Self-Supervised ECG Representation Learning for Emotion Recognition

Pritam Sarkar and Ali Etemad

Abstract—We exploit a self-supervised deep multi-task learning framework for electrocardiogram (ECG) -based emotion recognition. The proposed solution consists of two stages of learning a) learning ECG representations and b) learning to classify emotions. ECG representations are learned by a signal transformation recognition network. The network learns high-level abstract representations from unlabeled ECG data. Six different signal transformations are applied to the ECG signals, and transformation recognition is performed as pretext tasks. Training the model on pretext tasks helps the network learn spatiotemporal representations that generalize well across different datasets and different emotion categories. We transfer the weights of the self-supervised network to an emotion recognition network, where the convolutional layers are kept frozen and the dense layers are trained with labelled ECG data. We show that the proposed solution considerably improves the performance compared to a network trained using fully-supervised learning. New state-of-the-art results are set in classification of arousal, valence, affective states, and stress for the four utilized datasets. Extensive experiments are performed, providing interesting insights into the impact of using a multi-task self-supervised structure instead of a single-task model, as well as the optimum level of difficulty required for the pretext self-supervised tasks.

Index Terms—Self-supervised learning, ECG, emotion recognition, multi-task learning

---

1 INTRODUCTION

A FFECTIVE computing is a field of study that deals with understanding human emotions, intelligent human-machine interaction, and computer-assisted learning among others [1], [2]. The goal of affective computing is to equip machines with the ability to model and interpret the emotional states of humans. Emotion is considered a *physiological* and *psychological* expression associated with moods and personalities of individuals. As a result, sensing technologies integrated into wearable devices coupled with machine learning and deep learning techniques have recently been used to analyze physiological signals in order to classify or quantify human emotions [3], [4], [5]. Recent applications of affective computing include human monitoring systems [6], [7], stress or anxiety level detection [8], [9], emotion and personality recognition [4], [10], and others.

A wide variety of data sources have been utilized for affective computing purposes. These data sources include facial expressions [11], motion and gait [12], speech [13], electrocardiogram (ECG) [14], electroencephalogram (EEG) [5], electrooculogram (EOG) [15], and galvanic skin response (GSR) [16] among others. In particular, ECG, which is the focus of our study, has been explored in the past with works such as [14] showing a strong correlation between emotional attributes and the ECG waveform, and investigating the feasibility and limitations of using ECG signals for emotion detection. Further, recent advancements in machine learning and deep learning have proven ECG to be a reliable source of information for emotion recognition systems [3], [17].

The majority of machine learning or deep learning solutions for ECG-based emotion recognition utilize fully-supervised learning methods. Several limitations can be associated with this approach. First and foremost, in a typical setup of fully-supervised learning, the model needs to be trained from scratch for every classification or regression task [18], which requires considerable computational resources and time. Additionally, the learned representations from the trained fully supervised models are often very task-specific, which are not expected to generalize well for other tasks. Lastly, fully supervised learning generally requires training using large human-annotated datasets, since small datasets typically result in poor performance with deeper networks.

To tackle these problems in the context of ECG-based affective computing, inspired by the success of multi-task self-supervised learning in other domains notably [19], we use a deep learning approach based on self-supervised representation learning [20] (please see Fig. 1). In self-supervised learning, models are trained using automatically generated labels instead of human-annotated ones. There are a number of advantages to self-supervised learning. First, the representations learned using this approach are often invariant to inter- and intra-instance variations [21] due to learning rather generalized features as opposed to task-specific ones. As a result, these models can be reused for different tasks within the same domain. This property often improves the performance of the networks and also saves computation time compared to training a model from scratch for each task. Lastly, as self-supervised learning doesn't require human-annotated labels, larger datasets can be acquired and utilized, which in turn results in the ability to train deeper and more sophisticated networks. We use

● *The authors are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada.*
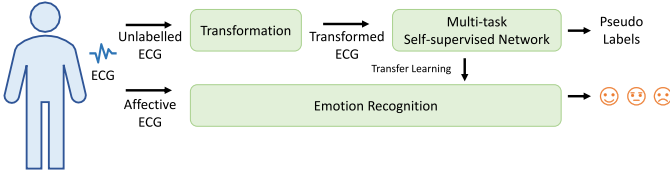*E-mail: {pritam.sarkar, ali.etemad}@queensu.ca.*

Fig. 1. An overview of the proposed framework for self-supervised emotion recognition is presented.

four publicly available datasets, AMIGOS [4], DREAMER [22], WESAD [23], and SWELL [9] to perform emotion recognition with ECG. First, a number of transformations are applied to the ECG signals. The automatically generated labels associated with the transformations are then utilized to carry out self-supervised learning of ECG representations through a deep multitask Convolutional Neural Network (CNN). Next, the convolutional layers responsible for learning the ECG representations are frozen and transferred to a second network. The dense layers of this new network are then trained with the labeled ECG data to perform emotion classification.

Our contributions can be summarized as follows:

- We exploit a self-supervised framework for emotion recognition based on multi-task ECG representation learning. While self-supervised techniques have been used for time-series representation learning in other domains (see [19], [24], [25]), we propose its use in ECG-based affective computing for the first time. The results show that the self-supervised network outperforms the same network for emotion recognition when trained in a fully-supervised fashion.

- We perform very thorough analyses on the parameters associated with the self-supervised learning tasks, providing interesting insights into the impact of different self-supervised tasks and their contribution towards learning of affective ECG representations. Additionally, we show that self-supervised multi-task learning results in better learned representations compared to single-task self-supervised learning. Our analysis also provides interesting insights into the relationship between the difficulty of self-supervised tasks and the original emotion recognition task, where there seems to be an optimum level of difficulty in which the network can learn strong representations in a self-supervised manner. The analysis shows that simpler tasks do not result in proper learning of ECG representations, whereas extremely hard tasks also prevent the network from learning generalized representations.

- We set new state-of-the-art results for all the undertaken emotion classification tasks namely arousal, valence, affective states, and stress recognition in the four utilized datasets AMIGOS, DREAMER, WESAD, and SWELL. We demonstrate that the ECG representations learned by the self-supervised model generalize very well across all four ECG datasets, consistently resulting in accurate emotion recognition.

This paper is an extension of our work [26], compared to which this paper additionally includes the following: *a)* Two additional publicly available datasets DREAMER and

WESAD are used in this paper to further investigate the benefits of the self-supervised learning framework compared to fully-supervised methods; *b)* The impact of different transformation tasks on the intended emotion recognition tasks are investigated in depth and presented in this paper; *c)* The relationship between the difficulty of self-supervised tasks and emotion recognition is investigated; *d)* The benefits of the multi-task architecture for the self-supervised transformation recognition network are compared to a single-task approach, and the results are presented; *e)* We analyze and discuss the impact of different parameters associated with each of the signal transformations with respect to the network performance; *f)* An analysis of the relationship between the learned representations and the depth of the proposed network is presented; *g)* Lastly, the effect of utilizing an aggregation of multiple datasets for self-supervised training as opposed to individual datasets is investigated and presented.

The rest of the paper is organized as follows. Section 2 discusses the background and related works on emotion recognition with ECG. We describe the methodology in Section 3. The datasets used in this paper and the experiment setup are described in Section 4. Section 5 presents the performance and results of the proposed architecture. A thorough analysis and discussion on the limitations of the framework is presented in Section 6. Section 7 presents the summary and concluding remarks.

2 BACKGROUND AND RELATED WORK

2.1 Electrocardiography

ECG signals contain information regarding cardiac electrical activity over a period of time, where an ECG beat is composed of a number of prominent waves: the P wave, the QRS complex, the T wave [14], and often the U wave [27]. Each waveform contains significant information that can be used to understand the cardiac state of individuals. To acquire ECG signals, electrodes are placed on the surface of the skin. The most common sensor configurations include 12-lead, 5-lead, 3-lead, or single-channel systems. ECG signals have been widely used to develop deep learning solutions, for example arrhythmia detection [28], emotion or stress detection [14], [29], biometric identification [30], and others. Additionally, research has shown strong correlations between cardiovascular activity and emotions, as emotions can influence the autonomic nervous system which controls the rhythm of heart beats [14]. To extract useful information from heart beats, different feature extraction techniques have been utilized [8], [31]. Most common approaches include either utilizing raw ECG signals to directly learn high-level representations [28], or extracting time-domain and frequency domain features successive to detection of fiducial points (P, Q, R, S, and T components) from the ECG waveform [14], [18]. Successive to learning of ECG representations, classifiers such as Multi-Layer Perceptrons are typically used learn the relationships between the high-level representations and the main task [3], [32].

2.2 ECG-Based Affective Computing

Systems capable of performing affective computing generally aim to classify or quantify the emotional state of individuals based on collected data and information. Recent

advancements in affective computing have shown ECG to be a reliable source of information to classify the emotional states of humans [3], [10], [17], [33]. The type of affective states often studied include stress, happiness, sorrow, and excitement among others.

