IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 7, 2022, accepted January 26, 2022, date of publication February 7, 2022, date of current version February 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149509

SigRep: Toward Robust Wearable Emotion Recognition With Contrastive Representation Learning

VIPULA DISSANAYAKE[1], SACHITH SENEVIRATNE[2], RAJIB RANA[3], (Member, IEEE), ELLIOTT WEN[1], THARINDU KALUARACHCHI[1], AND SURANGA NANAYAKKARA[1]

[1]Augmented Human Laboratory, Auckland Bioengineering Institute, The University of Auckland, Auckland 1010, New Zealand
[2]Transport, Health and Urban Design Research Laboratory, Faculty of Architecture Building and Planning, Melbourne School of Design, The University of Melbourne, Melbourne, VIC 3010, Australia
[3]Department of Computer Science, School of Science, University of Southern Queensland, Toowoomba, QLD 4301, Australia

Corresponding author: Vipula Dissanayake (vipula@ahlab.org)

This work was supported by the Assistive Augmentation Research Grant through the Entrepreneurial Universities (EU) initiative of New Zealand.

ABSTRACT Extracting emotions from physiological signals has become popular over the past decade. Recent advancements in wearable smart devices have enabled capturing physiological signals continuously and unobtrusively. However, signal readings from different smart wearables are lossy due to user activities, making it difficult to develop robust models for emotion recognition. Also, the limited availability of data labels is an inherent challenge for developing machine learning techniques for emotion classification. This paper presents a novel self-supervised approach inspired by contrastive learning to address the above challenges. In particular, our proposed approach develops a method to learn representations of individual physiological signals, which can be used for downstream classification tasks. Our evaluation with four publicly available datasets shows that the proposed method surpasses the emotion recognition performance of state-of-the-art techniques for emotion classification. In addition, we show that our method is more robust to losses in the input signal.

INDEX TERMS Emotion recognition, representation learning, self-supervised learning, wearable signals.

I. INTRODUCTION

Emotion recognition is becoming an increasingly important field in human-computer interaction. The common emotions displays are speech [1], facial expressions [2], gestures [3], and physiological signals [4]. Among them, physiological signals are one of the most reliable means as they originate from the activity of the Autonomous Nervous System (ANS) and can hardly be triggered/suppressed by any conscious or intentional control [4].

Before the emergence of smart wearable devices, physiological signals could only be obtained using medical sensing devices such as Electroencephalography (EEG) and Electrocardiograph (ECG) sensors. Such sensors are intrusive, non-portable, and cumbersome to use, making it challenging to embed emotion recognition technologies in real-life applications. Recent advancements in smart wearable devices have offered a paradigm shift in wearable sensing. Consumer-grade smart wearable devices such as smartwatches, fitness trackers are portable, non-invasive, and equipped with various sensors. They enable continuous monitoring of physiological signals and make affect detection technologies possible for daily usage.

Despite the merits of smart wearable devices, they are not as highly accurate as medical-grade devices. They also tend to get lossy due to users' activities or environmental interference. These could negatively impact the reliability of affect detection algorithms [5].

Deep Learning models are robust to lossy signals in general; therefore, they can be used to develop robust affect detection algorithms [6]. Deep learning models also make representation learning feasible, which fully or partially eliminates the need for feature engineering. Feature engineering is the method of designing features using domain knowledge. It is a complex task that requires significant human time and effort, which can take even decades for an entire community of researchers [7]. A representation learning algorithm can discover a good set of features for a task in a fraction of the time required by manual feature engineering. However, it requires an enormous amount of labelled data for deep

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

learning models to work effectively. It is challenging and labour-intensive to identify and assign emotion labels to sensor reading segments.

Self-supervised and unsupervised learning minimises the need for labelled data, making representation learning using deep models more feasible. Researchers have studied self-supervised representation learning for physiological signals. However, most of them are targeted for high-frequency signals such as EEG and ECG [8], [9], and little research has been done on using representation learning techniques for low-frequency signals such as heart rate, electrodermal activity that are generated from widely available wearable devices.

This paper presents a novel self-supervised representation learning mechanism, SigRep, that works well with low-frequency physiological sensor data generated from commodity wearable devices. These representations can be easily adapted for downstream emotion recognition tasks with limited numbers of labeled data. With SigRep, we introduce a signal encoder consisting of 1D convolutions. We use a block-like neural network architecture inspired by Inception. In the pre-training stage, our network learns to contrast signal samples with random augmentation. We train signal representations for individual signal modalities using a large set of unlabeled signals. Then we fuse pre-trained signal representations for emotion recognition tasks. In the emotion recognition phase, we keep the weights of representations frozen.

We extensively evaluate our method for 1) classification performance on intact datasets, ii) behaviour of the method when data is lossy, iii) performance of the system when a less amount of labelled data is available for training, iv) significance of our encoder component, and v) effect of the individual signal modality on the emotion recognition. The results show that our technique outperforms eight other state-of-the-art techniques in seven tasks out of 12 tasks. Our contributions are as follows:

1) We adopt contrastive learning, a self-supervised training technique, to learn signal representations from low-frequency physiological sensor data, which can be effectively used for downstream emotion classifiers.
2) We propose an improvement to the conventional contrastive learning framework by proposing a new inception-inspired lightweight encoder, which offers better performance than a conventional encoder for downstream emotion classification tasks.
3) To demonstrate our proposed technique's performance, we conduct a series of experiments on four datasets. Experimental results show that our proposed approach offers significantly better performance than state-of-the-art methods. Results also show that the proposed approach requires a significantly smaller amount of labelled data and is robust to data loss than a fully supervised model.

II. RELATED WORK

A. FEATURE ENGINEERING AND REPRESENTATION LEARNING FOR EMOTION RECOGNITION USING WEARABLE SENSING

Before the advent of deep learning, research on emotion recognition was based on feature engineering, which is essentially hand-crafting features based on domain knowledge. It is, however, challenging to design features from the high dimensional data captured from a multitude of wearable devices [10]. Recent deep learning methods aim to address this issue by representation learning, which automatically extracts features from the raw signal. These deep learning techniques are promising as they achieve higher accuracy than conventional approaches using hand-crafted features. For example, Santamaria-Granados et al. [11] compares deep CNN with several classical machine learning methods for emotion recognition tasks, where CNN learns features from raw electrocardiography (ECG) and electrodermal activity (EDA) signals and the classical methods use hand-crafted features. The comparisons show that representation learning outperforms feature engineering. Recently, Yang *et al.* [12] proposes a hybrid neural network architecture to learn human emotion using EEG signals. Authors propose a parallelly concatenated architecture of a CNN and a long-short term memory (LSTM) network to learn from raw electroencephalography (EEG) signals and validate their method using publicly available datasets. Again, experimental results show improved accuracy of CNN-LSTM over models using hand-crafted features.

Although deep learning methods outperform classical machine learning methods [11]–[13], they require a large amount of labelled data to learn representative features [11]. It is challenging and labor-intensive to identify and assign emotion labels to sensor reading segments. Unsupervised feature learning techniques in deep learning address the requirement of a large amount of labelled data. These techniques can learn representations from unlabelled data; then, the representations can be reused for multiple downstream tasks built around smaller labelled datasets [14]. However, in a recent review, Schmidt *et al.* [15] highlight the limited usage of unsupervised and semi-supervised learning methods for wearable-based affect recognition research.

Out of the studies in this research area, autoencoder is a widely used technique for unsupervised representation learning. Recently published CorrNet [16] uses autoencoder based automatic feature extraction in a wearable signal-based emotion recognition task and outperforms the state-of-the-art baseline for CASE dataset for arousal (74.03%) and valence (76.37%) detection. Martinez *et al.* [13] deploy a denoising autoencoder network to learn features from blood volume pulse (BVP) and electrodermal activity (EDA) signals and reuse the learned features to classify affective states. Tang *et al.* [17] also uses a denoising autoencoder to learn features from EEG and peripheral physiological signals, and achieve accuracies of 93.97% and 83.53% for binary arousal valence classification of SEED and DEAP datasets, respectively.

Recently, a few research studies have used self-supervised learning on ECG [18] and EEG signals [9], [19]. Banville *et al.* [9] proposed a self-supervised strategy to automatically learn features from unlabelled EEG signals and demonstrate that EEG features learned in a self-supervised manner outperforms traditional supervised features while performing similarly to a fully supervised model on sleep stage detection tasks. Furthermore, they demonstrate that self-supervised models outperform supervised methods in low data situations by extensive margins. Cheng *et al.* [20] propose a contrastive learning method for EEG and ECG signals. Their study shows that self-supervised representations yield comparable performance against a fully supervised counterpart.

B. EMOTION RECOGNITION USING LOSSY SENSOR DATA FROM WEARABLE DEVICES

Signal steams from wearable devices are inherently lossy, resulting in gaps in signal streams [21]. Traditionally, researchers use statistical values such as mean and median values to replace the gaps in data [22]. However, filling gaps with such values is problematic in time-series data as those values do not reflect the qualities of signals [23]. To overcome this problem, lately, researchers are looking into generating values to fill the gaps. For example, Che *et al.* [24] use a deep learning model, 'GRU-D', to impute missing data in multivariate time series. However, the lack of annotated data makes the 'GRU-D' technique less usable in wearable emotion recognition. Recently, Generative Adversarial Networks (GANs) based approaches have become popular for data imputation [25], [26]. Again, generative models are computationally heavy, but in this paper, we focus on lightweight methods that can be potentially used in resource-constrained environments.

C. EMOTION RECOGNITION USING LIMITED LABELLED DATA

Addressing limited labelled data problems is a popular research area in machine learning. Transfer learning [27], [28] has been a popular approach in addressing the challenge of limited labelled data. The technique focuses on transferring knowledge from a model trained on a similar task to a new task. In wearable sensing, it is common to transfer learn with the models trained initially for activity recognition tasks for emotion recognition. However, the signal modalities used in activity recognition (accelerometer and gyroscope) do not fully cover physiological signals captured with wearable sensors.

