Proceedings of Machine Learning Research 209:410–426, 2023 Conference on Health, Inference, and Learning (CHIL) 2023

Contrastive Learning of Electrodermal Activity Representations for Stress Detection

Katie Matton[*] KMATTON@MIT.EDU
CSAIL and MIT Media Lab, Massachusetts Institute of Technology, United States

Robert Lewis[*] ROBLEWIS@MIT.EDU
MIT Media Lab, Massachusetts Institute of Technology, United States

John Guttag GUTTAG@MIT.EDU
CSAIL, Massachusetts Institute of Technology, United States

Rosalind Picard PICARD@MIT.EDU
MIT Media Lab, Massachusetts Institute of Technology, United States

Abstract

Electrodermal activity (EDA) is a biosignal that contains valuable information for monitoring health conditions related to sympathetic nervous system activity. Analyzing ambulatory EDA data is challenging because EDA measurements tend to be noisy and sparsely labeled. To address this problem, we present the first study of contrastive learning that examines approaches that are tailored to the EDA signal. We present a novel set of data augmentations that are tailored to EDA, and use them to generate positive examples for unsupervised contrastive learning. We evaluate our proposed approach on the downstream task of stress detection. We find that it outperforms baselines when used both for fine-tuning and for transfer learning, especially in regimes of high label sparsity. We verify that our novel EDA-specific augmentations add considerable value beyond those considered in prior work through a set of ablation experiments.

Data and Code Availability This study uses the WESAD (Schmidt et al., 2018) and VerBIO datasets (Yadav et al., 2020), which are both publicly available. Code is available at: https://github.com/kmatton/contrastive-learning-for-eda.

1. Introduction

Electrodermal activity (EDA) measures electrical properties of the skin. It is most commonly recorded using a measure of electrical conductance on the surface of the skin (Boucsein, 2012; Tronstad et al., 2022). Conductance is affected by the sympathetic nervous system, and thus changes in EDA are representative of changes in psychological or physiological arousal (Boucsein, 2012; Dawson et al., 2017). EDA can be measured through wrist-worn smartwatches, which makes it a valuable signal for remote health monitoring. A leading application of EDA is stress estimation and numerous lab studies have validated its association with stress (Sharma and Gedeon, 2012; Dawson et al., 2017; Giannakakis et al., 2019; Can et al., 2019). Recent longitudinal studies have incorporated wrist-worn EDA measurements to deepen the understanding of how stress mediates conditions such as suicidal thinking (Kleiman et al., 2021), substance-use disorder (Carreiro et al., 2020), and post-traumatic stress disorder (McLean et al., 2020). Beyond stress, EDA is an important signal for remote monitoring of other conditions and constructs such as seizures, sleep, pain and depression (Johnson and Picard, 2020; Bhatkar et al., 2022; Ortega et al., 2022; Sarchiapone et al., 2018; Werner et al., 2022).

There are two significant limitations to the analysis of EDA data collected in ambulatory settings. First, identifying most EDA-related outcomes relies on self-reported patient labels (e.g., psychological stress level) or clinical labels (e.g., a diagnosis). Therefore, EDA datasets are usually sparsely labeled; there is a natural upper bound on the number of labels one can collect relative to the number of EDA measurements one can record (ambulatory EDA runs at sample rates ≥4Hz). Second, wrist-worn EDA

[*] These authors contributed equally

measurements are noisy (i.e., they contain artifacts) (Posada-Quintero and Chon, 2020; Tronstad et al., 2022). This noise often results from physical disruptions of the positioning of the watch, e.g., when the electrodes do not make consistent contact with the skin, or from environmental factors such as temperature and humidity.

Unsupervised contrastive learning can help to overcome these challenges. This is a self-supervised method for learning data representations from unlabeled data. It works by: (1) generating *positive* pairs of examples (examples that should have similar representations), and (2) training a model to push the representations of *positive* pairs of examples together and other pairs of examples (i.e., *negative* pairs) apart. Generating positive pairs is typically done by applying two different data augmentations to the same example (e.g., two different rotations to the same image). The choice of data augmentations (DA) has a large effect on the quality of the representations learned (Tian et al., 2020): it is important to select DAs that are both challenging (i.e., it is non-trivial to distinguish between positive and negative examples) and label-preserving.

Recently, there have been a number of studies that examine unsupervised contrastive learning applied to biosignal data, including ECG and EEG signals (Kiyasseh et al., 2021; Rabbani and Khan, 2022; Wagh et al., 2021). However, to our knowledge there is no work that examines approaches that are tailored specifically to the EDA signal. The EDA signal presents different challenges than the biosignals for which contrastive learning is well-studied. In particular, it is not quasi-periodic in the same sense as signals like ECG, which have well-defined repeating patterns over time. Further, it is subject to specific sources of noise related to the nature of the signal and where it is measured. These differences suggest that contrastive learning approaches developed for other signals might not generalize directly to EDA.

In this paper, we provide the first study of contrastive learning that focuses on the EDA signal. We focus on the choice of DAs used to generate positive examples. We develop a novel set of DAs that are designed to account for the particular properties of the EDA signal. We perform experiments on two data sets. WESAD (Schmidt et al., 2018) contains data from 15 subjects measured under laboratory conditions. VerBIO (Yadav et al., 2020) contains data from 55 subjects performing tasks related to public speaking. We find that our approach outperforms baselines on the downstream task of stress detection, yielding a ≈16% accuracy improvement on WESAD and a ≈6% improvement on VerBIO compared to fully-supervised learning in a setting of high label sparsity (1% of the data is labeled). Through a set of ablation experiments, we verify that our EDA-specific augmentations add considerable value beyond those considered in prior work on augmenting bio-signals. We also show that our approach can be used to generalize *across* datasets; it consistently beats baselines when performing transfer learning from one dataset to the other.

2. Related Work

A number of existing works have examined machine learning methods for stress detection from bio-signals (Gedam and Paul, 2021). Many of these studies compare the performance of supervised learning methods, including linear regression, SVMs, random forests, KNNs, and neural networks, working with either hand-crafted feature sets or the raw biosignals directly (Garg et al., 2021; Bobade and Vani, 2020; Schmidt et al., 2018). While supervised learning is well-studied in this space, there is little work on self-supervised learning.

Self-supervised learning using contrastive learning techniques such as Contrastive Predictive Coding, SimCLR and Wav2Vec has enabled significant progress on learning from noisy and sparsely labeled datasets in domains such as computer vision and speech recognition (Oord et al., 2018; Chen et al., 2020; Baevski et al., 2020; Jaiswal et al., 2020). More recently, contrastive methods have also been applied in health domains, for example to electrocardiogram (ECG) data for cardiac arrhythmia detection (Kiyasseh et al., 2021; Cheng et al., 2020) and stress detection (Rabbani and Khan, 2022), and to electroencephalogram (EEG) data for e.g., sleep stage scoring and eye state classification (open vs. closed) (Mohsenvand et al., 2020; Wagh et al., 2021; Cheng et al., 2020). Most relevant to our work, (Rabbani and Khan, 2022) explore the use of self-supervised contrastive learning for stress detection from ECG data. They find that this approach outperforms non-contrastive baselines on two stress datasets, including WESAD. Multimodal contrastive learning methods that use the signals from more than one sensor in a wearable have also been considered in studies to predict generic health and demographic characteristics such as VO_2 max and age (Spathis et al., 2021).

Regarding contrastive learning methods for electrodermal activity (EDA), EDA has been included as a modality in multimodal contrastive learning systems that apply standard transforms across all signals generically (i.e., using transforms that are agnostic to the nature of the sensor signals) (Dissanayake et al., 2022; Saeed et al., 2020). However, to the best of our knowledge, no previous work has focused explicitly on EDA in a way that carefully considers the nature of the EDA signal – which is non-periodic, non-stationary, and subject to specific sources of noise.

3. Methods

3.1. Contrastive Learning Task Formulation

In line with *SimCLR* (Chen et al., 2020), we formulate a self-supervised contrastive learning set up where we train a model to distinguish *positive examples* of the input signal from *negative examples*.

