Scaling Representation Learning From Ubiquitous ECG With State-Space Models

Kleanthis Avramidis, *Graduate Student Member, IEEE*, Dominika Kunc, Bartosz Perz, Kranti Adsul, Tiantian Feng, *Member, IEEE*, Przemysław Kazienko, *Senior Member, IEEE*, Stanisław Saganowski, and Shrikanth Narayanan, *Fellow, IEEE*

***Abstract*—Ubiquitous sensing from wearable devices in the wild holds promise for enhancing human well-being, from diagnosing clinical conditions and measuring stress to building adaptive health promoting scaffolds. But the large volumes of data therein across heterogeneous contexts pose challenges for conventional supervised learning approaches. Representation Learning from biological signals is an emerging realm catalyzed by the recent advances in computational modeling and the abundance of publicly shared databases. The electrocardiogram (ECG) is the primary researched modality in this context, with applications in health monitoring, stress and affect estimation. Yet, most studies are limited by small-scale controlled data collection and over-parameterized architecture choices. We introduce WildECG, a pre-trained state-space model for representation learning from ECG signals. We train this model in a self-supervised manner with 275 000 10 s ECG recordings collected in the wild and evaluate it on a range of downstream tasks. The proposed model is a robust backbone for ECG analysis, providing competitive performance on most of the tasks considered, while demonstrating efficacy in low-resource regimes.**

***Index Terms*—Electrocardiography, ubiquitous computing, self-supervised learning, state-space models.**

I. Introduction

ARTIFICIAL Intelligence (AI) has made significant inroads into human-centered signal modeling, notably in the fields of behavioral analysis [1] and health [2]. This progress benefits primarily from the algorithmic development of deep learning models and the substantial effort in curating publicly-shared datasets [3]. The rapid advances of deep learning in various application domains, such as computer vision (CV), speech, and natural language processing (NLP) are critically dependent on the availability of large datasets, allowing for designing and training large-scale neural networks. Within the medical domain of biosignal analysis, supervised learning algorithms have been employed to improve diagnostic performance and accelerate biomarker detection in many areas, including dermatology [4], ophthalmology [5], as well as in psychology, physical health and well-being [6], [7].

Driven by successful applications in multiple fields within health and well-being, AI technologies are increasingly demanded in ubiquitous modeling of human states in everyday settings. Studies have been employing either mobile phones and questionnaires [8], [9] or wearable sensors paired with phones [10], [11], [12] to track human states including stress [8], depression [9] and physical activity [10]. Of particular interest is the assessment of workplace stress and behavior patterns [13]. However, a significant portion of research has focused on modeling physiological responses to external stimuli in constrained interaction environments. These approaches typically consider signals derived from cardiac activity, respiration patterns, body temperature, electrodermal activity and even neural

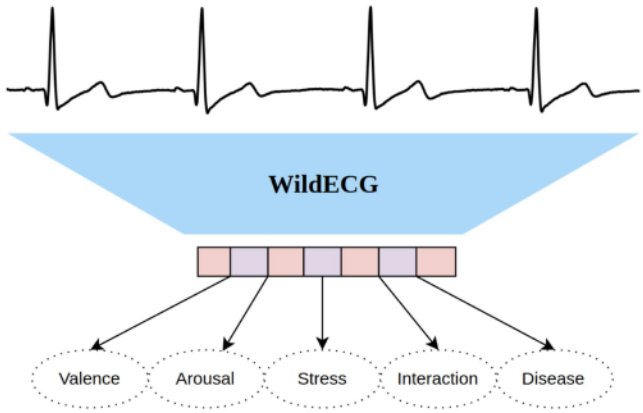

Fig. 1. Our proposed model extracts vector representations from input, single-lead ECG signals, and can be used both as a backbone encoder and feature extractor across multiple different sensing tasks.

Received 12 September 2023; revised 13 May 2024; accepted 12 June 2024. Date of publication 27 June 2024; date of current version 4 October 2024. This work was conducted at USC SAIL and supported in part by NSF SCH under Grant 2204942, in part by DARPA under Cooperative Agreement number N660012324006, in part by Toyota, in part by National Science Centre, Poland under Project 2020/37/B/ST6/03806, in part by the NAWA STER Programme Internationalisation of Wroclaw University of Science and Technology Doctoral School, and in part by the Polish National Agency for Academic Exchange (NAWA) – the Bekker Programme. *(Corresponding author: Kleanthis Avramidis.)*

Kleanthis Avramidis, Kranti Adsul, Tiantian Feng, and Shrikanth Narayanan are with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: avramidi@usc.edu; kadsul@usc.edu; tiantiaf@usc.edu; shri@ee.usc.edu).

Dominika Kunc, Bartosz Perz, Przemysław Kazienko, and Stanisław Saganowski are with the Department of Artificial Intelligence, Wrocław University of Science and Technology, 50-370 Wrocław, Poland (e-mail: dominika.kunc@pwr.edu.pl; bartosz.perz@pwr.edu.pl; kazienko@pwr.edu.pl; stanislaw.saganowski@pwr.edu.pl).

The code and pre-trained weights are shared publicly at github.com/klean2050/tiles_ecg_model.

Digital Object Identifier 10.1109/JBHI.2024.3416897

(brain) activity [14]. Cardiac activity, particularly the electrocardiogram (ECG), has been a prominent modality choice due to its well-recognized signal patterns and clinically-validated significance [15], [16]. While it requires low-cost recording equipment, ECG offers enormous diagnostic potential and hence has been densely researched, notably through the creation of shared databases and data-driven modeling approaches.

However, the transition from monitoring in clinical settings to sensing in the wild introduces novel challenges toward comprehending cardiac activity across diverse living contexts. Unlike text or image domains where abundant large datasets have enabled large (foundation) self-supervised models [17], the state of the art in biosignal models lags behind. One practical issue related to data acquisition is the need for long-term recording capabilities and the resulting cost of obtaining and monitoring an ambulatory data collection [18]. This has challenged the creation of datasets with high-quality recordings from a large number of participants and diverse backgrounds. Furthermore, the intricacy of extracting meaningful insights from vast and heterogeneous ECG recordings demands methods that can learn and infer from these data in self-supervised ways. Such methods are also motivated by the need to address inherent biases and subject heterogeneity in bio-behavioral responses that hinder model performance and reliability. Another critical challenge is engineering models that can adeptly capture the structure and temporal dependencies of ECG activity without the need of scaling to large and over-parameterized models [19]. Such models increase the risk of overfitting and are not practical for mobile deployment. This involves striking a delicate balance between model complexity and efficiency. Given the multifaceted nature of these challenges, there is an imminent need for developing methodologies toward practical, robust, and reliable solutions to perform ubiquitous ECG analysis.

II. Contributions

This study addresses those challenges by proposing a framework to pre-train a model on large-scale public data to extract general-purpose vector representations for the ECG signal (Fig. 1). Our contributions can be summarized as follows:

- Our model, called **WildECG**, is trained on TILES [20], one of the largest publicly available biosignal data collections recorded in the wild, manifesting a wide range of variability and subject heterogeneity.
- To minimize the impact of noisy and biased data annotations, WildECG is trained in a self-supervised manner to identify distortions automatically induced on ECG samples at training time. This enables us to test performance on a variety of downstream tasks related to ECG.
- Our model incorporates a lightweight architecture based on state-space models that are efficient in modeling sequences with long temporal dependencies. By using a small number of parameters, WildECG further reduces the risk of overfitting and is suitable for mobile deployment. This is particularly important for user privacy and data security by handling all computations locally.

In sum, the proposed framework offers an efficient way to extract robust ECG representations that perform competitively across multiple downstream tasks, including human interaction modeling, affect recognition, and disease prediction. WildECG outperforms multiple architectures and training algorithms, while retaining discriminative information in low-resource settings and during minimal fine-tuning.

III. Background

A. Electrocardiography

Electrocardiography is a non-invasive technique for recording the electrical activity of the heart. The resulting signal, called the electrocardiogram (ECG), provides information about the functioning and structure of the heart, including the timing and regularity of its rhythm, significant underlying conditions or abnormalities, along with psychological states such as stress or emotional arousal. ECG has a characteristic structure that consists of specifically documented signatures: the P wave, the R peak and broadly the QRS complex, and the T wave, each of which corresponds to a distinct phase of the cardiac cycle. In addition to these waves, one can also identify intervals on the ECG that hold disease information, e.g., PR and QT intervals, as well as the RR interval which is used to calculate heart rate and heart rate variability [21], [22], [23], [24].

