all $s^{th}$ subject's recording segments, and $\hat{y}_s$ is a majority vote on the $s^{th}$ subject, corresponding to the majority predicted class across the $s^{th}$ subject's recording segments.

## Machine learning models

We developed two types of baselines for the target task: 1) an E4-tailored deep-learning pipeline inputting raw recording segments (E4mer) and 2) CML models using handcrafted features extracted with FLIRT from

recording segments. We then assessed what boost in performance, if any, a self-supervised pre-training phase might deliver, where the SSL models share the same building blocks as the E4mer.

## Baseline Models

**E4mer** - This is an artificial neural network discriminative classifier modelling the probability of an MD acute episode given a recording segment. As shown in Figure 2, our E4mer has three sequential blocks: 1) channel embeddings (CE) set in parallel, consisting of 1D same Convolutions with a kernel size equal to the channel sampling frequency, followed by Gelu activation, 1D BatchNorm, and 1D MaxPooling using the channel sampling frequency as both kernel size and step size, so that each channel embedding output had the same dimensionality and could be conveniently concatenated with the others before being passed onto 2) a Transformer[33] representation module ($RM$), and 3) a Multi-Layer Perceptron (MLP) classification head ($H_{sl}$). The CE extracts features from the input E4 channels and are designed to handle channels sampled at different frequencies, the $RM$, powered by multi-head self-attention, learns contextual representations of the input tokens (timestamps in our case) within a recording segment, the $H_{sl}$, lastly, maps such representations onto a label space appropriate for a binary classification. The E4mer is trained to minimize the binary cross-entropy (BCE) loss between acute episode/euthymia predictions and the corresponding ground truth.

**Classical Machine Learning** – We experimented with the following algorithms, given their popularity and state-of-the-art performance in biomedical applications[60], including personal sensing[13,14]: Elastic Net

Logistic Regression (ENET), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost).

## Self-supervised learning schemes

SSL schemes rely on devising a pre-text task, for which a (relatively) large amount of unlabeled data is available, conducive to learning, during a pre-training phase, representations useful to solve the downstream target task[26]. What defines an SSL paradigm is thus its pretext task, consisting of a process, $P$, to generate pseudo-labels and an objective to guide the pre-training. An SSL model typically consists of (i) an encoder $EN(x;\theta): X \to V$, learning a mapping from input views $x \in X$ to a representation vector $v \in R^d$ and (ii) a transform head $H_{ssl}(v;\xi): V \to Z$ projecting the feature embedding into a label space $z \in R^{d'}$ compatible with the pretext task at hand. When solving the target task, the pre-trained encoder $EN$ is retained as a partial solution to the target problem, whereas the pre-trained transform head $H_{ssl}$ is discarded and replaced with a new one $H_{sl}$. Then, $EN$'s parameters $\theta$ may be kept fixed and only $H_{sl}$'s parameters may be learned on the target task. This approach, often referred to as *linear readout* amounts to treating $EN$ as a frozen feature extractor. Alternatively, instead of just training a new head, the entire network may be retrained on the target task, initializing $EN$'s parameters $\theta$ to the values learned during self-supervised pre-training, a paradigm known as *fine-tuning*. Our SSL models used the same architecture as the E4mer, that is an encoder $EN$, consisting of convolutional $CEs$ followed by a Transformer $RM$, and an MLP for the transform head $H_{ssl}$. The success of SSL methods largely comes from designing appropriate pretext tasks that will produce

representations useful for the downstream target task. This usually involves domain knowledge of the target task. We herewith investigated how different pretext tasks affected downstream performance, experimenting with two popular SSL routines that showed success in other applications: masked prediction and transformation prediction.

**Masked predictions** This family of SSL methods is characterized by training the model to impute data which was removed or corrupted by $P$. It relies on the assumption that context can be used to infer some types of missing information in the data if the domain is well-modelled. This strategy was popularized by the huge success of BERT [28] in NLP applications, and one of the first adaptations to multi-variate time-series classification was proposed by Zerveas et al.[31]. Similarly to their implementation, for each segment channel, we sampled a Boolean mask where the sequences of 0's and 1's were sampled from geometric distributions with means respectively of $l_0$ and $l_1$, with:

$$l_1 = \frac{1-r}{r} l_0$$

$r$ being the masking ratio. As in [31], the average length of the 0's sequences ($l_m$) and the proportion of masked values ($r$) were set to 3 (seconds) and 0.15 respectively. Each segment channel was then multiplied by its corresponding mask, effectively setting to 0 some of the channel recorded measurements, and inputted to a model which was tasked to recover the original channel values. This was done by minimizing the Root Mean Squared Error (RMSE) between the masked original value $x(t,c)$ and its reconstruction outputted by the network $\hat{x}(t,c)$ :

$$L_{RMSE} = \sqrt{\frac{1}{¿M \vee ¿ \sum_{t \in M} ❑ \sum_{c \in M} ❑ ¿¿}} ¿$$

where $c$ and $t$ respectively index the channels and the timestamps of the 0's values in the masks $M$ and $¿M \vee ¿$ is the total number of 0's sampled, i.e. the masks' cardinality.

**Transformation prediction** We followed the implementation by Wu et al. [32] which used SSL for a target task of emotion recognition with E4 recordings. In brief, for each channel one of six transformations (i.e. identity, Gaussian noise addition, magnitude-warping, permutation, time-warping, and cropping) was sampled uniformly at random and then applied. The transformed segment was then inputted into a model, which was tasked to guess, for each channel, which one of the six transformations was applied. This amounted to a multitask multi-class classification where the model was trained to minimize channel average categorical cross-entropy (CCE):

$$L_{CCE_{Multitask}} = \frac{1}{C} \sum_{c=1}^{C} ❑ \sum_{j=1}^{T} ❑ -1_{c,j} \cdot \log\left(p_{c,j}\right)$$

where $c$ indexes the channels, $j$ the transformations, $1_{i,j}$ is an indicator taking value 1 when $j$ is correct transformation for the channel $c$, 0 otherwise, and $p_{c,j}$ denotes the predicted probability that transformation $j$ was applied to channel $c$. By solving this task, the authors in[32] argue that the model learns representations robust to disturbances in the magnitude and time domains.

## Tuning

Hyperparameters search for all models was carried out with Hyperband Bayesian optimization [61]. For the target task, we selected the setting yielding the highest $ACC_{segment}$ in the validation set, whereas in self-supervised pre-training we selected hyperparameters associated with the lowest relevant loss in the validation pre-training set. Appendix A.1 shows the hyperparameters search space and the best configuration across all models. Deep-learning models were trained with AdamW optimizer for a maximum of 300 epochs, with a batch size of 256. Moreover, to speed up the training and search procedure, we employed an early stopping learning rate scheduler: we reduce the learning rate $\alpha_{LR}$ by a factor of 0.3 if the model has not improved in its validation performance after 10 consecutive epochs; we terminate the training procedure if the model has not improved after 2 learning rate reductions. Dropout[62] and weight decay were added to prevent overfitting.

## Post-hoc analyses

Towards elucidating key contributors to the viability of SSL, besides comparing different pretext task designs, we studied how a) progressively down-sampling unlabeled datasets or b) removing each dataset in turn from the unlabeled collection might impact the performance of our best SSL model. Thus, using the most performative self-supervised scheme, we re-trained from scratch the SSL model under configurations a) and b), and then tested it on the target task. Note that in both settings, the entire target

task training set was kept for pre-training; this is because pre-training on the training set can be always done at no extra cost in terms of data acquisition.

Lastly, we conducted statistical tests to better appreciate how the self-supervised E4mer compared against its fully-supervised counterpart and the best-performing CML algorithm and how it was affected by different ablations. Based on whether we consider either i) recording segments or ii) subjects as our basic analysis units, we have two different hypotheses. In i), we employed a linear mixed-effects model (LME) to analyze the difference in correct class probabilities between the SSL model and each comparator, considering subjects as a random effect. This accounts for the nested structure of the data, where segments are sampled from individual subjects. A fixed-effects intercept was included to test a zero-mean difference between the classifiers at the population level. Additionally, as the ML models we implemented, like most state-of-the-art algorithms[63], effectively treat segments as independent and identically distributed, we used a two-tailed paired t-test to assess a zero-mean difference in the probability assigned to the correct class is zero. In ii), we checked with a two-tailed paired t-test if the between-classifiers mean difference in the $\mathrm{ACC}_{\mathrm{segment}}$ by subject is different from zero. To account for multiple testing, within both i) and ii), a Bonferroni correction was applied. The number of tests was 19, that is 17 different ablation settings plus 2 tests comparing the best baselines (fully-supervised E4mer and the best CML) to SSL.

# Results

## Surrogate tasks used in self-supervised pre-training

The same model, using the E4mer architecture (Figure 2), was used across different pretext tasks. Figure 3 illustrates the surrogate tasks we experimented with. In masked prediction (Figure 3a) parts of the input segments are zeroed out by multiplication with a Boolean mask sampled as in Zerveas et al.[31] and the model is trained to recover the original input segments. While the model outputs entire segments, only the masked values are taken into account towards the loss computation, that is Root Mean Squared Error (RMSE). The assumption is that the model acquires good representations of the underlying structure of the data when learning to solve this task. Our best model had an error of 0.1347 on the test set (notice that input segments were channel-wise standardized).