We segment the EDA signal into windows of a fixed length ($|x_i| = M$ samples). We then generate transformed versions of each segment $\tilde{x}_i = t(x_i)$ by applying a transform t that is randomly sampled from a set of data augmentations T (which we discuss in Section 4). We define positive examples as those that have the same base segment (e.g., $\tilde{x}_i = t(x_i)$ and $\tilde{x}'_i = t'(x_i)$ where $t, t' \sim T$.), and negative examples as those that have different base segments (i.e., $\tilde{x}_j = t(x_j)$, $\forall j : j \neq i$).

We use a neural network model that maps from augmented samples $\tilde{x}_i \in \mathbb{R}^d$ to low-dimensional embeddings $z_i \in \mathbb{R}^k$, with $k \ll d$. To train the model, we sample a random mini-batch of N examples. For each sample, x_i, we generate two augmented versions, $\tilde{x}_i = t(x_i)$ and $\tilde{x}'_i = t'(x_i)$. We use the *InfoNCE* loss (Oord et al., 2018; Chen et al., 2020) to optimize model parameters during pre-training. For a postive pair of examples $(\tilde{x}_i, \tilde{x}'_i)$ this loss is defined as:

$$\ell_{i,i'} = -log \frac{exp(sim(z_i, z'_i)/\tau)}{\sum_{j=1}^{N} exp(sim(z_i, z'_j)/\tau)} \quad (1)$$

where $sim(z_i, z'_j)$ is the *cosine similarity* between embeddings and τ is a temperature parameter. The loss for the batch is computed as $L = \frac{1}{2N}\sum_{i=1}^{N}[\ell_{i,i'} + \ell_{i',i}]$.

3.2. Model Architecture

Our model architecture consists of an encoder, $f(\cdot)$, followed by a linear projection head, $g(\cdot)$. The encoder produces lower-dimensional embeddings of the input signal $h_i = f(\tilde{x}_i)$, which can be transferred to downstream prediction tasks.

We implement the encoder using a 1-D convolutional neural network (CNN), consisting of three convolutional blocks each with batch normalization, dropout and ReLU activation, followed by a single linear layer. This convolutional architecture was chosen as it is generally found to be a strong baseline for time-series classification while also being relatively less complex to train than e.g., recurrent architectures such as LSTM (Wang et al., 2017; Bai et al., 2018).

The projection head g further reduces the dimensionality of the input, producing $z_i = g(h_i)$. We implement this as a single linear layer. Further details on the model architecture and hyperparameters can be found in Appendix A.

4. Data Augmentations

We consider two sets of data augmentations (DA): (1) a comprehensive set of generic time-series augmentations, and (2) a set of augmentations we developed that are tailored to properties of the EDA signal. Figure 1 contains visualizations of these augmentations. We also provide our code with this paper to demonstrate how these DAs are implemented.

Most of the DAs have associated parameters that control how much they perturb the signal. In our main experiments, we use stochastic versions of each DA, where the parameters for the augmentation are randomly sampled from a fixed range of values. We selected these parameter ranges based on both prior knowledge about EDA properties and preliminary experiments on the WESAD *training* data. In these preliminary experiments, we used deterministic versions of each DA and swept over broad ranges of their parameter values. We then selected the optimal range of parameters based on their *validation* set accuracy.

This parameter selection approach helps us to select DAs that are both *challenging* and *label-preserving*, the two key properties needed for DAs that are used in contrastive learning (Tian et al., 2020). Details on the experiments to select the optimal parameter ranges are in Appendix B. From this, we see that several DAs are very sensitive to the strength of augmentation, and hence parameter selection is important to ensure DAs are well-calibrated.

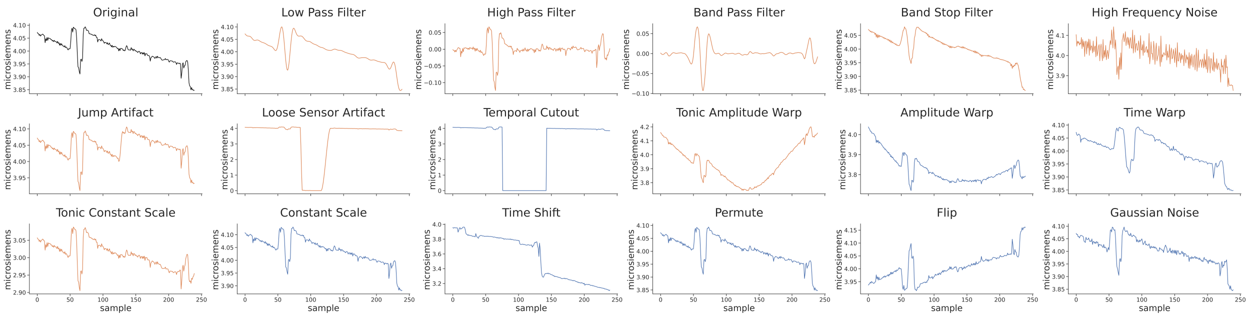


Figure 1: Example EDA segment with each data augmentation applied. Note: the range of the y-axis varies between subplots. EDA-Specific augmentations are in orange and others are in blue.

4.1. EDA-Specific Data Augmentations

4.1.1. Isolating & Altering Frequency Components of the Signal

EDA signals are characterized by two components: (1) a slow-varying *tonic* component, which represents baseline skin conductance level, and (2) a faster-varying *phasic* component, which represents abrupt changes in skin conductance that occur in response to external stimuli (Boucsein, 2012; Posada-Quintero and Chon, 2020). Both components can be informative of stress: tonic activity varies depending on an individual's psychological state, whereas phasic activity changes in the presence of external stressors (Turpin and Grandfield, 2007). In the datasets we consider, stress is induced via external conditions, and hence we expect that the phasic component is particularly relevant. The phasic component can be approximately extracted by applying a high-pass filter with a cutoff frequency of 0.05 Hz. The tonic component can be extracted with a low-pass filter also with cutoff at 0.05 Hz (Braithwaite et al., 2013). In the filtering-based augmentations described in this section, we consider parameter settings that correspond to tonic and phasic component extraction.

Our DAs are also informed by research aimed at identifying the frequency bands of EDA that are the most informative of sympathetic nervous system activity. Posada-Quintero et al. (2016) conducted a study in which subjects were exposed to a series of sympathetic-tone inducing stimuli. They found that frequencies in the range 0.08-0.24 Hz were the most responsive. Accordingly, we design augmentations that are aimed at isolating the information rich components of the signal and altering the parts of the signal where relevant information is *not* expected:

- *Low-Pass Filter:* We apply a filter that passes only signals with a lower frequency than a chosen cutoff $f \in [0.25, 1.0]$[1].
- *High-Pass Filter:* We apply a filter that passes only signals with a higher frequency than a chosen cutoff $f \in [0.05, 0.25]$.
- *Band-Pass Filter:* We apply a filter that passes only a specific frequency band $f \in [0.05, 0.25]$.
- *Band-Stop Filter:* We apply a filter that rejects / attenuates a specific frequency band, while letting all others pass. We sample from reject frequencies: $f \in [0.75, 1]$.
- *High Frequency Noise:* We add noise to the high frequency bands of the signal. To do this, we map the signal to the frequency-domain using an FFT, and add Gaussian noise to all frequency bands $\geq$ 1Hz, before running an iFFT to return the signal to the time-domain. We sample from $\sigma \in [0, 0.5]$ for the noise distribution.

4.1.2. Simulating Motion Artifacts

EDA data, even when collected in lab settings, are prone to artifacts, i.e., changes in the signal that are *not* a result of the electrodermal system (Boucsein, 2012; Taylor et al., 2015). Because artifacts do not reflect sympathetic nervous system activity, our model

[1] Note: All parameter ranges reported in these subsections are those selected from the initial parameter selection experiments described in Appendix B.

predictions should be invariant to them. To achieve this, we simulate adding two common types of motion artifacts. We designed these artifacts using mechanistic understanding of how physical disruptions affect the EDA measurement (Boucsein, 2012; Kleckner et al., 2018), as well as by visually comparing our simulated artifacts to authentic EDA artifacts that were labeled by experts in a large dataset (>100 hours of data) (Gashi et al., 2020):

- *Jumping / Motion Artifact:* We add simulated motion artifacts to the signal, i.e., artifacts that arise when the placement of wearable sensors is altered by movement. Motion artifacts often appear as abrupt drops or rises in the signal (a drop or rise of $\geq 0.1\mu S/sec$ is a common heuristic to identify them). We allow up to two jumping artifacts to occur within a 60s signal window, and we sample the size of the jump from $jump \in [0.01, 0.2]\mu S$. We allow each artifact to occur for a duration of $t \in [0.5, 3]$ seconds, where the jump may cause either a drop or a rise.