ECG is acquired through electrodes that are placed on the surface of the skin, usually at the chest or the wrists. The most common sensor configurations include 12-lead, 3-lead, or single-lead placements. The latter configuration, applied through wearable straps or wristbands, is a practical choice for measurements made in naturalistic settings, due to the ease of placement and minimal interference with the subject. To facilitate applications in both clinical and naturalistic domains, we restrict our study to single-lead ECG data.

B. Self-Supervised Learning

Self-supervised learning (SSL) is an emerging machine learning paradigm that provides an effective way to learn meaningful data representation without the need to acquire explicit labels. In contrast to supervised learning, which relies on labeled data, SSL leverages the intrinsic structure and relationships within the data to create pseudo-labels or tasks to learn from. As such, it holds several advantages over conventional supervised approaches for our task, as it avoids the need for annotations in large quantities, that would also constrain the scope of the model.

Most researchers distinguish two main types of SSL frameworks: 1) generative and 2) contrastive [25], [26], [27], [28], [29], [30]. The generative models (e.g., autoencoders) learn representations by reconstructing or generating the original data using masked or corrupted data as input, which defines their pretext task. Contrastive methods, on the other hand, train a model by contrasting the representations of semantically same data (e.g., two augmented views, positive samples) to other distant data (negative samples). Additional variants of SSL have also been proposed in the literature, including predictive [25], [28],

[29], property-based [27] or pretext learning [30] objectives, as well hybrid [27], [28] or cross-modal [26], [30] types.

Several promising approaches to SSL have been implemented, primarily in natural language processing (NLP) [31], [32] and computer vision [33], [34]. In the context of time series data, SSL has been used to learn representations for various tasks such as anomaly detection [35], frequency coupling [36], and masking [37]. Self-supervised learning of biosignals and ECG has already shown promising results in health applications and behavioral analysis [38]. We include a comprehensive review of related studies below.

C. Self-Supervised Learning on Biosignals

Authors of [35] used six signal transformations as pretext tasks to obtain ECG representation for downstream evaluation on affect recognition. They validated their obtained representation on four publicly available datasets. The results of this study in terms of model performance should be however interpreted with care since the introduced approach does not account for subject-specific biases in the training splits that could cause information leakage. In such an approach it is possible that two near-in-time samples will end up in training and test sets. The physiology across two near-in-time/consecutive samples (two 10-second long signal chunks) might be constant, especially in response to steady/constant/invariable stimuli.

In their analysis, the authors report that 1) the signal transformation parameters should be selected in such a way that the transformed signals are not too similar or too different from the original signal, because the model is unable to learn the representation, and that the most helpful transformations are scaling and time-warping; 2) multiple proxy tasks build better representation than a single proxy task; 3) the most prominent proxy-tasks are scaling and time-warping; 4) the representation learned on multiple datasets performs better than the representation learned on a single dataset.

Among other studies, Kachuee et al. [39] used ECG lead II signal of healthy and dysfunctional heart to recognize myocardial infarction. Their single-lead ECG SSL approach achieved results similar to a 12-lead ECG supervised approach. Dissanayake et al. [40] focused on representation of low-frequency signals from wearable devices. They utilized SimCLR [34] and simplified Inception [41] architectures in their pretext contrastive task. Their model performed competitively to the state-of-the-art results on publicly available datasets. Notably, their ablation studies showed that the model based on the representation is resistant to the loss of a large amount of training data, as well as to the loss of signal chunks.

Deldari et al. [42] proposed a cross-modal, contrastive, self-supervised learning approach to tackle several activity- and affect-related classification tasks. Using up to five modalities, their method outperformed several other approaches. According to their ablation analyses, as low as 10% of the labels in the downstream task is adequate for the model to provide satisfactory results. Zhang et al. [36] used contrastive learning with the assumption that time- and frequency-based embeddings are located near each other in the time-frequency space. On the other hand, [43] applied masking and attention to force the model to focus on relevant parts of the ECG signal. To build the signal representation, they trained on six public datasets and achieved competitive results on the AMIGOS [44] dataset for binary arousal and valence classification.

Recently, Wu et al. [45] introduced a multimodal SSL approach, based on transformers and common signal transformations, to recognize emotions from electrodermal activity, blood volume pulse, and temperature signals. Separate temporal convolutional encoders are used for each modality to extract low-level features. Then, a shared transformer-based encoder containing multi-head attention blocks combines these features to capture the complementary aspects of multimodal signals. The ablation studies showed that lack of electrodermal activity signal was reducing model performance the most, while temperature was less crucial. Jungo et al. [46] discovered that a multimodal transformer-based SSL model outperforms other techniques for imputing missing biobehavioral signals that change more frequently (but not in the case of monotonic signals). However, signals collected in naturalistic settings appear more challenging for the method than signals from in-lab studies. Ma et al. [47] also utilized transformer-based SSL, but their objective was a noninvasive blood pressure estimation. They designed seven transformations to learn the model hemodynamic information about the photoplethysmogram and four temporal-spatial transformations. Their model performed competitively to the SOTA, approaching clinical standards.

D. State-Space Models

State-space models (SSM) are a recently introduced category of deep neural networks [48] that were proposed to efficiently model long-term sequences, i.e., signals with either long duration or high sampling rate. Hence, the ECG modality constitutes a promising candidate for adopting a state-space model architecture. SSMs draw intuition from both convolutional and recurrent network architectures. The continuous-time SSM converts a 1-D input signal $u(t)$ into a latent state $x(t)$ before projecting it onto a 1-D output $y(t)$:

$$\begin{aligned} x'(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{1}$$

For discrete-time sequences that are sampled at step Δ, (1) can be mapped to the recurrence shown in (2), using the bilinear method [49] to convert A into an approximation $\bar{A}$:

$$\begin{aligned} x_k &= \bar{A}x_{k-1} + \bar{B}u_k \quad y_k = Cx_k + Du_k \\ \bar{A} &= (I - \Delta/2 \cdot A)^{-1}(I + \Delta/2 \cdot A) \\ \bar{B} &= (I - \Delta/2 \cdot A)^{-1}\Delta B \end{aligned} \tag{2}$$

Here $D = 0$ [50]. Equation (2) is a sequence-to-sequence map and the recurrence allows the discrete SSM to be computed like a recurrent network with hidden state $\bar{A}$. Equation (2) is also equivalent to a discrete convolution with kernel $\bar{K}$, as shown

in [50]:

$$\bar{K} = \left(C\bar{B}, C\overline{AB}, C\bar{A}^2\bar{B}, \ldots\right), \quad y = \bar{K} * u \tag{3}$$

Thus, SSMs can be viewed as special cases of convolutional and recurrent layers, inheriting their learning efficiency. Gu et al. [50] also contributed an efficient way of evaluating $\bar{K}$.

The *Structured State Space for Sequence Modeling* (S4) architecture was proposed in [50] to model sequences more efficiently than standard SSMs, also showing the capacity to capture long-range temporal dependencies. S4 is a particular instantiation of the SSM, where matrix A is parameterized as a diagonal plus low-rank (DPLR) that allows faster repeated computations. To capture long-range dependencies, S4 initializes A as HiPPO [51], so that the state x_k can memorize the history of the input u_k. At the same time, HiPPO preserves the DPLR form, as shown in [50]. Hence, the core S4 module is a linear, 1-D sequence mapping, however it handles high-dimensional features by defining independent copies of itself, and then mixing features with a position-wise linear layer. Nonlinear activations and dropouts in-between these layers provide the non-linearity of the whole architecture.

IV. Method

A. Pre-Processing

We adopt a universal approach in processing all ECG data used in pre-training and fine-tuning sessions. The following steps aim to alleviate the impact of discrepancies in different data collections, typically coming from the performed task, sampling rate, equipment noise, subject-specific and other artifacts. First, ECG signals are downsampled to 100 Hz and smoothed with a moving average kernel to remove powerline interference [52]. The specific sampling frequency provides a balance between preserving relevant information and reducing computational requirements. The majority of ECG datasets are recorded at 100 Hz or higher, and it has been reported [53], [54] that downsampling to 100 Hz does not compromise model performance. Next, we apply a high-pass Butterworth filter at 0.5 Hz. Finally, we perform subject-wise, z-score normalization. The signals are then segmented into non-overlapping windows of 10 seconds. During pre-training, where each sample is initially 15 seconds, 10-second samples are randomly extracted during training.