With transformation prediction (Figure 3b) one transformation is sampled from a set and applied to each channel independently and the model learns which transformation each channel underwent, minimizing the channel average categorical cross-entropy (CCE). We used the same transformations as Wu et al.[32], who experimented with an E4 for a downstream task of emotion recognition. The rationale is to encourage robustness against signal disturbances introduced with the transformations. The test loss of the selected model was 0.5000.

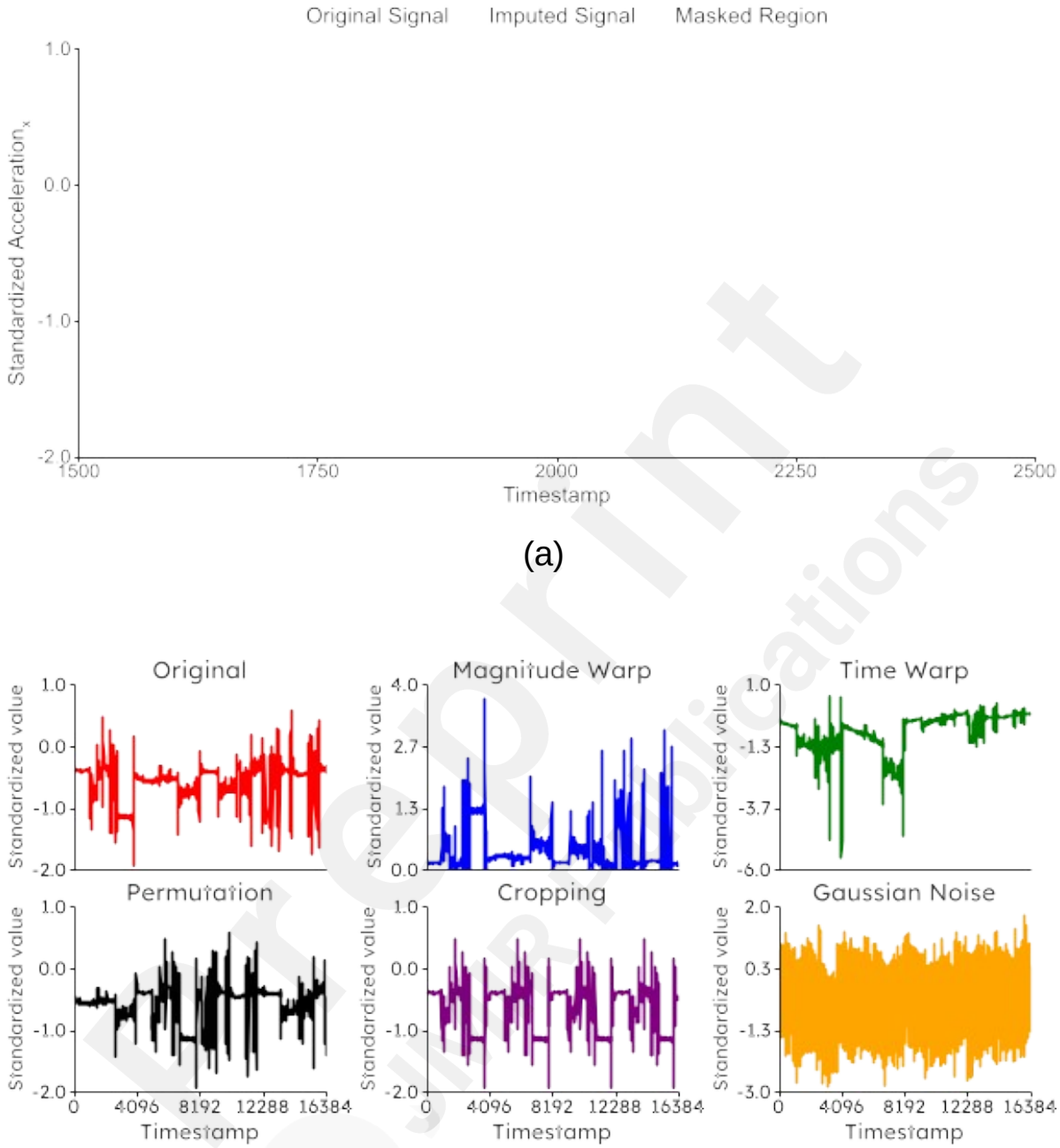



Figure 3: **Surrogate tasks used for self-supervised pre-training**. (a) *Masked prediction*: grey-shaded areas correspond to zeroed-out time-series portions; the model is tasked with minimizing the distance between the original time-series and the one imputed at the masked areas. (b)

*Transformation prediction*: the figure shows the type of transformations applied to the time-series; given the transformed channels, the model was trained to learn which transformation each channel underwent.

## Target Task Performance Comparison

Table 2 illustrates the performance under each model we developed. While they were all optimized for segment accuracy, since in a clinical scenario a decision needs to be taken at the subject level, we also reported subject accuracy. Note that while accuracy is a suitable metric in our use case as data is perfectly balanced, we also provide complementary metrics (precision, recall, $F_1$-score, and AUROC) both at the segment and the patient level.

| Model | | ACC | | Precision | | Recall | | $F_1$score | | AUROC | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | segment | subject | segment | subject | segment | subject | segment | subject | segment | subject |
| SL | ENET | 66.38 | 71.88 | 66.22 | 75 | 66.86 | 65.63 | 66.54 | 70 | 72.24 | 82.25 |
| | KNN | 70.37 | 82.81 | 69.09 | 80 | 73.74 | 81.2 | 71.34 | 80.6 | 73.27 | 83.26 |
| | SVM | 71.25 | 81.25 | 71.87 | 80 | 71.40 | 77.65 | 71.63 | 78.81 | 73.44 | 83.21 |
| | XGBoost | 72.02 | 82.81 | 71.33 | 83 | 72.11 | 81.1 | 71.72 | 82.03 | 72.44 | 83.17 |
| | E4mer | 75.35 | 81.25 | 73.46 | 80.55 | 75.34 | 82.14 | 74.39 | 81.33 | 75.68 | 82.22 |
| SSL | MP (LR) | 77.53 | 87.5 | 78.34 | 88.6 | 77.41 | 88 | 77.87 | 88.3 | 78.02 | 89.2 |
| | MP (FT) | **81.23** | **90.63** | **80.91** | **90.11** | **82** | **92.87** | **81.45** | **91.47** | **82.02** | **93.11** |
| | TP (LR) | 71.16 | 81.25 | 72.12 | 82.44 | 72.01 | 82.31 | 72.06 | 82.37 | 71.89 | 84.12 |
| | TP (FT) | 75.69 | 84.38 | 75.41 | 82.11 | 74.79 | 83.9 | 75.1 | 83 | 75.21 | 84.23 |

Table 2: **Masked prediction self-supervised pre-training comfortably outperformed end-to-end self-supervised learning while also surpassing other self-supervised approaches**. Performance in

differentiating a mood disorder acute episode from euthymia across different models. Note that data is perfectly balanced in terms of segment classes and subject classes. While this justifies the use of accuracy as a metric, we also herewith report segment and subject level precision, recall, $F_1$score, and area under the ROC curve. At the subject level, the predicted class was the result of a majority vote over that subject's segments, while the predicted probabilities under each class were derived by summing segments' predicted probabilities for that subject and normalizing by the corresponding segment number. FT: fine-tuning; LR: linear read-out; MP: masked prediction; SL: supervised learning; SSL: self-supervised learning; TP: transformation prediction.

The E4mer and the CML baselines performed to a similar level: while E4mer was superior to XGBoost in terms of $ACC_{segment}$ (75.35 vs 72.02), it was trumped by the CML on $ACC_{subject}$ (82.81 vs 81.25). Other CML baselines fared worse than XGBoost. Masked prediction pre-training led to a target task performance substantially higher than the baselines, under both metrics. While both linear read-out and tuning dominated over supervised learning, the latter scored the highest performance with an $ACC_{segment}$ and an $ACC_{subject}$ of 0.8123 and 0.9063 respectively. On the other hand, transformation prediction led to only modest improvement over E4mer. Statistical tests comparing the best SSL scheme, i.e. masked prediction with fine-tuning, against the fully-supervised E4mer and XGBoost, were significant at both the segment and the subject level. In particular, comparison with the E4mer yielded $p_{Bonferroni}$ values of 0.028 for LME, < 0.001 and 0.017 for the t-test at the segment and subject level respectively.

As for XGBoost, $p_{Bonferroni}$ values were 0.039 for LME, <0.012 and 0.012 for the t-test at the segment and subject level respectively.

Comparison of the best SSL with its supervised-learning counterpart in terms of $ACC_{segment}$ by subject (Figure 4) suggests that there are only two (euthymic) individuals (i.e. 3.13% of the sample) misclassified by SSL but correctly classified by the supervised E4mer. On the other hand, supervised learning mis-predicts eight individuals (i.e. 12.5%) that SSL gets right. Patients on an MD acute episode are shown as dots with a colour gradient proportional to their total score on the Hamilton Depression Rating Scale-17[35] (left half) and Young Mania Rating Scale[36] (right half), two clinician-administered questionnaires tracking depression and mania severity respectively. Subjects on an acute episode misclassified by supervised learning include patients with severe depressive (or manic) symptomatology. Notably, both SSL and supervised learning fail on four subjects (6.25%), including three patients on an acute episode with relatively moderate severity.