- *Loose Sensor Artifact:* We simulate the addition of artifacts that occur when sensor electrodes lose contact with the skin. This typically results in the signal values dropping to near-zero. We implement this by sampling a length of time that the sensor is loose for from $t \in [5, 20]$ seconds. We allow a smooth drop of the signal over a time of ≤ 5 seconds, as well as a smooth recovery over a time of ≤ 5 seconds, since this creates a more realistic looking loose sensor artifact relative to those documented in (Gashi et al., 2020). We also superimpose the absolute residual of the original signal minus its mean amplitude onto the signal in the window where the artifact occurs. This ensures that the signal is non-zero during this time, which again makes the artifact appear more authentic.

4.1.3. Simulating Thermoregulation Artifacts

Temperature and humidity affect the EDA signal because they influence the extent to which someone sweats through the process of *thermoregulation* (Bari et al., 2018; Qasim et al., 2022; Gashi et al., 2020; Boucsein, 2012). Temperature and humidity can change while EDA is being recorded, yet they are rarely directly related to a change in someone's level of psychological stress. We seek to learn representations that are invariant to these thermoregulation artifacts. To this end, we introduce two augmentations based on controlled EDA studies where participants are subjected to stress inducing tasks at different humidities (Bari et al., 2018) and temperatures (Qasim et al., 2022). These studies find that the slowly-changing tonic component of EDA is significantly increased by both increasing humidity and increasing temperature. In contrast, they find that the more rapidly-changing phasic component of EDA is *not* significantly changed by increases in humidity or temperature (though there is a non-significant increasing trend).

- *Tonic Amplitude Constant Scale:* We use Butterworth lowpass and highpass filters at 0.05Hz to extract the tonic and phasic components of the original signal. We then apply a constant scaling factor to the tonic component to mimic the effect of a constant change in temperature across the length of the recording. Finally, we recombine the signal as a simple sum of the tonic and phasic components. We sample scaling factors in the range $s \in [0.25, 2]$

- *Tonic Amplitude Warp:* We apply a time-varying scaling factor to the tonic component of the signal. This imitates when temperature or humidity is changing (for example, someone might move between rooms). We use a cubic spline to smoothly vary the scaling factor over time. The tonic and phasic components are extracted and recombined in the same way as the previous augmentation. We examine splines with 0 to 4 *knots* (where 0 represents a scaling factor that varies linearly over time) and knot heights (u_i) sampled from $u_i \sim N(1, \sigma^2)$ with $\sigma \in [0.01, 0.05]$.

4.2. Generic Time-Series Augmentations

We also include a set of standard time-series DAs that are commonly considered for supervised learning and contrastive learning on biosignals (Wen et al., 2021; Um et al., 2017; Iwana and Uchida, 2021; Saeed et al., 2019):

- *Amplitude Constant Scale:* We apply a constant scaling factor directly to the raw EDA signal. This augmentation is similar to the *Tonic Amplitude Constant Scale* augmentation described in Section 4.1, but it scales the whole signal rather than only the tonic component. We use the same scaling factors for both augmentations for comparison: $s \in [0.25, 2]$.

- *Amplitude Warp:* We apply a time-varying scaling factor to the raw EDA signal. We include this for comparison with the EDA-specific *Tonic Amplitude Warp* (see Section 4.1). The same range of spline parameters are considered.
- *Gaussian Noise:* We add random noise to the signal using independent samples from a Gaussian distribution for each sample, with $\sigma \in [0, 0.5]$. We scale σ for each signal to control the signal-to-noise ratio.
- *Time Shift:* We shift the signal forward or backward in time, with lengths of $t \in [5, 45]$ seconds.
- *Temporal Cutout:* We zero out / mask the signal over a subset of the window, with a sub-window of length $t \in [5, 15]$ seconds.
- *Time Warp:* We perturb the pattern of the signal in the temporal dimension. A cubic spline is used to warp the distance in time between samples in the signal, thus leading to local compression or stretching of the original signal. We examine splines with 1 to 4 *knots* (where 1 represents a linear scaling) and knot heights (u_i) sampled from $u_i \sim N(1, \sigma^2)$ with $\sigma \in [0.01, 0.1]$.
- *Permutation:* We cut the signal window into sub-windows and randomly permute their order. We vary the number of sub-windows $n \in [2, 6]$.
- *Signal Flip:* We flip / invert the signal over its amplitude dimension. There are no parameters for this augmentation.

5. Experiments

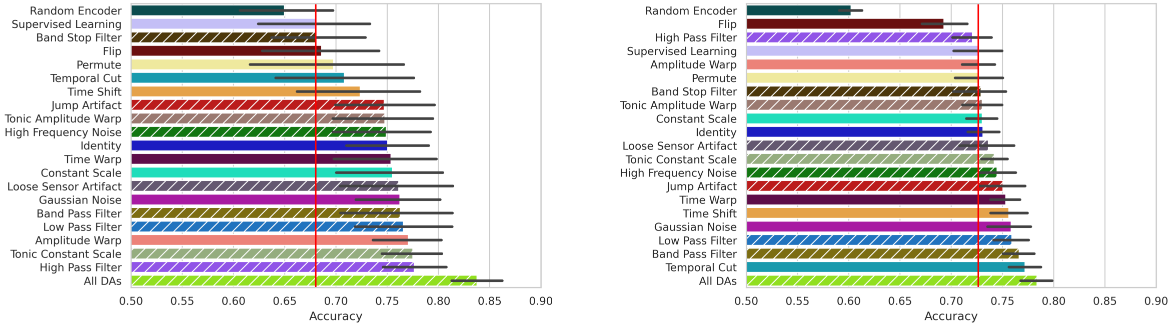
5.1. Datasets

We consider two datasets for our experiments. First, the WESAD dataset (Schmidt et al., 2018), a multimodal wearable dataset for stress and affect detection. WESAD includes data collected from 15 subjects in a laboratory setting. Subjects were exposed to experimental conditions that were designed to elicit different affective states. We focus on the binary classification task of distinguishing between the stress and baseline (i.e., low stress) conditions, and do not use the data from other parts of the protocol in our experiments (neither pre-training or the downstream evaluation). During the stress condition, the participants complete the Trier Social Stress Test (TSST) (Kirschbaum et al., 1993), which comprises public speaking and a mental arithmetic task. It lasts about 10 minutes on average. During the baseline, the participants sit or stand at a desk, and it lasts about 20 minutes on average. The EDA data we use for these experiments is collected using a wrist-worn device (Empatica E4). As in prior work (Schmidt et al., 2018), we segment the EDA data into 60-second windows, overlapping with a window shift of 0.25 seconds. This produces 103,037 segments with a stressed or non-stressed label (based on the experimental condition). There are more non-stressed segments (65%) than stressed segments (35%).

Second, we use the VerBIO dataset (Yadav et al., 2020), a multimodal wearable dataset to understand public speaking anxiety. VerBIO contains recordings from 55 participants, with 344 public speaking sessions across multiple days, where some sessions are in person and others are in a virtual reality (VR) environment. Participants are guided through a protocol for each session of a *relaxation* period (where they watch a nature video for 5 minutes), followed by a *preparation* period (where they prepare a speech about a news article for about 10 minutes), followed by a *presentation* period (where they present to either a live or virtual audience). We code the relaxation period as low stress and the presentation period as high stress. We do not assign a label to the preparation period, nor do we include it as unlabeled data during our pre-training phase. The EDA recordings in this dataset are again obtained using a wrist-worn device (Empatica E4). We create signal windows with the same length and overlap as WESAD. This produces 403,072 segments with a stressed or non-stressed label. Conversely to WESAD, there are more stressed segments (61%) than non-stressed segments (39%).