In an early work on ECG-based affective computing, stress detection was performed in [8], as participants went through a driving task. ECG signals were collected during the study, followed by extraction of time and frequency domain features which were then used to perform classification. In particular, Heart Rate Variability (HRV) features were calculated from power spectrum of ECG signals. The extracted features were used to perform classification using a linear discriminant classifier.

In a large study, in [31], emotion recognition was performed, where ECG signals from 391 subjects were captured, while individuals were shown a movie clip belonging to either the category of joy or sadness. Out of all the participants, 150 subjects were selected who had shown effective elicitation during this study. Next, continuous wavelet transform was performed to accurately detect P-QRS-T wave locations. A total of 79 features were extracted. To perform better and effective classification, feature selection was performed by means of a hybrid particle swarm optimization technique. Their study shows that the selected features can correctly classify emotions with a Fisher classifier, reporting an accuracy of 88.43 percent.

The relationship between personality and affect was studied in [10]. The study was conducted on 58 participants. While subjects were shown emotional movie clips, ECG data were recorded. Similar to [8], HRV features were calculated, along with heart rate and inter-beat intervals. These features were utilized to perform classification of arousal and valence scores. Classification was performed by a Support Vector Machine (SVM) and Naive Bayes classifier. The naive Bayes classifier outperformed the SVM, reporting an accuracy of 59 and 60 percent for arousal and valence respectively.

In a recent work, deep learning was used to classify cognitive load and expertise of learners in a medical simulation using ECG [3]. ECG data were recorded from 9 medical practitioners in 2 classes, expert and novice. First, baseline ECG was collected as participants were in rest. Next, ECG signals were collected during simulation where the subjects had to tend to an injured manikin. R-peaks were then detected, followed by the extraction of 11 time-domain features. Additionally, 9 frequency domain features were extracted successive to calculating the power of different frequency bands. Finally, the extracted features were normalized by corresponding baseline features. A deep neural network with 7 layers was employed to perform classification of expertise and cognitive load. The model achieved an accuracy of 89.4 and 96.6 percent for cognitive load and expertise respectively. The research shows an interesting relationships between level of expertise and cognitive load, indicating that under stressful conditions, experts tend show lower cognitive load compared to novice subjects.

A number of prior works such as [4], [17], [22], [23], [32], and [9] have used one or more of the four datasets utilized in this paper, AMIGOS, DREAMER, WESAD, and SWELL, for emotion recognition. In [4], a Gaussian Naive Bayes (GNB) classifier was used on AMIGOS and reported F1 scores of 54.5 and 55.1 percent for classifying arousal and valence respectively. To perform a similar study, a CNN was used in [32], reporting accuracies of 81 and 71 percent for arousal and valence classification. An initial study on DREAMER was performed in [22] where arousal and valence classification was performed using an SVM classifier and an accuracy of 62.37 percent was reported. A novel technique was employed in [17] to perform emotion recognition on both DREAMER and AMIGOS. Two different sets of features were fused to perform the classification. First HRV features were extracted by calculating RR intervals from the input ECG signals. Next, spectrogram images were generated from time-series ECG signals. A pre-trained VGG-16 was used to extract feature embeddings of size 4096, followed by dimensionality reduction to 30 dimensions. Finally, both sets of features were concatenated and used to perform classification by means of an Extreme Learning Machine (ELM) classifier. Three-level affect state detection was performed in [23], where a Linear Discriminant Analysis (LDA) classifier was used on WESAD and reported an accuracy of 66.29 percent. Later, [34] proposed a Mutlimodal-Multisensory Sequential Fusion (MMSF) model, which showed that the proposed model can be trained on different signals with different sampling rates in the same training batch. When utilizing only ECG, an accuracy of 83 percent was reported for detection of affect state on WESAD. In [35], stress detection was performed in SWELL. Feature analysis was performed on extracted time domain and frequency domain features. Time domain feature NN50 showed the best performance compared to other features and reported an accuracy of 86.36 percent by utilizing an SVM classifier.

2.3 Self-Supervised Representation Learning

Recent research in the area of machine learning and deep learning has shown the potential of self-supervised models in learning generalized and robust representations. This concept has been utilized in computer vision [21], [36], [37], [38], [39], speech [25], natural language processing [40], and others. Self-supervised learning is a machine learning paradigm in which a network is trained using automatically generated labels instead of human-annotated labels. Different techniques have been employed to generate automatic labels. For example, a self-supervised network was proposed in [41], where high-level spatiotemporal features were learned from unlabelled videos. The network was trained using different rotated video clips to predict rotations. Next, the trained model was fine-tuned on the action recognition datasets. Their research showed that activity recognition accuracy was improved by more than 15 percent on both utilized datasets, compared to a network trained in a fully-supervised manner. In another work in [36], 3D pose estimation was tackled using self-supervised learning. To overcome the need for large amounts of 3D ground truth data generally required for 3D pose estimation, epipolar geometry was employed to calculate 3D poses from the available 2D poses. Subsequently, a convolution neural network was trained using the available 3D poses and reported state-of-the-art results. Self-supervised learning has also been used in learning wearable data. For example, human
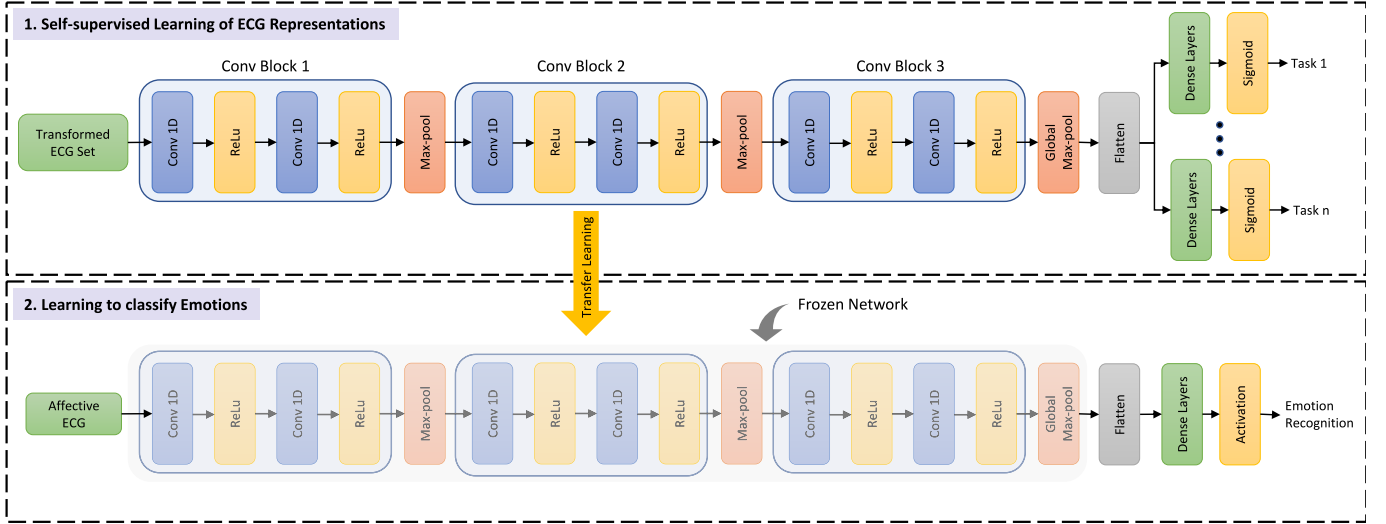
Fig. 2. The self-supervised architecture is presented. First, a multi-task CNN is trained using automatically generated labels to learn ECG representations. Then, the weights are transferred to the emotion recognition network, where the fully connected layers are trained to classify emotions.

activity recognition from Inertial Measurement Units (IMU) was performed in [19]. A set of different signal transformations were performed to train a convolutional neural network to predict transformations. Further, features were extracted from the trained network, and fully-connected layers were utilized to learn different activities such as walking, sitting, and jogging among others using the learned representations.

The works presented above, among others, demonstrate the advantages and promising results obtained using self-supervised learning in other domains. Nonetheless, such a framework has not been proposed for ECG representation learning, which we undertake in this paper in the context of emotion recognition.

3 SELF-SUPERVISED SOLUTION

Our goal in this paper is to learn ECG representations R_{ECG} capable of distinguishing between different classes of emotion. The proposed framework consists of two stages of learning: first, learning ECG representation, and second, using the network capable of extracting ECG representation to learn another network capable of classifying emotions. Accordingly, we define two sets of tasks, *pretext* and *downstream*, referred to as T_p and T_d respectively, where T_p is the set of signal transformation recognition tasks through which ECG representations are learned, while T_d learns to classify emotions. The goal of T_p is to learn robust generalized features from unlabelled ECG data through a self-supervised process, i.e., recognition of signal transformations. In the second stage in which T_d is learned, the original ECG data and the human-annotated emotion labels y_i are used. In this stage, the framework uses the frozen convolution layers of the first network to learn emotional classes upon supervised learning of the fully-connected layers at the end of the network. Fig. 2 shows the proposed self-supervised solution. In the following subsections, we further describe the architecture and details of each learning stage. The optimum architecture and hyperparameters for these two networks described below are selected either empirically or based on

searching among a large number of combinations for optimum performance.