Another way to address limited labelled data is by augmenting existing data to create new data points. The data augmentation method has been successfully used in computer vision. In wearable emotion recognition, a reflection of emotion has a personalised nature [29]. Given that existing wearable-based emotion recognition datasets consist of a limited number of subjects, data augmentation may not expand the inter-subject variability, leading to lower prediction performance. More studies are needed to understand how data augmentation can be used for emotion recognition using wearable devices.

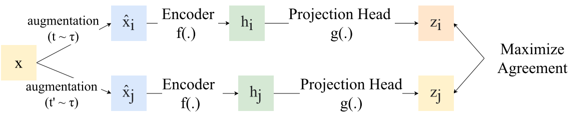

FIGURE 1. Contrastive learning framework. This framework borrows elements from SimCLR [30]. Initially, two separate transformation operations (t, t') selected from a set of transformations (τ) are applied to samples (x) in the training distribution. Then, transformed signals are used to train the encoder network $f(.)$ and projection head $g(.)$ to create latent vectors (z_i, z_j). Then, a contrastive loss is calculated between z_i and z_j to maximise the agreement. The calculated loss is propagated back through the network and weights are updated accordingly. After the training process, the projection head $g(.)$ is detached. The encoder network $f(.)$ and the latent representation h are used for downstream tasks.

In wearable sensing, datasets are usually large as most of the sensors run in the background as a daemon process producing enormous data points. However, due to the high cost of annotation, it is prohibitively expensive to label these large datasets. Self-supervised techniques can address these issues by learning meaningful representation from the data. However, studies using self-supervised techniques on wearable emotion recognition tasks are very limited. II-A.

To summarise: (1) A majority of the already limited unsupervised/self-supervised feature learning approaches are targeted for high-frequency signals such as EEG and ECG, and a little research work has used representation learning techniques for low-frequency signals from wearable devices. Our focus is to use physiological signals that can be captured with commodity smart wearable devices, and there is a clear gap in the literature regarding unsupervised/self-supervised methods that could be used for our purpose. (2) Advanced and computationally expensive techniques like GAN can be used to address missing data challenges. There is still a need for lightweight techniques to account for the data losses in the input signal in wearable sensing. (3) Limited labelled data is a universal challenge in machine learning and hence common in wearable sensing. Self-supervised learning and data augmentation are used to address the issue of limited labelled data. However, there is a gap in the literature regarding the suitability of these techniques in the wearable sensing platform.

III. MODEL ARCHITECTURE

Signal representations are core components of our research work. We propose a Self-Supervised Learning (SSL) paradigm to learn representations. In particular, we use contrastive learning [30], which learns an embedding space by minimising the distance between similar sample pairs while maximising the distance between dissimilar pairs. We use contrastive learning which has been show to be one of the most powerful self-supervised learning paradigms [31].

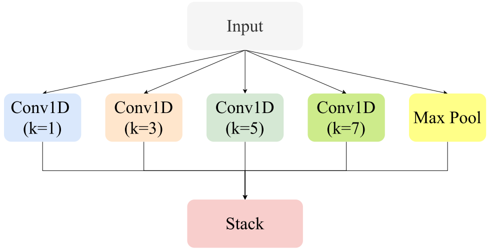

FIGURE 2. inception inspired block. We define an inception inspired block inspired by the inception block [34] with Conv1D layers with different kernel sizes [1, 3, 5, 7] and a max-pooling layer. Each convolution layer has two filters and uses rectified linear unit (ReLU) as the activation function. The max pool layer is configured to have a pool size of 3 and a stride of 1. All the parallel layers in this block use zero paddings to keep the output width similar to the input width. An input vector to a block goes through each layer parallelly. In the end, they are stacked together to construct the output.

We borrow elements from the SimCLR framework [30] for contrastive learning, which is originally proposed for visual representations. It simplifies the specialised architecture of contrastive learning yet outperforms previous self-supervised and semi-supervised learning methods on ImageNet. We bring SimCLR to the wearable sensor domain. We present our SimCLR Contrastive learning framework in Fig. 1 and describe its various components below.

A. DATA AUGMENTATION COMPONENT

In contrast to the conventional learning paradigms, SSL techniques do not require manual data labelling. They use data augmentation to generate labels. The data augmentation component transforms input data x into two views $(\widetilde{x}_i, \widetilde{x}_j)$ by applying transformations (t, t'). Informed by previous research, we randomly select a transformation τ from the following pool $\tau = \{$amplitude re-scaling, random DC shift, zero maskings, additive noise$\}$ [32], [33]. When $(\widetilde{x}_i, \widetilde{x}_j)$ are generated from the same input, we recognise them as a positive pair; otherwise, we consider them negative. We use following configurations for the signal transformations.

- **Amplitude re-scale:** This transformation selects a random scale factor *scale* from a uniform distribution $scale \in (0.1, 1.9)$ and multiplies it with the input signal.
- **Random DC shift:** For this transformation, we select a random shift value *shift* from a uniform distribution $shift \in (0.1, 0.9)$ and add it to the input signal.
- **Zero mask:** For this transformation, we select a random mask length w such that, $w \in (0.1 * l, 0.9 * l)$, where $l = signal\ length$; also a random starting point s_{id} such that $s_{id} < l/2$. Then, we mask the input signal with zeros starting from the s_{id} with a length of w. In case where $s_{id} + w > l$, zero mask is applied until the end of the signal starting from s_{id}.
- **Additive noise:** We generate a noise signal sampled from $(-1, 1)$ with the same length as the input signal for this transformation. Then we add the input signal with the noise signal to create the transformation.

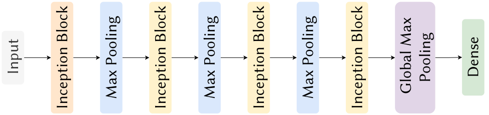

FIGURE 3. Proposed Encoder Network. We construct the encoder network with inception inspired blocks, max-pooling, global max pooling and dense layers. All inception blocks have the same hyperparameter values as mentioned in the description of the Fig. 2, while max pool layers in between inception like blocks are configured to have a pool size of two. The final dense layer consists of 40 units. As illustrated in the figure, a raw input signal to the encoder network goes through a sequence of blocks and layers. The dense layer at the end act as the feature embedding of the encoder network.

B. ENCODER

For the encoder we propose a Inception network block [34] as illustrated in *Fig. 3*. The Inception network uses convolutional layers with multiple kernel sizes on the same level in a CNN. By having multiple convolution kernel sizes, the network can learn patterns of different lengths from the input signal. We pass the input to multiple Conv1D layers with kernel sizes 1, 3, 5, 7 and a max-pooling layer before stacking as the output. The encoder is trained to learn a function $f(.)$, where $h = f(\widetilde{x})$; h denotes the latent representation of the transformed signal $\widetilde{x}$.

The Original inception network proposed by Szegedy et al. [34] uses many inception blocks resulting in approximately five million trainable parameters in the final network. However, our proposed signal encoder uses only four inception inspired blocks, resulting in an encoder network with less than 5,000 trainable parameters. Therefore our final encoder network can reduce resource consumption and avoid overfitting for smaller datasets.

C. PROJECTION HEAD

It is a neural network component in the SimCLR framework. It is designed to learn a function $g(.)$ on top of the representation h before calculating the contrastive loss. Chen *et al.* [30] experimented with the effect of having a non-linear, linear and no projection between the latent representation and contrastive loss calculation and reported that that having a non-linear projection on top of the representation outperforms the other two settings. Guided by this finding, we use a non-linear neural network for the projection head, consisting of two fully connected layers with 16 units each and use ReLU as an activation unit.

D. CONTRASTIVE LOSS

The contrastive loss function maximises the agreement between latent representations; positive pairs attract while negative pairs repel each other. In this work, we use the normalised temperature scaled cross-entropy loss *(NT-Xent)* as the loss function. Equation 1 defines the Contrastive loss, where $l(i, j)$ is defined in Equation 2, and $sim(i, j)$ is the cosine

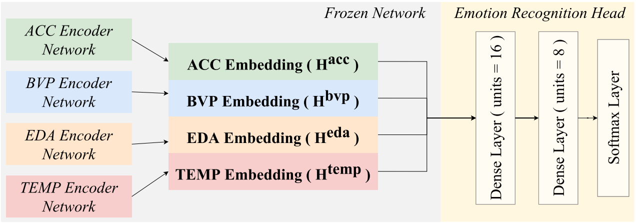

FIGURE 4. **Network for emotion recognition task. Neural network architecture is constructed for emotion recognition tasks by stacking the outputs of four pre-trained signal representations. A tiny fully-connected network is introduced on top of the stack of feature embeddings to learn the emotions. All the pre-trained representations are kept frozen during fine-tuning the network for emotion recognition.**

similarity of the i, j vectors.

$$Loss = \frac{1}{2N} * \sum_{k=1}^{N}[l(2k-1, 2k) + l(2k, 2k-1)] \quad (1)$$

$$l_{i,j} = -log \frac{exp(sim(z_i, z_j/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} exp(sim(z_i, z_k)/\tau)} \quad (2)$$

E. EMOTION CLASSIFICATION HEAD

After the training, the projection head $g(.)$ is detached. The encoder network $f(.)$ and the latent representation h are used for downstream tasks of emotion classification. The emotion classification head is a tiny, fully connected neural network component used in the downstream emotion classification tasks. As illustrated in Fig. 4, the classification head is built with two fully connected layers with 16 and 8 units each, followed by a softmax layer. Each fully connected layer uses 'ReLU' activation, and the number of units in the softmax layer is equal to the number of classes used in the classification task.