B. Signal Transformations

We base our proxy task for pre-training on predicting various signal transforms applied to ECG samples. To this end, we have implemented a Python module of ECG-tailored transformations that we share publicly.[1] The ECG-augmentations library [55] currently includes versatile transforms of multi-lead ECG signals. Implemented augmentations include:

- *Masking:* We support random masking or masking of PR and QRS intervals, whereas the user can also specify the ratio of intervals to be masked. Detection of R peaks is done using the NeuroKit2 library [52].
- *Cropping:* Random (r) crop of an ECG sub-sequence given the desired length λ: $s' = s[r : r + \lambda]$.
- *Noise:* We support both additive white noise and random wander, with adjustable signal-to-noise ratio (SNR).
- *Permutation:* Each ECG signal is divided into $m \leq 10$ segments, which are shuffled, i.e., by randomly perturbing their temporal order. Each segment has a set minimum length of 10% the total signal length.
- *Time Warping:* Randomly selected segments of the original ECG are stretched or squeezed along the temporal axis, through interpolation. The output signal is cropped or zero-padded when stretched or squeezed, respectively.
- *Scaling:* The ECG magnitude is multiplied by a random scalar $0 < \alpha < 5$: $s'[n] = -\alpha s[n], n = 0, \ldots, N$.
- *Inverting:* Implemented by negating the input signal along the temporal axis: $s'[n] = s[N - n], n = 0, \ldots, N$.
- *Reversing:* Simply implemented by scaling the input signal using $\alpha = -1$: $s'[n] = -s[n], n = 0, \ldots, N$.

C. Pre-Training Objective

Most SSL studies apply either masked sample reconstruction or contrastive learning objectives to pre-train their respective models. Since there is no established training algorithm for physiological signals, WildECG considers elements from both SSL approaches. Our objective aims to identify which signal transformations are applied to a sample ECG, where each signal is augmented randomly using at most four out of all the available transforms. Each transform is selected based on a set probability, so it is possible that some samples are input without any augmentation. We formulate this task as a multi-label classification task with nine classes (eight possible transformations plus the original signal).

This task draws from both predominant SSL approaches. Our first motivation comes from masked reconstruction objectives by including masking augmentations in our pre-training framework. These include both masking of random signal patches and of specific ECG intervals. Second, we follow the contrastive learning paradigm in the sense of modeling the impact of induced augmentations in the data. However, we intentionally choose to predict applied transformations over the conventional contrastive approach, in which we would identify the similarity of two distorted samples of the same input. The reason is that the latter objective would focus on invariant ECG features that are primarily subject-dependent. On the other hand, identifying distortions is intuitive for our scope, since the model focuses on ECG abnormalities that could potentially hold diagnostic information.

D. Model Architecture

Our proposed model inherits the S4 model as the backbone architecture, as it features critical elements that are desirable in ECG analysis. As mentioned before, S4 has demonstrated promising performance in modeling long-range sequences with dependencies over thousands of timesteps, which is a current limitation of state-of-the-art models like the Transformer [56]

[1] [Online]. Available: https://github.com/klean2050/ecg-augmentations

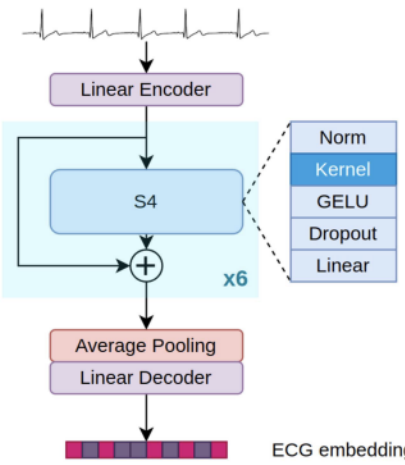

Fig. 2. The architecture of the proposed ECG backbone model, following a simple version of the original S4 [50]. The model consists of six S4 blocks, connected through residual connections. Linear classifiers are attached on top for both pre-training and fine-tuning tasks.

TABLE I
OVERVIEW OF THE STUDY'S DATASETS

Dataset	ECG Setting	# Subjects	# Classes
TILES [20]	24-h monitoring	200	N/A
PTB-XL [59]	clinical acquisition	18869	5
LUDB [60]	clinical acquisition	200	2
WESAD [61]	activity engagement	15	3
CASE [62]	video watching	30	⋆
AVEC-16 [21]	dyadic interaction	27	⋆
SWELL-KW [22]	workplace stress	25	2

⋆ Denotes regression.

architecture. ECG is a sequence of that type, with its sampling rate ranging from 100 to 1000 Hz. Also, S4 is implicitly a continuous-time model, making it well-suited to waveform signals. Indeed, prior work [57] has shown that variants of S4 provide excellent performance in classifying cardiovascular conditions based on controlled ECG data.

Here we employ a simplified version of the original S4, consisting of a linear encoder, six S4 blocks, and a linear decoder. Each block consists of a Layer Normalization module, the $\bar{K}$ estimation module, a GELU [58] activation, Dropout and an output projector. The blocks are connected with residual connections, as shown in Fig. 2. The input and output dimension is set to 256 and dropout layers of 20% are applied. For the pre-training phase, we adjust a linear layer to the decoder output and replace it during fine-tuning.

V. EXPERIMENTS

The training of the proposed framework consists of 2 stages: First, we pre-train our model using the large ECG database from the TILES study. Then, we fine-tune our model in each of the downstream tasks, in 2 modes: either tuning all parameters (full model), or just the classification head (projector).

Below we share details about the datasets used in this study (see Table I). Our experimentation covers 7 public ECG datasets and targets settings where the 1-lead ECG modality is prominent and the evaluation criteria are clearly defined. We thus omitted datasets such as DEAP [63], AMIGOS [44], or DREAMER [64] as EEG-oriented, and also medical datasets that depend heavily on full, 12-lead ECG recordings.

A. Pre-Training: TILES Dataset

Tracking Individual Performance with Sensors (TILES) [20], [65] is a research project that has collected multimodal datasets for the analysis of stress, task performance, behavior, and other factors pertaining to professionals in a high-stress workplace environment. Biological, environmental, and contextual data were collected from hospital nurses, staff, and medical residents both in the workplace and at home over a ten week period. Labels of human experiences, such as stress, anxiety, and affect, were collected using psychologically validated questionnaires which were administered at different times.

In the present study, we use the ECG data from the publicly available TILES 2018 dataset [20] to pre-train a general-purpose ECG model. Each participant had their ECG recorded for 15 seconds every 5 minutes during their work hours, for a total of 10 weeks. There were 213 participants in total, 200 of whom had agreed to wear a bioshirt that enabled high quality ECG data collection, making the aggregate number of samples conducive to pre-train a large ECG representation model. Since ECG was recorded in an ubiquitous manner, we apply a quality check by measuring the shape distance of heartbeats to the average heartbeat for each 15-second session. All days for which the total detection rate is lower than 90% are discarded, leading to approximately 275,000 ECG samples from 168 individuals for the following experiments.

B. Fine-Tuning: Ubiquitous Sensing

1) AVEC-16 Multimodal Affect Recognition Sub-Challenge: The multimodal affect recognition sub-challenge (MASC) of AVEC-16 [21] stems from the REmote COLlaborative and Affective interactions (RECOLA) dataset [66]. RECOLA included continuous multimodal recordings during dyadic interactions via video conferences. The complete dataset contains audio, visual, and physiological information from 27 French-speaking participants. The single-channel ECG data used in this work were sampled at 250 Hz and subsequently filtered using a band-pass filter at 3-27 Hz. The labels are continuous ratings for arousal and valence at 40 ms intervals, throughout the first five minutes of the complete recordings.

2) SWELL-KW: This dataset [22] aimed at analyzing employees' emotional states and workplace stress under three scenarios: *normal*, in which participants performed various office tasks for 45 minutes, *time-pressure*, in which participants had only 30 minutes to complete the same tasks, and *interruption*, in which they were also interrupted by emails and messages. ECG signals were collected from 25 participants using the TMSI MOBI device at a sampling rate of 2048 Hz. At the end of each scenario, participants were asked to report their valence, arousal, and also other states, such as stress.