3.1 Self-Supervised Learning of ECG Representations

We propose the use of a self-supervised deep multi-task CNN to learn R_{ECG}, through recognition of different transformations applied to input ECG signals (without utilizing any of human-annotated labels provided with the data). A similar approach was used in [19] for human activity recognition. Let's denote an example tuple of inputs and pseudo-labels for T_p as (X_j, P_j) where X_j is jth transformed signal, P_j is the automatically generated label according to the jth transformation, and $j \in [0, N]$ where N is the total number of signal transformations. To learn T_p, a signal transformation recognition network, $\gamma(X_j, P_j, \theta)$ is trained, where θ is the set of trainable parameters. The model learns θ by minimizing the total loss, L_{total}, composed of the weighted average of the individual losses of the signal transformation network L_j. The total network loss is defined by

$$L_j = [P_j \log \psi_j + (1 - P_j) \log (1 - \psi_j)], \quad (1)$$

$$L_{total} = \sum_{j=0}^{N} \alpha_j L_j, \quad (2)$$

where ψ_j and α_j are the predicted probability and loss coefficient of the jth transformation task respectively.

Six different signal transformation recognition tasks are performed as pretext tasks [19], [42]. The details of the parameters associated with these tasks (for instance signal-to-noise ratio, scaling factor, stretching factor, and others) are tuned based on a large number of experiments and are explored in depth in Sections 6.1 and 6.2. These transformations are described below and a sample of transformed signals with automatically generated labels is illustrated in Fig. 3.

Noise Addition. In this transformation, random noise from a Gaussian distribution, $N(t)$, is added to the original ECG signal $S(t)$. $N(t)$ is obtained from a random noise generator
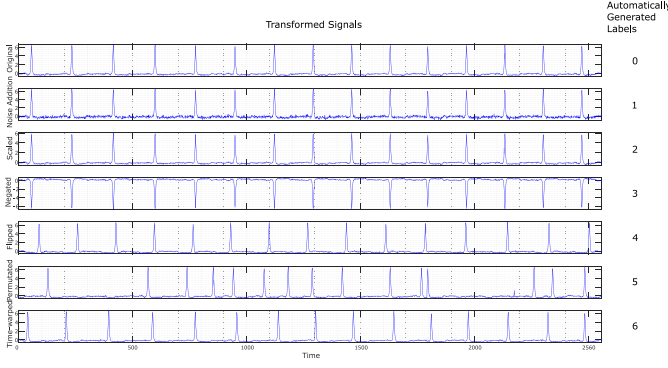
Fig. 3. A sample of an original ECG signal with the six transformed signals along with automatically generated labels are presented.

TABLE 1
The Architecture of the Signal Transformation Recognition Network is Presented

Module	Layer Details	Feature Shape
Input	–	2560×1
Shared Layers	$[conv, 1 \times 32, 32] \times 2$	2560×32
	$[maxpool, 1 \times 8, stride = 2]$	1277×32
	$[conv, 1 \times 16, 64] \times 2$	1277×64
	$[maxpool, 1 \times 8, stride = 2]$	635×64
	$[conv, 1 \times 8, 128] \times 2$	635×128
	global max pooling	1×128
Task-Specific Layers	$[dense] \times 2$ $\times$ *7 parallel tasks*	128
Output	–	2

where the mean of the distribution is set to 0, and the standard deviation to $\sqrt{E_{N_{avg}}}$. Here, the average power of $N(t)$, $E_{N_{avg}}$ is calculated as $10^{(E_{S_{avg}} - \alpha)/10}$, where $E_{S_{avg}}$ is the average power of $S(t)$ and α is the desired Signal to Noise Ratio (SNR). Finally the noise-added signal is generated as $S(t) + N(t)$.

Scaling. The original magnitude of the ECG $S(t)$ is transformed as $\beta \times S(t)$, where $\beta > 0$ and β is the manually assigned scaling factor.

Negation. The original amplitude of the ECG $S(t)$ is multiplied by -1, causing a spatial inversion of the time-series. The transformed signal can be mathematically expressed as $-S(t)$.

Temporal Inversion. Given the original ECG signal $S(t)$, where $t = 1, 2, \ldots, N$ and N is the length of the time-series, the temporally inverted version of the signal is expressed as $S'(t)$, where $t = N, N - 1, \ldots, 1$.

Permutation. In this transformation, the original ECG, $S(t)$, is divided into m segments and shuffled, randomly perturbing the temporal location of each segment. Let's assume $S(t)$ is expressed as $S(t) = [s_i(t)|i = 1, 2, \ldots, m]$, a sequence of segments with segment numbers $i = 1, 2, \ldots, m$. Accordingly, the permuted signal $S_p(t)$ can be obtained as $S_p(t) = [s_i(t)]$, where the sequence of $i = 1, 2, \ldots, m$ is randomly shuffled.

Time-Warping. In this transformation, randomly selected segments of the original ECG signals $S(t)$ are stretched or squeezed along the x (time) axis. Let's assume $\Phi(S(t), k)$ is an interpolation-based time-warping function where k is the stretch factor (with the corresponding squeeze factor represented as $1/k$). If the signal, $S(t)$, is expressed as $S(t) = [s_i(t)|i = 1, 2, \ldots, m]$, a sequence of segments with segment numbers $i = 1, 2, \ldots, m$, the time-warped (transformed) signal $T(t)$ is obtained by applying $\Phi(s_i(t), k)$ to half of the segments selected randomly, and $\Phi(s_j(t), 1/k)$ to the other half of the segments where $i \neq j$. Where $T(t)$ becomes longer or shorter than $S(t)$ due to the even or odd nature of m, $T(t)$ is clipped or zero-padded accordingly to maintain the original input length.

All of the above-mentioned transformations are applied to the original signals, the outcome of which are stacked to create the input matrix X, while the corresponding labels of the transformations $0, 1, \ldots, 6$ are stacked to create the corresponding output matrix P, with 0 denoting the original matrix and integers 1 to 6 indicating the 6 transformations.

The input-output matrices are then shuffled to re-order the transformations and their corresponding outputs.

The architecture of the proposed network for learning ECG representations using the created transformation dataset described above consists of 3 convolutional blocks (conv-block) as *shared layers*, and 7 branches, each with 2 dense layers, as *task-specific layers*. Each conv-block is made of $2 \times 1D$ convolution layers with ReLu activation functions followed by a max-pooling layer of size 8. The number of filters are increased from 32 to 64 and 128, whereas, the kernel size is decreased from 32 to 16 and 8 respectively. Global max-pooling operation is performed after the final conv-block. Next, successive to flattening the embedding, the output is fed to 2 fully connected layers with 128 hidden nodes and ReLu activation functions. To overcome possible overfitting, 60 percent dropout is introduced. Finally, the model output is generated through a sigmoid layer. The summary of the network architecture is presented in Table 1. Through learning the set of synthesized input-outputs described above, the network can learn strong ECG representations, which can be obtained by flattening the feature embeddings extracted after the shared convolutional blocks.

3.2 Emotion Recognition

In the second stage (T_d) of learning, the model is trained with the true emotion labels y_i for emotion classification. R_{ECG} obtained from T_p contain useful information regarding the original signals X_0, which are used to perform T_d. To utilize R_{ECG} towards performing T_d, a simple network $\rho = w(R_{ECG}, y_i, \theta')$ is employed, with ρ as the probability vector of T_d classes, and θ' as the set of trainable parameters. Finally, we calculate the optimum value of θ' by minimizing the cross entropy loss

$$L = \sum_{i=1}^{M} y_i \log \rho_i, \tag{3}$$

where M is total number of emotional classes.

The emotion recognition network contains convolution layers similar to those used in learning ECG representation, followed by fully connected layers with 512 hidden nodes. The weights of the shared layers of the signal transformation recognition network are frozen and transferred to the emotion recognition network. However, the fully connected layers of the emotion recognition network are not carried over from the transformation recognition network, and are instead trained using the labelled dataset (fully supervised)

TABLE 2
The Summary of the Four Datasets Used are Presented

Dataset	Participants	Attributes	Classes
AMIGOS	40	Arousal	9
		Valence	9
DREAMER	23	Arousal	5
		Valence	5
WESAD	17	Affect State	4
SWELL	25	Stress	3
		Arousal	9
		Valence	9

to perform emotion recognition. It should be noted that as we use different datasets in this study, the fully-connected layers are fine-tuned for each dataset. Specifically, 2 dense layers are utilized for SWELL and WESAD datasets, while 3 dense layers with L2 regularization (0.0001) and dropout of 40 and 20 percent are used for AMIGOS and DREAMER datasets respectively. The final output is taken from a sigmoid (binary classification) or softmax (multi-class classification) layer. We intentionally kept the fully-connected layers simple (single-task as opposed to multi-task) and somewhat shallow, in order to be able to evaluate the ability of the self-supervised approach to learn robust generalized ECG representations. The overview of the self-supervised solution is presented in Fig. 2, where successive to a self-supervised learning ECG representations, transfer learning [43] is used for emotion recognition with the fully-supervised model.