IV. EXPERIMENTAL SETUP

A. DATASETS

We use multiple publicly available datasets, which consists of physiological signals captured using wearable devices. We provide a brief description of the datasets below.

The **AffectiveROAD** [35] dataset consists of multi-model physiological and ambient sensor data captured during real-world driving. Data is collected from ten people across 14 driving sessions of 1.5 hours. Two wrist-worn devices, Empatica E4,[1] were used for data collection from both hands of the driver. A chest-worn device, BioHarness 3 ,[2] was also used to collect data, but we only consider data from the wrist-word devices for this work. Data streams have been annotated for stress from the perspective of an external party and later validated with the driver. We use the physiological signal streams from this dataset for representation training.

The continuously annotated signals of emotion (**CASE**) [36] dataset contains physiological signals (Electrocardiogram (ECG), Blood Volume Pulse (BVP), Electromyogram (EMG) and Electrodermal Activity (EDA)) captured from 30 participants while they were watching emotion stimulating videos. Data streams were annotated with arousal and valence values from the perspective of the participant. CASE dataset provides the arousal/valence rating in nine levels. However, in the literature [16], researchers have binned nine levels into two class and three class configurations for evaluation. For the comparison purpose, we follow the class configuration proposed by Zhang *et al.* [16] in our evaluations. We use the CASE for signal representation learning as well as the emotion recognition tasks.

The **CLAS** [37] dataset consists of physiological signals (Electrocardiogram (ECG), Photoplethysmography (PPG) and Electrodermal Activity (EDA)) with inertia signal (Accelerometer (ACC)) captured from 60 participants. Data were collected while participants engaged in various activities that elicit different cognitive load, affect and stress levels. A Shimmer 3 GSR+ and Shimmer 3 ECG units were used in the data collection process. Our study only uses PPG, EDA, and ACC signals from the dataset in representation learning and emotion recognition tasks.

The **K-EmoCon** [38] dataset contains multiple physiological signals (Electrocardiogram (ECG), Electroencephalogram (EEG), Blood Volume Pulse (BVP), Electrodermal Activity (EDA), Body Temperature (TEMP)) and inertia signals (Accelerometer (ACC)) from 32 participants during 16 debate sessions. Data of four participants were discarded due to sensor malfunctioning. The dataset contains annotations of arousal, valance, categorical emotions from multiple perspectives; first-person (self-report), second person (debate opponent) and third-person (external party). We use this dataset in booth representation learning and emotion recognition stages. In this dataset, arousal and valence values are reported at five different levels. However, due to heavy imbalance of class distribution, we binn arousal/valence levels into two and three binns and create binary and three-class classification problems for arousal and valence.

Further, authors have published intensity levels of the five categorical emotions (happy, sad, angry, cheerful and nervous). When we chunk the dataset for emotion prediction tasks, we treat the most intense emotion as the categorical emotion in that chunk. We turn those categorical emotions into a five-class classification problem. As previously mentioned, the dataset has been annotated from three different perspectives. For this research work, we select self-reported emotions for analysis.

The PPG dataset for motion compensation and heart rate estimation in daily life activities (**PPG-DaLiA**) [39] contains physiological signals (Blood Volume Pulse (BVP), Electrocardiogram (ECG), Electrodermal Activity (EDA), Body Temperature (TEMP)) and inertia signals (Accelerometer (ACC)) captured from 15 subjects while they engaged in a range of activities in daily life. The authors used a wrist-worn Empatica E4 device and a chest-worn RespiBAN device to capture signals. For our signal representation learning stage, we use signals captured from the wrist-worn device.

[1] *https://www.empatica.com/research/e4/*

[2] *https://wearabletech.io/zephyr-bioharness-3/*

TABLE 1. A summary of datasets used in the self-supervised training stage. We use data from six datasets contains signals captured from wearable devices. Each dataset has a different set of signal modalities with different capturing frequencies, as tabulated.

Dataset	Signal & Capturing Frequency	Unsupervised Examples
Affective Road	ACC (32Hz), BVP (64Hz), EDA (4Hz), TEMP (4Hz)	59,538
CASE	BVP (1kHz), EDA (1kHz), TEMP (1kHz)	24,480
CLAS	ACC (256Hz), BVP (256Hz), EDA (265Hz)	24,427
K-EmoCon	ACC (32Hz), BVP (64Hz), EDA (4Hz), TEMP (4Hz)	5,041
PPG Field Study	ACC (32Hz), BVP (64Hz), EDA (4Hz), TEMP (4Hz)	43,147
WESAD	ACC (32Hz), BVP (64Hz), EDA (4Hz), TEMP (4Hz)	28,940

Wearable stress and affect detection (**WESAD**) dataset [40] contains physiological signals (Blood Volume Pulse (BVP), Electrocardiogram (ECG), Electrodermal Activity (EDA), Body Temperature (TEMP)) and inertia signals (Accelerometer (ACC)) captured from 15 subjects during a controlled environment study. The dataset also contains signals captured using a RespoBAN device also ECG signals. The authors present the affective state in their dataset as a binary classification (stress vs non-stress) and three class classification (baseline vs stress vs amusement) problems. We use the partition of data recorded using the wrist-worn device in this paper's representation learning and emotion recognition stages. Our emotion recognition task tries to solve their three-class classification problem using the wearable signal partition. Further, the authors of the dataset present data with few other class labels not specified in the dataset description. We ignore those signals in the emotion evaluation; however, we use the physiological and inertia signals in our representation training stage as the labels are not required for SSL.

B. DATA PRE-PROCESSING

Datasets used in this work are captured with various devices with different sampling frequencies. To unify the signal frequency, we chunk the continuous signals into window size of four seconds with a one-second overlap. The window size is based on the findings from the literature [16]. Then, we reconstruct the signal within the signal chunk and resample it to the target signal frequency. To minimise the signal resampling, we chose the most common sampling frequency for each signal type as shown in *Table 1*. When we chunk signals, it is very important to have a proper convention to assign the correct class label to each chunk. We select the majority agreement protocol to select the class label. If there is more than one majority agreement on the class label, we discard that chunk from the emotion recognition tasks. *Table 1* summarises the signal chunks we use for representation learning while *Table 2* summarises the class distribution for each emotion recognition tasks.

C. MODEL TRAINING

We use two main stages for model training– (1) representation training and (2) emotion recognition. For each training stage, we use training parameters listed in the *Table 3. We implement our models using Tensorflow v2.3*[3] *in Python 3.8 environment. All the source code for data processing, model training and evaluation is open-sourced as a Github repository.*

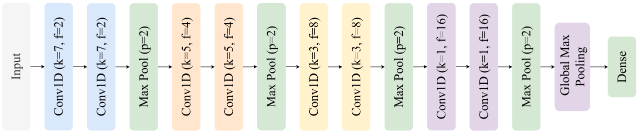

FIGURE 5. Basic Encoder Network. We construct the basic encoder network with 1D convolution, max-pooling, global max pooling and dense layers. We keep the structure of the network simple also the number of trainable parameters approximately similar to inception inspired encoder.

1) TRAINING SIGNAL REPRESENTATIONS

We train individual representations for each signal: ACC, BVP, EDA, TEMP, in a self-supervised manner. First, we preprocess datasets for representation learning (*see Table 1*). Second, we mix and shuffle datasets before using them as training data. We use the SimCLR framework illustrated in Fig. 1 for representation training, wherein we use the proposed inception inspired encoder architecture (*see Fig. 3*) as the encoder component of the SimCLR framework. We use 512 epochs and a batch size of 256 for encoder training. At the end of the training, the weights of each trained encoder network are saved for further usage in emotion recognition tasks.

We also train another set of signal representations with a basic encoder architecture illustrated in Fig. 5. As illustrated, the basic encoder is built with naively stacking Conv1D layers and MaxPool layers. In contrast, the inception inspired network is wider, with multiple Conv1D layers parallelly in each network level. In order to make the basic encoder and inspection inspire network comparable, we make the basic encoder deeper than the inception inspired encoder so that both networks have a similar amount of trainable parameters. We keep the training procedure identical to the procedure mentioned in the previous paragraph. We identify this encoder as 'Basic Encoder' in the rest of the paper. The purpose of this encoder is to experiment with the significance of the inception inspired encoder.

[3] *https://www.tensorflow.org/versions/r2.3/api_docs/python/tf*

TABLE 2. A summary of supervised tasks used in this research work. The table contains all twelve classification tasks in four datasets we address in our experiments. Each dataset contains a different number of subjects and multiple classification tasks, as mentioned. We use the Task-ID to identify each task throughout the paper.

Dataset	Subjects	Task	Task-ID	Class (Number of Examples)
CASE	30	arousal-2	*(a)*	low (18,865), high (5,615)
		arousal-3	*(b)*	low (5,526), neutral (16,735), high (2,219)
		valence-2	*(c)*	negative (11,944), positive (12,536)
		valence-3	*(d)*	negative (2,920), neutral (17,205), positive (4,355)
CLAS	60	arousal	*(e)*	low (12,177), high (12,299)
		valence	*(f)*	negative (12,226), positive (12,250)
K-EmoCon	28	arousal-2	*(g)*	low (3,432), high(1,609))
		arousal-3	*(h)*	low (1,609), neutral (1,866), high (1,566)
		valence-2	*(i)*	negative (4,106), positive (935)
		valence-3	*(j)*	negative (935), neutral (2,857), positive (1,249)
		categorical	*(k)*	angry (394), cheerful (2,747), happy (478), nervous (1,382), sad (40)
WESAD	15	categorical	*(l)*	stress (5,867), amusement (3,322), relax (1,857)

TABLE 3. A summary of training parameters. Neural network training parameters presented in the paper are listed in the Table. All the self-supervised training use parameter values listed under the representation training, while all the supervised models use the configurations listed under training for emotion recognition.