5.2. Evaluation Approach

In this study, our primary goal is to achieve strong stress detection performance in label sparse settings. Thus, to evaluate the utility of our proposed contrastive pre-training approach, we examine its ability to find a good initialization for an encoder that is later fine-tuned on a small amount of labeled data. To conduct this *fine-tuning* evaluation, we add a linear classification layer on top of the pre-trained en-



(a) WESAD

(b) VerBIO

Figure 2: **Fine-tuning evaluation**: test accuracy using 1% of labeled data. Methods that use EDA-specific data augmentations are denoted with white hatched lines. Almost all contrastive pre-training approaches outperform *Supervised Learning* (red line).

coder, and allow all model parameters to be updated during the downstream supervised learning stage (see diagram in Appendix D). To understand how well our approach generalizes across datasets, we also evaluate how the pre-trained encoders perform when used for *transfer learning*. In this experiment, we take the encoder pre-trained on one dataset, fine-tune it using the labeled data from the other dataset, and then examine the stress detection performance on the latter dataset.

Many contrastive learning studies also consider a *linear evaluation* scenario, in which the pre-trained encoder is frozen and only the linear classification layer is updated during the downstream supervised training. However, recent work on self-supervised learning has argued that *fine-tuning* performance should be the primary evaluation metric (He et al., 2022; Balestriero et al., 2023). This is because linear evaluation performance is often uncorrelated with that of fine-tuning and transfer learning, and it cannot be used to assess the ability of a method to generate useful non-linear features. We include linear evaluation results in Appendix C for completeness, but do not consider them to be our main focus.

As baselines, we consider three other models with the same architecture (i.e., the same encoder and linear classifier): (1) a model trained with constrative learning where positive examples are generated via the *Identity* transformation (i.e., no DA is applied); (2) a linear classifier trained on top of a frozen, randomly initialized encoder; and (3) a model trained end-to-end with supervised learning.

During the contrastive learning phase we train with all training examples (and assume no access to labels). During the supervised learning phase, we train on $x\%$ of the labeled data (we vary x from 0.1% to 100%, randomly selecting the subset of the data to keep labeled). This artificial downsampling of the available labels allows us to examine how each method performs when we make different assumptions about the sparsity of labels and is a standard evaluation approach in the contrastive learning literature. We split the data by subject into 5 folds and perform leave-N-subjects-out (LNSO) cross-validation to assess how well each method performs on *held-out* subjects. All results are reported as the average over 5 random experimental seeds. The details of our model architecture and hyperparameter selection are provided in Appendix A.

5.3. Results and Discussion

5.3.1. Fine-tuning Evaluation

We evaluate each method on its ability to initialize the parameters of an encoder that leads to good performance when updated by *fine-tuning* on labeled data. We examine versions of the contrastive pre-training encoder trained with only one DA applied (i.e., for positive pairs of examples, one segment is transformed $\tilde{x}_i = t(x_i)$, while the other is not $\tilde{x}'_i = x'_i$). We also examine the performance of an encoder pre-trained using all of the DAs. For this approach, we generate positive pairs of examples by applying two different DAs to the original sample, which are randomly sampled from the full set of DAs (i.e., $\tilde{x}_i = t(x_i)$ and $\tilde{x}'_i = t'(x_i)$ where $t, t' \sim T$). Figure 2 shows the test accuracy for each experiment, where 1% of the labeled data is used. Across both datasets, contrastive pre-training outperforms the supervised learning baseline for almost all of the DAs. This highlights the general utility of contrastive learning as an approach to address label sparsity for stress detection using EDA data.

We see that the *Identity* transformation (no augmentation) is a surprisingly strong baseline, especially for the WESAD dataset, where it outperforms *Supervised Learning* by a large margin. To understand this result, it is helpful to consider the contrastive loss function in Equation 1. When the *Identity* transformation is used to generate positive examples, the numerator – which encourages positive examples to be close together – will be a fixed constant. Thus, optimization will focus on the denominator, which pushes negative samples apart. As discussed in (Wang and Isola, 2020; Wang and Liu, 2021), this translates into encouraging the resulting features to be uniformly distributed. Uniformity is a desirable property because it encourages information preservation. In addition, we suspect that a more uniform feature distribution may help to counteract the subject heterogeneity that is typical of EDA data (due to variations in skin properties across individuals), leading to more generalizable features. We provide empirical evidence to support this idea in Section 5.3.3.

The set of DAs that performs best is not identical across the two datasets, but there is a fairly large set of DAs that outperform the *Identity* for both. This includes four EDA-Specific DAs (*Band Pass Filter*, *Low Pass Filter*, *Tonic Constant Scale*, and *Loose Sensor Artifact*) and two generic DAs (*Time Warp* and *Gaussian Noise*). The high utility of the band pass and low pass filtering DAs suggests that they are able to help build invariance to irrelevant frequency components of the signal; we further discuss this point in Appendix B. We find that using *All DAs* is the best performing approach, leading to 15.7% and 5.7% gains in accuracy over supervised learning on the WESAD and VerBIO datasets, respectively.

While contrastive learning provides a clear performance benefit for both datasets, the accuracy gain is more pronounced on WESAD than on VerBIO. We suspect that a few factors may contribute to this. First, VerBIO is about four times larger than WESAD, so training on 1% of labeled examples results in 4,000 datapoints (versus 1,000 for WESAD). This likely contributes to the higher accuracy of supervised learning on VerBIO. Second, in the VerBIO study, the stress detection task is more complex: VerBIO asks participants to deliver different public speeches in both real-life and VR environments, whereas WESAD uses the standard TSST stress test. This is likely to introduce more variance in the nature of the stress that is elicited, making the prediction task more challenging.

Table 1: **EDA-specific transform ablation**: *fine-tuning* test accuracy using pre-trained contrastive encoders at 1% label fraction on downstream task. Rows below the *All DAs* model show classification accuracy reduction when removing each EDA-specific DA from the contrastive pre-training. Note: all values in percentage points and negative values represent reduction in accuracy.

	WESAD	VerBIO
All DAs	83.77	78.38
- Band Pass Filter	-5.20	-0.97
- Band Stop Filter	-1.24	-0.91
- High Freq. Noise	-2.41	-0.48
- High Pass Filter	-8.17	-0.04
- Jump Artifact	-0.29	-1.17
- Loose Sensor	-1.91	-1.30
- Low Pass Filter	-1.68	-0.75
- Tonic Ampl. Warp	-1.03	-0.91
- Tonic Constant Scale	-6.17	-0.52

To further understand the utility of each EDA-specific DA, we conduct an ablation analysis where we pre-train with the full set of DAs except for one. The results of this ablation are in Table 1. We see that removing EDA-specific DAs leads to a 0.29-8.17 percentage point reduction in test accuracy for WESAD, and a 0.04-1.30 reduction for VerBIO. For WESAD, these results clearly show that no EDA-specific DA is dispensable in the contrastive pre-training, as results consistently worsen when any individual augmentation is excluded. For VerBIO, the trend is the same, though we note that the performance drops are generally smaller than those for WESAD, with the exclusion of some DAs resulting in only a slight change in performance. This may be partially due to the more challenging nature of VerBIO.

