Author/s:
Yang, K;Tag, B;Gu, Y;Wang, C;Dingler, T;Wadley, G;Goncalves, J

Title:
Mobile Emotion Recognition via Multiple Physiological Signals using Convolution-augmented Transformer

Date:
2022-06-27

Citation:
Yang, K., Tag, B., Gu, Y., Wang, C., Dingler, T., Wadley, G. & Goncalves, J. (2022). Mobile Emotion Recognition via Multiple Physiological Signals using Convolution-augmented Transformer. ICMR 2022 - Proceedings of the 2022 International Conference on Multimedia Retrieval, pp.562-570. Association for Computing Machinery. https://doi.org/10.1145/3512527.3531385.

Persistent Link:
https://hdl.handle.net/11343/313640

Mobile Emotion Recognition via Multiple Physiological Signals using Convolution-augmented Transformer

Kangning Yang
The University of Melbourne
Australia
kangning.yang@student.unimelb.edu.au

Benjamin Tag
The University of Melbourne
Australia
benjamin.tag@unimelb.edu.au

Yue Gu
Rutgers University
USA
yg202@scarletmail.rutgers.edu

Chaofan Wang
The University of Melbourne
Australia
chaofanw@student.unimelb.edu.au

Tilman Dingler
The University of Melbourne
Australia
tilman.dingler@unimelb.edu.au

Greg Wadley
The University of Melbourne
Australia
greg.wadley@unimelb.edu.au

Jorge Goncalves
The University of Melbourne
Australia
jorge.goncalves@unimelb.edu.au

ABSTRACT

Recognising and monitoring emotional states play a crucial role in mental health and well-being management. Importantly, with the widespread adoption of smart mobile and wearable devices, it has become easier to collect long-term and granular emotion-related physiological data passively, continuously, and remotely. This creates new opportunities to help individuals manage their emotions and well-being in a less intrusive manner using off-the-shelf low-cost devices. Pervasive emotion recognition based on physiological signals is, however, still challenging due to the difficulty to efficiently extract high-order correlations between physiological signals and users' emotional states. In this paper, we propose a novel end-to-end emotion recognition system based on a convolution-augmented transformer architecture. Specifically, it can recognise users' emotions on the dimensions of arousal and valence by learning both the global and local fine-grained associations and dependencies within and across multimodal physiological data (including blood volume pulse, electrodermal activity, heart rate, and skin temperature). We extensively evaluated the performance of our model using the K-EmoCon dataset, which is acquired in naturalistic conversations using off-the-shelf devices and contains spontaneous emotion data. Our results demonstrate that our approach outperforms the baselines and achieves state-of-the-art or competitive performance. We also demonstrate the effectiveness and generalizability of our system on another affective dataset which used affect inducement and commercial physiological sensors.

ICMR '22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531385

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

emotion recognition, physiological signals, off-the-shelf mobile devices, convolution-augmented transformer

ACM Reference Format:
Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2022. Mobile Emotion Recognition via Multiple Physiological Signals using Convolution-augmented Transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512527.3531385

1 INTRODUCTION

The emotions we feel every day shape our behaviours and guide our decisions. Emotions can, however, go awry and lead to negative repercussions to an individual's well-being. For instance, long periods under stress and anxiety may not only impair mental health but also induce related diseases. Recent studies have demonstrated a high correlation between affective instability and psychosis [18]. Thus, automatic and accurate emotion recognition has increasingly become an important research topic as it can assist with early diagnosis, continuous monitoring, and can inform interventions for mental health and well-being.

In recent decades, researchers have explored different ways to empower machines with human-like perception of emotional states. Current automatic emotion recognition approaches can be categorized into two main types according to the signals used. One approach entails using human behavioral signals, such as facial expressions [4, 29, 43], voice [19], and gestures [28]. While these signals are typically easier to collect, the reliability of such methods cannot be guaranteed since these signals are semi-voluntary or voluntary responses [8, 26]. This means that people can easily disguise inner emotions by controlling their behaviors. For example, people

can conceal their real emotions by showing a “poker face” in social communications [44]. The other approach entails using physiological signals, such as electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG), galvanic skin response (GSR), and body temperature. Unlike behavioral signals, physiological signals originate from the activity of the central nervous system (CNS) and the autonomic nervous systems (ANS), which are involuntary responses and quite challenging to be controlled or hidden intentionally [31]. Therefore, approaches based on physiological signals are more reliable and universally applicable to discern the inner emotional feelings of human beings [14, 46].

However, emotion recognition based on physiological signals is still challenging. On the one hand, existing methods mainly rely on experience-based feature engineering to manually quantify numerical attributes that can characterize affective information contained in different physiological signals. Typically, the extracted features concentrate on the time and frequency domains, for example, the commonly used time-domain statistical features — minimum, maximum, mean, standard deviation [7], and frequency-domain features — energy distribution at different frequency bands [40]. Despite being widely adopted, these handcrafted features are essentially low-level feature representations, which do not generalize well in different scenarios and can not efficiently model the complex and non-linear spatio-temporal relationships among multiple physiological signals. Recent research has found that multiple physiological signals can more effectively reflect emotional changes than a single signal [31]. On the other hand, most affective databases (e.g., DEAP [14], MAHNOB-HCI [33]) used in physiological signals-based emotion studies are acquired by using artificial induction methods, i.e., by using specific music, pictures, or video clips as stimuli to induce the generation of certain emotional states from participants. Although such emotional portrayals could be considered as “spontaneous" emotional expressions, they do not represent natural emotional responses collected from in-the-wild social interactions and lack adequate contextual information. Moreover, the data acquisition in these databases typically relies on medical-level equipment because of its affordances, such as high sensitivity and high sampling rate. For example, the acquisition equipment used in the DEAP database supports a 512 Hz sampling rate for physiological responses, while the acquisition equipment used in the MAHNOB-HCI database supports a 1024 Hz sampling rate. Although high-quality physiological signals can be collected in this way, these facilities tend to be expensive and intrusive. For example, an EEG-based system requires a large number of electrodes to be attached to an individual’s scalp [10]. These hinder emotion recognition techniques from being integrated into everyday devices and being deployed into real-world contexts.