Training Stage	Hyper-parameter	Value
Representation Training	Loss Function	NT-Xent
	Optimiser	adam
	Learning Rate	0.001
	Batch Size	256
	Epochs	512
Training for Emotion Recognition	Loss Function	Categorical cross-entropy
	Optimiser	adam
	Learning Rate	0.001
	Batch Size	64
	Epochs	64

With this SimCLR based approach, we expect the encoder network to achieve a comprehensive understanding of the raw input signal. For the contrastive loss to get minimised, the encoder should be able to create similar latent representations for positive pairs regardless of the random augmentation added to the signal. To achieve that encoder should either learn to decode the augmentation or learn how to extract information about the underline signal. Since the augmentation added is random in each run, and each augmentation has randomness within the method of augmentation, it is unlikely the encoder network learns to decode the applied augmentation. Therefore the only way the contrastive loss get minimise would be the encoder learning qualities of the underline signal. For the same reason, the trained encoder should be able to retrieve information from a lossy signal, improving the robustness of downstream tasks.

2) TRAINING FOR EMOTION RECOGNITION

We use representations learned in the previous step for the downstream task of emotion recognition. We build a new neural network by stacking the outputs of individual signal representation networks. On top of the representation embeddings stack, we implement a smaller neural network for emotion recognition task, as shown in the *Fig. 4*. The emotion recognition network is created with two fully connected layers with ReLU activation and a Softmax layer. We keep the trained parameters of the representations frozen in this phase of training. We train the emotion recognition network in a fully supervised manner for tasks and datasets listed in *Table 2*. We evaluate our emotion recognition model with the leave one user out method and report the average accuracy and F1 scores.

3) BASELINE MODEL TRAINING

To compare the performance of our model, we benchmark it against a fully supervised model. In this paper, we refer to it as the 'baseline model'. The only difference between the proposed model and the baseline model is that the encoder component in the proposed model is trained in a self-supervised manner, whereas that in the supervised model is trained in a supervised manner. We use the same amount of training data like that used for the representation-based emotion recognition model. Also, we keep the similar training parameters as tabulated in Table 3.

D. SELF-SUPERVISED BENCHMARK

Although self-supervised learning is heavily used in computer vision and natural language processing tasks, only a few explorations have been conducted with time series sensor data from wearable devices. In addition, the majority of the existing self-supervised learning methods for wearable sensor signals are focused on downstream tasks such as activity recognition. Despite the lack of comparable works, we benchmark our work with Sense & Learn framework [41] given that it the most recent state-of-the-art self-supervised representation learning work with wearable sensor signals. In the Sense & Learn framework, authors have proposed a generic representation learning framework for heterogeneous sensor signals. Saeed *et al.* [41] evaluate eight self-supervised tasks to train signal representations and evaluated on multiple downstream tasks (activity recognition, sleep stage detection, stress detection and WiFi-sensing) and provided insights on choosing representation learning techniques for different downstream tasks.

We replicated Sense & Learn framework [41] with the parameters used for stress detection, as it is the closest task to emotion recognition. We train representations using all eight proxy tasks and use them in the downstream emotion recognition task. Initially, we use data chunks with 30 seconds following the stress detection task proposed in the Sense & Learn framework. Representations based on all eight proxy tasks result in poor performance in emotion recognition. Prior work [16] suggests sampling with smaller window sizes results in better emotion recognition accuracy in the context of the wearable signal-based emotion recognition. Therefore we attempt to evaluate with a smaller sample size. However, small sample windows are theoretically impossible with the encoder architecture suggested in Sense & Learn framework by Saeed *et al.* [41]. Therefore, we use the encoder architecture proposed in the current work to train representations with proxy tasks defined in Sense & Learn framework [41].

The eight proxy tasks we adapted can be summarised as follows.

1) T1: BLEND DETECTION

The blend detection task is defined as a three-class classification. The classification task's data samples and labels are generated by blending two signal samples with a random weight. The original sample without blending is labelled as class A. If two signal samples are selected from different modalities, it is marked as class B. If two signal samples are from the same modality, they are labelled as class C. The random weight for blending is selected from a uniform distribution in range (1,0). Finally, negative log-likelihood is used as the loss function to train the classification task on these three classes.

2) T2: FUSION MAGNITUDE PREDICTION

In this task, signals are blended in a similar strategy as the previous task (T1). In the learning phase, the objective of the network is to predict the random weight used for blending. For a clean sample, weight is considered zero.

3) T3: FEATURE PREDICTION FROM A MASKED WINDOW

In this task, a random segment is selected from an input sample. Eight statistical values (mean, standard deviation, maximum, minimum, median, kurtosis, skewness, number of peaks) are generated from the selected segment. Then mask the segment with zeros. Later, a model is trained to predict the statistics of the masked segment.

4) T4: TRANSFORMATION RECOGNITION

The transformation recognition task is based on previous work of Saeed *et al.* [32]. One transformation from eight pre-defined transformations (permutation, channel shuffle, time-warp, scale, noise, rotation, flip, negation) is applied to the input sample per instance. Each transformation is labelled with a class-index. Then the representation learning model is trained to classify the respective class of the transformation.

5) T5: TEMPORAL SHIFT PREDICTION

An input sample is circularly shifted with a random interval in the temporal domain. The random shifting interval is divided into seven classes based on the shifting period. Then the representation learning model is trained to predict the seven classes of shifts.

6) T6: MODALITY DENOISING

This task has a similarity with a denoising autoencoder. A clean input sample is blended with a random sample from a different signal modality to generate the noisy signal. The blending process uses a random weight selected from a uniform distribution. Then a model is trained to re-generate the clean sample given the blended sample.

7) T7: ODD SEGMENT RECOGNITION

In odd segment recognition task, an input sample is split into four similar length segments. One of the segments is replaced with a similar length signal segment chosen from a random sample from a different modality. Then the representation learning model is trained as a four-class classification problem to predict the replaced segment id.

8) T8: METRIC LEARNING WITH TRIPLET LOSS

For this task, a triplet (anchor, positive, negative) of samples is used as the input. The original sample is chosen as the anchor. While the positive is generated by applying a transformation to the anchor. The negative is selected from a different signal modality. Finally, the representation learning model is trained with triplet loss to minimise the distance between the anchor and the positive while increasing the distance between the anchor and the negative.

V. EVALUATIONS AND RESULTS

We evaluate our proposed emotion recognition model with four public datasets (CASE, CLAS, K-EmoCon, WESAD). As shown in *Table 2*, each dataset has different emotion and affective state labels. We evaluate them using the Leave One Subject Out (LOSO) method. We report average categorical prediction accuracy and average macro F1 scores.

TABLE 4. SigRep emotion recognition performance with state of the art results in literature. CorrNet [16] CLAS paper benchmark [37] WESAD paper benchmark [40]. We could not find any existing benchmark for emotion recognition tasks in K-EmoCon dataset.

Dataset	Task	Method	Accuracy	F1-Scores
CASE	arousal - 2	CorrNet	0.7403	**0.72**
CASE	arousal - 2	SigRep	**0.7630**	0.7127
CASE	arousal - 3	CorrNet	0.5822	0.55
CASE	arousal - 3	SigRep	**0.6507**	**0.6108**
CASE	valence - 2	CorrNet	0.7403	**0.76**
CASE	valence - 2	SigRep	**0.7408**	0.7064
CASE	valence - 3	CorrNet	0.6015	0.53
CASE	valence - 3	SigRep	**0.6483**	**0.6025**
CLAS	arousal	CLAS-video	0.7160	-
CLAS	arousal	CLAS-image	**0.7760**	-
CLAS	arousal	SigRep	0.7581	0.7526
CLAS	valence	CLAS-video	0.7170	-
CLAS	valence	CLAS-image	0.7420	-
CLAS	valence	SigRep	**0.7453**	0.7247
K-EmoCon	arousal - 2	SigRep	0.7754	0.7478
K-EmoCon	arousal - 3	SigRep	0.6493	0.6291
K-EmoCon	valence - 2	SigRep	0.8131	0.7680
K-EmoCon	valence - 3	SigRep	0.5906	0.5615
K-EmoCon	categorical	SigRep	0.5121	0.5028
WESAD	categorical	WESAD-wrist	0.7521	0.6412
WESAD	categorical	SigRep	**0.7813**	**0.7735**

A. EXPERIMENT 1: EVALUATION OF EMOTION RECOGNITION MODELS

1) EXPERIMENT

In this experiment, we evaluate the performance of SigRep emotion recognition models. We train the emotion recognition models for each classification task from each dataset. As tabulated in *Table 2*, we have 12 classification tasks from four different datasets. Because the class labels are heavily imbalanced in most tasks, the accuracy metric alone does not reflect model performance. Therefore we report the macro F1 score along with the prediction accuracy metric.

As discussed previously, current literature has very little work on using self-supervised techniques for wearable signal based emotion recognition task. The majority of existing supervised work is based on classic machine learning approaches. Therefore benchmarking only against work published in the literature may not reflect the advantages of using SigRep in emotion recognition tasks. On the other hand, benchmarking only against self-supervised learning methods may not properly position SigRep within existing literature. Therefore we benchmark performance of SigRep in two different scenarios. i) benchmark against current state-of-the-art for each emotion recognition task from the literature, ii) benchmark against other self-supervised learning methods.

2) RESULTS

For the CASE dataset, CorrNet [16] provides the state-of-the-art emotion recognition performance. The CASE dataset consists of arousal and valence levels in nine intensities. Due to the heavy class imbalance, CorrNet [16] uses only two and three-class configurations for evaluation. Following that, we also evaluate our method using only two and three-class configurations. Although the two-class results are on par with each other, results of the three-class problem clearly demonstrate the superior performance of the proposed method over CorrNet.