3) WESAD: The dataset for WEarable Stress and Affect Detection (WESAD) [61] contains ECG data from 15 participants.

RespiBAN Professional sensors were used to collect ECG at a sampling rate of 700 Hz. The goal was to study four different affective states (neutral, stressed, amused, and meditated). First, 20 minutes of neutral condition data were collected, during which participants were asked to do normal activities. Then participants watched 11 funny video clips (amusement) and went through public speaking and arithmetic tasks (stress). Finally, they went through a guided meditation session of 7 minutes. Upon completion of each trial, labels for the affect states were collected using 9-scale PANAS.

C. Fine-Tuning: Controlled Sensing

1) PTB-XL: The PTB-XL dataset [59] is a set of 21799 clinical, full 12-lead ECGs from 18869 patients of 10 s length. The raw waveform data were annotated by up to two cardiologists, who assigned potentially multiple ECG statements to each record. The waveform data underlying the PTB-XL ECG dataset were collected with devices from Schiller AG over the course of nearly seven years between October 1989 and June 1996. In total 71 different ECG statements conform to the SCP-ECG standard and cover diagnostic, form, and rhythm statements. The dataset is complemented by extensive metadata on demographics, infarction characteristics, diagnostic statements, and annotated signal properties.

2) LUDB: The Lobachevsky University Electrocardiography Database (LUDB) [60] is a 12-lead ECG dataset with annotated boundaries and peaks of P, T and QRS waves. It consists of 200 10-second ECG signals at 500 Hz, representing different morphologies, out of which we only use the first lead, to comply with our framework. The ECG records were collected from healthy volunteers and patients of the Nizhny Novgorod City Hospital during 2017–2018. The patients had various cardiovascular diseases while some of them had pacemakers. Cardiologists annotated each record with the corresponding diagnosis. For this study, we consider the task of identifying sinus rhythm against a super-set of different abnormalities.

3) CASE: The Continuously Annotated Signals of Emotion (CASE) dataset [62] contains data from 30 participants collected in laboratory conditions. During the experiment, participants watched a series of 8 video stimuli and continuously annotated their emotions in a two-dimensional arousal-valence space using a joystick interface developed by the researchers. Additionally, a two-minute long blue-screen video served as an in-between resting phase. The 1-lead ECG data were collected at 1000 Hz using Thought Technology SA9306 sensors, and affect annotations were collected at 20 Hz.

D. Implementation Details

We pre-train WildECG for 100 epochs on the TILES data using a batch size of 256 samples and an AdamW optimizer with a 0.001 learning rate. A linear layer is used to map the ECG embeddings to the transform classes. We checkpoint the resulting model of the last epoch and apply it to a set of downstream tasks to evaluate the learned representations. For each task, the respective ECG data are extracted and processed akin to the TILES data (see Section IV-A), whereas the additive linear layer is replaced by a 2-layer MLP classifier that maps the pre-trained embeddings to the target space.

We evaluate each task with 5-fold cross-validation in primarily subject-agnostic settings. *Subject-agnostic* refers to the setting where test splits do not contain samples from subjects of the training splits, whereas *mixed-subject* denotes the opposite. We use a batch size of 256 samples in all experiments except LUDB (32 samples). The learning rate is tuned to each dataset separately, within $\{0.0001, 0.0005, 0.001\}$. All datasets are trained for a maximum of 200 epochs with early stopping based on validation loss. Model checkpoints are selected based on the highest F1-macro or correlation coefficient in cross-validation, and lowest validation loss for PTB-XL and AVEC-16, which have specified validation and test splits.

TABLE II
OVERVIEW OF COMPARABLE STUDIES ON ECG MODELING

Study	Dataset	Features	Classifier	Evaluation
[21]	AVEC-16	ECG statistics	Linear SVM	Arousal: 0.271
[21]	AVEC-16	ECG statistics	Linear SVM	Valence: 0.153
[61]	WESAD	HR, HRV	LDA	0/1 Stress: 0.813
[61]	SWELL	HR, HRV	LDA	0/1 Stress: 0.560
[67]	AVEC-16	ECG statistics	Linear SVM	Arousal: 0.118
[67]	AVEC-16	ECG statistics	Linear SVM	Valence: 0.085
[68]	WESAD	Spectrogram	2D-CNN	0/1 Stress: 0.794
[69]	WESAD	1-lead raw ECG	Transformer	0/1 Stress: 0.697
[69]	SWELL	1-lead raw ECG	Transformer	0/1 Stress: 0.588
[70]	WESAD	1-lead raw ECG	ECGNet	0/1 Stress: 0.857
[70]	SWELL	1-lead raw ECG	ECGNet	0/1 Stress: 0.688
[71]	WESAD	HR, HRV	RBF SVM	0/1 Stress: 0.818
[72]*	SWELL	*Multimodal*	Kernel SVM	0/1 Stress: 0.641

Tasks on regression are CCC, Classification F1-macro (*Accuracy).

TABLE III
DOWNSTREAM PERFORMANCE ON **AVEC-16** DATASET (EVAL SPLIT)

Model	Training	Arousal CCC	Valence CCC
Linear SVM [21]	–	0.271	0.153
Linear SVM [67]	–	0.118	0.085
Baseline S4	full model	0.328	0.162
WildECG (ours)	full model	**0.356**	**0.303**
WildECG (ours)	projector	0.346	0.289

VI. RESULTS

Below we present the downstream evaluation of WildECG. Our objective is to highlight the performance of the proposed model when employed both as a backbone and as a feature extractor, compared to training supervised classifiers. Wherever possible, we compare our model with available studies, a summary of those is shown in Table II.

A. Ubiquitous Sensing

AVEC-16: Table III includes results for the AVEC-16 dataset, quantified using the Concordance Correlation Coefficient (CCC). S4 persistently outperforms the scores reported from all prior studies, which rely on knowledge-based ECG features and conventional classifier architectures. For arousal estimation, WildECG achieves a state-of-the-art CCC of 0.356

TABLE IV
DOWNSTREAM PERFORMANCE ON **WESAD** DATASET (3-WAY ACTIVITY)

Model	Training	Mixed-subject		Subject-agnostic	
		Accuracy	F1-macro	Accuracy	F1-macro
AdaBoost [61]	–	–	–	0.617 (–)	0.525 (–)
LDA [61]	–	–	–	**0.663** (–)	0.560 (–)
1D-CNN [35]	projector *	0.969 (–)	0.963 (–)	–	–
Baseline S4	full model	0.956 (0.031)	0.955 (0.032)	0.489 (0.089)	0.410 (0.097)
WildECG (ours)	full model	**0.978** (0.028)	**0.978** (0.028)	0.644 (0.044)	**0.592** (0.058)
WildECG (ours)	projector	0.742 (0.044)	0.721 (0.064)	0.600 (0.089)	0.524 (0.075)

* Pre-training includes WESAD. Standard deviation among folds is included in parentheses. The bold values indicate the best numerical result.

TABLE V
DOWNSTREAM PERFORMANCE ON **WESAD** DATASET (STRESS VS NORMAL–SUBJECT-AGNOSTIC)

Model	Training	Accuracy	F1-macro
LDA [61]	–	0.854	0.813
2D-CNN [68]	full model	0.824	0.794
Transformer [69]	full model	0.804	0.697
ECGNet [70]	full model	0.908	0.857
SVM [71]	–	0.811	0.818
Baseline S4	full model	0.900	0.899
WildECG (ours)	full model	**0.967**	**0.966**
WildECG (ours)	projector	0.900	0.891

The bold values indicate the best numerical result.

when fully fine-tuned and 0.346 when only the projector is trained. In both cases it outperforms the S4 variant that is trained from scratch. Similar results are obtained for valence, where our proposed model surpasses 0.3 CCC.

WESAD: Table IV presents detailed results for WESAD, evaluated in both mixed-subject and subject-agnostic settings. Here, the objective is to identify the type of activity the subject performs out of three scenarios: baseline, stress, and amusement. For the mixed-subject setting, we compare our performance with Sarkar and Etemad [35] where we observe marginal improvements of 1% to 1.5% in F1-macro. We should note that the results are close to absolute correct accuracy, which is attributed to the temporal correlation that each subject and each recording inherits. This is evident by observing the drop of 23-25 percentage points (pp) when we freeze the pre-trained encoder, as well as the drop of more than 30 pp that the subject-agnostic setting induces. Nonetheless, WildECG outperforms the literature in obtained F1-macro, reaching 59.2%, with an accuracy of 64.4%. In this case, the pre-training mechanism is critical, since the S4 baseline cannot reach an accuracy better than random chance.