To address these issues, we design a deep multimodal architecture with a convolution-augmented transformer (conformer [11]) mechanism to classify emotions on different categories of arousal and valence. Specifically, it first aligns physiological signals by using the nearest-neighbor interpolation method to synchronize different sensor inputs. Then, it extracts high-level informative features from different signals through individual convolutional neural networks (CNNs). Lastly, by leveraging a conformer encoder, it can learn the latent local and global associations among different physiological signals. Moreover, with the development of mobile technology, modern mobile and wearable devices are increasingly low-cost, sensor-rich, and lightweight. This makes them well-suited to detect multimodal physiological signals and emotional responses in an unobtrusive manner, overcoming the aforementioned inherent limitations from medical-grade acquisition equipment in traditional physiological emotion recognition research. Therefore, in this study, we evaluated our system using the K-EmoCon dataset [21], a publicly available multimodal sensor dataset acquired with off-the-shelf wearable devices in naturalistic conversations. We used four kinds of physiological signals collected from an Empatica E4 Wristband as inputs, blood volume pulse (BVP), electrodermal activity (EDA), heart rate (HR), and skin temperature (SKT), and conducted five classification tasks, two-category and five-category classification on arousal and valence respectively, and a four-category classification on combined arousal-valence emotion space. We also tested the generalizability of our system by conducting two additional experiments on the ASCERTAIN dataset [35].

Thus, the contribution of our work is two-fold:

(1) We propose a novel conformer-based deep learning structure that detects human emotional states based on multimodal physiological signals from off-the-shelf devices. To the best of our knowledge, we are the first to apply a conformer mechanism in the field of affective computing.
(2) We conduct extensive experiments to evaluate and test our proposed system. The results demonstrate that our model outperforms previous techniques, achieving state-of-the-art or comparable performance.

2 RELATED WORK

Despite a large number of studies on emotion recognition, most of them have focused on audio-visual signals [22] (e.g., facial expressions and speech), since they are seen as the most direct channels for human emotional expressions. However, these outward emotional signals can easily be modified or suppressed in social settings, reducing their reliability as signs of inner emotional feelings [31, 46]. In addition, recent work has argued that emotional facial expressions vary across races and cultures, and facial configurations can not act as a reliable signal for particular emotional states [1]. Furthermore, as faces and voices are directly linked to identity, these can raise significant privacy concerns, especially if they have to be uploaded to the cloud for processing [30].

For these reasons, the link between emotions and a range of physiological signals has recently received increased attention in the field of affective computing. For example, previous work has leveraged BVP, EDA, and SKT to recognise emotions [44]. They adopted several different types of film clips as stimuli to evoke participants’ emotions, and utilized an Empatica E4 wristband to collect the induced physiological responses. In another example, researchers proposed a method to recognise emotions relying on frontal EEG signals [41]. Specifically, they first selected four kinds of VR affective scenes containing happiness, fear, peace, and disgust. Then by using textile dry electrodes, they synchronously recorded three-channel frontal EEG signals when participants watched each VR stimulus. Unlike facial or acoustic data containing biometric information, physiological data as biosensed information may be

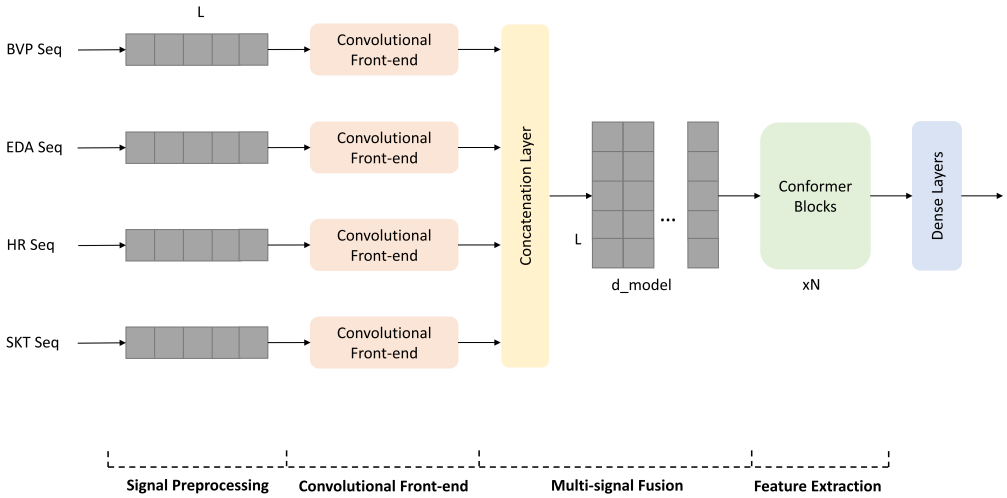

Figure 1: Overall structure of our proposed emotion recognition system.

used to describe our health status, behaviors, or emotions, but are not enough to identify us uniquely [30].