We used prediction results reported by Markova *et al.* [37] as the benchmark for the CLAS dataset. The CLAS dataset contains emotion data elicited in two ways, using (1) image and (2) video stimuli. The results are presented in *Table 5*.

For the WESAD dataset, most of the works in the literature are focused on using ECG and EMG signals. The best performance for three-class affective state classification using ACC, BVP, EDA and TEMP signals is achieved by Schmidt *et al.* [40]. As we focus on commodity sensors (such as sensors built into a smartwatch), we only use the wrist-based signals and compare our performance with results of wrist-based signals reported by Schmidt *et al.* [40]. Similarly, our method shows superior performance, as shown in the *Table 5*.

Except for CorrNet, which is based on a representation learning approach, other state-of-the-art results are based on classic machine learning approaches. In order to compare SigRep with other self-supervised learning-based methods, as mentioned before, we have re-implemented the self-supervised methods proposed in the Sense & Learn framework [41]. We benchmark the emotion recognition performance of SigRep against all eight proxy tasks proposed in the Sense & Learn framework [41]. *Table 5* presents classification accuracy, and *Table 6* presents the F1-Score for all 12 classification tasks.

As the results reflect, SigRep has demonstrated the top accuracy for 7 out of 12 emotion recognition tasks and second best accuracy for 3 out of the remaining 5 tasks. In the case of F1-Scores, SigRep has achieved the top two F1 scores in 10/12 tasks. Overall, SigRep has demonstrated better emotion recognition performance.

3) DISCUSSION

Diving deeper into the emotion recognition performance, we observe that out of the eight proxy tasks in the Sense & Learn framework [41], Tasks 3, 4 and 8 have achieved one of the top two accuracies and F1 scores frequently. To explain this observation, we analyse those proxy tasks and the proxy task proposed in SigRep.

The proxy task proposed in the SigRep contrasts samples after adding a random augmentation to the signal components. The proxy task 8 in the Sense & Learn framework [41] is to contrast samples from different modalities. Both proxy tasks have a common element of learning how to contrast distinct elements and identify similar elements at a higher level. SigRep uses random data augmentations before learning to contrast them. Those augmentations are re-scaling amplitude, random DC shift, additive noise and random zero masking.

TABLE 5. **Emotion Recognition Model Performance: Accuracy. The table benchmarks performance of emotion recognition models trained using the pre-trained in different approaches. We adopt all methods proposed in Sense & Learn framework [41] to benchmark SigRep. Best results for each task is presented in bold text while the second best result is presented in italic.**

Dataset	Emotion Task	SigRep	S&L: T1	S&L: T2	S&L: T3	S&L: T4	S&L: T5	S&L: T6	S&L: T7	S&L: T8
CASE	arousal - 2	*0.7630*	0.7312	0.7585	**0.7706**	0.7543	0.7583	0.7431	0.7583	0.7502
CASE	arousal - 3	0.6507	0.6493	0.6567	**0.6836**	0.6529	0.6334	0.6522	0.6386	*0.6726*
CASE	valence - 2	*0.7408*	0.6294	0.6990	**0.8090**	0.6842	0.7024	0.6310	0.6637	0.6190
CASE	valence - 3	0.6483	0.6782	0.6564	**0.7028**	0.6689	0.6788	0.6776	*0.6880*	0.6864
CLAS	arousal	**0.7581**	0.7434	0.7385	0.7117	*0.7483*	0.6800	0.6961	0.6654	0.6800
CLAS	valence	**0.7453**	0.6790	*0.7141*	0.6234	0.6946	0.6873	0.6678	0.6824	0.6873
K-EmoCon	arousal - 2	**0.7754**	0.5233	*0.6643*	0.6536	0.5973	0.6013	0.5557	0.5900	0.6562
K-EmoCon	arousal - 3	**0.6493**	0.5083	0.5092	*0.6144*	0.5165	0.5381	0.4982	0.4832	0.5547
K-EmoCon	valence - 2	*0.8131*	0.7724	0.8112	**0.8135**	0.7750	0.8001	0.7746	0.7591	**0.8135**
K-EmoCon	valence - 3	**0.5906**	0.4709	0.5458	*0.5564*	0.4183	0.4852	0.4570	0.4491	0.5528
K-EmoCon	categorical	**0.5121**	0.3421	0.4445	0.3909	*0.4656*	0.4164	0.3876	0.3353	0.3411
WESAD	categorical	**0.7813**	0.6961	0.6984	0.6984	*0.7460*	0.6866	0.7224	0.7305	0.7212

TABLE 6. **Emotion Recognition Model Performance: F1 Scores. The table benchmarks performance of emotion recognition models trained using the pre-trained in different approaches. We adopt all methods proposed in Sense & Learn framework [41] to benchmark SigRep. Best results for each task is presented in bold text while the second best result is presented in italic.**

Dataset	Emotion Task	SigRep	S&L: T1	S&L: T2	S&L: T3	S&L: T4	S&L: T5	S&L: T6	S&L: T7	S&L: T8
CASE	arousal - 2	**0.7127**	0.6790	0.6901	0.6861	0.6951	*0.6981*	0.6884	0.6940	0.6816
CASE	arousal - 3	**0.6108**	0.5802	0.5771	0.5822	0.5842	0.5678	0.5857	0.5751	*0.5903*
CASE	valence - 2	*0.7064*	0.6150	0.6822	**0.7257**	0.6664	0.6602	0.6460	0.6502	0.6632
CASE	valence - 3	**0.6025**	0.5948	0.5798	*0.5987*	0.5872	0.5933	0.5912	0.5977	0.5968
CLAS	arousal	*0.7526*	0.7116	0.7069	0.6974	0.7223	0.6974	0.7103	0.6978	**0.7664**
CLAS	valence	**0.7247**	0.6915	0.6974	0.6583	*0.6990*	0.6931	0.6880	0.6894	0.6922
K-EmoCon	arousal - 2	*0.7478*	0.6553	0.7452	0.7085	0.7208	0.7043	0.6940	0.7421	**0.7535**
K-EmoCon	arousal - 3	*0.6291*	0.5061	0.5355	**0.6839**	0.5231	0.5499	0.5009	0.5267	0.5230
K-EmoCon	valence - 2	**0.7680**	0.7217	0.7408	*0.7418*	0.7191	0.7388	0.7254	0.7162	*0.7418*
K-EmoCon	valence - 3	0.5615	0.5278	0.5653	*0.5720*	0.5045	0.5315	0.5251	0.5189	**0.5808**
K-EmoCon	categorical	0.5028	0.4435	**0.5632**	0.4678	*0.5194*	0.5089	0.5104	0.4910	0.4978
WESAD	categorical	**0.7735**	0.6920	0.6690	0.7386	0.7220	0.6659	0.7272	0.7274	*0.7704*

The first three augmentations are similar to transformations added in the task 4; zero masking is similar to task 5. At a higher level, the proxy task in SigRep contains the essence of proxy tasks 3,4 and 8 of the Sense & Learn framework [41]. Based on that, we suggest that the combined effect of the proxy task in SigRep has resulted in better emotion recognition performance. Further, the findings of this experiment support the argument that the pre-training proxy task has an effect on the downstream prediction task. Also, we recommend using a combined proxy task consisting of signal transformations, zero masking and a contrastive learning approach for wearable signal base emotion recognition.

B. EXPERIMENT 2: EVALUATION OF ROBUSTNESS

1) EXPERIMENT

In real-life usage, signals captured from consumer-grade wearable devices can be lossy due to various reasons such as user movements, software errors, and malfunctioning sensors. These signal losses have been identified as a technical limitation by researchers using wearable devices in the wild [5].

Algorithm 1 The Process of the Lossy Data Evaluation

```
for $p \leftarrow 0$ to $0.9$ by $0.1$ do
    $S$ = [data record with all signal modalities]
    shuffle $S$
    $dropCounter = 0$
    foreach $s$ in $S$ do
        $p' = random()$
        if $p' < p$ then
            replace $s$ with zeros
            $dropCounter += 1$
            if $dropCounter == len(S) - 1$ then
                break
            end
        end
    end
    evaluate(S)
end
```

To evaluate the robustness of our method to signal losses, we randomly drop data frames from every evaluation record. An evaluation record is a set of data frames from each