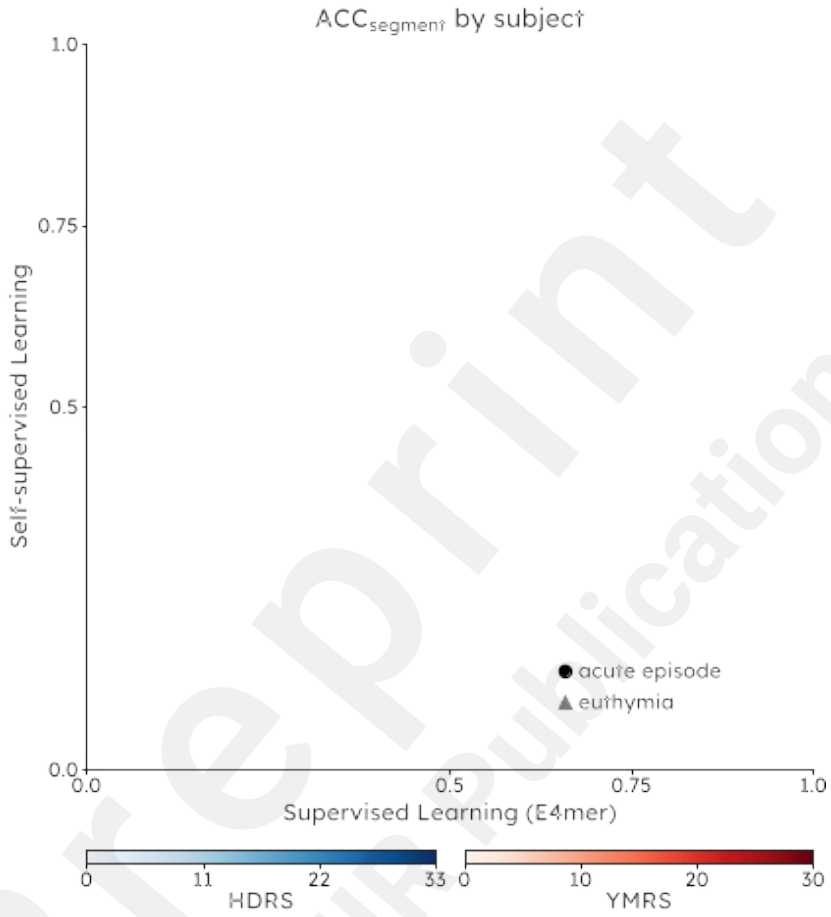



Figure 4: **Self-supervised learning beats supervised learning by six (i.e. 9.38%) more subjects correctly classified**. Segment Accuracy ( $ACC_{segment}$ ) under self-supervised learning and supervised learning (E4mer) within each subject's test segments. Subjects in euthymia are represented as triangles while subjects on an acute episode are shown as circles with the left (right) half colored in blue (red) with a gradient proportional to total sum on the Hamilton Depression Rating Scale-17 (Young Mania Rating Scale), a doctor-administered questionnaire gauging depression (mania) severity. Subjects' position on the x (y) axis corresponds to their

proportion of recording segments correctly classified by supervised (self-supervised) learning. Note that a subject's majority vote over their segments is in agreement with the subject's true mood state when the proportion of correctly classified segments from that subject is greater than 0.5. HDRS (YMRS) range shown on the color bar refers to values scored in the TIMEBASE/INTREPIBD sample, while the total score, in general, can range between [0-52] ([0-60]).

## Ablation analyses & learned representations

Table 3 shows the difference in target task $ACC_{segment}$ and $ACC_{subject}$ resulting from pre-training the best SSL on parts of the unlabeled data collection and then fine-tuning it onto the target task.

(a)

| **Resampling ratio** | **80%** | **60%** | **40%** | **20%** | **0%** |
|---|---|---|---|---|---|
| $ACC_{segment}$ **difference** | -0.23 | -2.14 | -6.07 | -6.35 | -7.07 |
| $ACC_{subject}$ **difference** | -1.57 | -1.57 | -4.70 | -4.70 | -7.82 |
| **LME p** | 0.087 | 0.072 | 0.059 | 0.045 | 0.041 |
| **t-test (segment) p** | < 0.001 | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| **t-test (subject) p** | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |

(b)

| **Dataset** | **Relative Size** | $ACC_{segment}$ **difference** | $ACC_{subject}$ **difference** | **LME p** | **t-test (segment) p** | **t-test (subject) p** |
|---|---|---|---|---|---|---|
| **ADARP** | 12.34 | -2.44 | -1.57 | 0.010 | < 0.001 | 1 |
| **Stress Predict** | 0.30 | -0.21 | -1.57 | 0.228 | 1 | 1 |
| **Toadstool** | 0.04 | 0.52 | -3.13 | 1 | 1 | 1 |
| **UE4W** | 2.32 | -1.93 | -4.70 | 1 | 0.003 | 1 |
| **WEEE** | 0.18 | 1.19 | 1.57 | 1 | 0.9 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **WESAD** | 0.42 | -0.51 | -1.57 | 1 | 1 | 1 |
| **WESD** | 0.72 | 1.90 | 1.57 | 0.058 | 0.054 | 1 |
| **In-GaugeEn-Gage** | 17.55 | -4.44 | -4.70 | 1 | < 0.001 | 0.631 |
| **Nurse Stress Detection** | 11.82 | -0.81 | -1.57 | 1 | 1 | 1 |
| **BIG IDEAs Lab** | 19.38 | -2.09 | -1.57 | 1 | 0.082 | 1 |
| **PPG-DaLiA** | 0.69 | 1.93 | 4.70 | 0.525 | < 0.001 | 1 |
| **TIMEBASE/ INTREPIBD** | 34.24 | -4.32 | -3.13 | 1 | 1 | 0.379 |

Table 3: **Ablation analyses show a positive trend between unlabeled data availability and target-task performance, but dataset-specific unobserved factors likely have a role**. The difference in $ACC_{segment}$ and $ACC_{subject}$ from pre-training on just parts of the entire unlabeled data collection is shown. Green (red) highlighted cells represent deterioration (improvement) in performance upon retraining on the ablated unlabeled data collection. A linear mixed-effects model (LME) and a two-tailed paired t-test assess if the mean difference in predicted probabilities for the segment's correct class differs from zero, with the former correcting for subjects as a random effect. A two-tailed paired t-test assesses if the mean difference in the number of correctly classified segments by subject differs from zero. In each test, the comparator is the best-performing self-supervised model. P-values are corrected with Bonferroni's method. Note that a majority vote over a subject's segments is used to issue subject-level predictions and $ACC_{subject}$ is simply the fraction of correct majority votes in the test set. $ACC_{subject}$ therefore does not consider the proportion of votes over a subject's segments in favour of the subject's correct class but just whether a majority, no matter how small or large, is reached in agreement with the correct class. On the other hand, the t-test (subject) assesses a zero average difference in the proportion of votes, within subjects, for the correct class. (a) The unlabeled collection was down-sampled, stratifying

by datasets. Self-supervised pre-training, preceding fine-tuning on the target task, used therefore only a fraction of the total unlabeled collection. A resampling ratio of 0% means that self-supervised pre-training was done on the target training set only. (b) Self-supervised pre-training was conducted leaving out each dataset in turn from the unlabeled collection.

The Pearson correlation coefficient (PCC) between unlabeled data down-sampling ratios and difference in $ACC_{segment}$ and $ACC_{subject}$ is 0.9401 and 0.9449 respectively, indicating a strong dependence between performance and unlabeled data availability. Similarly, excluding individual datasets from pre-training impacted $ACC_{segment}$ and $ACC_{subject}$ proportionally to their relative size (PCC of -0.8185 and -0.4083, respectively). Notably, however, TIMEBASE/INTREPIBD, despite being collected at the same site as the target task data and making up the largest share of the unlabeled data collection, did not leave the largest dent in performance when excluded from training. Furthermore, excluding some datasets resulted in a performance improvement. Differences in $ACC_{segment}$ and $ACC_{subject}$ do not always have the same sign because of the way they are defined. Indeed, it is for example possible that the absolute number of correctly classified segments decreases but enough previously misclassified segments within a subject are now correctly classified so that the majority vote for that subject flips. Statistical analyses show that the ablation of a single dataset is associated with non-significantly different performance in terms of correctly classified segments within subjects. At the level of the probability assigned to the correct class for each segment, LME was significant only for a dataset whereas results were mixed for t-tests. Stratified resampling gave positive results but significance for LME was

reached only at lower down-sampling ratios.