Researchers also proposed different approaches to analyze physiological data in order to detect an individual's emotions. For example, Zhao et al. [44] first used adaptive band-pass filter and low-pass filter to eliminate the noise contained in the collected signals. Then, they extracted the distinct emotion-related features from the time domain, frequency domain, and nonlinear analysis. By adopting the *sequential forward floating selection method* to select the best feature set, they classified different emotions based on support vector machine (SVM). Similarly, Xu et al. [41] extracted three kinds of handcrafted features including the time domain, frequency domain, and space domain from collected frontal EEG signals. Then by using the *model stacking method*, they combined three models, Gradient Boosting Decision Tree, Random Forest, and SVM, to achieve EEG-based emotion recognition.

Researchers have, however, recently noted the limitations of handcrafted feature representations, and explored ways of designing deep learning-based architectures to extract fine-grained and higher-level features by taking full advantage of neural networks' powerful feature abstraction ability. For instance, Wang et al. [39] presented a deep learning method to learn high-level feature representations from the raw EEG signals. They built a three-dimensional CNN and classified emotions on valence and arousal scales. Zitouni et al. [47] proposed a bidirectional long short-term memory (LSTM) neural network to extract informative features from four kinds of physiological signals, and predicted emotions into binary levels and quadrants of the arousal-valence space.

More recently, transformers were used to model long-range global context. For example, Wang et al. [38] designed a heartbeat-aware attention mechanism, and added it into transformer structure to enhance the alignment between encoded and decoded sequences. On this basis, they made arrhythmia classification from ECG signals. Similarly, Behinaein et al. [2] placed a convolutional front-end before the transformer encoder to extract more informative representations, and achieved stress detection from ECG signals. Despite models with these structures being able to capture either local spatial dependencies, or temporal dependencies, or global information, they lack the capability to learn both local and global interactions simultaneously.

Our architecture is inspired by the conformer mechanism used in recent automatic speech recognition research [11] that combines the advantages of convolutions and self-attention mechanisms. In this paper we extend it to multimodal environments for emotion recognition, and design the conformer encoder to learn both position-wise local features and content-based global associations within and across different modalities.

3 METHOD

The overall structure of our proposed emotion recognition system is illustrated in Figure 1. There are four major parts of the system: signal preprocessing, convolutional front-end, multi-signal fusion, and feature extraction.

3.1 Signal Preprocessing

Our system accepts raw data retrieved from different sensors as sequence inputs. Through the data preprocessing module, the heterogeneous inputs will be formatted into specific representations, which can be effectively used in the following modules.

Specifically, in this study, we focused on four different kinds of commonly used physiological signals: (1) BVP (blood volume pulse), is related to the changes in blood volume in arteries and capillaries, and can be measured by a non-invasive optical sensor that detects changes in light absorption density of the localized tissue (e.g., skin) when illuminated [15]; (2) EDA (electrodermal activity), measures the variations in electric characteristics of the skin resulting from changes in sweat production and fluid concentration

in the sweat ducts, and can be monitored by the voltage changes between electrodes [15]; (3) HR (heart rate), is a measure of the functional activity of the heart, and can be estimated using the ECG or BVP signals; (4) SKT (skin temperature), can be measured by a thermopile infrared sensor.

Following the preprocessing operations stated in [47], we first normalized the raw signals of each subject separately to overcome individual differences which may vary due to age, gender, and personality. The normalization is performed based on the signals collected in the final 1.5 minutes of the relaxation period prior to each trial. Then, we leveraged the nearest-neighbor method and interpolated the lower frequency signals based on the highest sampling frequency to synchronize different physiological signals. Subsequently, each bio-signal can be represented as an array with the same length L, where $L = t \times f$, t is the time interval of emotional annotation, and f is the highest sampling frequency of four sensor inputs.

3.2 Convolutional Front-end

To extract more expressive features from physiological inputs, we applied the CNNs structure for each signal, which is able to capture fine-grained feature patterns by local-perceiving convolutional kernels and weight-sharing translation equivariance [3].

Inspired by [2, 38], the convolutional front-end module consists of two parts (shown in Figure 2). Each part starts with a 1D convolutional layer (1x3 padded convolution), then followed by a batch normalization and a rectified linear unit (ReLU).

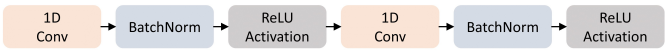

Figure 2: Convolutional front-end structure.

We set the out-channels produced by the convolution to 8 and 16 respectively. After feeding the bio-signal outputted from the preprocessing phase, it will be converted into a feature matrix with $L \times 16$ dimensions.

3.3 Multi-signal Fusion

We considered the four physiological inputs (BVP, EDA, HR, and SKT) as a whole. It is crucial to leverage several physiological inputs as emotions are subjective feelings generated by a complex coordination of multiple neurophysiological systems [23], in other words, the emotional clues should be simultaneously reflected in multiple physiological signs.

Thus, in this module, we combined the four physiological signals together and treated them as an entire physiological embedding to represent the hidden affective information. Specifically,

$$P = [P_{bvp}, P_{eda}, P_{hr}, P_{skt}] \tag{1}$$

where P_{bvp}, P_{eda}, P_{hr}, and P_{skt} are feature representations of BVP, EDA, HR, and SKT retrieved from the convolutional front-end. We concatenated them over the time steps, and got the output with $L \times d_{model}$ dimensions (d_{model} is 64=16*4 in this case). Similar to natural language processing, here L can be regarded as the number of words in one sentence, and d_{model} can be viewed as the size of the embedded word vector by *word2vec*.

3.4 Feature Extraction

We applied the conformer structure as feature extraction encoder to extract local and global associations within and across physiological signals.