4 EXPERIMENTS

4.1 Datasets

We used four publicly available datasets to evaluate the proposed solution in depth on a large number of different subjects, in different circumstances and under different data collection protocols, and using different hardware. The important characteristics of the datasets are summarized in Table 2 and a brief description of each dataset is provided bellow.

4.1.1 AMIGOS

The AMIGOS dataset [4] was collected from 40 participants to study personality, mood, and affective responses, while engaging with multimedia content in two different contexts: a) alone, and b) in a group of four people. To perform this study, participants were shown short and long emotional video clips to elicit affective states. The short video clips had a length of 250 seconds, whereas the long video clips were longer than 14 minutes. ECG data were recorded using Shimmer sensors [44] at a sampling frequency of 256 Hz. A total of 16 video clips were shown to each participant, and self-assessed arousal and valence scores were recorded after each trial on a scale of 1 to 9.

4.1.2 DREAMER

The DREAMER dataset [22] was collected from 23 participants while they were presented with audio and video stimuli in the form of movie clips in order to elicit emotional reactions. A total of 18 video clips were shown, inducing 9 different emotions namely amusement, excitement, happiness, calmness, anger, disgust, fear, sadness, and surprise. Each clip was 60 seconds long. In addition, neutral clips were shown before each session to help subjects return to a neutral emotional state. ECG signals were collected using shimmer sensors [44] at a sampling rate of 256 Hz. Subjective experience of arousal and valence scores were collected using self-assessment manikins [45]. At the end of each session, both arousal and valence scores were recorded for the entire session on a scale of 1 to 5, ranging from uninterested/bored to excited/alert for arousal and unpleasant/stressed to happy/elated for valence.

4.1.3 WESAD

The dataset for WEarable Stress and Affect Detection (WESAD) contains ECG data from 17 participants. A RespiBAN Professional [46] sensors were used to collect ECG at a sampling rate of 700 Hz. The goal was to study four different affective states namely neutral, stressed, amused, and meditated. To perform this study, four different test scenarios were created. First, 20 minutes of neutral data were collected, during which participants were asked to do *normal* activities such as reading a magazine and sitting/standing at a table. During the *amusement* scenario, participants watched 11 funny video clips for a total length of 392 seconds. Next, participants went through public speaking and arithmetic tasks for a total of 10 minutes as part of the *stress* scenario. Finally, participants went through a guided *meditation* session of 7 minutes in duration. Upon completion of each trial, the ground truth labels for the affect states were collected using the Positive and Negative Affect Schedule (PANAS) scheme [47].

4.1.4 SWELL

This dataset was created with the goal of understanding employees' mental stress or emotional attributes in a typical office environment under different conditions [9]. Three types of scenarios namely, normal, time pressure, and interruptions were designed. Under the *normal* condition, participants were allowed to work on different tasks, for example preparing reports, making presentations, and so on, which were carried out over a maximum duration of 45 minutes. During the *time pressure* condition, the allowed time was reduced to 30 minutes to induce pressure. In the *interruption* session, participants were interrupted by sending emails and messages, with the participants being asked to respond to these messages. ECG signals were collected from 25 participants using a TMSI MOBI device [48] with self-adhesive electrodes at a sampling rate of 2048 Hz. Finally, at the end of each scenario, the participants' self-reported affect scores were collected on a scale of 1 to 9.

4.2 Data Pre-Processing

Since the above-mentioned datasets have been collected using different hardware, the signals have different spatio-temporal properties such as spatial range and sampling rate. To minimize the effects of such inter-dataset variations and discrepancies, three simple pre-processing steps were taken. First, SWELL and WESAD ECG signals are
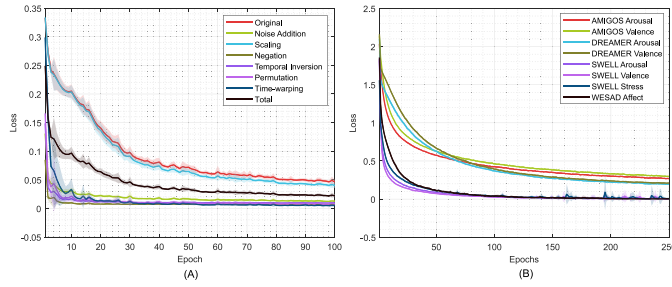
Fig. 4. Training losses versus training epochs are presented for both networks, the of signal transformation network (left) and the emotion recognition network (right). In (A), the average and standard deviation of individual losses and the total output loss of the signal transformation recognition network are obtained from 10-fold training, while in (B), the average and standard deviation of individual losses for the different emotion recognition tasks obtained with 10-fold training for each dataset are presented.

downsampled to 256 Hz to be consistent with AMIGOS and DREAMER. Next, we remove baseline wander from all the four datasets by applying a high-pass IIR filter with a pass-band frequency of 0.8 Hz. Lastly, we perform user-specific z-score normalization. While a number of other pre-processing operations such as feature extraction could have been done, we intentionally kept the pre-processing minimal and simple in order to better understand the impact of the proposed model on learning important ECG representations based on almost raw input. Finally, the filtered ECG signals are segmented into a fixed window size of 10 seconds and stacked into an array. No overlap is designated between segments to avoid any potential data leakage between training and testing data. It should be noted that the selection of the window size was empirical. Prior works utilizing these datasets have selected a wide range of different window sizes. For example, [4] and [32] selected a window length of 20 seconds for AMIGOS, whereas [22] and [35] selected a window of 60 seconds for both DREAMER and SWELL. As other examples, [23] has selected 5-second windows for WESAD, while [34] has selected 1-second windows for the same dataset.

4.3 Implementation and Training

In order to train the model, successive to pre-processing, each segment is used to generate the 6 transformations described earlier. Finally, the original ECG signals along with the transformed signals are used to train the signal transformation recognition network. We implement the proposed architecture using TensorFlow. We share the implementation of the self-supervised network.[1]

Similar to other works in this area [4], [22], [23], [32], [34], [35], we use 10-fold cross-validation to evaluate the performance of the model successive to shuffling of the pre-processed dataset. We randomly select 90 percent of the data for training and the remaining 10 percent is used for testing (this process is repeated 10 times without repeating the shuffling step). To train both the signal transformation recognition network and the emotion recognition network, Adam optimizer [49] is used with a learning rate of 0.001

TABLE 3
The Average and Standard Deviations of the Accuracies and F1 Scores for Signal Transformation Recognition Across the Four Datasets are Presented

Transformation	All datasets combined	
	Acc.	F1
Original	0.980 ± 0.003	0.927 ± 0.007
Noise Addition	0.995 ± 0.000	0.979 ± 0.003
Scaling	0.982 ± 0.003	0.932 ± 0.010
Temporal Inversion	0.998 ± 0.000	0.992 ± 0.004
Negation	0.998 ± 0.000	0.990 ± 0.000
Permutation	0.998 ± 0.000	0.989 ± 0.003
Time-warping	0.997 ± 0.003	0.992 ± 0.006
Average	0.992 ± 0.001	0.972 ± 0.005

and batch size of 128. The signal transformation recognition network is trained for 100 epochs, while the emotion recognition network is trained for 250 epochs. The number of training epochs for each network is selected to enable the training reach a steady state. Figs. 4 A and 4 B show the loss versus training epoch for the transformation recognition and emotion recognition networks respectively during training. Fig. 4 A shows that the loss for temporal inversion, negation, permutation, and time-warping reach steady states earlier than the other transformations. Therefore, the network is trained until all the individual losses reach their respective steady states. Fig. 4 B shows that the different datasets reach their training steady states at different epochs. For example, the loss of WESAD and SWELL become stable well before 100 epochs, while the steady states of AMIGOS and DREAMER are achieved after 200 epochs. In summary, Fig. 4 shows that both pretext tasks and emotion recognition tasks train well and reach steady states using the proposed solution.

5 RESULTS

5.1 Signal Transformation Recognition

Tables 3 and 4 presents the F1 scores and accuracies of our proposed network for transformation recognition. Results shown in Table 3 are calculated on all the datasets combined. The results show very high F1 scores and accuracies for all the transformations as well as the original signal. F1 scores 0.927 and 0.932 are achieved when detecting the original signal and the scaling transformations. Next, relatively higher F1 score of 0.979 is achieved for recognition of noisy signals. We notice few tasks consistently report very high scores (≈ 0.99), namely temporal inversion, negation, permutation, and time-warping. For accuracy, very high values (greater than 0.98) are achieved for every transformation. Beside high F1 scores and accuracies, we also achieve very low standard deviations for all the transformation recognition tasks, which indicates a consistent performance by the network in all the training folds. Finally, an average F1 score of 0.972 and an average accuracy of 0.992 are reported for all the tasks combined.