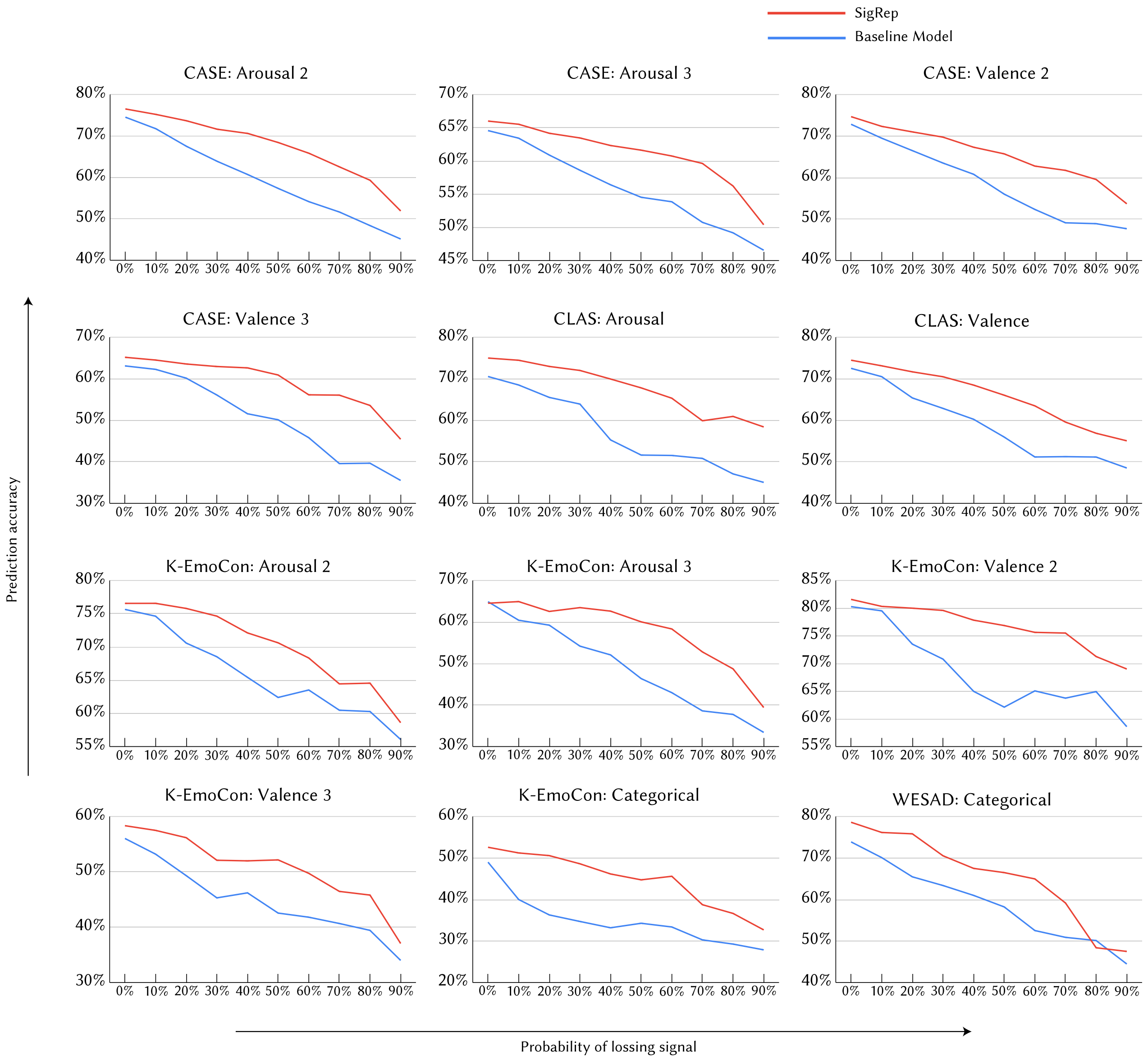

FIGURE 6. Results of the evaluation of lossy signals. Classification accuracy of each task in model training setting is visualised against the probability of losing a signal frame.

signal modality and the target emotion label. To identify the threshold of noise robustness, we define a variable p, which corresponds to the probability of dropping a data frame. We gradually increase the value of p from 0 to 0.9 with a step of 0.1 for each evaluation round. We simulate the signal loss by replacing the corresponding data frame with a vector of zeros. We demonstrate our strategy to drop data frames in *Algorithm 1*. We ensure that at least one data frame has non-zero values. To avoid bias, we randomise the selection of signal frame dropping for each evaluation record. To compare the performance of robustness, we benchmark our proposed method against the baseline model, which is a fully supervised model (please see description in section IV-C3). Further, we evaluate the emotion recognition models based on eight proxy tasks presented in the Sense & Learn framework [41]. In this evaluation, we consider a scenario where there is a 50% chance of losing a signal frame.

2) RESULTS

We report the observed accuracy for each classification task in *Fig. 6*). Interestingly, the SigRep model achieves higher accuracy than the baseline models for almost every p value. To quantify the robustness, we conduct a post-hoc test using the Tukey Honest Significant Difference test (HSD) on each scenario to determine which p value makes the significant loss of accuracy. We identify p values, where the drop of accuracy starts significantly in each task for both SigRep and baseline models. We then average those p values for each

TABLE 7. Emotion Recognition Model Performance with Lossy Signal (50%): Accuracy. The best accuracy for each classification task is highlighted in bold text while the second best accuracy is marked in italic format. S&L: T# refers to each proxy task proposed in the Sense & Learn framework [41].

Dataset	Emotion Task	SigRep	S&L: T1	S&L: T2	S&L: T3	S&L: T4	S&L: T5	S&L: T6	S&L: T7	S&L: T8
CASE	arousal - 2	**0.6844**	0.6415	0.6308	0.6206	0.6619	0.6386	*0.6747*	0.6420	0.6737
CASE	arousal - 3	**0.6165**	0.5111	0.6014	*0.6036*	0.5514	0.5429	0.5621	0.5768	0.5971
CASE	valence - 2	**0.6574**	0.4671	*0.5007*	0.3439	0.4929	0.4922	0.4875	0.4916	0.4556
CASE	valence - 3	**0.6095**	0.5446	0.5683	0.5928	0.5544	0.5848	0.5921	0.5905	*0.5936*
CLAS	arousal	**0.6784**	0.4561	*0.4581*	0.4557	0.4556	0.4565	0.4579	0.4558	0.4501
CLAS	valence	**0.6607**	0.4632	0.4642	0.4474	*0.4688*	0.4618	0.4645	0.4636	0.4682
K-EmoCon	arousal - 2	**0.7064**	0.4877	*0.6075*	0.6040	0.5442	0.5867	0.5018	0.5458	0.5894
K-EmoCon	arousal - 3	**0.6008**	0.3926	0.3858	*0.4044*	0.3741	0.3821	0.3709	0.3541	0.3946
K-EmoCon	valence - 2	**0.7689**	0.7160	0.7598	*0.7635*	0.7019	0.7481	0.7187	0.6817	*0.7633*
K-EmoCon	valence - 3	**0.5219**	0.4199	0.4933	*0.5062*	0.3726	0.4433	0.4149	0.3841	0.4796
K-EmoCon	categorical	**0.4477**	0.2759	0.3759	0.3411	*0.3946*	0.3706	0.3412	0.2645	0.2744
WESAD	categorical	**0.6652**	0.5307	0.5000	*0.5689*	0.5307	0.4811	0.5317	0.5486	0.5418

TABLE 8. Emotion Recognition Model Performance with Lossy Signal (50%): F1-Score. The best F1-Score for each classification task is highlighted in bold text while the second best F1-Score is marked in italic format. S&L: T# refers to each proxy task proposed in the Sense & Learn framework [41].

Dataset	Emotion Task	SigRep	S&L: T1	S&L: T2	S&L: T3	S&L: T4	S&L: T5	S&L: T6	S&L: T7	S&L: T8
CASE	arousal - 2	**0.6381**	0.6119	0.6084	0.6361	0.6294	0.6207	*0.6308*	0.6108	0.6257
CASE	arousal - 3	**0.5792**	0.4981	0.5254	0.5322	0.5114	0.4992	0.5344	0.5271	*0.5423*
CASE	valence - 2	**0.5264**	0.5130	0.5056	0.2404	0.5121	0.5134	0.5149	*0.5180*	0.5040
CASE	valence - 3	**0.5647**	0.5339	0.5405	0.5487	0.5379	0.5474	0.5488	0.5487	*0.5496*
CLAS	arousal	**0.4662**	0.4376	0.4163	0.4061	0.4322	0.4224	0.4386	*0.4403*	0.3835
CLAS	valence	**0.4874**	0.4398	0.4381	0.3361	*0.4499*	0.4467	0.4488	0.4485	0.4425
K-EmoCon	arousal - 2	**0.5863**	0.4880	*0.5440*	0.5384	0.5259	0.5357	0.4959	0.5255	0.5266
K-EmoCon	arousal - 3	**0.2905**	*0.2865*	0.2475	0.2470	0.2697	0.2662	0.2614	0.2566	0.2653
K-EmoCon	valence - 2	**0.7343**	0.6768	0.6900	*0.6918*	0.6612	0.6865	0.6797	0.6577	0.6917
K-EmoCon	valence - 3	**0.4275**	0.3758	0.3921	0.3903	0.3737	0.3721	0.3728	0.3576	*0.3912*
K-EmoCon	categorical	**0.2651**	0.1752	0.2456	0.1432	0.2378	*0.2592*	0.2212	0.1884	0.1478
WESAD	categorical	**0.4970**	0.4888	0.4545	*0.4894*	0.4588	0.4620	0.4795	0.4894	0.4817

setting and identify that when the average p value is greater than 0.27, the accuracy drop in the baseline setting gets significant. In contrast, models in SigRep settings demonstrate a significant drop in accuracy when the average p is greater than 0.55. This result indicates that our proposed method is significantly more robust compared to a model with similar architecture trained in an end-to-end manner.

Table 7 and Table 8 show the accuracy and F1-Score of SigRep and Sense & Learn framework [41] at a 50% signal loss probability. Overall results suggest that SigRep has shown better accuracy and F1-Scores for all 12 emotion classification tasks. Further, for nine out of 12 tasks, Sense & Learn proxy tasks 3,4 and 8 have achieved the second-best results based on prediction accuracy. Which is consistent with the results of previous experiment.

3) DISCUSSION

Prior work indicates that representation learning can achieve a better understanding of the underline data [14]. Also, as we discussed in Section IV-C, contrastive learning inherently offers robustness to noise and losses. Due to these aspects of our model, we conjecture that we achieve higher robustness than the baseline model.

C. EXPERIMENT 3: IMPACT OF THE AMOUNT OF LABELLED DATA

1) EXPERIMENT

The cost of data annotation is one major issue in physiological signal-based emotion recognition. Our proposed method addresses this challenge by adapting to the downstream task with less labelled data leveraging on the learned representations. We experiment by reducing the amount of labelled data used in the downstream task to quantify the performance. Since we use leave-one-subject-out evaluation, we control the training data as a fraction of available subjects for training in this experiment. Especially for each evaluation round, we leave out the evaluation subject and then drop 50% of subjects from the leftover set for training. We keep the training parameters similar to experiment 1. Similar to the previous experiment, we compare the performance of our model with that of the baseline model.