Table 2: **Transfer evaluation:** test accuracy for 1% of labeled data. Models pre-trained with *All DAs* and transferred across datasets significantly outperform *Supervised Learning.*

	WESAD→VerBIO	VerBIO→WESAD
Supervised	72.65 ± 6.00	68.03 ± 14.42
All DAs	76.86 ± 3.26	82.98 ± 7.29

5.3.2. Transfer Learning Evaluation

In this section, we seek to understand how well our approach generalizes *across* datasets. We consider the contrastive pre-training approach that uses all DAs, as this performed best in our earlier experiments (see Section 5.3.1). We examine the performance of an encoder pre-trained on one dataset and fine-tuned for the downstream task of stress detection on the other, using 1% of the labeled data for fine-tuning. We compare the accuracy of this approach to training a model end-to-end on labeled data from the target domain (i.e., *Supervised Learning*). As shown in Table 2, we find that transfer learning in both directions results in significantly better performance compared to supervised learning. We see a 4.21% gain in accuracy when transferring from WESAD to VerBIO, and a 14.95% gain when transferring from VerBIO to WESAD. Remarkably, this performance approaches the accuracy that we saw in the in-domain (i.e., fine-tuning) evaluation scenario. For the WESAD to VerBIO experiment, the accuracy obtained is only 1.5% less than that obtained when using a model pre-trained on VerBIO, and for VerBIO to WESAD, the accuracy is within 1% of that obtained with a model pre-trained on the WESAD data.

The WESAD and VerBIO datasets are considerably different from each other; they contain data from different subjects and were collected in the context of different types of stress elicitation tasks. The strong performance of our approach when used for transfer learning across these two datasets therefore suggests that (1) our data augmentations and pre-training approach are effective at learning representations that are useful for generalized stress detection, and (2) our approach has a high degree of robustness to dataset shift, which is a challenging problem that often arises when working with physiological data.

5.3.3. Representation Quality Analysis

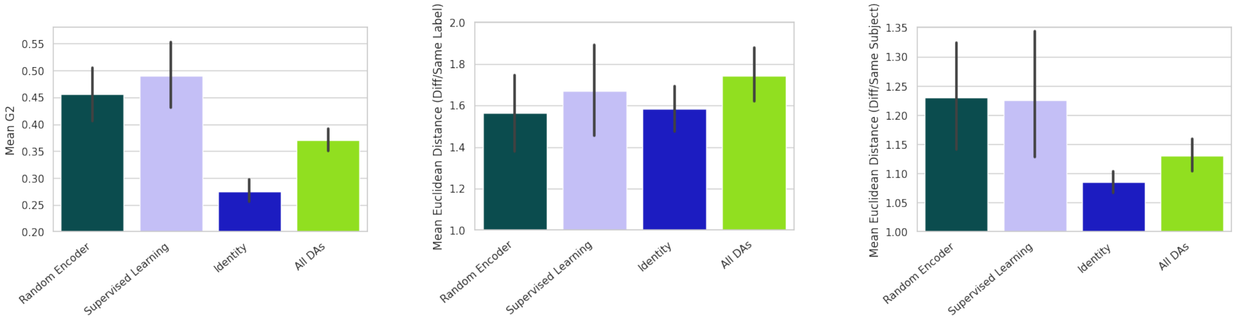
Our evaluation of downstream performance provided an understanding of the overall quality of the representations learned by each method. In this section, we seek to better understand the reason for the performance differences by analyzing the quality of the learned representations along three dimensions:

1. *Uniformity:* We desire representations that are both invariant to unnecessary details and that maximally preserve relevant information. Measuring the uniformity of the learned representation space is a way of capturing the latter of these two objectives. We adopt the uniformity metric proposed in (Wang and Isola, 2020), which measures how well the (normalized) representation space matches the uniform distribution on the unit hypersphere. This is computed as the mean Gaussian potential between pairs of embeddings, where a *lower* value implies *greater* uniformity. The Gaussian potential between two embeddings u, v is given by: $G_t(u, v) = e^{-t\|u-v\|_2^2}$, where t is a non-negative parameter. We set $t = 2$ as in (Wang and Isola, 2020).

2. *Label Separability:* An ideal encoder maps examples of the same label close together and examples of different labels far apart. We assess this by computing the ratio of the mean Euclidean distance between pairs of examples with different labels to that of examples with the same labels. A representation space that better separates the two label classes will receive a *higher* score.

3. *Subject Separability:* EDA signals are known to exhibit strong inter-individual differences due to variations in skin properties across individuals (Sarchiapone et al., 2018), which can manifest as subject-specific clusters in the data. We expect that a representation space that reduces differences across subjects will be useful for generalization. To measure this, we compute the ratio of the mean Euclidean distance between pairs of examples that belong to the same subject to that of examples that belong to different subjects. A representation space that better reduces subject heterogeneity will receive a *lower* score.

We use these metrics to compare the quality of the representations obtained with the *All DAs* method relative to the baselines, as shown in Figure 3. Examining the *Identity* method, we find that it has the



(a) Uniformity ($G_2 \downarrow$) (b) Label Separability ($\uparrow$) (c) Subject Separability ($\downarrow$)

Figure 3: Measures of representation quality on WESAD test data. Uniformity, as measured by the mean Gaussian Potential (G_2) between pairs of embeddings, captures how well the learned representation space preserves information (*lower* G_2 is better). Label separability is computed as the ratio of mean pairwise Euclidean distance for examples with different versus same labels (*higher* is better). Subject separability is computed as the ratio of mean pairwise Euclidean distance for examples from different versus same subjects (*lower* is better, since we desire representations that generalize across people). The *All DAs* contrastive pre-training method beats the *Random Encoder* and *Supervised Learning* baselines across all three metrics.

best score among all methods for the uniformity and subject separability metrics. As discussed in Section 5.3.1, the *Identity* method induces a fixed constant for the numerator of the contrastive loss (Equation 1) and thereby amounts to directly optimizing for uniformity of the (normalized) embedding space. Our finding that the *Identity* achieves the best uniformity (lowest mean G_2) is consistent with this understanding. We expect that uniformity is desirable not only because it can help with information preservation, but also because it can counteract subject heterogeneity. Our finding that the *Identity* obtains the smallest value of subject separability provides evidence to support this idea, and helps to explain why it obtained fairly strong performance in the downstream evaluation experiments.

While the *Identity* is strong in terms of the uniformity and subject separability metrics, it achieves relatively poor label separability. This aligns with the intuition that the *Identity* method encourages all representations to be equally distant from each other, *including* those with the same label. In contrast, the *All DAs* method obtains high values for all three metrics. It achieves the highest label separability out of all the methods, and obtains uniformity and subject separability scores that are much closer to the *Identity* compared to the *Random Encoder* and *Supervised Learning*. The observation that the *All DAs* method is able to effectively balance performance along these different dimensions helps to explain its superior performance in the downstream evaluation experiments.

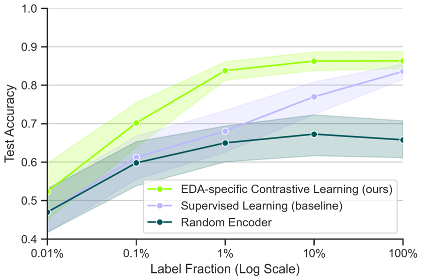


Figure 4: Label sparsity ablation for WESAD: Test performance of fine-tuned contrastive model compared to baselines. The fine-tuned contrastive method outperforms the baselines at all label fractions.

5.3.4. Impact of Label Sparsity

In this experiment, we examine the impact of the amount of labeled data available on the performance of contrastive pre-training and baseline methods. We focus on the contrastive pre-training method that uses all DAs, and consider the fine-tuning evaluation scenario. In Figure 4, we show the test accuracy of each method on the WESAD dataset for labeled data fractions ranging from 0.1% to 100%. We see that the contrastive learning approach consistently outperforms the two baselines. It achieves greater performance than supervised learning at all label fractions, including when all the labels are used.

6. Summary and Conclusion

We introduced a set of novel data augmentations tailored to the EDA signal and the task of stress detection. These include modifying the frequency bands of the EDA signal and incorporating priors on the sources of noise that affect it. Our experiments show that when used to generate positive examples for contrastive learning, these novel EDA-specific DAs add value beyond the generic time-series DAs considered in prior work. Our resulting approach achieves strong performance on the downstream task of stress detection, outperforming supervised learning by a considerable margin in regimes of high label sparsity. All of our experiments are conducted on held-out subjects; thus, our results show that our approach is able to generalize across people. By conducting a set of transfer learning experiments, we further show that our approach generalizes across datasets.