In addition to evaluating the 3-way condition in WESAD, we also assess the binary task of stress versus the two other conditions and report related results in Table V. WildECG achieves very high accuracy, reaching 96.6% F1-macro and outperforming all previous studies by a large margin. Both frozen and fully fine-tuned models outperform convolutional networks and transformer encoders, achieving 11–17% better F1-macro despite having fewer parameters.

SWELL-KW: Finally, we present our results on SWELL-KW in Table VI. Here we only share results for the subject-agnostic setting, since the mixed-subject one quickly overfits to perfect 100% accuracy, with similar results shown in [35]. For SWELL-KW, we evaluate three different binary cases: valence and arousal estimation, both binarized at the mean of the obtained values, and stress, as indicated again by the activity performed by the subject. For this task, N refers to the normal condition whereas T and I represent the stress conditions. WildECG performs on par with the S4 trained from scratch, with smaller variations among folds in arousal. However, most studies evaluate their methods on the latter task of stress estimation, achieving more than 75% accuracy and close to 69% F1-macro. Our models provide competitive performance with these studies, with the pre-trained model reaching state-of-the-art 71.1% F1-macro.

B. Controlled Sensing

Here we report our performance scores for the datasets described in Section V-C. We begin with the CASE dataset which incorporates target variables of affect. Table VII contains results for regression on arousal (A), valence (V), and anxiety (N) levels, where anxiety is defined as $N = A(1 - V)$, as proposed in [71]. The prediction results show strong performance for WildECG, which outperforms both the S4 baseline and the frozen variant with a substantial margin in both subject-agnostic and mixed-subject settings. While the baseline shows higher performance for arousal, WildECG is better on valence, gaining 0.21 CCC from baseline S4. This difference is also reflected and magnified in the anxiety measures. We further observe that, even though WildECG embeddings on their own offer limited performance improvements, our method leads to more consistent predictions in all tasks, with substantially lower variance than the baseline. To the best of our knowledge, no previous study provides continuous estimates of affect variables from ECG in CASE.

Although our system is not trained on clinical settings and data, we evaluate its performance as an out-of-distribution task on PTB-XL, one of the largest clinical ECG testbeds that is publicly available. Unfortunately, there is no extensive research assessing single-lead ECG systems for disease diagnosis. Moreover, many studies in cardiology report prominent disease biomarkers on several leads of a clinical ECG recording [73], [74]. With that premise, we compare our single-lead results with 12-lead ECG systems (Table VIII). An S4 architecture similar to ours recently reported state-of-the-art performance [57] for 12-lead PTB-XL. Here we effectively benchmark the performance drop of 1-lead S4 to 10% AUROC, reaching 83.2%. We also demonstrate that when pre-trained on TILES, our model can improve upon this baseline by about 1.3 pp in AUROC and 2.3 pp in F1-macro. We also highlight that, possibly due to the scale of the dataset, the frozen model substantially under-performs in this 5-way classification task, with a 13% drop in F1-macro. Further, in Table IX, we evaluate our model in LUDB, a much smaller medical dataset of various cardiac conditions. Despite the low-resource setting, WildECG distinguishes between healthy and non-healthy recordings, with a mean accuracy of 91.5% and F1-macro of about 90%, demonstrating data efficiency.

TABLE VI
DOWNSTREAM PERFORMANCE ON **SWELL-KW** DATASET FOR THE SUBJECT-AGNOSTIC SETTING (BINARY CLASSIFICATION)

Model	Training	Valence > mean		Arousal > mean		Stress (N vs T/I)	
		Accuracy	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro
Kernel SVM [72]	–	–	–	–	–	0.641 (–)	–
Transformer [69]	full model	–	–	–	–	0.581 (–)	0.588 (–)
Deep ECGNet [70]	full model	–	–	–	–	**0.755** (–)	0.688 (–)
S4 baseline	full model	0.598 (0.098)	0.560 (0.105)	0.743 (0.153)	**0.731** (0.148)	0.680 (0.075)	0.643 (0.107)
WildECG (ours)	full model	**0.629** (0.050)	**0.623** (0.050)	**0.751** (0.064)	0.704 (0.077)	0.660 (0.080)	0.637 (0.094)
WildECG (ours)	projector	0.607 (0.118)	0.560 (0.157)	0.731 (0.088)	0.698 (0.089)	0.740 (0.102)	**0.711** (0.127)

Standard deviation among folds is included in parentheses. * Not only ECG. The bold values indicate the best numerical result.

TABLE VII
DOWNSTREAM PERFORMANCE ON **CASE** DATASET (REGRESSION TASK)

Model	Training	Mixed-subject			Subject-agnostic		
		Arousal CCC	Valence CCC	Anxiety CCC	Arousal CCC	Valence CCC	Anxiety CCC
Baseline S4	full model	0.249 (0.121)	0.231 (0.133)	0.292 (0.168)	0.198 (0.085)	0.162 (0.066)	0.310 (0.159)
WildECG (ours)	full model	**0.391** (0.047)	**0.439** (0.081)	**0.565** (0.089)	**0.253** (0.064)	**0.351** (0.086)	**0.424** (0.049)
WildECG (ours)	projector	0.226 (0.098)	0.114 (0.066)	0.219 (0.112)	0.202 (0.034)	0.123 (0.053)	0.205 (0.058)

Standard deviation among folds is included in parentheses. The bold values indicate the best numerical result.

TABLE VIII
DOWNSTREAM PERFORMANCE ON **PTB-XL** DATASET (SUP-DIAG TASK)

Model	Training	AUROC	F1-macro
12-lead LSTM [54]	full model	0.927	–
12-lead Inception-1D [54]	full model	0.921	–
12-lead Transformer [57]	full model	0.887	–
12-lead S4 [57]	full model	0.931	–
1-lead S4 baseline	full model	0.832	0.457
WildECG (ours)	full model	**0.845**	**0.480**
WildECG (ours)	projector	0.815	0.346

The bold values indicate the best numerical result.

TABLE IX
DOWNSTREAM PERFORMANCE ON **LUDB** DATASET (BINARY TASK)

Model	Training	Accuracy	F1-macro
1-lead S4 baseline	full model	0.770 (0.040)	0.618 (0.141)
WildECG (ours)	full model	**0.915** (0.064)	**0.894** (0.082)
WildECG (ours)	projector	0.855 (0.062)	0.792 (0.094)

Standard deviation among folds is included in parentheses. The bold values indicate the best numerical result.

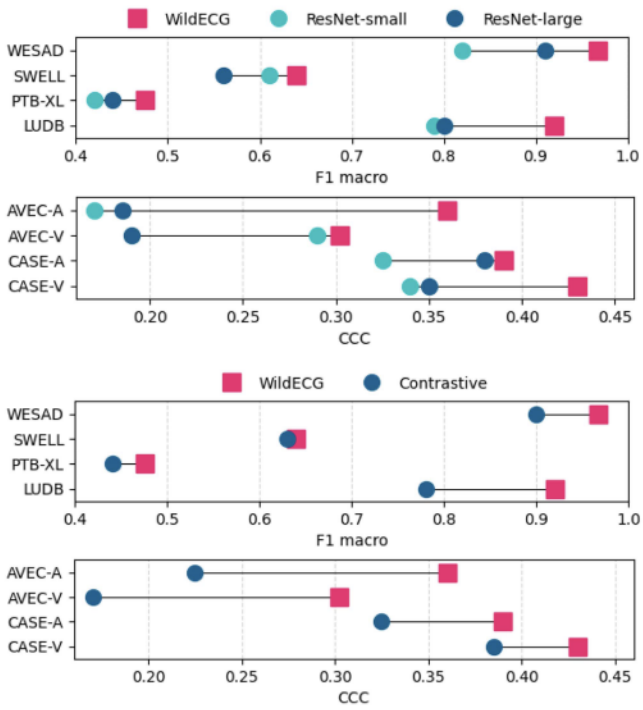

Fig. 3. Top two plots: Backbone architecture comparison. Bottom two plots: Pre-training algorithm comparison. V denotes valence and A arousal. WESAD, SWELL refer to the binary experiment. Classification tasks are measured with F1 macro and regression tasks with CCC.