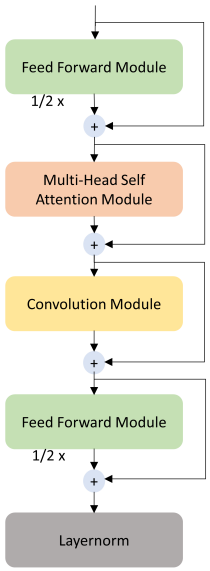

Figure 3: Conformer block structure [11].

The conformer mechanism presented by Gulati et al. [11] is a variant of the transformers. As shown in Figure 3, a conformer block consists of four modules, two feed forward modules, one multi-head self-attention module, and one convolution module. Firstly, the multi-head self-attention module uses a scaled dot-product attention mechanism to capture the dynamic global dependencies in the feature sequence [37]. It creates three vectors from each of the input feature vectors (the 64-dimensional embedding of each time step in this case): a query vector, a key vector, and a value vector. By computing the dot products of one query with all keys and applying a softmax function, it obtains the attention weights on all values, which determines how much focus to place on other parts of the input vectors as encoding one feature vector at a certain position. As defined by [38], given an input sequence $S = (S_1, S_2, ..., S_L)$, the output sequence $\tilde{S}_h = (\tilde{S}_{h1}, \tilde{S}_{h2}, ..., \tilde{S}_{hL})$ of the self-attention for a single head h can be computed by:

$$e_{ij} = \frac{(S_i W_q)(S_j W_k)^T}{\sqrt{d_k^h}}, j \in [1, L] \tag{2}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{L} \exp(e_{ik})} \tag{3}$$

$$\tilde{S}_{hi} = \sum_{j=1}^{L} \alpha_{ij}(S_j W_v) \tag{4}$$

where $W_q, W_k \in \mathbb{R}^{d_{model} \times d_k^h}$ and $W_v \in \mathbb{R}^{d_{model} \times d_v^h}$ are learnable linear transformation matrices that generate the query, key, and value vectors, and $S_i \in \mathbb{R}^{d_{model}}$. The output of all heads are then concatenated together and condensed by another linear transformation:

$$\tilde{S} = Concat(\tilde{S}_1, ..., \tilde{S}_n)W^o \tag{5}$$

where $W^o \in \mathbb{R}^{nd_v^h \times d_{model}}$ and n is the number of heads.

Secondly, the convolution module is responsible for enforcing locality, which contains a pointwise convolution with an expansion factor of 2, a gated linear unit (GLU) activation layer, a 1D depthwise convolution, a batch normalization, a swish activation layer, and another pointwise convolution projecting the number of channels back. By leveraging the channel-wise and spatial-wise learning capabilities, the convolution module is able to capture fine-grained local features, which largely complements the weakness of the self-attention mechanism. In this study, we concatenated features from four different physiological signals to form an entity. With the help of the convolution module, the system can better capture associations across signal-specific features.

For our feature encoder, we used 4 conformer blocks, and applied $n = 4$ parallel attention heads. For each head, we set $d_k^h = d_v^h = d_{model}/n = 16$. Results from feature extraction were flattened and passed into two dense layers to get the final representation for further decision making.

4 EVALUATION

4.1 Dataset

We first evaluated the performance of our proposed system on the K-EmoCon dataset [21]. We chose this particular dataset based on two criteria: (1) the dataset contains *spontaneous* emotional expressions in *naturalistic* environments or social interactions; (2) the data collection apparatus are *off-the-shelf, low-cost, mobile and wearable* devices. To the best of our knowledge, the K-EmoCon is the only publicly available multimodal affective dataset that is suitable for this experiment. Other widely used emotion datasets either consist of posed or induced emotions, or rely on expensive and intrusive data acquisition equipment that is not suitable for daily use. For example, the MAHNOB-HCI [33] and the DEAP [14] were considered to contain induced emotions and apply a number of intrusive electrodes [35]. Instead, the K-EmoCon dataset was collected via natural communication between individuals without professional training in acting. Specifically, the K-EmoCon contains two kinds of visual data (i.e., face and gesture), five kinds of physiological data (i.e., EEG, ECG, BVP, EDA, and SKT), speech audio, and accelerometer data recorded from 32 participants (20 males and 12 females). In our experiments, we only used four kinds of physiological signals captured by the Empatica E4 wristband, BVP, EDA, HR (which was derived from BVP), and SKT, since compared with chest-worn (used to detect ECG) and head-mounted (used to detect EEG) devices, a wristband is more ubiquitous and unobtrusive in everyday life and is without spatio-temporal limitations.

Moreover, the annotations for emotions in the K-EmoCon were measured on arousal and valence affective dimensions from the circumplex model of affect by James Russell [27]. Each affective dimension was rated on five-point Likert scales. Accordingly, annotated emotion labels can be categorized into either five levels directly, or two levels based on the median value (i.e., 2.5) in which annotations with values ranging from 1 to 2 can be converted into *low* (L), and annotations with values ranging from 3 to 5 can be converted into *high* (H) [47]. To conduct a comprehensive assessment, we formalized the emotion recognition as five classification tasks: two-category arousal, two-category valence, five-category arousal, five-category valence, and four-category arousal-valence space (i.e., high arousal-high valence (HAHV), high arousal-low valence (HALV), low arousal-high valence (LAHV), and low arousal-low valence (LALV)).