In future work, there are several limitations that we will address. First, we did not consider methods for choosing the optimal subset of DAs to include in the contrastive learning model (we include all of them in our *All DAs* model), nor do we apply augmentations to a signal *in composition* (i.e., applying > 1 augmentation to create a data view $\tilde{x}_i$). Second, the datasets that we used were collected under controlled experimental conditions and contained a limited number of subjects. While these are sufficient to demonstrate the proof of concept of contrastive learning on EDA data, they do not encompass all of the EDA noise profiles or stress responses that would be observed in naturalistic settings, and they are not representative of the EDA signals produced by all individuals. Furthermore, they contain a definition of stress induced by experimental conditions, rather than self-reported stress. Therefore, future work should consider how these methods generalize to additional datasets. Finally, we only consider EDA as a modality. Stress is likely best categorized by multiple physiological modalities, so multimodal approaches may yield further performance improvements.

Despite these limitations, we believe that our results demonstrate 1) the potential of using unsupervised contrastive learning to learn useful representations of EDA for stress detection, and 2) the utility of EDA-specific augmentations. While in this study we focused only on EDA, we believe that the approach we take of carefully considering signal properties to curate a set of signal-specific augmentations could be generalized to other biosignals.

Institutional Review Board (IRB)

We use two public, de-identified datasets in this research that both sought IRB approval when they ran their experiments and data collection (Schmidt et al., 2018; Yadav et al., 2020). In the work of this paper we do not have any interactions with participants, nor do we attempt to re-identify them from the deidentified data. Thus, this work is IRB exempt.

Acknowledgments

We are grateful for the thoughtful insights and feedback provided by our colleagues. We'd especially like to thank Szymon Fedor and Neska Elhaouij, whose advice and expertise benefited the project greatly. We also thank Renato Berlinghieri for his help creating the diagram visualizing our approach (Figure 7), Shkurta Gashi for her input on understanding the causes of artifacts in EDA data, and Ehsanul Haque Nirjhar for his support on preparing the VerBIO data for our experiments. Finally, we thank the anonymous reviewers for their useful feedback.

This research was supported in part by the Kerr Fellowship.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Dindar S Bari, Haval Y Yacoob Aldosky, Christian Tronstad, Håvard Kalvøy, and Ørjan G Martinsen. Influence of relative humidity on electrodermal levels and responses. *Skin pharmacology and physiology*, 31(6):298–307, 2018.

Viprali Bhatkar, Rosalind Picard, and Camilla Staahl. Combining electrodermal activity with the peak-pain time to quantify three temporal regions of pain experience. *Frontiers in Pain Research*, 3, 2022. ISSN 2673-561X. doi: 10.3389/fpain.2022.764128. URL https://www.frontiersin.org/articles/10.3389/fpain.2022.764128.

Pramod Bobade and M Vani. Stress detection with machine learning and deep learning using multimodal physiological data. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 51–57. IEEE, 2020.

Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012.

Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments. *Psychophysiology*, 49(1):1017–1034, 2013.

Yekta Said Can, Bert Arnrich, and Cem Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, 92:103139, 2019.

Stephanie Carreiro, Keerthi Kumar Chintha, Sloke Shrestha, Brittany Chapman, David Smelson, and Premananda Indic. Wearable sensor-based detection of stress and craving in patients during treatment for substance use disorder: A mixed methods pilot study. *Drug and Alcohol Dependence*, 209:107929, 2020. ISSN 0376-8716. doi: https://doi.org/10.1016/j.drugalcdep.2020.107929. URL https://www.sciencedirect.com/science/article/pii/S0376871620300946.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. 2017.

Vipula Dissanayake, Sachith Seneviratne, Rajib Rana, Elliott Wen, Tharindu Kaluarachchi, and Suranga Nanayakkara. Sigrep: Toward robust wearable emotion recognition with contrastive representation learning. *IEEE Access*, 10:18105–18120, 2022.

Prerna Garg, Jayasankar Santhosh, Andreas Dengel, and Shoya Ishimaru. Stress detection by machine learning and wearable sensors. In *26th International Conference on Intelligent User Interfaces-Companion*, pages 43–45, 2021.

Shkurta Gashi, Elena Di Lascio, Bianca Stancu, Vedant Das Swain, Varun Mishra, Martin Gjoreski, and Silvia Santini. Detection of artifacts in ambulatory electrodermal activity data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–31, 2020.

Shruti Gedam and Sanchita Paul. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, 9:84045–84066, 2021.

Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7): e0254841, 2021.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Kristina T Johnson and Rosalind W Picard. Advancing neuroscience through wearable devices. *Neuron*, 108(1):8–12, 2020.

C Kirschbaum, KM Pirke, and DH Hellhammer. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76—81, 1993. ISSN 0302-282X. doi: 10.1159/000119004. URL https://doi.org/10.1159/000119004.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

Ian R. Kleckner, Rebecca M. Jones, Oliver Wilder-Smith, Jolie B. Wormwood, Murat Akcakaya, Karen S. Quigley, Catherine Lord, and Matthew S. Goodwin. Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. *IEEE Transactions on Biomedical Engineering*, 65(7):1460–1467, 2018. doi: 10.1109/TBME.2017.2758643.

Evan M Kleiman, Kate H Bentley, Joseph S Maimone, Hye-In Sarah Lee, Erin N Kilbury, Rebecca G Fortgang, Kelly L Zuromski, Jeff C Huffman, and Matthew K Nock. Can passive measurement of physiological distress help better predict suicidal thinking? *Translational psychiatry*, 11(1): 1–6, 2021.

Samuel A McLean, Kerry Ressler, Karestan Chase Koenen, Thomas Neylan, Laura Germine, Tanja Jovanovic, Gari D Clifford, Donglin Zeng, Xinming An, Sarah Linnstaedt, et al. The AURORA study: a longitudinal, multimodal library of brain biology and function after traumatic stress exposure. *Molecular psychiatry*, 25(2):283–296, 2020.

Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*, pages 238–253. PMLR, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Marta Casanovas Ortega, Elisa Bruno, and Mark P Richardson. Electrodermal activity response during seizures: A systematic review and meta-analysis. *Epilepsy & Behavior*, 134:108864, 2022.

Hugo F Posada-Quintero and Ki H Chon. Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors*, 20(2): 479, 2020.

Hugo F Posada-Quintero, John P Florian, Álvaro D Orjuela-Cañón, and Ki H Chon. Highly sensitive index of sympathetic activity based on time-frequency spectral analysis of electrodermal activity. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 311(3): R582–R591, 2016.

Masood S Qasim, Dindar S Bari, and Ørjan G Martinsen. Influence of ambient temperature on tonic and phasic electrodermal activity components. *Physiological Measurement*, 43(6):065001, 2022.

Suha Rabbani and Naimul Khan. Contrastive self-supervised learning for stress detection from ECG data. *Bioengineering*, 9(8):374, 2022.

Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30, 2019.

Aaqib Saeed, Flora D Salim, Tanir Ozcelebi, and Johan Lukkien. Federated self-supervised learning of multisensor representations for embedded intelligence. *IEEE Internet of Things Journal*, 8(2): 1030–1040, 2020.

Marco Sarchiapone, carla maria Gramaglia, Miriam Iosue, Vladimir Carli, Laura Mandelli, Alessandro Serretti, Debora Marangon, and Patrizia Zeppegno. The association between electrodermal activity (eda), depression and suicidal behaviour: A systematic review and narrative synthesis. *BMC Psychiatry*, 18, 01 2018. doi: 10.1186/s12888-017-1551-4.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 400–408, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356923. doi: 10.1145/3242969.3242985. URL https://doi.org/10.1145/3242969.3242985.

Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine*, 108(3):1287–1301, 2012. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2012.07.003. URL https://www.sciencedirect.com/science/article/pii/S0169260712001770.

Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 69–78, 2021.

Sara Taylor, Natasha Jaques, Weixuan Chen, Szymon Fedor, Akane Sano, and Rosalind Picard. Automatic identification of artifacts in electrodermal activity data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1934–1937. IEEE, 2015.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Christian Tronstad, Maryam Amini, Dominik R Bach, and Ørjan G Martinsen. Current trends and opportunities in the methodology of electrodermal activity measurement. *Physiological Measurement*, 43(2):02TR01, 2022.