VII. DISCUSSION

A. Pre-Training Settings and Complexity

Our proposed framework incorporates several design parameters that contribute to positive experimental performance. In this section, we conduct a close inspection of each of these design elements by probing and comparing alternative approaches in the literature. Specifically, we compare WildECG to a network that uses a 1D ResNet [19] backbone, in order to assess the additive value of the selected architecture. ResNets have shown to be superior to other modeling approaches in a recent review on ECG signals [75]. We create two versions, a ResNet-large of 14.7 M parameters that includes 10 residual blocks and an input size of 64 filters, and a ResNet-small of 923 K parameters by reducing filter sizes to $1/4$ of ResNet-large. We note that WildECG holds 313 K parameters in total. Both networks are trained like WildECG and the obtained results are in Fig. 3 (top). We observe that our proposed S4 model clearly outperforms both ResNet variants, except for SWELL-KW where the accuracy is similar. On the other hand, increasing the parameters of the ResNet model provides limited benefits to performance boost.

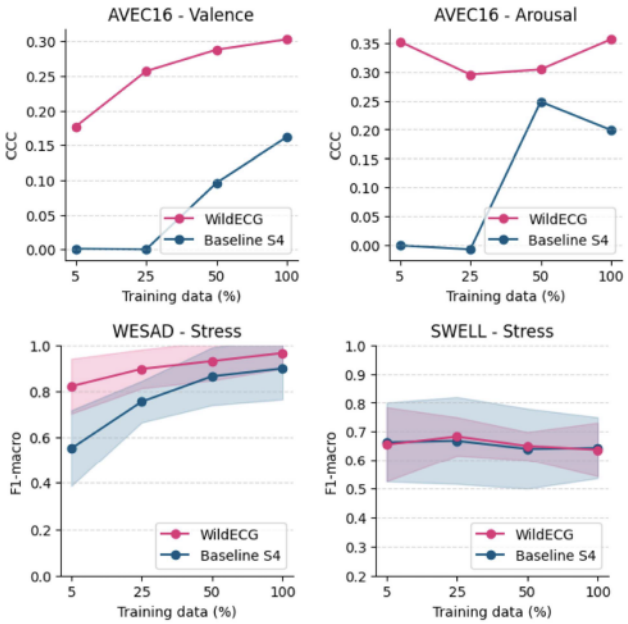

Fig. 4. Low-resource model performance for AVEC-16, WESAD, and SWELL-KW: **Ubiquitous sensing**. Horizontal axis is not drawn to scale.

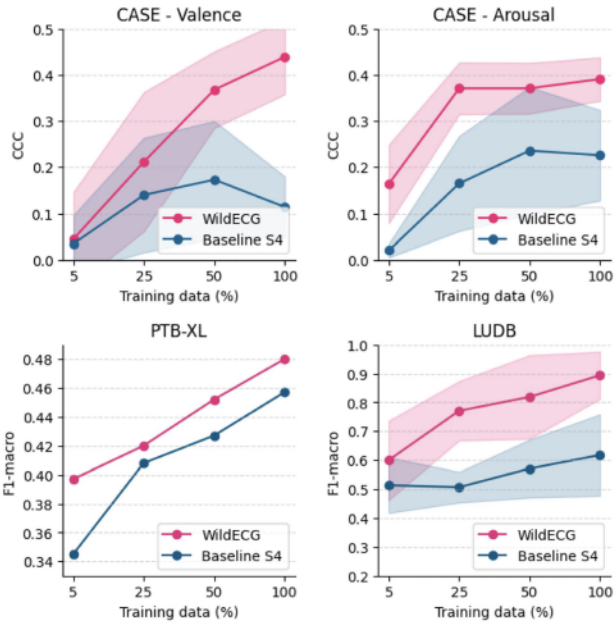

Fig. 5. Low-resource model performance for CASE, PTB-XL, and LUDB: **Controlled sensing**. Horizontal axis is not drawn to scale.

We further compare the ECG representations of our pre-training algorithm with those obtained using one of the standard contrastive learning approaches [34]. For this purpose, we train an identical to WildECG network with the alternative objective. The results on the same downstream tasks are shown in Fig. 3 (bottom). The chosen objective shows clear advantages over the contrastive one, as seen by the persistently improved performance in most evaluation cases. These results support our choice of pre-training objective as it is more suitable as and intuitive to ECG signal analysis. In future work, it is worth investigating tuning the baseline contrastive objective in order to reduce subject bias in the representations, for which Cheng et al. [76] have provided proof of concept by configuring subject-specific negative pairing.

B. Low-Resource Scenarios

Thus far we have quantified our model's superiority against other architecture choices and across diverse applications. Herein we evaluate WildECG in low-resource settings, where we randomly restrict the number of training samples in each of the downstream tasks. Figs. 4 and 5 contain a respective graph for each dataset, with the horizontal axis denoting the percentage of the training samples that were actually used. In this section, we include the same subset of labels that we used previously. We observe that WildECG achieves small or even negligible performance loss in most cases compared to the baseline S4 network. For AVEC16, we are able to retain state-of-the-art performance even with 5% of the training data, while the baseline fails to converge. Similarly, model pre-training alleviates the performance drop for WESAD and PTB-XL, whereas for LUDB the baseline never reaches performance substantially above chance. As mentioned earlier, SWELL-KW is associated with unnoticeable performance differences between WildECG and the S4 baseline, while our model does exhibit smaller variance in its predictions.

C. Method Interpretability

Although the proposed model shows state-of-the-art performance across different tasks, it is still unclear how the ECG is processed within the model and whether the derived representations have any feature-wise or semantic correspondence, which is critical for adoption in clinical practice. The field of explainable AI [77] has already made substantial progress in deciphering what deep learning models can learn. For the scope of this study, we provide post-hoc interpretability measures by mapping the high-dimensional ECG embeddings to a 2D space and analyzing between-sample distances. To reduce the dimensionality we use the t-SNE algorithm [78]. Since the TILES dataset incorporates a relatively large number of samples, we randomly select 10% of its data (every 10th sample of each subject) to avoid over-plotting.

We first investigate whether WildECG incorporates subject-specific biases in its representations. To that end, we provide t-SNE visualizations colored by subject ID, in Fig. 6. As for the pre-training data, due to the number of samples TILES embeddings appear rather mingled, without visible clusters. However, the mean euclidean distance for intra-subject samples is found to be lower than the mean inter-subject distance, i.e., 7.02 ± 1.24 vs. 10.55 ± 2.59, respectively; p-value $1.77 \cdot 10^{-27}$ for paired t-Student test. This indicates that samples related to the same subject are closer to each other rather than to samples from other subjects. On the other hand, for CASE, WESAD, and SWELL-KW, the embeddings form well-separated, participant-related clusters. This comes in spite of the subject-wise standardization

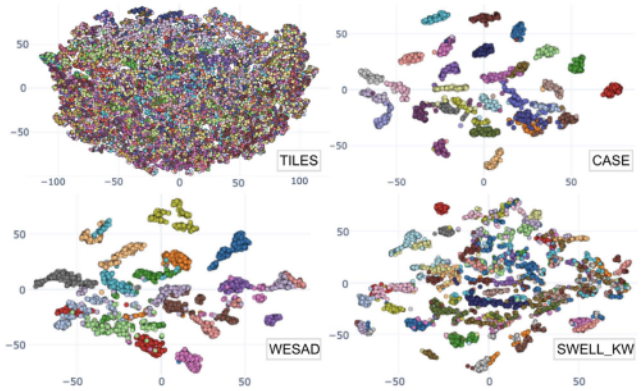

Fig. 6. T-SNE visualizations of WildECG embeddings on the 2D space for TILES (downsampled to 10%), CASE, WESAD and SWELL-KW datasets, colored by subject ID to reveal subject-specific bias.

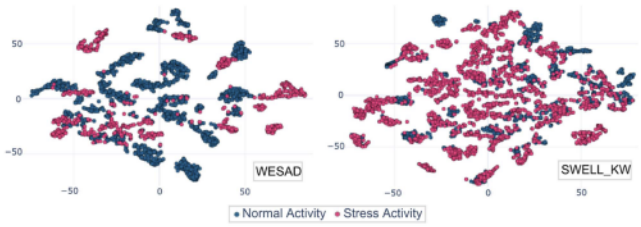

Fig. 7. T-SNE visualizations of WildECG embeddings on the 2D space for WESAD and SWELL-KW colored by type of activity.