Next, we further investigate the consistency and generalization of the performance of the transformation recognition network across different datasets. Table 4 presents the results on each of the four datasets, showing very high F1

TABLE 4
Results for the Individual Transformation Recognition Tasks in Each of the Four Datasets are Presented

Transformation	AMIGOS		DREAMER		WESAD		SWELL	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Original	0.973 ± 0.003	0.901 ± 0.010	0.986 ± 0.005	0.953 ± 0.020	0.981 ± 0.008	0.950 ± 0.014	0.984 ± 0.003	0.941 ± 0.009
Noise Addition	0.987 ± 0.002	0.954 ± 0.008	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.000	0.999 ± 0.000
Scaling	0.975 ± 0.003	0.904 ± 0.020	0.986 ± 0.005	0.955 ± 0.024	0.984 ± 0.008	0.952 ± 0.012	0.987 ± 0.001	0.948 ± 0.006
Temporal Inversion	0.997 ± 0.001	0.988 ± 0.004	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.001	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Negation	0.996 ± 0.001	0.987 ± 0.005	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.002	0.995 ± 0.007	1.000 ± 0.000	0.998 ± 0.004
Permutation	0.996 ± 0.001	0.984 ± 0.005	0.999 ± 0.000	0.999 ± 0.003	0.999 ± 0.001	0.994 ± 0.007	0.998 ± 0.001	0.991 ± 0.009
Time-warping	0.998 ± 0.001	0.993 ± 0.007	0.998 ± 0.003	0.999 ± 0.003	0.993 ± 0.009	0.991 ± 0.006	0.996 ± 0.004	0.990 ± 0.005
Average	0.989 ± 0.002	0.959 ± 0.008	0.996 ± 0.002	0.987 ± 0.004	0.994 ± 0.004	0.983 ± 0.007	0.995 ± 0.001	0.981 ± 0.005

scores and accuracies. This is a strong indicator of the generalizability of our approach as the model performs consistently on all the datasets with F1 scores of 0.959 to 0.987 and accuracies of 0.989 to 0.996, with low standard deviations in every case.

5.2 Emotion Recognition

Results obtained from the emotion recognition network are presented in Table 5. Arousal and valence detection are performed with three datasets AMIGOS, DREAMER, and SWELL, stress detection is performed with SWELL, and classification of different affective states are performed in WESAD, as per the availability of output labels. As presented in Table 2, the four datasets contain different number of classes for each target output. Therefore, the classification tasks are performed accordingly, the accuracies and F1 scores of which are presented in Table 5. It should be noted that to the best of our knowledge, no previous work has utilized AMIGOS, DREAMER and SWELL datasets for multi-class classification. this is the first time, multi-class classification of arousal and valence score is attempted on. With AMIGOS, we achieve accuracies of 79.6 and 78.3 percent for arousal and valence respectively, where 9-class classification is performed. In DREAMER, 5-class classification is performed, achieving accuracies of 77.1 and 74.9 percent for arousal and valence respectively. Next, we perform detection of 4 affective states using WESAD, reporting an accuracy of 95.0 percent. Lastly, arousal, valence, and stress detection is performed on SWELL, achieving accuracies of 92.6, 93.8, and 90.2 percent respectively, where arousal and valence scores have 9 classes while stress has 3 classes.

TABLE 5
The Emotion Categories, Number of Classes for Each Class, and Multi-Class Emotion Recognition Results are Presented for Each of the Four Datasets

Dataset	Attribute	Classes	Acc.	F1
AMIGOS	Arousal	9	0.796	0.777
	Valence	9	0.783	0.765
DREAMER	Arousal	5	0.771	0.740
	Valence	5	0.749	0.747
WESAD	Affect State	4	0.950	0.940
SWELL	Arousal	9	0.926	0.930
	Valence	9	0.938	0.943
	Stress	3	0.902	0.900

Comparison. To further evaluate the proposed framework, we compare our model's performance to past work as shown in Table 6. It should be noted that most of the prior works using the four studied datasets have performed two-class or three-class classifications [4], [9], [22], [23] instead of utilizing all the multiple classes available for each dataset. Therefore, in order to perform a fair comparison, in addition to the multi-class classifications presented in Table 5, in Table 6 we present the results of the model applied to two/three-class classifications in accordance to the prior works and state-of-the-art methods presented in this table. In Table 6, AMIGOS, DREAMER, and SWELL have all been classified in two-class format by setting a threshold at the mean output value dividing the outputs to high and low, while WESAD has been classified in three classes by ignoring the meditated output class. It should be noted that the works mentioned in Table 6 have also adopted a 10-fold cross-validation scheme similar to the validation scheme used in this paper. Moreover, for each dataset, we have also carried out *fully-supervised* classification by skipping the self-supervised step and directly training the emotion recognition network using the input ECG and emotion labels. In [32], a CNN was implemented to perform emotion classification on AMIGOS. It reported accuracies of 81 and 71 percent for classification of arousal and valence respectively. The results in Table 6(A) show that the self-supervised CNN achieves accuracies of 88.9 and 87.5 percent for classification of arousal and valence respectively, outperforming the past works as well as the fully-supervised baseline CNN. In [22], classification of arousal and valence was performed with DREAMER. An SVM classifier was implemented, reporting an accuracy of 62.4 percent for both arousal and valence. Table 6(B) shows that the proposed solution achieves considerable improvement over the state-of-the-art and the baseline CNN with accuracies of 85.9 and 85 percent for arousal and valence respectively. In a study on WESAD, [34] proposed a multimodal CNN that was able to achieve state-of-the-art with an accuracy of 83 percent while using only ECG. As shown in the Table 6(C), our proposed model is able to outperform the CNN in [34] as well as the baseline CNN with accuracy of 96.9 percent. Lastly, a study on SWELL performed stress detection [35], in which two classifiers, *k*NN and SVM, were utilized. The results in Table 6(D) show that the baseline CNN slightly outperforms the SVM, whereas the self-supervised model considerably outperforms both with an accuracy of 93.3 percent. Additionally, we perform two-class classification of arousal and valence on SWELL for the first time, achieving very accurate results.

TABLE 6
The Results of the Self-Supervised Model on all the Datasets are Presented and Compared With
Prior Works Including the State-of-the-Art, as well as a Fully-Supervised CNN as a Baseline

A: AMIGOS

Ref.	Method	Arousal		Valence	
		Acc.	F1	Acc.	F1
[4]	GNB	–	0.545	–	0.551
[32]	CNN	0.81	0.76	0.71	0.68
Ours	Fully-Supervised CNN	0.844	0.835	0.811	0.809
	Self-Supervised CNN	**0.889**	**0.884**	**0.875**	**0.874**

B: DREAMER

Ref.	Method	Arousal		Valence	
		Acc.	F1	Acc.	F1
[22]	SVM	0.624	0.580	0.624	0.531
Ours	Fully-Supervised CNN	0.707	0.708	0.666	0.658
	Self-Supervised CNN	**0.859**	**0.859**	**0.850**	**0.845**

C: WESAD

Ref.	Method	Affect State	
		Acc.	F1
[23]	kNN	0.548	0.478
	DT	0.578	0.517
	RF	0.604	0.522
	AB	0.617	0.525
	LDA	0.663	0.560
[34]	CNN	0.83	0.81
Ours	Fully-Supervised CNN	0.932	0.912
	Self-Supervised CNN	**0.969**	**0.963**

D: SWELL

Ref.	Method	Stress		Arousal		Valence	
		Acc.	F1	Acc.	F1	Acc.	F1
[35]	kNN	0.769	–	–	–	–	–
	SVM	0.864	–	–	–	–	–
Ours	Fully-Supervised CNN	0.894	0.874	0.956	0.962	0.961	0.956
	Self-Supervised CNN	**0.933**	**0.924**	**0.967**	**0.964**	**0.973**	**0.969**

It should be noted that in order to perform a fair comparison, Table 6 excludes prior works such as: i) [50] and [17] which have used different validation schemes; ii) [51], [52], [53] which have used multiple modalities (ECG with EEG or GSR) available in some of the datasets; and iii) [54], which has formulated the problem as regression instead of classification.