Table 1: Sample distribution

	Two-class	Five-class	Four-class
Arousal	L: 1023 H: 2157	1: 104 2: 919 3: 1159 4: 679 5: 319	-
Valence	L: 554 H: 2626	1: 69 2: 485 3: 1815 4: 728 5: 83	-
Arousal-Valence	-	-	LALV: 172 LAHV: 851 HALV: 382 HAHV: 1775

Specifically, we used the self-reported emotion labels as the ground truth. Since annotations in the K-EmoCon were labeled at every 5 seconds, the raw input size of BVP sequence, EDA sequence, HR sequence, and SKT sequence was respectively 1×320, 1×20, 1×5, and 1×20 (BVP sampled at 64Hz, EDA at 4Hz, HR at 1Hz, and SKT at 4Hz), and $L = 320$ in this case. Additionally, the E4 data of 6 participants (Person IDs: 2, 3, 6, 7, 17, and 20) is missing due to the device malfunction or human error during data collection. Thus, the final evaluation dataset consists of 3180 data samples. Table 1 shows the sample distribution. We note that while there are imbalanced distributions of emotion labels in all five tasks, as stated in [21], *this imbalance is expected as emotion data is commonly imbalanced by its nature in the wild (i.e., people are more often neutral than angry or sad).*

4.2 Baselines

To compare with the state-of-the-art physiological signals-based emotion recognition approaches, we chose the following methods as baselines: (1) Naive Bayes (NB) [47], (2) Support Vector Machine with linear kernel (SVM-LR) [35], (3) Support Vector Machine with radial basis function kernel (SVM-RBF) [35], (4) eXtreme Gradient Boosting (XGBoost) [47], (5) bidirectional LSTM (BiLSTM) [47]; and (6) transformer [2]. For the first four methods, we extracted

both the time-domain and frequency-domain features based on the proposed features in the literature [14].

4.3 Implementation Details

We implemented our proposed system using the PyTorch framework. We initialized the learning rate as 0.0001 and decayed the learning rate by $\gamma = 0.985$ every epoch. We used Adam optimizer, and applied batch normalization and dropout function to overcome overfitting and internal covariate shift [12].

To avoid data leakage, we did not use data re-sampling approaches [24] to solve the imbalanced data problem as they will change the original dataset itself. Instead, we chose focal loss [16] as the loss function of our model, which can automatically down-weight the contribution of easily classified samples and focus on hard misclassified samples by applying a modulating term to the cross-entropy loss. We used a 80–20 training and testing split, and further split 20 percent of the data from the training set as validation. We trained the model with 500 epochs and used early stopping with patience equal to 10. We also used 256 as the batch size and evaluated the model with 5-fold cross-validation. To make a fair comparison between the proposed system and baselines, we re-trained all models on the same training-testing set split.

4.4 Evaluation Metric

Considering that the classes of arousal and valence are heavily imbalanced, we did not use recognition accuracy as the evaluation metric like most studies. Instead, following [17, 24], we chose the average F1 score (Macro-F1) and unweighted average recall (UAR) as our validation metrics. These metrics give the same importance to each class, and are defined as the mean of class-wise F1 scores and recall scores respectively. Unlike accuracy or weighted F1 or weighted recall, their values will not be overwhelmed by the vast number of easily classified samples, which makes them suitable as the validation metrics of our experiment. Consider the prediction of two-class valence as an example. Its ground truth consists of 554 *low* and 2626 *high* samples. If there is a classifier that predicts *high* for all samples, it will have an accuracy of more than 82%, a weighted F1 score around 0.75, and a weighted recall score around 0.83. Despite having high values, we cannot say it is not a good classifier since it does not learn any informative features from the samples. Macro-F1 score and UAR can avoid this situation (only 0.45 and 0.50 in this example) because their values will be low if the model only performs well on the large number of common classes while performing poorly on the rare classes.

4.5 Results

The results of the emotion recognition from physiological signals on the K-EmoCon dataset are shown in Table 2. Our proposed system achieved better performance compared with all baseline approaches. For example, for five-class classification tasks, our model outperformed NB, SVM-LR, SVM-RBF, XGBoost, BiLSTM, and Transformer methods respectively by (1) 32.50%, 34.76%, 28.35%, 13.59%, 12.27%, 7.76% Macro-F1 increase, and 35.25%, 41.38%, 36.32%, 25.94%, 12.59%, 5.35% UAR increase on arousal; and (2) 26.17%, 30.57%, 27.42%, 7.76%, 6.52%, 3.78% Macro-F1 increase, and 43.91%, 44.94%, 43.46%, 29.73%, 10.22%, 2.93% UAR increase on valence. Additionally, the recognition results were tested for statistical significance using a Wilcoxon signed-rank test, by comparing the Macro-F1 and UAR scores from different approaches over cross-validation folds. The test results show statistically significant improvements of our model over the baselines on all five tasks ($p < 0.05$) except two-class valence comparison on Macro-F1 with BiLSTM and five-class valence comparison on UAR with Transformer.

From the results, we can also observe that (1) BiLSTM performed slightly better than Transformer on 2-class tasks, while performing slightly worse than Transformer on multi-class tasks; (2) the major difference between our system and Transformer architecture is the conformer encoder, which further shows the importance of convolution modules that can effectively extract the local interactions

Table 2: Comparison of emotion recognition performance using different approaches. A: arousal, V: valence, A-V: arousal-valence emotion space.