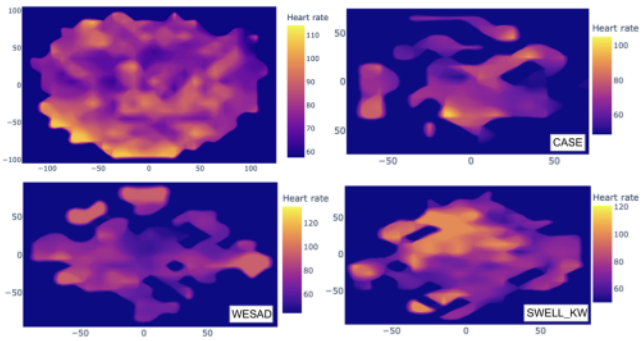

Fig. 8. Distribution of heart rate values in 2D t-SNE space for WildECG embeddings of TILES (downsampled), CASE, WESAD, and SWELL-KW samples. Brighter colors indicate higher heart rate.

and implies that the model indeed learns strong subject- or sensor-specific characteristics from the ECG. As for WESAD, each participant typically expands two clusters, grouped by heart rate value of the respective ECG. As shown in Fig. 7, these clusters indeed correspond to the *stress* and *no-stress* activities performed and in some cases form super-clusters across subjects.

Next, we focus on whether the WildECG representations retain cardiac information, in this case heart rate (HR). Ground truth HR values were obtained from the filtered 10-second ECG samples using the NeuroKit2 [52] library. The mean HR values were grouped into bins of 10 within the acceptable range of human HR [79], i.e., from 40 to 210 bpm. In Fig. 8, for TILES

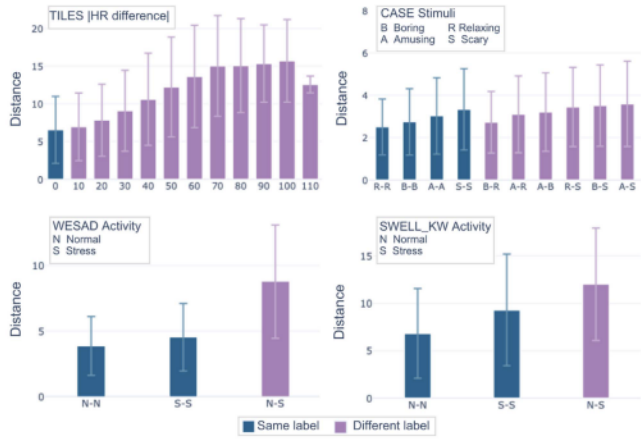

Fig. 9. Average Euclidean distance of within-participant samples with respect to: Absolute heart rate difference for TILES (downsampled), stimuli for CASE, and activity for WESAD and SWELL-KW. Colors indicate distances between samples with same or different labels.

dataset we observe a left-to-right transition from low to higher HR values. On the contrary, CASE embeddings, which reflect strong subject biases (Fig. 6), do not present any generalized HR patterns. The transition across HR levels happens though within subject-specific clusters. Visualization for WESAD shows that the mixed-subjects' groups represent higher mean HR values, which are also associated with stress activity (Fig. 7). A similar pattern is also present in SWELL-KW, where samples from different subjects related to higher HR values and stress activity are close to each other.

To assess the model's capability of capturing physiology that is characteristic of the context (activity or stimulus) we compute the Euclidean distance between high-dimensional embeddings of different context (Fig. 9). For each subject, we determine the average distance between samples with the same label and the average distance between samples with different labels. The most notable difference between within-context and across-context samples can be observed for WESAD and SWELL-KW, where samples related to stressful activities are distant from other activities. As for CASE, the distinction between within-scary sample distance and distance to other-context samples is not significant, but remains noticeable. Further analysis of TILES data reveals that samples with similar HR values are situated closer to each other, while those with more significant differences in HR are positioned farther apart. This observation verifies that WildECG embeddings effectively preserve the cardiac information.

D. Limitations and Challenges

A robust and general-purpose representation model for the ECG is an important step towards expanding scientific research and clinical translation, notably for broad dissemination of smart and ubiquitous health applications. Equal importance however should be given to the limitations of these models, from the methodological aspects of evaluation, to fundamental questions regarding the applicability of physiological measures in estimating complex human conditions. Computational modeling of

human behavior and physiology currently lacks standardized protocols and metrics that would ensure the reproducibility and validity of the obtained results. Consequently, our study also diverged from other comparable studies on how each dataset is set up for evaluation. However, we hope that it contributes a cohesive and comprehensive testbed for studies of similar scope to evaluate their approaches.

An important characteristic of physiological signals that influences the evaluation protocol is the inter-subject variability, which challenges the application of transfer learning. Indeed, multiple studies report that machine learning models trained on a specific dataset rarely generalize to other datasets and settings [70], [80]. Even within a single dataset, subject bias could prevent evaluation on unseen subjects [35]. In our study, we demonstrate that pre-training with a large-scale ECG dataset in a self-supervised way can help alleviate these issues. However, our feature visualizations reveal that the learned features still reflect such biases, e.g., by forming subject-specific clusters. Hence, adopting specialized objectives to eliminate this bias is an important direction for future work.

Taking a step back, it is crucial to underscore the limitations of standalone measures like the ECG to solely estimate the range of human conditions. It is well known that emotional states are heavily influenced by the social and environmental context [81], [82], in a way that a single-dimensional signal cannot reflect. We highlight that models like WildECG should be adopted in a holistic perspective that takes into consideration multiple views of human behavior and contextual information. For example, fusing information from multiple physiological and behavioral signal measures like electrodermal activity, speech [21], and human activity [72] has provided better performance than using ECG alone on the same datasets. Incorporating WildECG in a multimodal sensing framework is another direction of future work to be pursued.

VIII. Conclusion

Ubiquitous sensing and monitoring are already transforming digital health and well-being with new, on-demand services. Hence there is an unmet need to address challenges related to the analysis of the resulting human bio-behavioral states. In this study, we propose WildECG, a versatile AI framework for ECG representation learning. By utilizing a large, diverse corpus of biosignals collected in the wild, along with a state-of-the-art state-space network and pre-training algorithm, we demonstrate competitive performance on the tasks of estimating human affect, dimensional emotion, stress levels, as well as pathological markers. We further quantify the contributions of our design factors and verify model robustness in low-resource settings. The conducted qualitative analysis reveals that WildECG indeed incorporates explainable and tractable insights related to the ECG structure and features that could prove beneficial for researchers as well as clinicians.

At the same time, the potential of ubiquitous sensing in digital health and well-being is advancing rapidly. As the field continues to expand, several directions for future research could emerge from the proposed framework. Despite the competitive performance demonstrated by WildECG, there is still room to further investigate the model biases and hence its generalization across populations, demographics, and cultural contexts. Research could also develop adaptive learning techniques to fine-tune large models to personalized needs, especially since the lightweight architecture of WildECG can be easily deployed in mobile devices and run offline. Lastly, a promising direction is the combination of pre-trained ECG representations with other behavioral modalities. Several technical challenges are to be addressed in this realm, most importantly the sparsity of target information across signals.

Acknowledgment

The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning for mental health research and clinical applications [Perspectives]," *IEEE Signal Process. Mag.*, vol. 34, no. 5, pp. 196–195, Sep. 2017.

[2] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, 2018.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[5] L. Dai et al., "A deep learning system for detecting diabetic retinopathy across the disease spectrum," *Nature Commun.*, vol. 12, 2021, Art. no. 3242.

[6] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Med.*, vol. 28, pp. 31–38, 2022.

[7] T. Feng, B. M. Booth, B. Baldwin-Rodríguez, F. Osorno, and S. Narayanan, "A multimodal analysis of physical activity, sleep, and work shift in nurses with wearable sensor data," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 8693.

[8] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Proc. IEEE Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 671–676.

[9] S. Saeb et al., "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study," *J. Med. Internet Res.*, vol. 17, no. 7, 2015, Art. no. e175.

[10] G. Thatte et al., "KNOWME: An energy-efficient, multimodal body area network for physical activity monitoring," *ACM Trans. Embedded Comput. Syst.*, vol. 11, 2012, Art. no. 48.