6 ANALYSIS AND DISCUSSION

6.1 Single-Task Self-Supervision

We study the impact of individual transformations on the emotion recognition performance. For this study, we create a single-task CNN, instead of a multi-task CNN. For this study, we select a range of possible values for the parameters that control the significance of each individual transformation. For noise addition, the noise amplitude parameter is changed to result in signal-to-noise ratios of 2 to 45. For the scaling transformation parameter, scaling factors of 0.1 to 10 are selected. For permutation, the number of segments is varied between 2 to 40. For time-warping, the stretch factor is varied between 1.05 to 4, while the number of segments varies from 2 to 40. Finally, temporal inversion and negation, which do not have any controllable parameters are also studied. We should point out that these ranges of parameters are selected to create a wide range of transformed signals, almost identical to the original signal in one end and considerably different in the other. Fig. 5 shows the emotion recognition performance for each dataset versus the controllable parameters for each individual transformation task. To simplify the search, a single 90-10 split for training-testing was performed instead of 10-fold. This analysis provides in-depth insight into the effect of the parameters associated with each self-supervised recognition task (transformation) on the emotion recognition outcome. Furthermore, this analysis helps us narrow down the set of suitable transformation parameters in order to achieve the best performance. From Fig. 5 A, we notice that for SNR values of 15 and 20 the model shows highest performance for

emotion recognition compared to SNR values greater than 30 or less than 10. Moreover, from Fig. 5 B, we notice that for the scaling factor of 1, where the scaled signal and the original signals are identical, the model performance on emotion recognition is poor. However, for slightly greater or smaller scaling values, the model shows significantly better performance. Interestingly, the performance drops when the scaling factor is greater than 1.4 or less than 0.6. Fig. 5 C illustrates that when utilizing permutations, poor performance is achieved when the input is divided into only 2 segments or more than 20 segments. Otherwise, when the number of segments varies in the range of 3 to 20, a relatively steady performance is observed. Time-warping analysis is presented in a 3D plot in Fig. 5 F where dark red regions show higher accuracies. It can be seen that number of segments in the range of 6 to 9 and a stretch factor in the range of 1.05 and 1.35 consistently result in better performance compared to other combinations.

In summary, the analysis above shows that for all the different transformation tasks, when the transformed signal is highly distorted compared to the original input, for example with high amplitude noise (SNR < 10), large number of segments (> 20), very high or very low scaling factor (< 1.5 or > 0.6), or large stretch factor (> 1.4), emotion recognition accuracy drops. This is due to the fact that distinguishing between the original signal and the highly distorted versions of the signal becomes simple. This results in the network not learning useful ECG representations. Similarly, when the transformed signal remains almost similar to the original signal, for instance when the SNR > 35, or a scaling factor in the range of 0.95 to 1.05, or the number of permutated segments equal to 2, the performance also drops. This is due to the fact that the recognition tasks in such cases become too difficult for the network to properly learn the required ECG representations from. Hence, we conclude that there is a range of values for the parameters associated with the self-supervised tasks, for which these tasks are located in an suitable difficulty range, resulting in optimum learning.
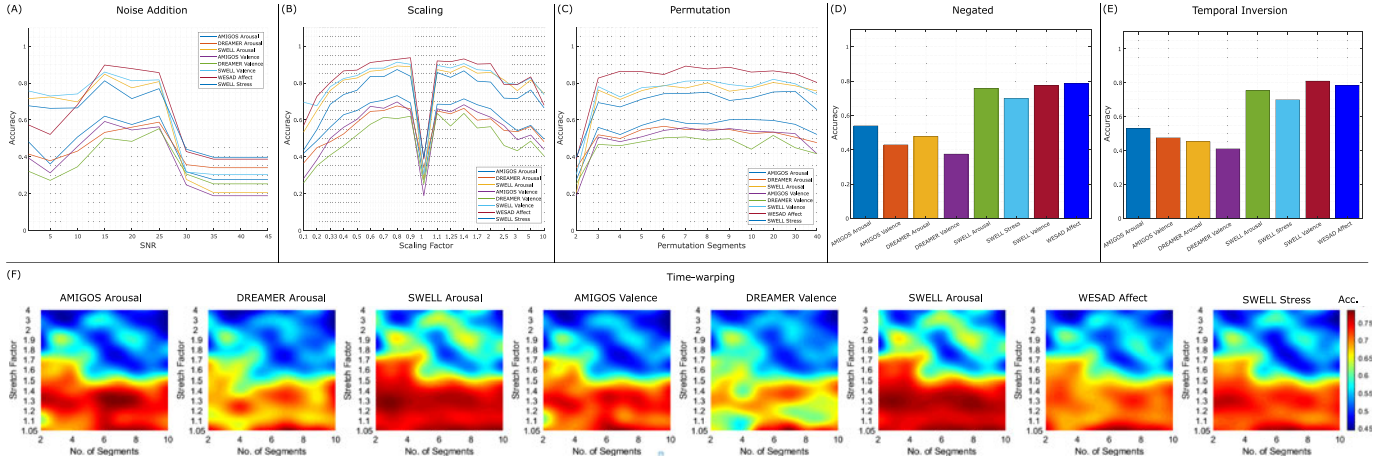
Fig. 5. The impact of transformation parameters on emotion recognition accuracy is illustrated. Parts (A), (B), and (C) are 2D graphs as noise addition, scaling, and permutation transformations only have one parameter. Parts (D) and (E) are presented in bar-plot format as temporal inversion and negation transformations are operations with no controlling parameter. Lastly, part (F) shows 3D plots for time-warping as this operation has two parameters, with x and y axes representing the number of segments and stretch factor respectively, and the accuracy is denoted by the color intensity.

6.2 Multi-Task Self-Supervision

This section discusses the emotion recognition performance when multiple signal transformations (multi-task) are performed to learn ECG representations. The reason we believe this analysis is necessary despite already analyzing individual transformations in the previous subsection is that the use of a multi-task network may have an impact on the overall performance which might be different than the aggregation of several individual self-supervised tasks. To perform this study, we use the results from the previous section (Section 6.1) to narrow down the search given the large degree of freedom in the multi-task parameter space. Also, to further simplify the search, similar to Fig. 5, a single 90-10 split for training-testing was performed instead of a complete 10-fold. The results obtained from different combination of signal transformation parameters are presented in a six-dimensional plot as shown in Fig. 6. Our analysis shows that $(15, 0.9, 20, 9, 1.05)$ is the optimal vector of parameters for SNR, scaling factor, number of permutation segments, number of time-warping segments, and time-warping stretch factor respectively. Our analysis further confirms that the results obtained from single-task self-supervision are in compliance with results obtained from multi-task self-supervision. For example, Fig. 5 A shows that an SNR = 15 results in better emotion recognition compared to SNR = 25. Very similar results are also observed in Fig. 6, as SNR = 15 results in the best accuracy in 6 out of 8

instances. As another example, in both single-task and multi-task settings, a stretch factor of 1.05 shows superior performance compared to 1.35.

Fig. 7 compares the emotion recognition results when ECG representations are learned using self-supervised single-tasks to when multi-task representation learning is used. These results are reported for the optimum parameters for each case. The figure shows that different transformation recognition tasks help the self-supervised model learn better high-level representations of ECG. As the multi-task network consistently outperforms single-task setups (more prominently in AMIGOS and DREAMER), it can be concluded that different transformations help the model in learning different aspects of ECG representations. For instance while noise addition, scaling, and negation may be contributing more towards learning spatial representations of ECG, others such as temporal inversion, permutation, and time-warping may be contributing more to learning the temporal relations. Moreover, with respect to the relationship between different transformation recognition and emotion recognition tasks, specific commonalities are observed. In particular, we see that scaling (T2) has greater impact on emotion recognition compared to noise addition (T1) and negation (T6). On the other hand, time-warping (T4) has greater impact compared to permutation (T3) and temporal inversion (T5).

In the multi-task network, the total loss is calculated as a weighted average of individual losses (see Eq. (2)). We
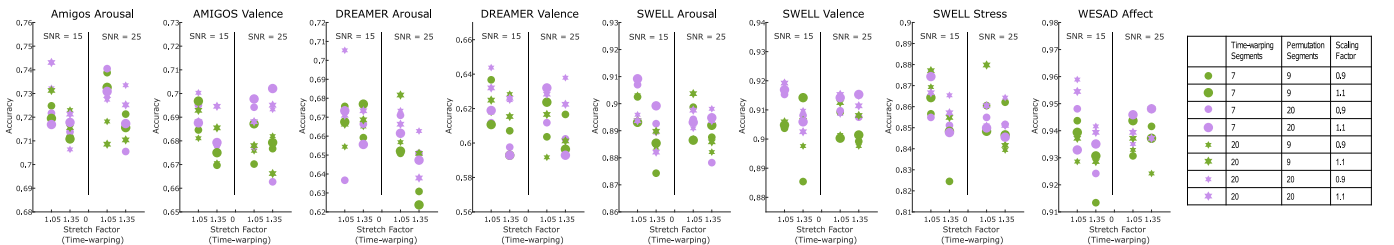