Approach	Task	Macro-F1(%)	UAR(%)
NB	Two-class A	36.16	53.78
	Two-class V	47.76	50.91
	Five-class A	23.43	30.28
	Five-class V	20.01	21.56
	Four-class A-V	23.54	31.96
SVM-LR	Two-class A	40.42	50.00
	Two-class V	46.48	50.56
	Five-class A	21.17	24.15
	Five-class V	15.61	20.53
	Four-class A-V	19.38	25.51
SVM-RBF	Two-class A	54.18	56.60
	Two-class V	45.94	50.30
	Five-class A	27.58	29.21
	Five-class V	18.76	22.01
	Four-class A-V	26.18	29.19
XGBoost	Two-class A	59.58	59.80
	Two-class V	54.88	54.79
	Five-class A	42.34	39.59
	Five-class V	38.42	35.74
	Four-class A-V	36.65	35.95
BiLSTM	Two-class A	74.50	76.53
	Two-class V	67.31	77.14
	Five-class A	43.66	52.94
	Five-class V	39.66	55.25
	Four-class A-V	47.33	56.34
Transformer	Two-class A	72.39	75.37
	Two-class V	60.36	73.11
	Five-class A	48.17	60.18
	Five-class V	42.40	62.54
	Four-class A-V	51.84	62.78
Our model	Two-class A	**77.37**	**79.42**
	Two-class V	**69.11**	**80.60**
	Five-class A	**55.93**	**65.53**
	Five-class V	**46.18**	**65.47**
	Four-class A-V	**56.35**	**65.45**

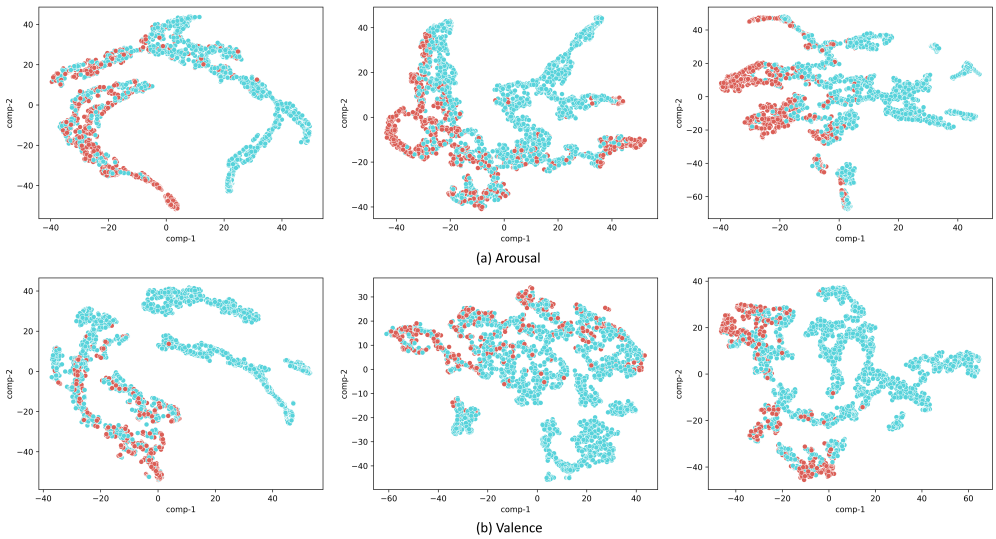

Figure 4: t-SNE visualisations of trained models for two-class tasks. Left is for the baseline BiLSTM approach; Middle is for baseline Transformer approach; Right is for our proposed model.

across different signal-specific features through the convolution mechanism. This also demonstrates the combination of convolutions with self-attention can improve the system's performance.

We also presented a visualization of the extracted feature representation from the BiLSTM, the Transformer, and our model. For this, we used the t-distributed stochastic neighbor embedding (t-SNE) technique [36] which allows visualizing high-dimensional data in a two or three-dimensional map. Figure 4 shows the 2D t-SNE plots for two-class tasks.

As illustrated in the figure, the extracted representations of our proposed method are visibly more distinguishable by classes on both arousal and valence dimensions compared to the extracted representations from baseline methods. Moreover, this trend can also be found for other tasks, which further indicates the stronger ability of our model on learning the high-level and latent features and dependencies.

4.6 Generalizability Tests

To further test the generalizability of our proposed model, we performed additional experiments on another multimodal affective dataset, the ASCERTAIN dataset [35]. Despite emotional inducement being applied during the data collection (instead of relying on spontaneous affective responses like the K-EmoCon dataset), this dataset used wearable and commercial sensors for examining users' physiological behavior, partially fulfilling the requirements of a more naturalistic and pervasive affective dataset. We designed two binary emotion recognition experiments to test our model's generalizability.

As both the K-EmoCon and the ASCERTAIN datasets contain EDA sensor data, we first conducted a cross-datasets generalizability test. Specifically, we trained on the K-EmoCon dataset and then tested on the ASCERTAIN dataset. Considering that there is only one input modality, we removed the BVP, HR, and SKT branches of our original structure and converted our model into a single modality emotion recognition system. Moreover, due to the different sampling frequencies (4Hz in the K-EmoCon versus 100 Hz in the ASCERTAIN), we used a downsampling method with an anti-aliasing filter to replace the interpolation in the preprocessing phase when testing the performance. On this basis, our model achieved 64.37% accuracy on arousal dimension and 61.88% accuracy on valence dimension, which is comparable to the unimodal performance of previous work that conducted both training and testing on the same ASCERTAIN dataset [45]. This result demonstrates that our proposed model is able to learn generalized and transferable physiological features rather than overfitting the data distribution of one particular dataset.

Table 3: Performance comparison on the ASCERTAIN dataset (accuracy in percentage)

Approach	Arousal	Valence
VM2HL-P [45]	**72.54**	68.53
Our model	71.79	**69.11**

We then followed-up with a second experiment to test the generalizability of our proposed model on other kinds of physiological signals. Considering the diversity of commercial physiological sensors and their recorded signals, our system should have good expansibility for different input modalities. Therefore, we further replaced the original input modalities of our model with the ECG, GSR, and EEG modalities from the ASCERTAIN dataset, and retrained our model from scratch. Following the work by Zhao et al. [45], we re-trained our model under the same implementation details and applied the same evaluation metric. As can be seen in

Table 3, our model achieved 71.79% recognition accuracy on arousal and 69.11% accuracy on valence, which is equivalent to a 0.75% accuracy decrease and a 0.58% improvement respectively compared to the system proposed by Zhao et al. [45]. Despite leading to a 0.58% accuracy decrease on arousal dimension, it is important to note that we only used three modalities (we excluded facial landmarks because our work focused on physiological signals), while the system from the literature used all four modalities [45]. In other words, our system can achieve competitive performance with less data input, which is significant in real-world scenarios due to the common data missing problem [32].