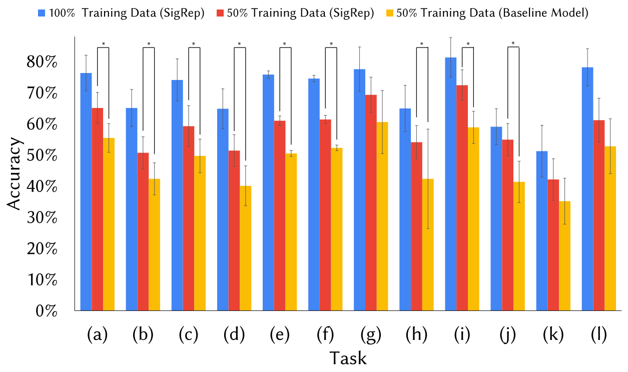

FIGURE 7. Performance comparison of the SigRep with different amounts of training data. The Blue column represents the performance of the SigRep with all available training data. The Red column represents the performance of SigRep with 50% of training data. The Yellow column represents the performance of the baseline model (*see SectionIV-C3*) with 50% of training data. Error bars show the 95% confidence intervals.

2) RESULTS

Classification accuracy for each task for each scenario is plotted in *Fig. 7*. As anticipated, with limited training data, classification accuracy drops significantly for all classification tasks. On average, for the baseline, for a 50% drop of training data, accuracy drop around 20% (calculated by comparing with 100% training data used for baseline); however, with the proposed method with learned representation, the average accuracy drop is around 10%. t-test shows that the difference between the baseline and the SigRep method is significant in nine out of twelve classification tasks ($p < 0.05$). For the remaining three tasks (g), (k) and (l), although the SigRep method demonstrates a higher accuracy, we do not find a statistical significance.

D. EXPERIMENT 4: SIGNIFICANCE OF INCEPTION INSPIRED ENCODER

1) EXPERIMENT

Our proposed encoder architecture is built with Conv1D layers inspired by the inception architecture. To test the effect of the proposed architecture, we compare it with a simple stacked convolutions architecture built with Conv1D layers with a similar number of trainable parameters (see Fig. 5). We denote it as the ‘‘basic encoder’’. We train the basic encoder with the proposed SSL method with the same datasets and training configurations as the proposed inception inspired encoder. Then we train emotion classifiers for all 12 tasks using the learned representations with the basic encoder and evaluate emotion classification performance. We keep the evaluation conditions identical to our Experiment 1 (see section V-A).

2) RESULTS

The results of this experiment are plotted in Fig. 8. For all classification tasks, the average accuracy of the proposed inception inspired encoder is higher than the basic encoder. We conduct t-tests for each classification results for an in-depth analysis. We observe that for all twelve classification tasks, the inception inspired encoder performs better than the basic encoder, where for seven tasks, the inception inspired encoder significantly ($p < 0.05$) outperforms the basic encoder.

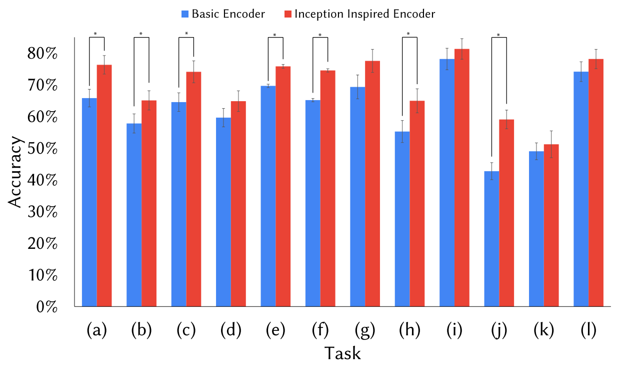

FIGURE 8. Encoder Architecture Comparison. Prediction accuracy by Inception inspired encoder architecture (see Fig. 3) compared to a basic Conv1D architecture (see Fig. 5) for all 12 classification tasks. 95% confidence intervals are marked on each column. Overall, inception inspired encoder show higher average accuracy. In seven tasks the accuracy gain is statistically significant.

E. EXPERIMENT 5: PERFORMANCE AND ROBUSTNESS COMPARISONS OF INDIVIDUAL MODALITIES

1) EXPERIMENT

Our proposed model makes use of four types of signals (ACC, BVP, EDA, TEMP). Each type of signal carries independent and correlated pieces of information. In this experiment, we investigate the performance of individual modalities, also their robustness to data losses. Some sensors are more reliable than others. This experiment can potentially assist researchers in selecting sensors for their applications. In this experiment, we train emotion classification models for all twelve classification tasks using only a single signal modality in each run. All the evaluation rounds use the leave-one-subject-out evaluation method and used similar training parameters as Experiment 1 *(see Section V-A)*. We evaluate models in two settings, (1) without data losses and (2) with 50% of data loss. Evaluation process with data loss is similar to our Experiment 2 *(see Section V-B)*.

2) RESULTS

Fig. 9 shows the average accuracy of each signal modality as well as the combined modalities. Fig. 9(a) shows the results without data loss setting, while Fig. 9(b) shows the lossy signal scenario. As one would expect, combined modalities should offer better performance than individual modalities, which is what we observe in Fig. 9. While comparing individual modalities, the BVP signal and ACC signal show higher accuracy than the EDA and TEMP signals. This observation can be explained based on the findings reported in the literature: when someone experiences an emotion, bodily

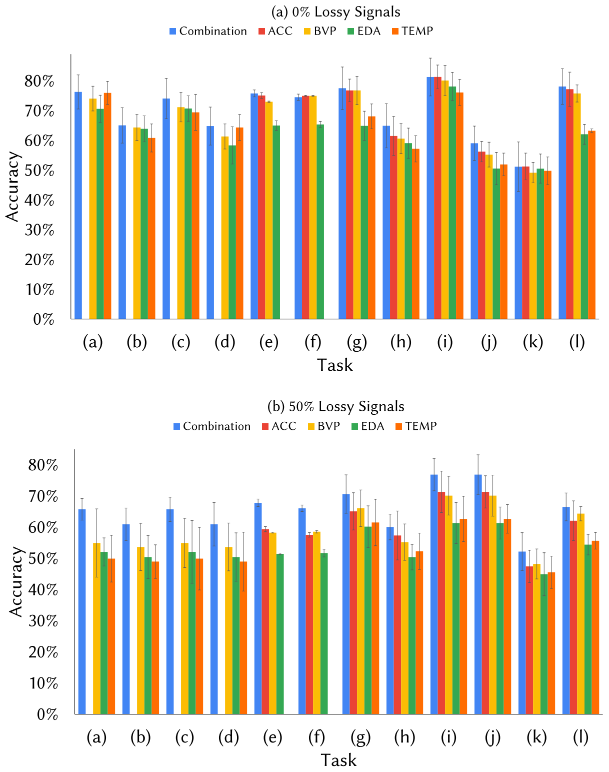

FIGURE 9. **Ablation results: 0% vs. 50% Lossy Signals. Plots show the effect of individual signal modality for the combined results with and without losses to the signals. Error bars shows 95% confidence interval. (a) effect without data losses. (b) effect with 50% of data loss.**

reaction reflects faster with heartbeat compared to body temperature variations and skin conductance changes [42]. Also, literature [43], [44] suggest a higher correlation between heart pulse and wrist accelerometer readings, providing better accuracy for the accelerometer.

Interestingly, when there is no data loss, the prediction accuracy of the combined model is not significantly ($p > 0.05$) higher than the prediction accuracy of any individual signal. However, when the signals are lossy, seven out of 12 tasks, combined models demonstrate significantly higher prediction accuracy than individual modalities. This result attests that combined modalities can offer higher robustness compared to individual modalities.

VI. CONCLUSION AND FUTURE WORK

This paper presents a novel contrastive representation learning approach for emotion recognition using wearable signals. We achieve the following key results:

- We excel the state-of-the-art methods for emotional classification performance over three widely used datasets (CASE, CLAS and WESAD) and create benchmark performance for the K-EmonCon dataset.
- We benchmark SigRep with state of the art self-supervised methods for signal representation learning and show that SigRep outperforms.
- We demonstrate that our self-supervised model using augmented data achieves significantly higher robustness to data losses than a fully supervised baseline. We also observe that while combined modalities do not achieve significantly higher accuracy than individual modalities without data loss; but with data loss combined modalities provides significantly better performance than that of individual modalities.
- We demonstrate that we can reduce the requirement of labelled data for downstream emotion classification tasks by learning representation.

In future work, we aim (1) to explore the effect of different fusion techniques on downstream task performance, and (2) to investigate the feasibility of using different self-supervised learning methods for on-device learning. Understanding the effect of fusion would help build better wearable signal representation based systems optimal for downstream tasks. On-device learning could improve representation based models on the go and personalise models after the deployment.

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2017.

[3] K. Schindler, L. Van Gool, and B. de Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural Netw.*, vol. 21, no. 9, pp. 1238–1246, Nov. 2008.

[4] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: A review," in *Proc. 7th Int. Colloq. Signal Process. Appl.*, 2011, pp. 410–415.

[5] S. W. T. Chan, S. Sapkota, R. Mathews, H. Zhang, and S. Nanayakkara, "Prompto: Investigating receptivity to prompts based on cognitive load from memory training conversational agent," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–23, Dec. 2020.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[7] L. Oakden-Rayner, G. Carneiro, T. Bessen, J. C. Nascimento, A. P. Bradley, and L. J. Palmer, "Precision radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, Art. no. 1648.