Fig. 6. Emotion recognition results versus 5D vector of parameters controlling all 6 tasks simultaneously are presented. The resulting 6 dimensional plots are represented as follows: Y axis presents the emotion recognition accuracy, the two plots separated by the vertical line correspond to SNR, the time-warping stretch factor is represented by the X axis, the scaling factor, permutation segments, and time-warping segments are represented by marker size, color, and shape respectively.
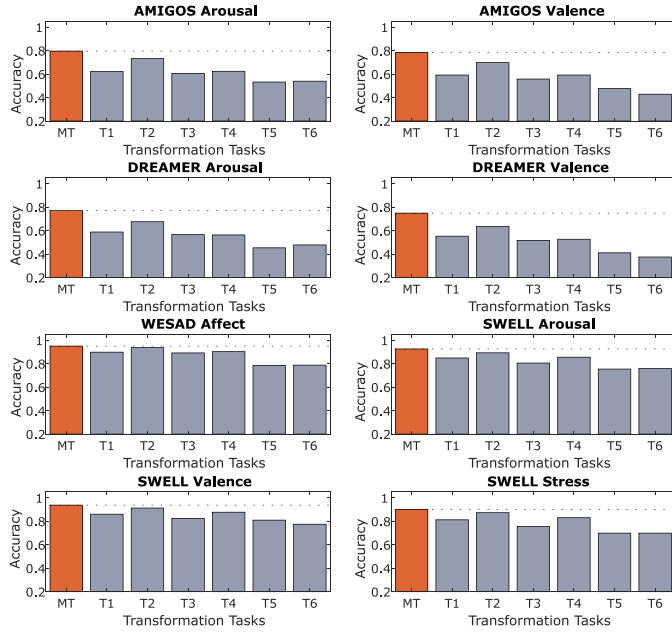
Fig. 7. Comparison between single-task and multi-task self-supervision is presented. MT denotes multi-task, while T1 through T6 denote noise addition, scaling, permutation, time-warping, temporal inversion, and negation respectively.
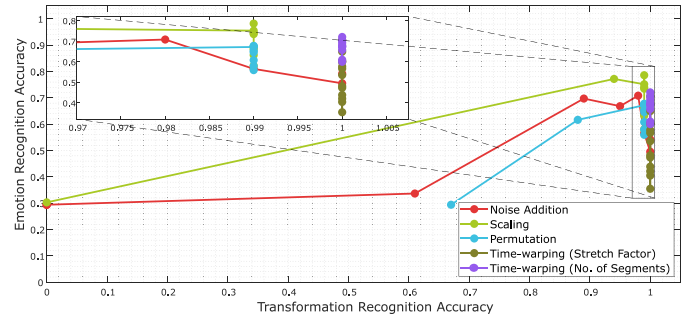

Fig. 8. The relationship between emotion recognition accuracy and transformation recognition is presented. Time-warping (Stretch Factor) denotes that the stretch factor is varied while the number of segments is set to the optimum value. Similarly, Time-warping (No. of Segments) indicates that the number of segments is varied while the stretch factor is set to the optimum value.

realize that these weights (loss coefficients) play an important role in learning the ECG representations since during training of the self-supervised network, different losses reach their steady states at different epochs. Specifically, the losses of the temporal inversion and spatial negation saturate very quickly and report almost perfect $F1$ scores just after 5 to 7 epochs. The possible reason for this phenomena is the clear difference between inverted and negated signals with respect to the original signal, making it very easy for the model to correctly classify these transformations. As a result, we utilize smaller weights for temporal inversion and negation losses compared to the other transformations. All the weights were set empirically with the goal of maximizing performance. The loss coefficients of 0.0125 are used for temporal inversion and negation, while for the rest of the 5 losses, the weights are set to 0.195.

6.3 Relationship Between Downstream and Pre-Text Tasks

To further understand the relationship between the accuracy of classification of the pre-text tasks and downstream emotion recognition, we present this relationship in Fig. 8. In this figure, the transformation recognition accuracies are obtained based on varying degrees of task difficulty as discussed in Section 6.1. Corresponding emotion recognition accuracies are then calculated for this analysis. Since similar patterns are noticed across different emotions and datasets, the average accuracies are utilized in the figure. For time-warping, there are two transformation parameters: number of segments and stretch factor. Therefore, first we keep the number of segments at optimum value according to the result achieved in Section 6.1, and vary the other transformation parameter (stretch factor), and vice versa. Temporal inversion and negation transformation are not included

here as there is no controllable parameters to vary the transformation recognition difficulty.

The figure indicates that in the case of noise addition, scaling, and permutation, when the transformation recognition accuracy is poor, the emotion classification performance is also poor. This behaviour occurs when the transformation recognition tasks are too difficult. On the other hand, when the transformed signals are highly distorted compared to the original signal, the model reports high accuracy in transformation recognition, while poor performance is acquired for classifying emotions. This observation further confirms that for an optimum amount of difficulty in the pretext tasks, the self-supervised model shows the best performance for downstream tasks.

6.4 Network Embeddings

Here, we evaluate the quality of the learned representations versus the depth of the self-supervised model. The representations learned with each conv-block are separately extracted and utilized to perform emotion recognition as shown in Fig. 9 A. This analysis helps us understand whether representations at different depths of the network are informative and valuable for the downstream tasks. First, the learned representations from conv-block 1, conv-block 2, and conv-block 3 are separately extracted and utilized to perform emotion recognition. As shown in Fig. 9 B, significant improvement in performance is noticed when features are extracted from conv-block 3 compared to conv-block 1 and conv-block 2. In addition to using the individual embeddings, we also stacked the three embeddings to create a new embedding, which we then used for the downstream tasks. However, the embedding obtained from conv-block 3 still showed the best performance. This analysis indicates that learning representations obtained from the last convolutional layer is better and more generalizable for the downstream emotion recognition.

6.5 Impact of Multiple of Datasets

One of the major advantages of the self-supervised solution is that it allows for the self-supervised network to be trained using an aggregation of the existing datasets by removing the main barrier of having different types of output labels for each dataset. To further analyze this aspect of the proposed
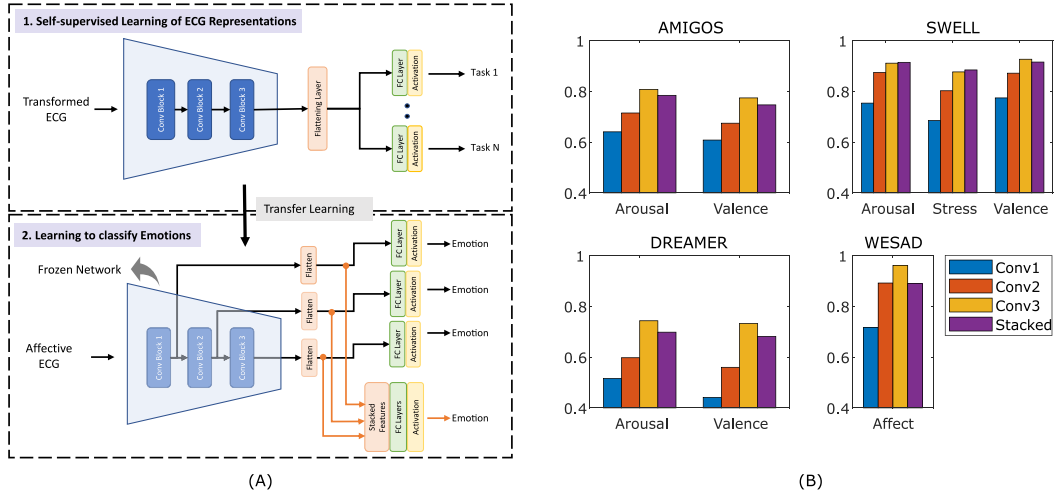
Fig. 9. The feasibility of using embeddings obtained through the network at different depths for emotion recognition is investigated. In (A) the schematic of the experiment is presented, while (B) presents the results.

solution, we investigate the impact of using all four datasets for training the network compared to when individual datasets are used. To perform this study, we use the four datasets separately to train the self-supervised signal transformation recognition network, which we then use for emotion recognition. Fig. 10 shows that using all 4 datasets combined, the model learns more generalized and robust features as evident by the better performance in classifying emotions compared to learning ECG representations from individual datasets. In the case of AMIGOS, DREAMER, and WESAD datasets, the performance improves significantly while marginal improvement is noticed for SWELL dataset.

6.6 Limitations and Future Work

In this paper we performed an extensive analysis of the proposed framework and highlight the benefits of the self-supervised approach over a fully-supervised technique. However, the limitations of the self-supervised solution should also be noted. We notice that the proposed solution performs poorly in subject-independent emotion recognition. We hypothesize the main possible reasons for this: *1)* subject-*invariant* features may require specific hand-crafted

features or a combination of hand-crafted and deep learning features such as those used in [3] and [17] respectively, which might not be learnable solely with convolutions applied to the raw ECG; *2)* Moreover, in order to properly normalize user-specific ECG features linked to factors such as genetics, physiology, health, and others, user-specific calibrations with respect to baseline sessions were carried out in [3], which enabled subject-independent validation to perform well. We avoided such calibration steps in the pipeline since: *a)* a calibration-free framework is generally more convenient and hence preferred, and *b)* some of the datasets did not include the information required to perform such calibrations.

For future work, multi-modal emotion recognition will be explored, exploiting other modalities such as EEG available in some datasets. Furthermore, we will employ the proposed solution for other ECG-related domains such as arrhythmia detection, ECG-based activity recognition, and others. Lastly, further research towards utilization of the proposed architecture for cross-subject and cross-corpus schemes will be carried out.