5 DISCUSSION

With the rapid development of smart mobile and wearable devices, there has been a growing interest in ecological validity [13] and real-world application of emotion recognition techniques [25]. In this paper, we presented a novel emotion recognition system based on physiological signals from low-cost, off-the-shelf mobile and wearable devices, and evaluated its performance with spontaneous and natural emotion data collected in natural interactions.

It is worth noting that our proposed system has good generalizability in terms of different physiological signals. The system can efficiently extract high-level and generalized affective features from physiological signals, and perform well on spontaneous emotion recognition. Considering that smart mobile and wearable devices are increasingly sensor-rich, less intrusive, and becoming an essential and integral part of daily life, we can expect to see an increase in the need for robust mobile emotion recognition systems.

5.1 Towards Robust Multimodal Mobile Emotion Recognition

It is undisputed that human emotions are expressed through a multitude of signals. Thus, as Delplanque and Sander [5] argue, it is risky to rely on a single component to infer true emotional states. For example, while fear might be correlated to heart rate changes, and smiling can be an expression of happiness, neither signal can be used as a marker for one emotion, as the relation between emotions and such signals is "many-to-one" [5]. A smile can also be used to communicate non-emotional messages [1]. Consequently, as each of these signals can be influenced by numerous non-emotional phenomena, it is important that automated emotion detection systems rely on multiple signals directly related to expressions of emotion. However, it is a complicated task for computers to disentangle the multitude of signals that define emotion. Our proposed detection model was validated using a labeled dataset that contains multiple streams of physiological data recorded in naturalistic settings and outperformed different models.

While automated emotion detection and recognition promise to contribute to support human well-being and mental health, there are discussions about the potential risks and ethical concerns of this type of data, e.g., in medicine [34], where automated emotion detection has been used for mental health assessments. The criticism focuses mainly on the validity, the training, and ethical scaffolding necessary for a qualified assessment of the produced outputs [34]. Furthermore, automated pervasive sensing can raise serious privacy concerns, including the potential misuse of data, e.g., in the workplace, where the detection of boredom by the employer could potentially lead to increased workload and stress for employees.

In comparison to vocal and facial features - while not impossible - it is more difficult to use physiological signals, such as the ones used in this study, to identify individuals, especially in naturalistic settings [9, 42]. Nevertheless, it is crucial to develop systems that process data in-situ without requiring the data to be sent through networks and cloud-services, as leaks of any kind of physiological or biometric data can put users' security and privacy at risk. Our system has the potential to run entirely on smartphones, whose processing power and memory are continuously increasing. Furthermore, a phone-based system can more easily give the user control over which specific data are being collected at any moment in time.

5.2 Limitations

Our work has several limitations. First, the amount of data samples in the final evaluation dataset is relatively small, which, to a certain extent, limited the learning ability of our proposed model. Compared with some well-known larger-scale datasets (e.g., ImageNet [6], LibriSpeech [20]) used in computer vision and speech recognition, our evaluation dataset only consists of 3180 samples. There is a need in emotion recognition research for larger natural and spontaneous emotion datasets based on commercial mobile and wearable devices. Second, we only considered the dimensional model (i.e., arousal and valence) to describe human emotions available in the used datasets. Further work is needed to test our approach using the discrete emotion model (i.e., anger, disgust, fear, happiness, sadness, and surprise). Third, we did not implement our system in a real-time field deployment as we first aimed to test its robustness and usefulness using publicly datasets. In the future, we aim to develop a mobile application version and install our system on users' personal devices to conduct a long-term user study.

6 CONCLUSION

In this paper, we propose a novel emotion recognition system based on multimodal physiological signals from off-the-shelf, low-cost mobile and wearable devices, which offers greater data privacy and the potential to be deployed into real-world contexts. Our system uses a convolutional front-end to embed each physiological signal, and employs the conformer structure to extract the local and global dependencies within and across different signals. We evaluated our system on a natural and spontaneous emotion dataset acquired with off-the-shelf devices. Our results show that our system outperformed the baselines, and achieved state-of-art or competitive performance. In addition, our generalizability experiment further established that our system has the ability to scale and handle different sensor signals. We discuss the potential of our system to be deployed into real-world scenarios for daily emotion monitoring and management. We also reflect on the potential risks and concerns related to pervasive emotion recognition approaches.

7 ACKNOWLEDGMENTS

This work was supported by an Australian Research Council Discovery Project (DP190102627).

REFERENCES

[1] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.

[2] Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. 2021. A Transformer Architecture for Stress Detection from ECG. In *2021 International Symposium on Wearable Computers*. 132–134.

[3] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. 2019. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3286–3295.

[4] Ira Cohen, Ashutosh Garg, Thomas S Huang, et al. 2000. Emotion recognition from facial expressions using multilevel HMM. In *Neural information processing systems*, Vol. 2. Citeseer.

[5] Sylvain Delplanque and David Sander. 2021. A fascinating but risky case of reverse inference: From measures to emotions! *Food Quality and Preference* November 2020 (2021), 104183.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

[7] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2019. Laughter recognition using non-invasive wearable devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 262–271.