G. Turpin and T. Grandfield. Electrodermal activity*. In George Fink, editor, *Encyclopedia of Stress (Second Edition)*, pages 899–902. Academic Press, New York, second edition edition, 2007. ISBN 978-0-12-373947-6. doi: https://doi.org/10.1016/B978-012373947-6.00139-2. URL https://www.sciencedirect.com/science/article/pii/B9780123739476001392.

Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 216–220, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450355438. doi: 10.1145/3136755.3136817. URL https://doi.org/10.1145/3136755.3136817.

Neeraj Wagh, Jionghao Wei, Samarth Rawal, Brent Berry, Leland Barnard, Benjamin Brinkmann, Gregory Worrell, David Jones, and Yogatheesan Varatharajah. Domain-guided self-supervision of eeg data improves downstream classification performance and generalizability. In *Machine Learning for Health*, pages 130–142. PMLR, 2021.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017.

Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4653–4660. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/631. URL https://doi.org/10.24963/ijcai.2021/631. Survey Track.

Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind W. Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, 13(1):530–552, 2022. doi: 10.1109/TAFFC.2019.2946774.

Megha Yadav, Md Nazmus Sakib, Ehsanul Haque Nirjhar, Kexin Feng, Amir H Behzadan, and Theodora Chaspari. Exploring individual differences of public speaking anxiety in real-life and virtual presentations. *IEEE Transactions on Affective Computing*, 13(3):1168–1182, 2020.

Appendix A. Model Architecture and Hyperparameters

A.1. Model architecture

The following modules are used to create the different models in our experiments:

1. **Encoder,** $f(\cdot)$: 1-D convolutional layers followed by 1-D max-pooling are used to transform the M-dimensional input signal, x_i, to a k-dimensional embedding, h_i. *Batch normalisation* and *dropout* are used in each convolutional *block* and *ReLU* is used as the activation function. For the experiments of this paper, we use 3 convolutional blocks in the encoder, transforming the input from a 240 dimensional input segment (60 seconds at 4Hz) to a 64 dimensional embedding. The input and output channels of these blocks are Conv_1(in=1, out=4), Conv_2(in=4, out=16) and Conv_3(in=16, out=32). A kernel size of 7 and a stride of 1 are used in the convolutional layers and a kernel size of 2 is used in the max-pooling operations. A single linear layer with *ReLU* activation then maps the output to 64 dimensions.

2. **Projection head,** $g(\cdot)$: is implemented as a single linear layer that maps the 64-D embedding, h_i, to a 32-D embedding, z_i. The loss function (1) is computed on these 32-dimensional embeddings.

3. **Linear classifier for task prediction**: a linear classifier is used for supervised learning on the downstream prediction task (binary classification in the case of WESAD). This is implemented as a single linear layer mapping from the 64-D embeddings, h_i, to 1-D for the binary prediction. This model is trained for this task with binary cross entropy loss.

For the **contrastive learning models** in the *pre-training* phase, we compose the encoder module (a) and the projection head module (b). Then, for the downstream prediction phase, the pre-trained encoder is again used, but the linear classifier (c) replaces the projection head.

For the **supervised learning (SL)** and **random encoder (RE)** models, there is only the downstream prediction phase, and these models comprise modules (a) and (c). In the case of SL, the model is trainable *end-to-end* (i.e., all parameters in (a) and (c) can be optimized), whereas for RE only the linear classifier module (c) can be trained.

A.2. Hyperparameters

Different hyperparameters are used for the contrastive pre-training and the supervised learning phases of the modeling.

- **Contrastive pre-training hyperparameters**: A temperature of 0.1 is used within the InfoNCE loss. Adam is used as the optimizer with batch size of 256, learning rate of 0.001, and 400 epochs. The dropout probability is 0.1. Early stopping is implemented using the training InfoNCE loss.

- **Supervised learning fine-tuning hyperparameters**: Adam is used as the optimizer with batch size of 32, learning rate of 0.0001, weight decay of 0.01, and 200 epochs. The dropout probability is set to 0. Early stopping is implemented using the validation loss.

- **Supervised learning linear evaluation hyperparameters**: Adam is used as the optimizer with batch size of 32, learning rate of 0.001, weight decay of 0.01, and 200 epochs. The dropout probability is set to 0. Early stopping is implemented using the validation loss.

Appendix B. Selection of Data Augmentation Parameters

To select an optimal range of parameters for each data augmentation, we conducted experiments on the WESAD data in which we considered a *deterministic* version of each data augmentation, and then swept over a broad range of possible parameter settings for this augmentation. We examined the *validation* set accuracy of each parameter value. The results of this experiment are presented Figure 5. We see that several augmentations are very sensitive to the strength of augmentation. For example, the validation accuracy obtained with the *Band Pass Filter* augmentation varies by as much as 10% percent depending on the chosen frequency parameter, and the *Low Pass Filter*, *Gaussian Noise*, and *Temporal Cutout* augmentations all have as much as a 7% percent change in validation accuracy depending on the parameter setting. These results showcase the importance of carefully selecting data augmentation parameters when using them for contrastive pre-training.

We used these results to inform our selection of an optimal range to sample from for the *stochastic* versions of each data augmentation module that are used in the main contrastive learning experiments. The selected range is presented alongside the design of each augmentation in Section 4.

In addition to informing our parameter selection, these results also provide insights into how the design of each augmentation impacts downstream stress detection performance. Looking at the results of the filtering DAs (plots a-d in Figure 5), we see that performance varies considerably depending on which frequency bands the filter is applied to. Generally, we see that performance is strong when the DA parameters filter out parts of the signal where information related to sympathetic nervous system activity is *not* expected (i.e., outside of the 0.08-0.24 Hz range (Posada-Quintero et al., 2016)). For example, we see that the performance of the *Low Pass Filter* DA peaks when all signal content above 0.25 Hz is filtered, and that there is a steep drop in performance when filtering frequency components of the signal that are lower than this. Similarly, we see that the *Band Pass Filter* augmentation performs best when frequencies in the range 0.05-0.25 Hz are passed, and exhibits a considerable drop in performance when high frequencies ($\geq$ 1.25 Hz) are passed.