Lastly, we visualized the representations learned by the encoder EN part of our best-performing models, to gain further insights. As the EN's output had dimensionality (*B*=segments #, *N* =timestamps #, *D*=Transformer's model dimension), for visualization purposes we averaged out the *D* axis and then employed UMAP[64], a powerful nonlinear dimensionality reduction technique, to embed the resulting *N*-dimensional data points into three dimensions. The top-left plot of Figure 5 shows the representations learned during self-supervised pre-training with masked prediction. The segments herewith shown are the target task test segments along with an equal number of segments belonging to the same sessions but taken from sleep state, which the SSL model was never exposed to during training. Wake and sleep segments have different embeddings, suggesting that the model captured this structure in the physiological data: a Gaussian mixture model indeed recovered two clusters, one with predominantly sleep segments (82.66%) and the other with a majority of wake segments (95.58%). It should be noted that sleep and wake naturally have quite different semantics with respect to physiological data and the algorithm we employed for sleep/wake differentiation (*Van Hees*[52]) uses a simple heuristic defining sleep as a sustained lack of significant changes in the acceleration angle. The top-right and bottom plots of Figure 5 illustrate the representations from the SSL model upon fine-tuning on the target task. The top-right scatter plot displays the target task test segments as well as pre-training validation set segments (except for the pre-training segments from the TIMEBASE/INTREPIBD collection). The latter group of segments we assumed as taken from subjects without an MD acute episode and, arguably, most even without any MD historical diagnosis, since

the open-access datasets we found did not select for patients with an MD. The plot shows three clusters whose composition, as recovered with a Gaussian mixture model, is as follows: 1) 79.26% acute episode, 20.73% euthymia, 2) 74.16% euthymia, 25.84% acute episode, and 3) 91.01% unlabeled segments, 7.96% euthymia, and 1.02% acute episode. The bottom plots in Figure 5 show target task segments test segments only (no unlabeled segment), colored with a gradient proportional to symptoms' severity, as assessed with Hamilton Depression Rating Scale-17[35] and Young Mania Rating Scale[36]. Embeddings would seem to suggest a progression in symptoms' severity across the two clusters of segments on the right of the scatter plot.

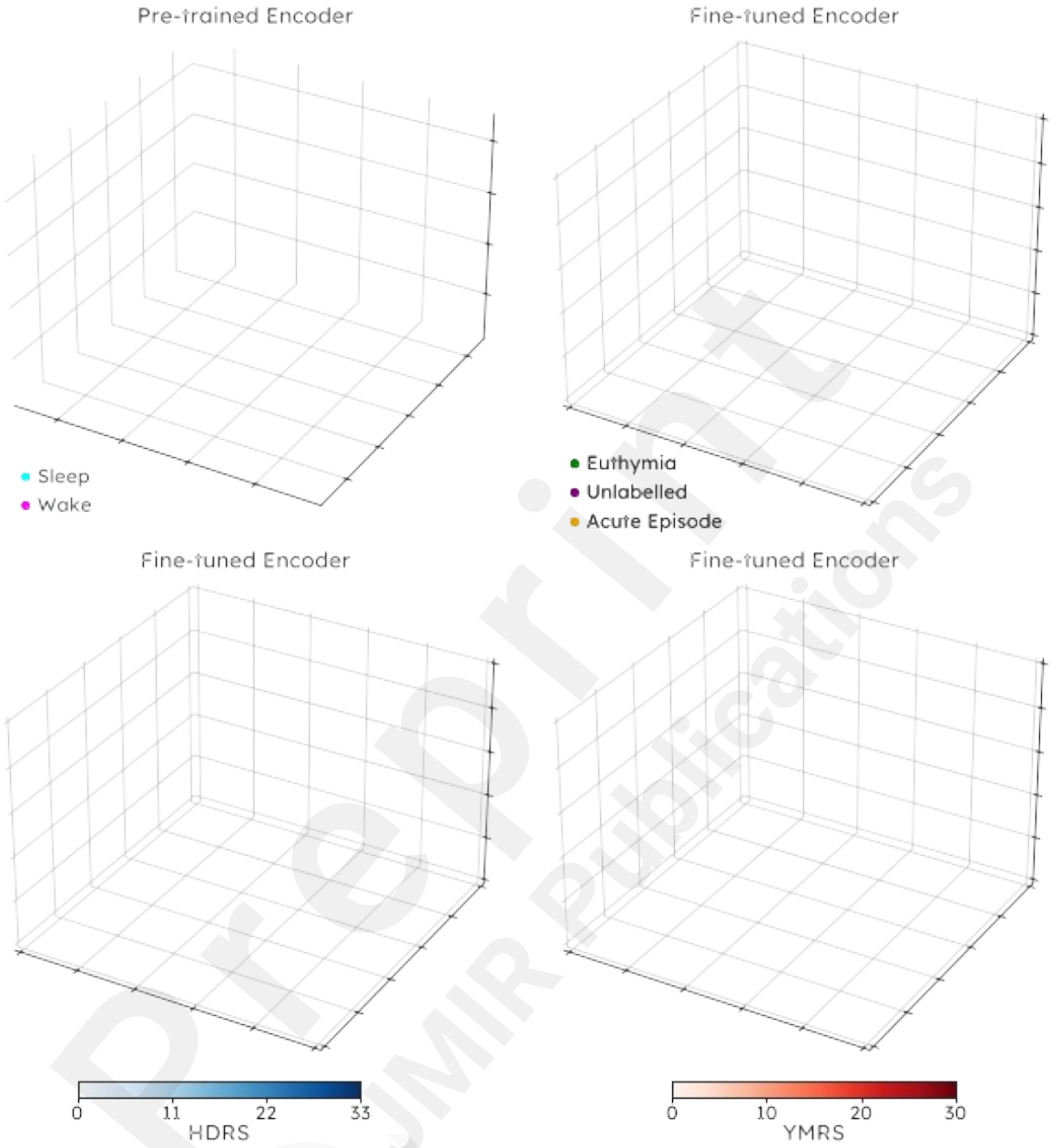



Figure 5: **Reassuringly, the learned embeddings seem to have captured meaningful semantics about the underlying context**. Top left: embeddings from the encoder pre-trained on mask prediction map sleep and wake segments to different parts of the latent space. Top right: embeddings from the encoder fine-tuned on the target task show that segments from the unlabeled open-access datasets, which presumably do not contain subjects on an acute mood episode, tend to cluster with part of the

segments from patients in euthymia. Bottom left (right): embeddings from the fine-tuned encoder show a gradient in symptoms' severity across target task segments, as revealed by Hamilton Depression Rating Scale-17 (Young Mania Rating Scale) total score. Note that unlabeled segments are not shown in the bottom left (right) plot and that the HDRS (YMRS) range showed on the color bar refers to values scored in the TIMEBASE/INTREPIBD sample, while the total score, in general, can range between [0-52] ([0-60])

## Discussion

Personal sensing is likely to play a key role in healthcare supply, creating unprecedented opportunities for patient monitoring and just-in-time adaptive interventions[65]. Towards delivering on this promise, expert annotation is a major obstacle; this is especially the case with MDs, wherein data annotation is particularly challenging and time-consuming, considering the nature of the disorder.

To the best of our knowledge, we are the first to show that SSL is a viable paradigm in personal sensing for MDs, mitigating the annotation bottleneck thanks to the repurposing of existing unlabeled, data collected in settings as different as subjects playing Super Mario[42], taking university exams[44] , or performing physical exercise[43].

We took on a straightforward yet fundamental task, i.e. distinguishing acute episodes from euthymia. Timely recognition of an impending mood episode in someone with a historical MD diagnosis, regardless of the episode polarity (depressive, manic, or mixed), may indeed enable pre-emptive interventions and better outcomes[5]. Our results suggest that, with

a sample size on the order of magnitude that is typical of studies into personal-sensing for MD, a modern deep-learning fully-supervised pipeline (E4mer) may offer no substantial improvements over simpler CML algorithms (e.g., XGBoost), despite higher development and computational costs. On the other hand, the accumulation and repurposing of existing unlabeled datasets for an SSL pre-training phase leads to a confident margin of improvement: $ACC_{segment}$ ($ACC_{subject}$) improves by 7.8% (11.54%) relative to the fully-supervised E4mer, with 6 (out of 64) more subjects correctly classified.

Our findings further show that careful choice of the pretext task, as well-documented in the literature on SSL[30], is key towards learning useful representations for the downstream target task. Unlike masked prediction, improvement, if any at all, from transformation prediction was only modest. This is not to say that such pretext task may in general fail to deliver on acute episode vs. euthymia differentiation. Indeed, the specific transformations we implemented, borrowed from Wu et al.[32], may have been suboptimal for our downstream task, pointing to the importance of domain knowledge (including clinical expertise) in pretext task design. Lastly, while SSL relaxes dependence on large, annotated datasets, our results indicate that its success relies on the size of unlabeled data. Ablation analyses indeed showed a positive correlation between target task performance and size of the corpus available for pre-training. Dataset-idiosyncratic factors accounting for the non-perfect correlation between the relative size and impact on target task performance may be present. Speculatively, these may include noise in the data, (dis)similarity of recording conditions, or (ir)relevance for the target task of the representations learned modelling the domain of the unlabeled dataset.

Statistical analyses showed that excluding from pre-training any of the individual unlabeled datasets, while keeping all others, was not associated with a significant change in performance on the proportion of correctly classified segments within subjects. The lack of a significant effect in either direction (improvement or deterioration), along with a significantly superior performance of SSL over fully supervised schemes, indicate that pre-training on big data collections leads to higher performance than taking on the target task from scratch. Of importance, adding datasets for pre-training from domains not immediately related to the target task did not undermine the model. Pre-training under progressively lower down-sampling ratios lent further support to the importance of data size. This is consistent with the deep-learning recipe where the bigger the pre-training corpus the better the results [66]. Results from tests at the level of segment-predicted probabilities are consistent with the view above. Of the datasets comprising less than 1% of the entire unlabeled collection, only one reached statistical significance. LME has more flexibility to explain the data since, rather than pooling all segments together in a unique (bigger) population, it treats them as embedded within subjects. This explains the lack of statistical significance relative to the t-tests, under various data ablation regimes.