[8] G. Singh, K. Verma, N. Sharma, A. Kumar, and A. Mantri, "Emotion recognition using deep convolutional neural network on temporal representations of physiological signals," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. Netw. Technol. (ICMLANT)*, Dec. 2020, pp. 1–6.

[9] H. Banville, I. Albuquerque, A. Hyvarinen, G. Moffat, D.-A. Engemann, and A. Gramfort, "Self-supervised representation learning from electroencephalography signals," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.

[10] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Inf. Fusion*, vol. 49, pp. 46–56, 2019.

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

[12] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2018, pp. 1–7.

[13] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, ''Learning deep physiological models of affect,'' *IEEE Comput. Intell. Mag.*, vol. 8, no. 2, pp. 20–33, May 2013.

[14] Y. Bengio, A. Courville, and P. Vincent, ''Representation learning: A review and new perspectives,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[15] P. Schmidt, A. Reiss, R. Dárichen, and K. V. Laerhoven, ''Wearable-based affect recognition—A review,'' *Sensors*, vol. 19, no. 19, p. 4079, Sep. 2019.

[16] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, ''CorrNet: Fine-grained emotion recognition for video watching using wearable physiological sensors,'' *Sensors*, vol. 21, no. 1, p. 52, Dec. 2020.

[17] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, ''Multimodal emotion recognition using deep neural networks,'' in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2017, pp. 811–819, doi: 10.1007/978-3-319-70093-9_86.

[18] P. Sarkar and A. Etemad, ''Self-supervised ECG representation learning for emotion recognition,'' *IEEE Trans. Affect. Comput.*, early access, Aug. 6, 2020, doi: 10.1109/TAFFC.2020.3014842.

[19] M. N. Mohsenvand, M. R. Izadi, and P. Maes, ''Contrastive representation learning for electroencephalogram classification,'' in *Proc. Mach. Learn. Health*, 2020, pp. 238–253.

[20] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, ''Subject-aware contrastive learning for biosignals,'' 2020, *arXiv:2007.04871*.

[21] P. Siirtola, E. Peltonen, H. Koskimäki, H. Mönttinen, J. Röning, and S. Pirttikangas, ''Wrist-worn wearable sensors to understand insides of the human body: Data quality and quantity,'' in *The 5th ACM Workshop Wearable Syst. Appl.*, New York, NY, USA, 2019, pp. 17–21, doi: 10.1145/3325424.3329663.

[22] R. Houari, A. Bounceur, A. K. Tari, and M. T. Kecha, ''Handling missing data problems with sampling methods,'' in *Proc. Int. Conf. Adv. Netw. Distrib. Syst. Appl.*, Jun. 2014, pp. 99–104.

[23] Z. Liu, Y. Yang, W. Huang, Z. Tang, N. Li, and F. Wu, ''How do your neighbors disclose your information: Social-aware time series imputation,'' in *Proc. World Wide Web Conf.*, New York, NY, USA, 2019, pp. 1164–1174, doi: 10.1145/3308558.3313714.

[24] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, ''Recurrent neural networks for multivariate time series with missing values,'' *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 6085.

[25] D. Snow, ''MTSS-GAN: Multivariate time series simulation generative adversarial networks,'' in *Proc. SSRN Electron. J.*, 2010, pp. 1603–1614.

[26] J. Yoon, J. Jordon, and M. Schaar, ''GAIN: Missing data imputation using generative adversarial nets,'' in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.

[27] A. Akbari and R. Jafari, ''Transferring activity recognition models for new wearable sensors with deep generative domain adaptation,'' in *Proc. 18th Int. Conf. Inf. Process. Sensor Netw.*, New York, NY, USA, 2019, pp. 85–96, doi: 10.1145/3302506.3310391.

[28] W.-S. Chien, H.-C. Yang, and C.-C. Lee, ''Cross corpus physiological-based emotion recognition using a learnable visual semantic graph convolutional network,'' in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, 2020, pp. 2999–3006, doi: 10.1145/3394171.3413552.

[29] P. Barros, E. Barakova, and S. Wermter, ''Adapting the interplay between personalized and generalized affect recognition based on an unsupervised neural framework,'' *IEEE Trans. Affect. Comput.*, early access, Jun. 15, 2020, doi: 10.1109/TAFFC.2020.3002657.

[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, ''A simple framework for contrastive learning of visual representations,'' 2020, *arXiv:2002.05709*.

[31] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, ''Self-supervised learning: Generative or contrastive,'' *IEEE Trans. Knowl. Data Eng.*, early access, Jun. 22, 2021, doi: 10.1109/TKDE.2021.3090866.

[32] A. Saeed, T. Ozcelebi, and J. Lukkien, ''Multi-task self-supervised learning for human activity detection,'' *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 2, pp. 1–30, Jun. 2019.

[33] T. T. Um, F. M. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, ''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.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[35] N. E. Haouij, J.-M. Poggi, S. Sevestre-Ghalila, R. Ghozi, and M. Jaïdane, ''Affectiveroad system and database to assess driver's attention,'' in *Proc. 33rd Annu. ACM Symp. Appl. Comput.*, 2018, pp. 800–803.

[36] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, ''A dataset of continuous affect annotations and physiological signals for emotion analysis,'' *Sci. Data*, vol. 6, no. 1, pp. 1–13, Dec. 2019.

[37] V. Markova, T. Ganchev, and K. Kalinkov, ''CLAS: A database for cognitive load, affect and stress recognition,'' in *Proc. Int. Conf. Biomed. Innov. Appl. (BIA)*, Nov. 2019, pp. 1–4.

[38] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, ''K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations,'' *Sci. Data*, vol. 7, no. 1, p. 293, Dec. 2020.

[39] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, ''Deep PPG: Large-scale heart rate estimation with convolutional neural networks,'' *Sensors*, vol. 19, no. 14, p. 3079, Jul. 2019.

[40] 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 ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 400–408.

[41] A. Saeed, V. Ungureanu, and B. Gfeller, ''Sense and learn: Self-supervision for omnipresent sensors,'' *Mach. Learn. Appl.*, vol. 6, Oct. 2021, Art. no. 100152.

[42] T. Zhang, ''Multi-modal fusion methods for robust emotion recognition using body-worn physiological sensors in mobile environments,'' in *Proc. Int. Conf. Multimodal Interact.*, New York, NY, USA, 2019, pp. 463–467, doi: 10.1145/3340555.3356089.

[43] J. Hernandez, D. McDuff, and R. W. Picard, ''Biowatch: Estimation of heart and breathing rates from wrist motions,'' in *Proc. 9th Int. Conf. Pervasive Comput. Technol. Healthcare*, 2015, pp. 169–176.

[44] V. Dissanayake, D. S. Elvitigala, H. Zhang, C. Weerasinghe, and S. Nanayakkara, ''CompRate: Power efficient heart rate and heart rate variability monitoring on smart wearables,'' in *Proc. 25th ACM Symp. Virtual Reality Softw. Technol.*, Nov. 2019, pp. 1–8.

VIPULA DISSANAYAKE received the B.Sc. degree (engineering) in computer science and engineering from the University of Moratuwa, Sri Lanka, and the Master of Engineering degree from The University of Auckland, in 2019, where he is currently pursuing the Ph.D. degree with the Augmented Human Laboratory, Auckland Bioengineering Institute. His research interests include ubiquitous computing, machine learning, and human–computer interactions.

SACHITH SENEVIRATNE received the B.Sc. degree in computer science and engineering from the University of Moratuwa, Sri Lanka, and the Ph.D. degree in machine learning from Monash University, Australia. Currently, he is working as a Research Fellow at The University of Melbourne. His current research interests include deep learning, with a focus on contrastive representation learning and applications. He is broadly interested in self-supervised deep learning approaches across various disciplines, such as computer vision, NLP, and reinforcement learning.

RAJIB RANA (Member, IEEE) received the B.Sc. degree in computer science and engineering from Khulna University and the Ph.D. degree in computer science and engineering from the University of New South Wales, Sydney, Australia, in 2011. He received the Postdoctoral Training with the Autonomous System Laboratory, CSIRO, before joining the University of Southern Queensland, as a Faculty Member, in 2015. He is currently a Senior Advance Queensland Research Fellow and an Associate Professor with the University of Southern Queensland, where he is also the Director of the IoT Health Research Program. His current research interests include unsupervised representation learning, adversarial machine learning, re-enforcement learning, federated learning, emotional speech generation, and domain adaptation. He received the Prime Minister and the President's Gold Medal of Outstanding Achievements for his B.Sc. degree.

ELLIOTT WEN received the Ph.D. degree from The University of Auckland. He is currently working as a Research Fellow at The University of Auckland. His research interests include mobile sensing, software engineering, and computer networking.

THARINDU KALUARACHCHI received the bachelor's degree in electronic and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 2016. He is currently pursuing the Ph.D. degree with the Auckland Bioengineering Institute, The University of Auckland. His research interests include human-centered machine learning, emotion recognition, unsupervised learning, and human–computer interaction.

SURANGA NANAYAKKARA received the B.Eng. and Ph.D. degrees from the National University of Singapore, in 2005 and 2010, respectively. Later, he was a Postdoctoral Researcher with the Pattie Maes's Fluid Interfaces Group, MIT Media Lab. In 2011, he founded the "Augmented Human Laboratory" to explore ways of creating novel human–computer interfaces as natural extensions of our body, mind, and behaviour. He is currently working as an Associate Professor and leading the Augmented Human Laboratory, Auckland Bioengineering Institute, The University of Auckland. For the totality and breadth of achievements, he has won many awards, including young inventor under 35 (TR35 award) in the Asia–Pacific Region by MIT TechReview, the Outstanding Young Persons of Sri Lanka (TOYP), and the INK Fellowship, in 2016