[11] S. Saganowski, "Bringing emotion recognition out of the lab into real life: Recent advances in sensors and machine learning," *Electronics*, vol. 11, no. 3, 2022, Art. no. 496.

[12] D. Kunc, J. Komoszyńska, B. Perz, S. Saganowski, and P. Kazienko, "Emognition system - wearables, physiology, and machine learning for real-life emotion capturing," in *Proc. IEEE 11th Int. Conf. Affect. Comput. Intell. Interaction Workshops*, 2023, pp. 1–3.

[13] B. M. Booth, T. Feng, A. Jangalwa, and S. S. Narayanan, "Toward robust interpretable human movement pattern analysis in a workplace setting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2019, pp. 7630–7634.

[14] B. Rim, N.-J. Sung, S. Min, and M. Hong, "Deep learning in physiological signal data: A survey," *Sensors*, vol. 20, no. 4, 2020, Art. no. 969.

[15] A. Mincholé and B. Rodriguez, "Artificial intelligence for the electrocardiogram," *Nature Med.*, vol. 25, no. 1, pp. 22–23, 2019.

[16] Y. Awni et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, 2019.

[17] T. Brown et al., "Language models are few-shot learners," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[18] M. Brandon et al., "Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: Framework for sensor selection, deployment, and management," *J. Med. Internet Res.*, vol. 21, no. 8, 2019, Art. no. e12832.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[20] K. Mundnich et al., "TILES-2018: A longitudinal physiologic and behavioral data set of hospital workers," *Sci. Data*, vol. 7, 2020, Art. no. 354.

[21] M. Valstar et al., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, 2016, pp. 3–10.

[22] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 291–298.

[23] K. Nashiro et al., "Increasing coordination and responsivity of emotion-related brain regions with a heart rate variability biofeedback randomized trial," *Cogn., Affect., Behav. Neurosci.*, vol. 23, no. 1, pp. 66–83, 2023.

[24] K. Nashiro et al., "Effects of a randomised trial of 5-week heart rate variability biofeedback intervention on cognitive function: Possible benefits for inhibitory control," *Appl. Psychophysiol. Biofeedback*, vol. 48, no. 1, pp. 35–48, 2023.

[25] A. Mohamed et al., "Self-supervised speech representation learning: A review," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022.

[26] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[27] Y. Liu et al., "Graph self-supervised learning: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5879–5900, Jun. 2023.

[28] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, and Z. Huang, "Self-supervised learning for recommender systems: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 335–355, Jan. 2024.

[29] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, 2022.

[30] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–37, 2023.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*. 2019, pp. 4171–4186.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[33] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

[34] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[35] P. Sarkar and A. Etemad, "Self-supervised ECG representation learning for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1541–1554, Jul.–Sep. 2022.

[36] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 3988–4003, 2022.

[37] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Front. Hum. Neurosci.*, vol. 15, 2021, Art. no. 653659.

[38] T. Mehari and N. Strodthoff, "Towards quantitative precision for ECG analysis: Leveraging state space models, self-supervision and patient metadata," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 11, pp. 5326–5334, Nov. 2023.

[39] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG heartbeat classification: A deep transferable representation," in *Proc. IEEE Int. Conf. Healthcare Inform.*, 2018, pp. 443–444.

[40] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara, "SigRep: Toward robust wearable emotion recognition with contrastive representation learning," *IEEE Access*, vol. 10, pp. 18105–18120, 2022.

[41] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[42] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. D. Salim, "COCOA: Cross modality contrastive learning for sensor data," *Proc. the ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 3, pp. 1–28, 2022.

[43] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Transformer-based self-supervised learning for emotion recognition," in *Proc. IEEE 26th Int. Conf. Pattern Recognit.*, 2022, pp. 2605–2612.

[44] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr.–Jun. 2018.

[45] Y. Wu, M. Daoudi, and A. Amad, "Transformer-based self-supervised multimodal representation learning for wearable emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 157–172, Jan.–Mar. 2024.

[46] J. Jungo, Y. Xiang, S. Gashi, and C. Holz, "Representation learning for wearable-based applications in the case of missing data," 2024, *arXiv:2401.05437*.

[47] C. Ma et al., "STP: Self-supervised transfer learning based on transformer for noninvasive blood pressure estimation using photoplethysmography," *Expert Syst. Appl.*, vol. 249, 2024, Art. no. 123809.

[48] A. Gu et al., "Combining recurrent, convolutional, and continuous-time models with linear state space layers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 572–585.

[49] A. Tustin, "A method of analysing the behaviour of linear systems in terms of time series," *J. IEE-Part IIA: Autom. Regulators Servo Mechanisms*, vol. 94, no. 1, pp. 130–142, 1947.

[50] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2022, *arXiv:2111.00396*.

[51] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "HIPPO: Recurrent memory with optimal polynomial projections," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1474–1487.

[52] D. Makowski et al., "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, pp. 1689–1696, 2021.

[53] O. Kwon et al., "Electrocardiogram sampling frequency range acceptable for heart rate variability analysis," *Healthcare Inform. Res.*, vol. 24, no. 3, pp. 198–206, 2018.

[54] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1519–1528, May 2021.

[55] K. Avramidis, "ECG-augmentations," Zenodo, v1.0.0-beta, 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8053043

[56] M. Zaheer et al., "Big bird: Transformers for longer sequences," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 17283–17297.

[57] M. Zhang, K. K. Saab, M. Poli, T. Dao, K. Goel, and C. Ré, "Effectively modeling time series with simple discrete state spaces," 2023, *arXiv:2303.09489*.

[58] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[59] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 154.

[60] A. I. Kalyakulina et al., "LUDB: A new open-access validation tool for electrocardiogram delineation algorithms," *IEEE Access*, vol. 8, pp. 186181–186190, 2020.

[61] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interaction*, 2018, pp. 400–408.

[62] K. Sharma, C. Castellini, E. v. d. Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Sci. Data*, vol. 6, no. 1, 2019, Art. no. 196.

[63] S. Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[64] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[65] J. C. Yau et al., "TILES-2019: A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit," *Sci. Data*, vol. 9, no. 1, 2022, Art. no. 536.

[66] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[67] F. Ringeval et al., "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. Audio/Vis. Emotion Challenge Workshop*, 2018, pp. 3–13.

[68] L. Liakopoulos, N. Stagakis, E. I. Zacharaki, and K. Moustakas, "CNN-based stress and emotion recognition in ambulatory settings," in *Proc. IEEE 12th Int. Conf. Inf., Intell., Syst. Appl.*, 2021, pp. 1–8.

[69] B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad, "A transformer architecture for stress detection from ECG," in *Proc. ACM Int. Symp. Wearable Comput.*, 2021, pp. 132–134.
[70] P. Prajod and E. André, "On the generalizability of ECG-based stress detection models," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2022, pp. 549–554.
[71] J. Henry, H. Lloyd, M. Turner, and C. Kendrick, "On the robustness of machine learning models for stress and anxiety recognition from heart activity signals," *IEEE Sensors J.*, vol. 23, no. 13, pp. 14428–14436, Jul. 2023.
[72] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 227–239, Apr.–Jun. 2018.
[73] M. Green, M. Ohlsson, J. L. Forberg, J. Björk, L. Edenbrandt, and U. Ekelund, "Best leads in the standard electrocardiogram for the emergency detection of acute coronary syndrome," *J. Electrocardiol.*, vol. 40, no. 3, pp. 251–256, 2007.
[74] M. A. Reyna et al., "Will two do? Varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021," in *Proc. IEEE Comput. Cardiol.*, 2021, pp. 1–4.
[75] N. Nonaka and J. Seita, "In-depth benchmarking of deep neural network architectures for ECG diagnosis," in *Proc. Mach. Learn. Healthcare Conf.*, 2021, pp. 414–439.
[76] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," 2020, *arXiv:2007.04871*.
[77] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–9, 2020.
[78] L. V. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[79] S. Saganowski, B. Perz, A. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1876–1897, Jul.–Sep. 2023.
[80] V. Mishra et al., "Evaluating the reproducibility of physiological stress detection models," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, pp. 1–29, 2020.
[81] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Int.*, vol. 20, no. 1, pp. 1–68, 2019.
[82] I. Pikoulis, P. P. Filntisis, and P. Maragos, "Leveraging semantic scene characteristics and multi-stream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2021, pp. 01–08.