We acknowledge the following limitations to this study. We deliberately chose the simplest task that has clinical relevance in personal sensing for MDs since our focus was on SSL; however, we appreciate that a more fine-grained MD description, beyond a simple acute episode vs euthymia binary classification, may add further clinical value[24]. As the literature on SSL is expanding at a fast pace, a thorough search of different approaches was beyond the scope of this work. We acknowledge that other pretext tasks

can be deployed and while the architectural choice may have an impact on SSL, we settled for just one reasonable, modern model design with a Transformer[33] as a workhorse for representation learning. Lastly, given the naturalist design of the study, reflective of the intended use of personal sensing in a clinical setting, we could not exclude the effect of confounders, including medications, on the physiological variables. However, we reported medication classes administered in the cohort and verified a lack of any significant association between target classes (euthymia vs acute episode) and being on a given medication class.

**Conclusion** - This work shows that self-supervised learning is a promising paradigm for mitigating the annotation bottleneck, one of the major barriers towards the development of AI-powered clinical decision support systems using personal sensing to help monitor mood disorders, thus enabling early interventions. The collection and pre-processing of open-access unlabeled datasets that we curated (E4SelfLearning) can foster future research into self-supervised learning, therefore advancing the translation of personal sensing into the clinical practice.

**Future directions** - As our findings indicate that the choice of pretext task has a significant impact on target task performance, further efforts should be put into pretext task design. Indeed, while masked prediction is a general-purpose strategy inspired by the great success of BERT[28] in Natural Language Processing, literature on self-supervised learning[30] suggests that domain knowledge may help tailor the pretext task to the specific use case. A promising approach we have not explored is contrastive learning[67], which indeed relies on domain knowledge of how augmented views of the input are created, especially since most experience today is in Computer Vision and Natural Language Processing while physiological

multivariate time-series are relatively unexplored.

## Code availability

Python 3.10 programming language was used where deep-learning and classical machine learning models were implemented in PyTorch[68] and Scikit-learn[69]/XGBoost[70] respectively, while hyperparameter tuning was performed in both cases with Weight & Biases [71]. The best hyperparameters' setting found during tuning for each model is reported in Supplementary Material. All deep-learning models were trained on a single Nvidia A100 GPU.

## Data availability

The E4SelfLearning collection is available at link to be released upon acceptance for publication. Data in de-identified form from the TIMEBASE/ INTREPIBD study may be made available from the corresponding author upon reasonable request.

## Ethics and confidentiality

The TIMEBASE/INTREPIBD study was conducted in accordance with the ethical principles of the Declaration of Helsinki and Good Clinical Practice and the Hospital Clinic Ethics and Research Board (HCB/2021/104). All participants provided written informed consent prior

to their inclusion in the study. All data were collected anonymously and stored encrypted in servers complying with all GDPR and HIPAA regulations. As regards other studies included in the present work, we refer to the relevant publications.

## Acknowledgments

We acknowledge the contribution of all the participants of the study.

F.C. and B.M.L. are supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any author-accepted manuscript version arising.

G.A. is supported by a Rio Hortega 2021 grant (CM21/00017) and M-AES mobility fellowship (MV22/00058), from the Spanish Ministry of Health financed by the Instituto de Salud Carlos III (ISCIII) and co-financed by the Fondo Social Europeo Plus (FSE+).

I.G. thanks the support of the Spanish Ministry of Science and Innovation (MCIN) (PI23/00822) integrated into the Plan Nacional de I+D+I and cofinanced by the ISCIII-Subdirección General de Evaluación y confinanciado por la Unión Europea (FEDER, FSE, Next Generation EU/Plan de Recu- peración Transformación y Resiliencia PRTR); the Instituto de Salud Carlos III; the CIBER of Mental Health (CIBERSAM); and the Secretaria d’Universitats i Recerca del Departament d’Economia i Coneixement

(2021 SGR 01358), CERCA Programme / Generalitat de Catalunya as well as the Fundació Clínic per la Recerca Biomèdica (Pons Bartran 2022-FRCB PB1 2022).

A.H.Y.’s independent research is funded by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Accepted Author Manuscript version arising from this submission.

D.H.M. is supported by a Juan Rodés JR18/00021 granted by the Instituto de Salud Carlos III (ISCIII).

A.V. is supported by the "UNREAL" project (EP/Y023838/1) selected by the ERC and funded by UKRI EPSRC.

## Authors contributions

F.C. conceived of the study, proposed the methodology, developed the software codebase for the analyses, prepared the manuscript, and curated data collection. B.M.L. contributed to codebase development and manuscript writing. G.A., C.V.P., A.M., I.P., M.V., I.G.F., A.B., and M.G. collected the data for the TIMEBASE/INTREPIBD study. E.V., A.H.Y., S.L., and H.W. critically reviewed the manuscript and provided feedback on the clinical side. D.H.M. is the co-ordinator of the TIMEBASE/INTREPIBD study and critically reviewed the manuscript.

A.V. supervised this study and contributed to the study design, methodology development, and manuscript writing.

## Competing interests

G.A. has received CME-related honoraria, or consulting fees from Janssen-Cilag, Lundbeck, Lundbeck/Otsuka, and Angelini, with no financial or other relationship relevant to the subject of this article.

I.G. has received grants and served as a consultant, advisor or CME speaker for the following identities: ADAMED, Angelini, Casen Recordati, Esteve, Ferrer, Gedeon Richter, Janssen Cilag, Lundbeck, Lundbeck-Otsuka, Luye, SEI Healthcare, Viatris outside the submitted work. She also receives royalties from Oxford University Press, Elsevier, Editorial Médica Panamericana.

All authors report no financial or other relationship relevant to the subject of this article.

## References

1 American Psychiatric Association D, Association AP, others. *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013.

2 Santomauro DF, Herrera AMM, Shadid J, Zheng P, Ashbaugh C, Pigott DM *et al.* Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet* 2021; **398**: 1700–1712.

3 Greenberg PE, Fournier A-A, Sisitsky T, Simes M, Berman R, Koenigsberg SH *et al.* The economic burden of adults with major depressive disorder in the United States (2010 and 2018). *Pharmacoeconomics* 2021; **39**: 653–665.

4 Brådvik L. Suicide risk and mental disorders. International journal of environmental research and public health. 2018; **15**: 2028.

5 Joyce K, Thompson A, Marwaha S. Is treatment for bipolar disorder more effective earlier in illness course? A comprehensive literature review. *International journal of bipolar disorders* 2016; **4**: 1–9.

6 Buchman-Wildbaum T, Váradi E, Schmelowszky Á, Griffiths MD, Demetrovics Z, Urbán R. The paradoxical role of insight in mental illness: The experience of stigma and shame in schizophrenia, mood disorders, and anxiety disorders. *Archives of Psychiatric Nursing* 2020; **34**: 449–457.

7 Rimmer A. Mental health: Staff shortages are causing distressingly long waits for treatment, college warns. 2021.

8 Satiani A, Niedermier J, Satiani B, Svendsen DP. Projected workforce of psychiatrists in the United States: a population analysis. *Psychiatric Services* 2018; **69**: 710–713.

9 Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 2017; **13**: 23–47.

10 Faurholt-Jepsen M, Brage S, Kessing LV, Munkholm K. State-related differences in heart rate variability in bipolar disorder. *Journal of psychiatric research* 2017; **84**: 169–173.

11 Sarchiapone M, Gramaglia C, Iosue M, Carli V, Mandelli L, Serretti A *et al.* The association between electrodermal activity (EDA), depression and suicidal behaviour: A systematic review and narrative synthesis. *BMC psychiatry* 2018; **18**: 1–27.

12 Tazawa Y, Wada M, Mitsukura Y, Takamiya A, Kitazawa M, Yoshimura M *et al.* Actigraphy for evaluation of mood disorders: A systematic review and meta-analysis. *Journal of affective disorders* 2019; **253**: 257–269.

13 Tazawa Y, Liang K, Yoshimura M, Kitazawa M, Kaise Y, Takamiya A *et al.* Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. *Heliyon* 2020; **6**: e03274.

14 Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. *NPJ digital medicine* 2019; **2**: 3.

15 Côté-Allard U, Jakobsen P, Stautland A, Nordgreen T, Fasmer OB, Oedegaard KJ *et al.* Long–Short Ensemble Network for Bipolar Manic-Euthymic State Recognition Based on Wrist-Worn Sensors. *IEEE Pervasive Computing* 2022.

16 Nguyen D-K, Chan C-L, Li A-HA, Phan D-V, Lan C-H. Decision support system for the differentiation of schizophrenia and mood disorders using multiple deep learning models on wearable devices data. *Health Informatics Journal* 2022; **28**: 14604582221137537.

17 Ghandeharioun A, Fedor S, Sangermano L, Ionescu D, Alpert J, Dale C *et al.* Objective assessment of depressive symptoms with machine learning and wearable sensors data. In: *2017 seventh international conference on affective computing and intelligent*

*interaction (ACII)*. IEEE, 2017, pp 325–332.

18 Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena D *et al.* Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in psychiatry* 2020; **11**: 584711.