[8] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *Comput. Surveys* 47, 3 (2015), 1–36.

[9] Jorge Goncalves, Pratyush Pandab, Denzil Ferreira, Mohammad Ghahramani, Guoying Zhao, and Vassilis Kostakos. 2014. Projective Testing of Diurnal Collective Emotion. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. 487–497.

[10] Hector A Gonzalez, Shahzad Muzaffar, Jerald Yoo, and Ibrahim M Elfadel. 2020. BioCNN: A hardware inference engine for EEG-based emotion detection. *IEEE Access* 8 (2020), 140896–140914.

[11] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).

[12] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.

[13] John F Kihlstrom. 2021. Ecological validity and "ecological validity". *Perspectives on Psychological Science* 16, 2 (2021), 466–471.

[14] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[15] Azadeh Kushki, Jillian Fairley, Satyam Merja, Gillian King, and Tom Chau. 2011. Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. *Physiological measurement* 32, 10 (2011), 1529.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[17] Terrance Liu, Paul Pu Liang, Michal Muszynski, Ryo Ishii, David Brent, Randy Auerbach, Nicholas Allen, and Louis-Philippe Morency. 2020. Multimodal privacy-preserving mood prediction from mobile data: A preliminary study. *arXiv preprint arXiv:2012.02359* (2020).

[18] Steven Marwaha, Matthew R Broome, Paul E Bebbington, Elizabeth Kuipers, and Daniel Freeman. 2014. Mood instability and psychosis: analyses of British national survey data. *Schizophrenia bulletin* 40, 2 (2014), 269–277.

[19] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. 2003. Speech emotion recognition using hidden Markov models. *Speech communication* 41, 4 (2003), 603–623.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

[21] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7, 1 (2020), 1–16.

[22] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[23] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.

[24] Jingyu Quan, Yoshihiro Miyake, and Takayuki Nozawa. 2021. Incorporating Interpersonal Synchronization Features for Automatic Emotion Recognition from Visual and Audio Data during Communication. *Sensors* 21, 16 (2021), 5317.

[25] Mika Raento, Antti Oulasvirta, and Nathan Eagle. 2009. Smartphones: An emerging tool for social scientists. *Sociological methods & research* 37, 3 (2009), 426–454.

[26] Erika L Rosenberg and Paul Ekman. 1994. Coherence between expressive and experiential systems in emotion. *Cognition & Emotion* 8, 3 (1994), 201–229.

[27] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[28] Sriparna Saha, Shreyasi Datta, Amit Konar, and Ramadoss Janarthanan. 2014. A study on emotion recognition from body gestures using Kinect sensor. In *2014 International Conference on Communication and Signal Processing*. IEEE, 056–060.

[29] Zhanna Sarsenbayeva, Gabriele Marini, Niels van Berkel, Chu Luo, Weiwei Jiang, Kangning Yang, Greg Wadley, Tilman Dingler, Vassilis Kostakos, and Jorge Goncalves. 2020. Does Smartphone Use Drive Our Emotions or Vice Versa? A Causal Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[30] Elaine Sedenberg and John Chuang. 2017. Smile for the camera: privacy and policy implications of emotion AI. *arXiv preprint arXiv:1709.00396* (2017).

[31] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.

[32] Yangyang Shu and Shangfei Wang. 2017. Emotion recognition through integrating EEG and peripheral signals. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2871–2875.

[33] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* 3, 1 (2011), 42–55.

[34] Isabel Straw. 2021. Ethical implications of emotion mining in medicine. *Health Policy and Technology* 10, 1 (2021), 191–195.

[35] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. 2016. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* 9, 2 (2016), 147–160.

[36] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[38] Bin Wang, Chang Liu, Chuanyan Hu, Xudong Liu, and Jun Cao. 2021. Arrhythmia Classification with Heartbeat-Aware Transformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1025–1029.

[39] Yi Wang, Zhiyi Huang, Brendan McCane, and Phoebe Neo. 2018. EmotioNet: A 3-D Convolutional Neural Network for EEG-based Emotion Recognition. In *2018 International Joint Conference on Neural Networks (IJCNN)*. 1–7. https://doi.org/10.1109/IJCNN.2018.8489715

[40] Zhu Wang, Zhiwen Yu, Bobo Zhao, Bin Guo, Chao Chen, and Zhiyong Yu. 2020. EmotionSense: An Adaptive Emotion Recognition System Based on Wearable Smart Devices. *ACM Transactions on Computing for Healthcare* 1, 4 (2020), 1–17.

[41] Tianyuan Xu, Ruixiang Yin, Lin Shu, and Xiangmin Xu. 2019. Emotion recognition using frontal eeg in vr affective scenes. In *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, Vol. 1. IEEE, 1–4.

[42] Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition. *IEEE Transactions on Affective Computing* (2021), 1–17.

[43] Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The Visual Computer* 37 (2021), 1447–1466.

[44] Bobo Zhao, Zhu Wang, Zhiwen Yu, and Bin Guo. 2018. EmotionSense: Emotion recognition based on wearable wristband. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 346–355.

[45] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. 2018. Personality-Aware Personalized Emotion Recognition from Physiological Signals.. In *IJCAI*. 1660–1667.

[46] Junjie Zhu, Yuxuan Wei, Yifan Feng, Xibin Zhao, and Yue Gao. 2019. Physiological Signals-based Emotion Recognition via High-order Correlation Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 3s (2019), 1–18.

[47] M Sami Zitouni, Cheul Young Park, Uichin Lee, Leontios Hadjileontiadis, and Ahsan Khandoker. 2021. Arousal-Valence Classification from Peripheral Physiological Signals Using Long Short-Term Memory Networks. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 686–689.