7 CONCLUSION

In this work we present an ECG-based emotion recognition solution using self-supervised deep multi-task learning. To the best of our knowledge, this is the first time self-supervised learning is utilized to perform emotion recognition using ECG. Four public datasets AMIGOS, DREAMER, WESAD and SWELL are used in this study to perform emotion recognition. We set new state-of-the-art results for the four datasets in classification of arousal, valence, affective states, and stress. We show that the proposed approach significantly improves classification performance compared to a fully-supervised solution. We explore different self-supervised transformation recognition tasks for learning the ECG representations and analyze their impacts. Our analysis shows that for an optimum amount of difficulty in the pretext tasks, the network learns better ECG representations which can used for emotion classification. Our experiments also show that for learning of ECG representations, a
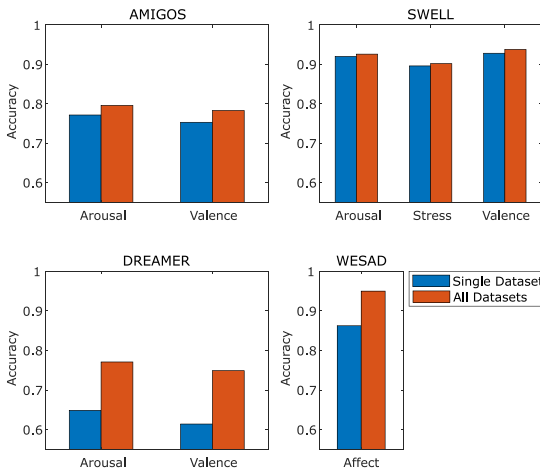

Fig. 10. The impact of using multiple datasets with the self-supervised architecture is investigated and presented.

multi-task CNN improves the performance compared to a single-task network. Lastly, our analysis on the use multiple datasets for training the self-supervised network shows that utilizing multiple datasets benefits the final downstream classification.

REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[2] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

[3] P. Sarkar, K. Ross, A. J. Ruberto, D. Rodenbura, P. Hungler, and A. Etemad, "Classification of cognitive load and expertise for adaptive simulation using deep multitask learning," in *Proc. 8th IEEE Int. Conf. Affect. Comput. Intell. Interaction*, 2019, pp. 1–7.

[4] 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.*, to be published, doi: 10.1109/TAFFC.2018.2884461.

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

[6] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Trans. Affective Comput.*, vol. 11, no. 2, pp. 200–213, Apr.-Jun. 2020.

[7] S. P. Shashikumar, A. J. Shah, Q. Li, G. D. Clifford, and S. Nemati, "A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology," in *Proc. IEEE Int. Conf. Biomed. Health Informat.*, 2017, pp. 141–144.

[8] J. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.

[9] 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.

[10] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affective Comput.*, vol. 9, no. 2, pp. 147–160, Apr.-Jun. 2018.

[11] A. Sepas-Moghaddam, A. Etemad, P. L. Correia, and F. Pereira, "A deep framework for facial emotion recognition using light field images," in *Proc. 8th IEEE Int. Conf. Affect. Comput. Intell. Interaction*, 2019, pp. 1–7.

[12] A. Etemad and A. Arya, "Classification and translation of style and affect in human motion using RBF neural networks," *Neurocomputing*, vol. 129, pp. 585–595, 2014.

[13] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003.

[14] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG pattern analysis for emotion detection," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 102–115, Jan.-Mar. 2012.

[15] G. Zhang and A. Etemad, "Capsule attention for multimodal EEG and EOG spatiotemporal representation learning with application to driver vigilance estimation," 2019, *arXiv: 1912.07812*.

[16] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *Proc. IEEE Int. Conf. Eng. Medicine Biol. Soc.*, 2006, pp. 1355–1358.

[17] S. Siddharth, T. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2019.2916015.

[18] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Trans. Affective Comput.*, vol. 6, no. 4, pp. 385–394, Jan.-Mar. 2015.

[19] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," *Proc. Interactive Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 2, 2019, Art. no. 61.

[20] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 759–766.

[21] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1329–1338.

[22] 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.

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

[24] M. Ravanelli et al., "Multi-task self-supervised learning for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6989–6993.

[25] M. Tagliasacchi, B. Gfeller, F. de Chaumont Quitry, and D. Roblek, "Self-supervised audio representation learning for mobile devices," 2019, *arXiv: 1905.11796*.

[26] P. Sarkar and A. Etemad, "Self-supervised learning for ECG-based emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 3217–3221.

[27] A. L. Goldberge, *Basic ECG Waves*. Maryland Heights, MO, USA: Mosby, 2006.

[28] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, 2019, Art. no. 65.

[29] Y. Hsu, J. Wang, W. Chiang, and C. Hung, "Automatic ECG-based emotion recognition in music listening," *IEEE Trans. Affective Comput.*, vol. 11, no. 1, pp. 85–99, Jan.-Mar. 2020.

[30] A. Page, A. Kulkarni, and T. Mohsenin, "Utilizing deep neural nets for an embedded ECG-based biometric authentication system," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, 2015, pp. 1–4.

[31] Y. Xu and G.-Y. Liu, "A method of emotion recognition based on ECG signal," in *Proc. IEEE Int. Conf. Comput. Intell. Natural Comput.*, pp. 202–205, vol. 1, 2009.

[32] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. J. I. A. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2018.

[33] K. Ross et al., "Toward dynamically adaptive simulation: Multimodal classification of user expertise using wearable devices," *Sensors*, vol. 19, no. 19, 2019, Art. no. 4270.

[34] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An explainable deep fusion network for affect recognition using physiological signals," in *Proc. 28th Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 2069–2072.

[35] S. Sriramprakash, V. D. Prasanna, and O. V. Ramana Murthy, "Stress detection in working people," *Procedia Comput. Sci.*, vol. 115, pp. 359–366, 2017.

[36] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3D human pose using multi-view geometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019.

[37] K. Wang, L. Lin, C. Jiang, C. Qian, and P. Wei, "3D human pose machines with self-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1069–1082, May 2020.

[38] M. Rastgar-Jazi and X. Fernando, "Detection of heart abnormalities via artificial neural network: An application of self learning algorithms," in *Proc. IEEE Canada Int. Humanitarian Technol. Conf.*, 2017, pp. 66–69.

[39] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu, "Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4006–4015.

[40] F. Wu and D. S. Weld, "Open information extraction using wikipedia," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 118–127.

[41] L. Jing, X. Yang, J. Liu, and Y. Tian, "Self-supervised spatiotemporal feature learning via video rotation prediction," 2018, *arXiv: 1811.11387*.

[42] T. T. Um et al., "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 216–220. [Online]. Available: https://doi.org/10.1145/3136755.3136817

[43] H. Shin et al.,"Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[44] Shimmer ECG, Accessed: Sep. 17, 2019. [Online]. Available: http://www.shimmersensing.com/products/shimmer3-ecg-sensor

[45] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[46] Respiban professional, Accessed: Sep. 17, 2019. [Online]. Available: https://www.biosignalsplux.com/index.php/respiban-professional

[47] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *J. Pers. Soc. Psychol.*, vol. 54, no. 6, 1988, Art. no. 1063.

[48] Tmsi-mobi, Accessed: Sep. 17, 2019. [Online]. Available: https://www.tmsi.com/products/mobi/

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[50] R. Harper and J. Southern, "A Bayesian deep learning framework for end-to-end prediction of emotion from heartbeat," *IEEE Trans. Affect. Comput.*, 2020.

[51] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep canonical correlation analysis," 2019, *arXiv: 1908.05349*.

[52] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Trans. Affective Comput.*, vol. 9, no. 2, pp. 227–239, Apr.-Jun. 2018.

[53] P. Verma and S. K. Sood, "A comprehensive framework for student stress monitoring in fog-cloud IoT environment: M-health perspective," *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 231–244, 2019.

[54] A. Alberdi, A. Aztiria, A. Basarab, and D. J. Cook, "Using smart offices to predict occupational stress," *Int. J. Ind. Ergonom.*, vol. 67, pp. 13–26, 2018.

Pritam Sarkar received the bachelor's degree in electrical engineering from the West Bengal University of Technology, Kolkata, India, in 2015. He is currently working toward the MASc degree in Electrical and Computer Engineering Department, Queen's University, Kingston, Canada. Prior to joining Queen's, he worked in the industry as a software engineer. Currently, his research interests include deep learning, affective computing, computer vision among others.

Ali Etemad received the MASc and PhD degrees in electrical and computer engineering from Carleton University, Ottawa, Canada, in 2009 and 2014, respectively. As of 2017, he has been an assistant professor at the Department of Electrical and Computer Engineering, Queen's University, Canada, where he leads the Ambient Intelligence and Interactive Machines (Aiim) Lab. His main area of research is machine learning and deep learning focused on human-centered applications with wearables and smart devices. Prior to joining Queen's, he held several industrial positions. He has co-authored more than 80 articles and granted/applied patents, and has delivered more than 15 invited talks regarding his work. He has been a member of program committees for several conferences in the field. His lab and research program have been funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Ontario Centers of Excellence (OCE), Canadian Foundation for Innovation (CFI), Mitacs, and other organizations, as well as the private sector.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.