19 Lee H-J, Cho C-H, Lee T, Jeong J, Yeom JW, Kim S *et al.* Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: a prospective nationwide cohort study. *Psychological Medicine* 2022; : 1–9.

20 Li BM, Corponi F, Anmella G, Mas A, Sanabra M, Hidalgo-Mazzei D *et al.* Inferring mood disorder symptoms from multivariate time-series sensory data. In: *NeurIPS 2022 Workshop on Learning from Time Series for Health*. 2022https://openreview.net/forum?id=awjU8fCDZjS.

21 Empatica. E4 wristband technical specifications – Empatica Support. E4 wristband technical specifications. 2020.https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications.

22 Ronca V, Martinez-Levy AC, Vozzi A, Giorgi A, Aricò P, Capotorto R *et al.* Wearable Technologies for Electrodermal and Cardiac Activity measurements: A Comparison between Fitbit Sense, Empatica E4 and Shimmer GSR3+. *Sensors* 2023; **23**: 5847.

23 Shani C, Zarecki J, Shahaf D. The lean data scientist: recent advances toward overcoming the data bottleneck. *Communications of the ACM* 2023; **66**: 92–102.

24 Corponi F, Li BM, Anmella G, Mas A, Pacchiarotti I, Valentí M *et al.* Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number. *Transl Psychiatry* 2024; **14**: 1–9.

25 Rani V, Nabi ST, Kumar M, Mittal A, Kumar K. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering* 2023; **30**: 2761–2775.

26 Ericsson L, Gouk H, Loy CC, Hospedales TM. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine* 2022; **39**: 42–62.

27 Ohri K, Kumar M. Review on self-supervised image recognition using deep neural networks. *Knowledge-Based Systems* 2021; **224**: 107090.

28 Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805* 2018.

29 Huang S-C, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine* 2023; **6**: 74.

30 Zhang K, Wen Q, Zhang C, Cai R, Jin M, Liu Y *et al.* Self-Supervised Learning for

Time Series Analysis: Taxonomy, Progress, and Prospects. *arXiv preprint arXiv:230610125* 2023.

31 Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C. A transformer-based framework for multivariate time series representation learning. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp 2114–2124.

32 Wu Y, Daoudi M, Amad A. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. *IEEE Transactions on Affective Computing* 2023.

33 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN *et al.* Attention is all you need. *Advances in neural information processing systems* 2017; **30**.

34 Anmella G, Corponi F, Li BM, Mas A, Sanabra M, Pacchiarotti I *et al.* Exploring digital biomarkers of illness activity in mood episodes: hypotheses generating and model development study. *JMIR Mhealth and Uhealth* 2023.

35 Hamilton M. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry* 1960; **23**: 56.

36 Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *The British journal of psychiatry* 1978; **133**: 429–435.

37 Tohen M, Frank E, Bowden CL, Colom F, Ghaemi SN, Yatham LN *et al.* The International Society for Bipolar Disorders (ISBD) Task Force report on the nomenclature of course and outcome in bipolar disorders. *Bipolar disorders* 2009; **11**: 453–473.

38 Reiss A, Indlekofer I, Schmidt P, Van Laerhoven K. Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors* 2019; **19**: 3079.

39 Sah RK, McDonell M, Pendry P, Parent S, Ghasemzadeh H, Cleveland MJ. ADARP: A Multi Modal Dataset for Stress and Alcohol Relapse Quantification in Real Life Setting. In: *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2022, pp 1–4.

40 Schmidt P, Reiss A, Duerichen R, Marberger C, 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*. 2018, pp 400–408.

41 Iqbal T, Simpkin AJ, Roshan D, Glynn N, Killilea J, Walsh J *et al.* Stress Monitoring Using Wearable Sensors: A Pilot Study and Stress-Predict Dataset. *Sensors* 2022; **22**: 8135.

42 Svoren H, Thambawita V, Halvorsen P, Jakobsen P, Garcia-Ceja E, Noori FM *et al.* Toadstool: A dataset for training emotional intelligent machines playing Super Mario Bros. In: *Proceedings of the 11th ACM Multimedia Systems Conference*. 2020, pp 309–314.

43 Gashi S, Min C, Montanari A, Santini S, Kawsar F. A multidevice and multimodal dataset for human energy expenditure estimation using wearable devices. *Scientific Data* 2022; **9**: 537.

44 Amin MR, Wickramasuriya DS, Faghih RT. A Wearable Exam Stress Dataset for Predicting Grades using Physiological Signals. In: *2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*. IEEE, 2022, pp 30–36.

45 Hosseini S, Gottumukkala R, Katragadda S, Bhupatiraju RT, Ashkar Z, Borst CW *et al.* A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Scientific Data* 2022; **9**: 255.

46 Gao N, Marschall M, Burry J, Watkins S, Salim FD. Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. *Scientific Data* 2022; **9**: 261.

47 Hinkle LB, Metsis V. Unlabeled Empatica E4 Wristband Data (UE4W) Dataset (1.0). 2022. doi:10.5281/zenodo.6898244.

48 Bent B, Cho PJ, Henriquez M, Wittmann A, Thacker C, Feinglos M *et al.* Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. *NPJ Digital Medicine* 2021; **4**: 89.

49 Vieluf S, Amengual-Gual M, Zhang B, El Atrache R, Ufongene C, Jackson MC *et al.* Twenty-four-hour patterns in electrodermal activity recordings of patients with and without epileptic seizures. *Epilepsia* 2021; **62**: 960–972.

50 Nasseri M, Nurse E, Glasstetter M, Böttcher S, Gregg NM, Laks Nandakumar A *et al.* Signal quality and patient experience with wearable devices for epilepsy management. *Epilepsia* 2020; **61**: S25–S35.

51 ETH Zurich. Emaptics User Manual. 2023.https://archive.arch.ethz.ch/esum/downloads/manuals/emaptics.pdf.

52 Van Hees VT, Sabia S, Anderson KN, Denton SJ, Oliver J, Catt M *et al.* A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PloS one* 2015; **10**: e0142533.

53 Patterson MR, Nunes AA, Gerstel D, Pilkar R, Guthrie T, Neishabouri A *et al.* 40 years of actigraphy in sleep medicine and current state of the art algorithms. *NPJ Digital Medicine* 2023; **6**: 51.

54 Cui Z, Chen W, Chen Y. Multi-scale convolutional neural networks for time series classification. *arXiv preprint arXiv:160306995* 2016.

55 Panagiotou M, Zlatintsi A, Filntisis PP, Roumeliotis AJ, Efthymiou N, Maragos P. A comparative study of autoencoder architectures for mental health analysis using wearable sensors data. In: *2022 30th European Signal Processing Conference (EUSIPCO)*. 2022, pp 1258–1262.

56 Föll S, Maritsch M, Spinola F, Mishra V, Barata F, Kowatsch T *et al.* FLIRT: A feature generation toolkit for wearable data. *Computer Methods and Programs in Biomedicine* 2021; **212**: 106461.

57 Özdenizci O, Wang Y, Koike-Akino T, Erdoğmuş D. Learning invariant representations from EEG via adversarial inference. *IEEE access* 2020; **8**: 27074–27085.

58 Cheng JY, Goh H, Dogrusoz K, Tuzel O, Azemi E. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:200704871* 2020.

59 Sabry F, Eltaras T, Labda W, Alzoubi K, Malluhi Q, others. Machine learning for healthcare wearable devices: the big picture. *Journal of Healthcare Engineering* 2022; **2022**.

60 Strzelecki M, Badura P. Machine Learning for Biomedical Application. *Applied Sciences* 2022; **12**: 2022.

61 Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research* 2017; **18**: 6765–6816.

62 Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 2014; **15**: 1929–1958.

63 Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: a review. *Data mining and knowledge discovery* 2019; **33**: 917–963.

64 McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv 2018. *arXiv preprint arXiv:180203426* 1802.

65 Birk RH, Samuel G. Digital phenotyping for mental health: Reviewing the challenges of using data to monitor and predict mental health problems. *Current Psychiatry Reports* 2022; **24**: 523–528.

66 El-Nouby A, Izacard G, Touvron H, Laptev I, Jegou H, Grave E. Are Large-scale Datasets Necessary for Self-Supervised Pre-training? 2021. doi:10.48550/arXiv.2112.10740.

67 Kumar P, Rawat P, Chauhan S. Contrastive self-supervised learning: review, progress, challenges and future research directions. *International Journal of Multimedia Information Retrieval* 2022; **11**: 461–488.

68 Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp 8024–8035.

69 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; **12**:

2825–2830.