Appendix C. Linear Evaluation Results

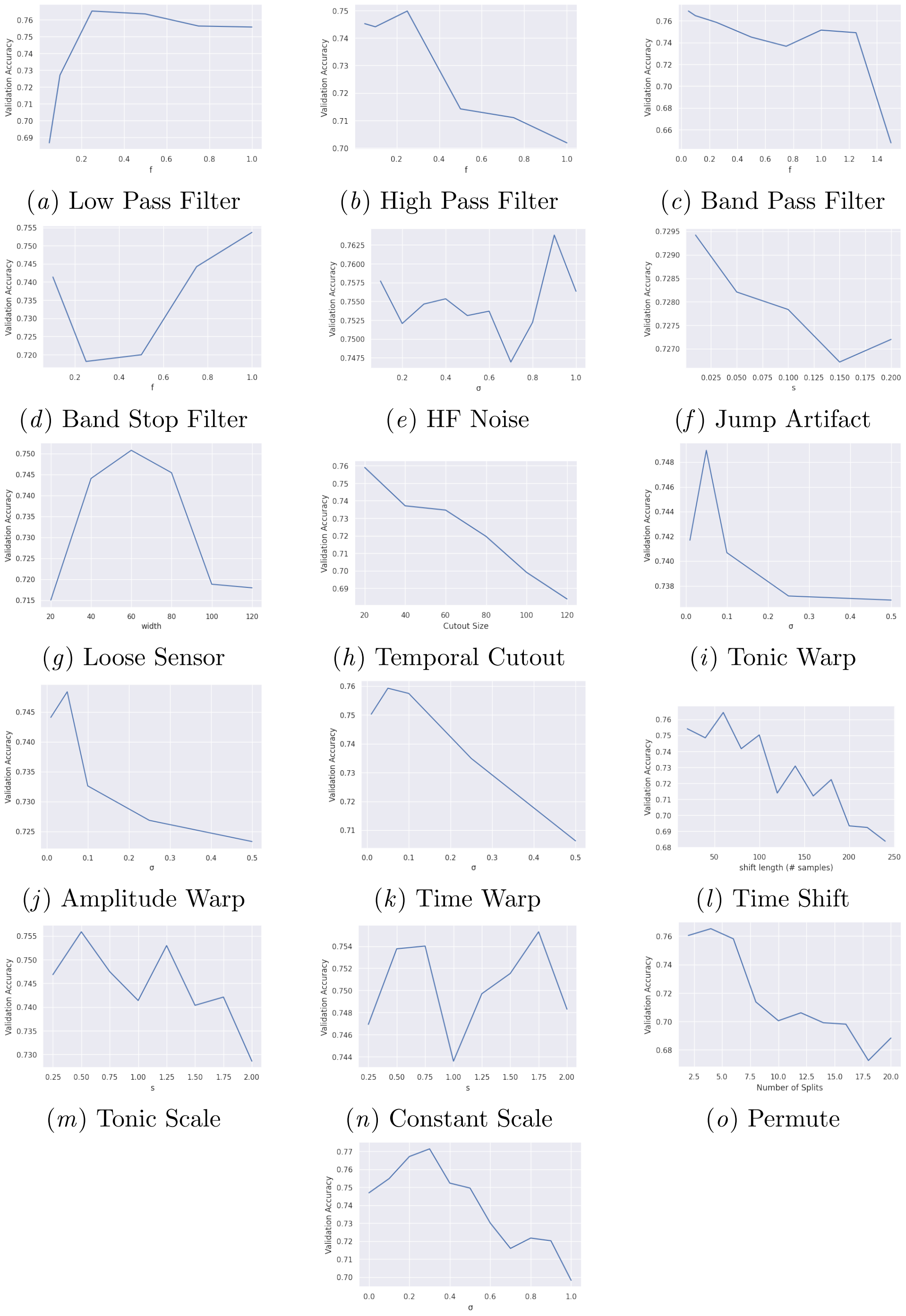
We consider the *linear evaluation* setting, to understand how well the pre-trained encoders perform when used as fixed feature extractors. We compare contrastive pre-training encoders that use a single DA, as well as an approach where we pre-train by sampling from the set of all DAs. Figure 6 displays the test accuracy for each experiment, where 1% of the labeled data is used in the downstream task. We see that across both datasets, almost all of the contrastive learning approaches outperform the randomly initialized encoder, indicating that the features learned are more meaningful than a random transformation of the data.

We see that the *Time Warp* DA is the best peforming single augmentation for both datasets for this evaluation scenario. *Gaussian Noise* is the only other DA that outperforms the *Identity* baseline for both datasets. On the WESAD dataset, many of the DAs outperform the *Identity* baseline. In contrast, on the VerBIO dataset, relatively few do. The VerBIO dataset contains considerably more subjects than the WESAD dataset, and the tasks performed by each subject are less structured. Thus, we suspect that there may be a greater degree of heterogeneity in the VerBIO dataset, which may both make optimization more challenging and increase the relative utility of the *Identity* baseline (which helps to counteract heterogeneity, as shown Section 5.3.3).

We find that training with all DAs leads to considerable improvement on the WESAD dataset, achieving a more than 3% accuracy lift over all the other methods. On VerBIO, this multi-transform method performs about as well as the top-performing single DA, but does not yield improvement beyond this. We suspect that the best performing set of transforms may be some subset of the full set, but since finding the optimal subset is not trivial, we leave this for future work.

Appendix D. Approach Diagram

We provide a visualization of our approach for pretraining and downstream task evaluation in Figure 7.



(a) Low Pass Filter

(b) High Pass Filter

(c) Band Pass Filter

(d) Band Stop Filter

(e) HF Noise

(f) Jump Artifact

(g) Loose Sensor

(h) Temporal Cutout

(i) Tonic Warp

(j) Amplitude Warp

(k) Time Warp

(l) Time Shift

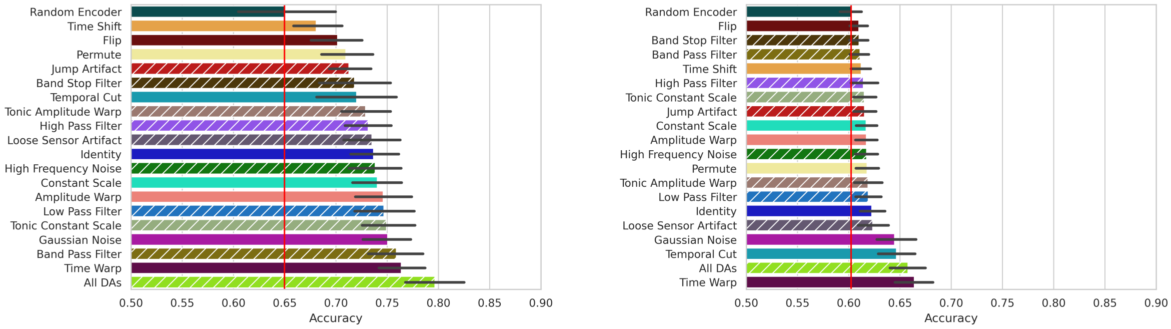
(m) Tonic Scale

(n) Constant Scale

(o) Permute

(p) Gaussian Noise

Figure 5: Hyperparameter tuning of individual data augmentation parameters for WESAD dataset. All reported accuracies are on the *validation* set. We see that several augmentations are very sensitive to the strength of augmentation. Parameter ranges for each data augmentation were selected from this experiment and used in the stochastic augmentation modules for the main experiments. Parameters selected are reported in Section 4.



(*a*) WESAD

(*b*) VerBIO

Figure 6: Test accuracy for linear classifier trained on 1% of labeled data, using frozen pre-trained encoders as features. Methods that use EDA-specific data augmentations are denoted with white hatched lines. Most of the encoders pre-trained with a single data augmentation outperform both the *Random Encoder* baseline (red line).

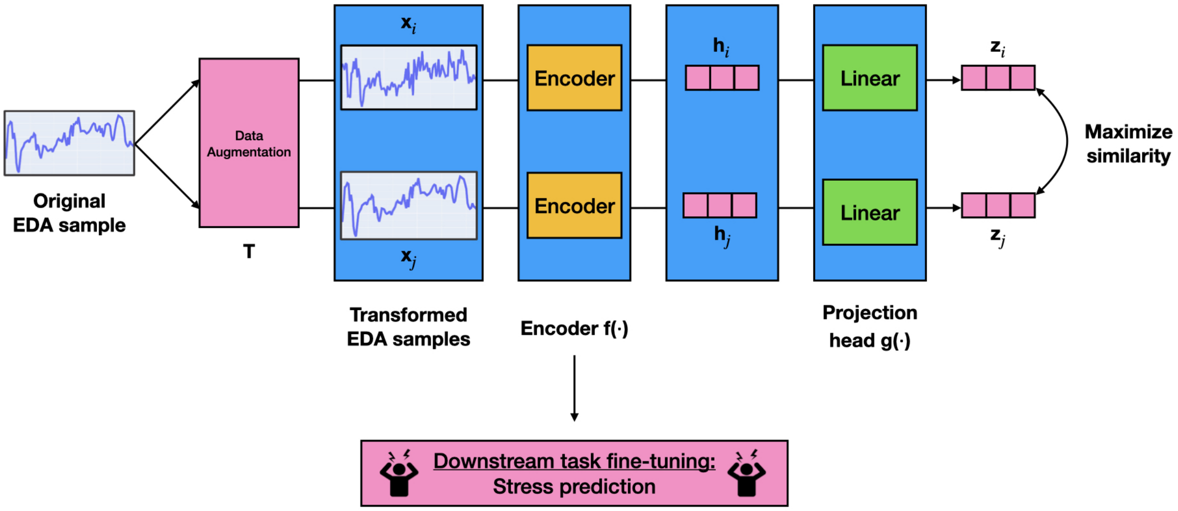


Figure 7: A visualization of our approach to pre-training and evaluation. We adopt the SimCLR contrastive learning framework proposed by Chen et al. (2020) and evaluate our pre-trained encoders on the downstream task of stress detection.