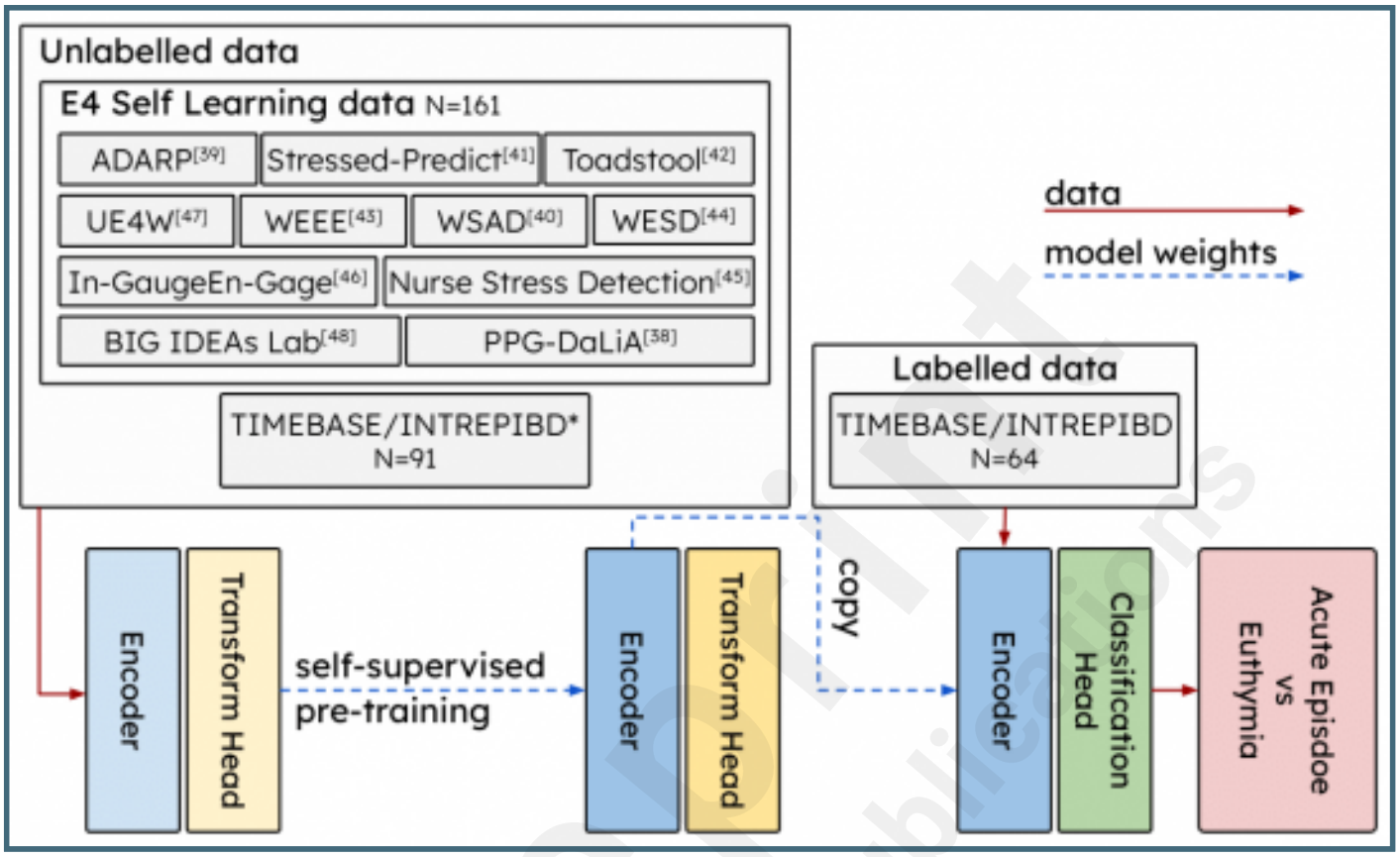
70 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM: New York, NY, USA, 2016, pp 785–794.

71 Biewald L. Experiment Tracking with Weights and Biases. 2020.https://www.wandb.com/.

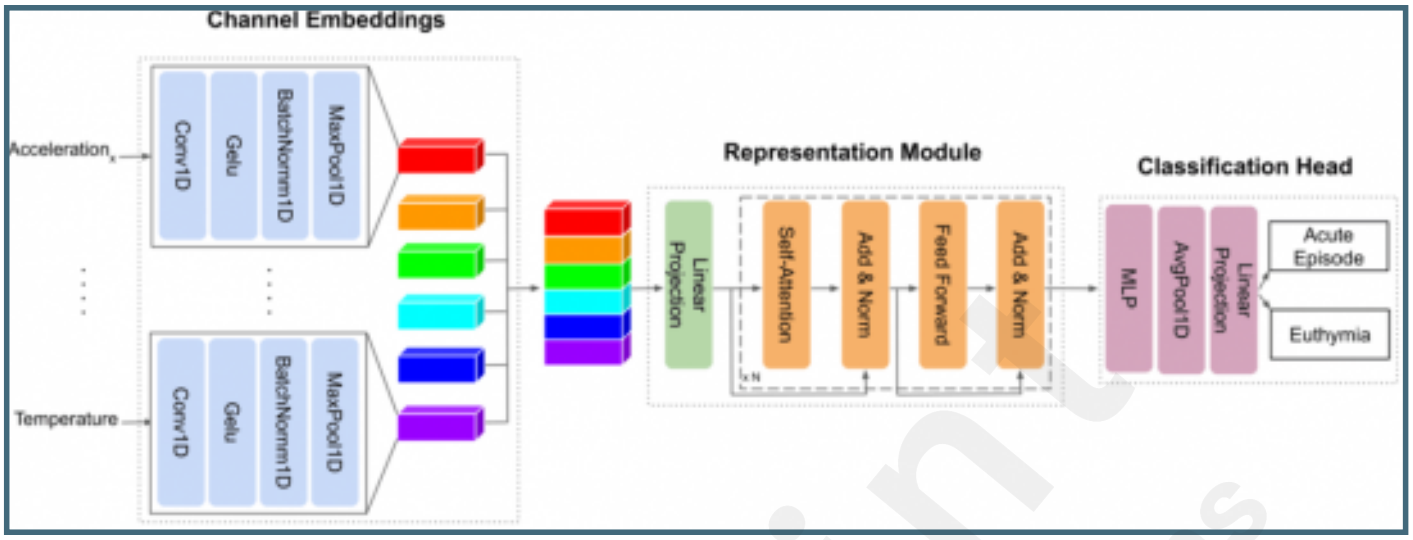
# Supplementary Files

# Figures

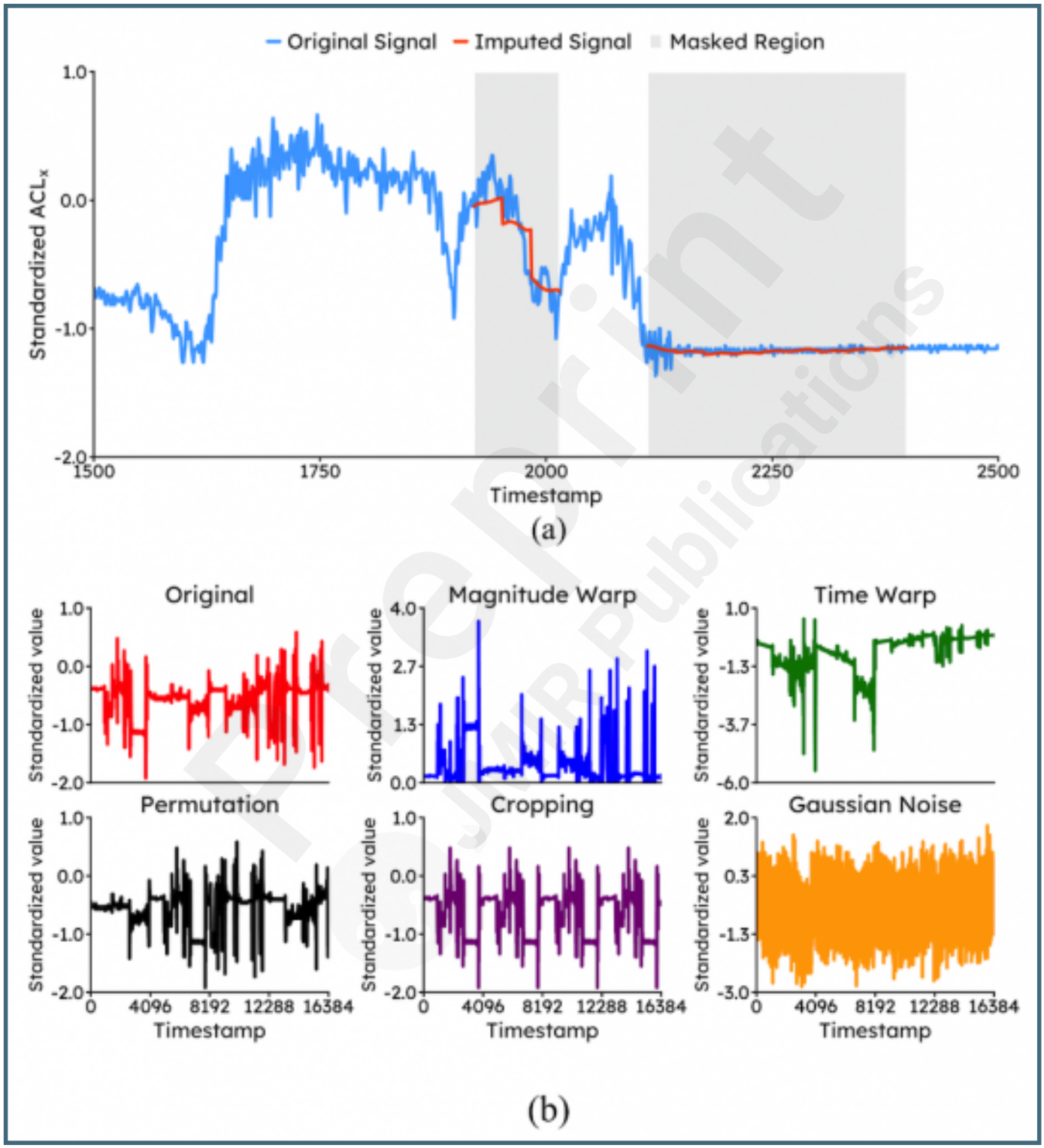
A total of ?6254 hours (261 days) of unlabeled recordings from 252 subjects while awake were used for self-supervised pre-training.



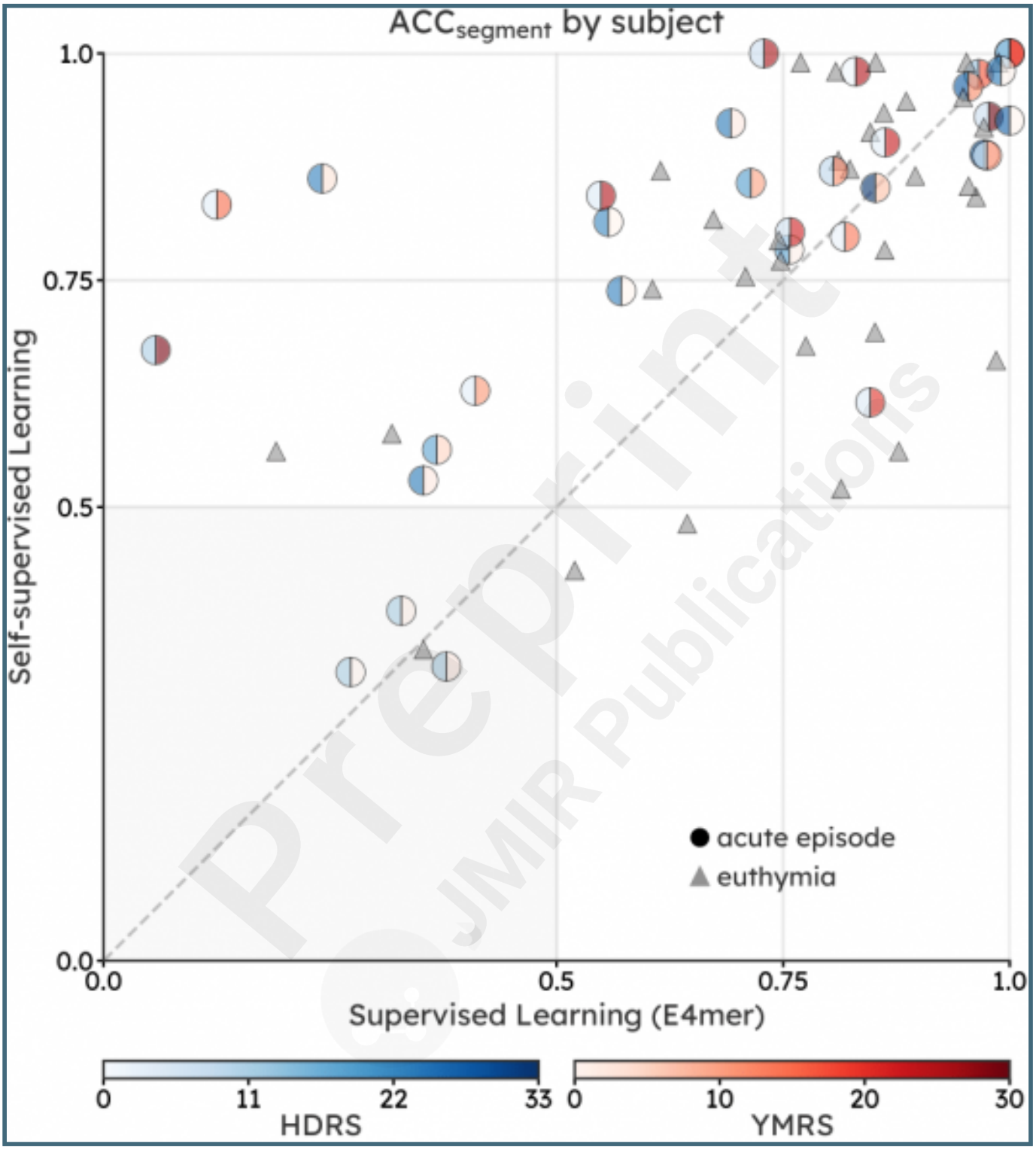
E4mer is a Transformer model tailored to the Empatica E4 input data.



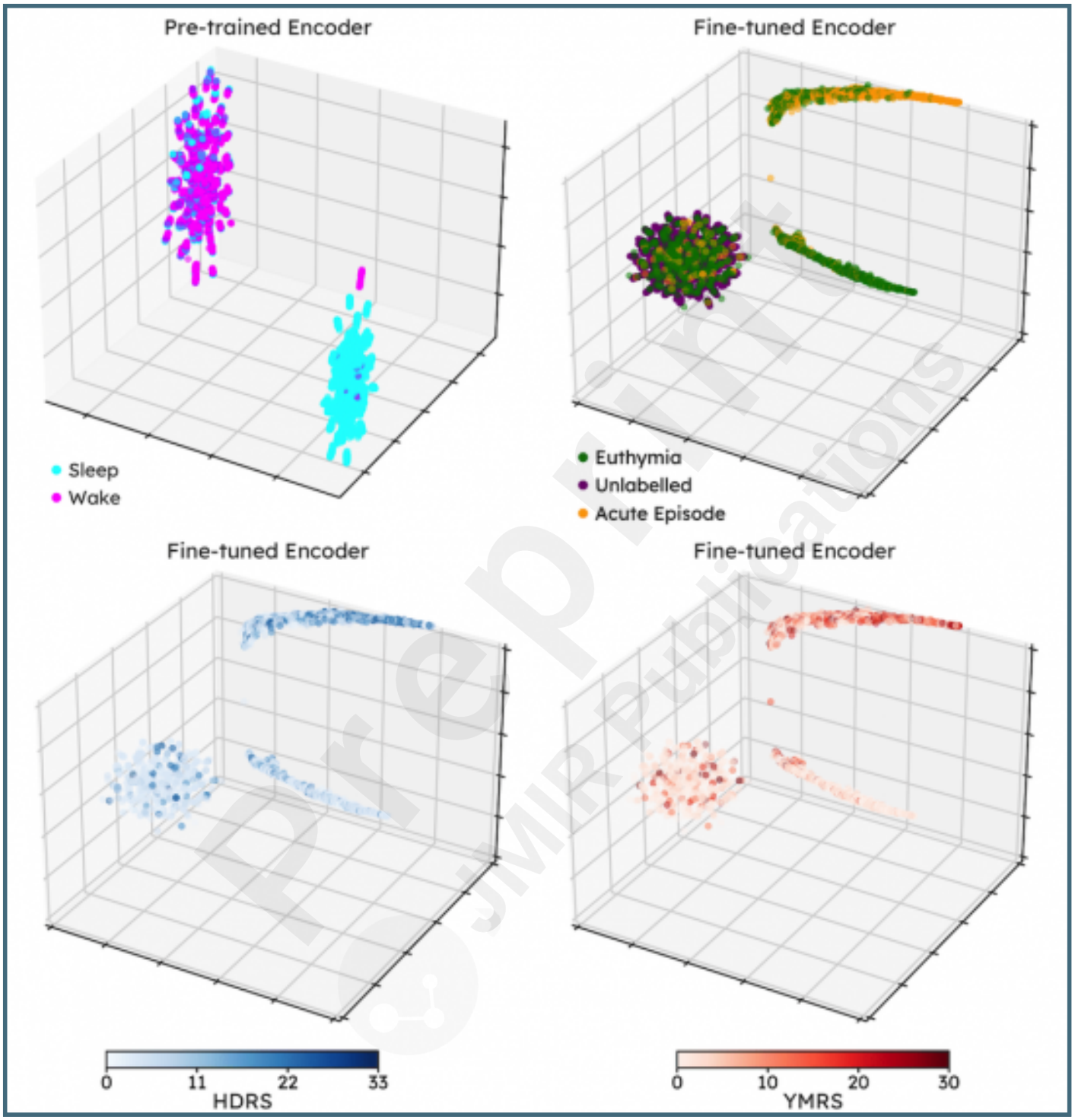
Surrogate tasks used for self-supervised pre-training. (a) Masked prediction: grey- shaded areas correspond to zeroed-out time-series portions; the model is tasked with minimizing the distance between the original time-series and the one imputed at the masked areas. (b) Transformation prediction: the figure shows the type of transformations applied to input time-series; given transformed channels, the model was trained to learn which transformation each channel underwent.



Self-supervised learning beats supervised learning by six (i.e. 9.38%) more subjects correctly classified.



Reassuringly, the learned embeddings seem to have captured meaningful semantics about the underlying context.



# Multimedia Appendixes

Supplementary Material.
URL: http://asset.jmir.pub/assets/bb61b2a717fccccb7a958d16a772ea4d.doc

# Related publication(s) - for reviewers eyes onlies

Revised version with tracked changes.
URL: http://asset.jmir.pub/assets/c92be1459b0bb80ec7e42096c2ba17f4.pdf

Revised manuscript marked-up.
URL: http://asset.jmir.pub/assets/18974e4000b4b309f0d45660e7f5719b.pdf