LSTM-Modeling of Emotion Recognition Using Peripheral Physiological Signals in Naturalistic Conversations

M. Sami Zitouni, *Member, IEEE*, Cheul Young Park, Uichin Lee, Leontios J. Hadjileontiadis, and Ahsan Khandoker

***Abstract*—The automated recognition of human emotions plays an important role in developing machines with emotional intelligence. Major research efforts are dedicated to the development of emotion recognition methods. However, most of the affective computing models are based on images, audio, videos and brain signals. Literature lacks works that focus on utilizing only peripheral signals for emotion recognition (ER), which can be ideally implemented in daily life settings. Therefore, this paper present a framework for ER on the arousal and valence space, based on using multi-modal peripheral signals. The data used in this work were collected during a debate between two people using wearable devices. The emotions of the participants were rated by multiple raters and converted into classes in correspondence to the arousal and valence space. The use of a dynamic threshold for ratings conversion was investigated. An ER model is proposed that uses a Long Short-Term Memory (LSTM)-based architecture for classification. The model uses heart rate (HR), temperature (T), and electrodermal activity (EDA) signals as its inputs with emotional cues. Additionally, a post-processing prediction mechanism is introduced to enhance the recognition performance. The model is implemented to study the use of individual and different combinations of the peripheral signals, as well as utilizing annotations from different ratings. Additionally, it is employed for classification of valence and arousal in an independent and combined fashion, under subject dependent and independent scenarios. The experimental results have justified the efficient performance of the proposed framework, achieving classification accuracy >96% and >93% for the independent and combined classification scenarios, accordingly. The comparison of the achieved performance against the baseline methods shows the superiority of the proposed framework and the ability to recognize arousal-valance levels with high accuracy from peripheral signals, in real-life scenarios.**

***Index Terms*—Affective computing, arousal, emotion recognition, LSTM, physiological signals, valence.**

I. Introduction

HUMAN emotions are complex processes that are caused by physiological and psychological reaction to an interaction (with human, object, or machine) or to a situation. Moreover, emotions can lead to having bodily changes, feelings, thoughts, and behaviors and can be affected by the human personality, mood, motivation, and temperament [1]. Emotions are essential for humans' communication in their daily life. They can be expressed through facial expressions, gestures, and vocal traits, or verbally by using emotional vocabulary [2]. Moreover, they have a direct effect on human cognition including decision-making, human interaction, perception, and human intelligence [3].

Emotion recognition (ER) plays a major role in many areas, such as healthcare, education, rehabilitation, and robotics. ER is a demanding process, as emotions' mechanisms and origin are still mysteries, and the different feelings that can be experienced can not be clearly defined [4]. The increasing presence of mobile and wearable devices in our daily lives and their use for shopping, gaming, social media, and healthcare leads to more human-computer interactions. However, the deficiency of the current human-computer interaction systems in understanding and processing the human's emotional cues is evident. This lack of emotional intelligence causes them to be unreliable in identifying the human emotional state and execute the proper actions, accordingly [5].

Conventional approaches mainly depend on visible manifestations, such as facial expressions, gestures, and speech to recognize emotional states in human-computer interactions systems. This facilitates emotion annotations since most humans respond to emotional stimuli with similar manifestations. However, a major uncertainty arises from using methods that rely on such external manifestations, since they can be consciously regulated, as humans can conceal their feeling, or are naturally suppressive [1]. This makes such methods of ER somehow subjective, which leads to inconsistent performance and serious implications in certain applications. Privacy also could be a major concern to many people, and therefore they may object to sharing their images, videos, or audios with local machines or remote cloud databases [6].

Manuscript received 1 February 2022; revised 28 August 2022 and 2 October 2022; accepted 20 November 2022. Date of publication 29 November 2022; date of current version 6 February 2023. This work was supported by a funded project [8474000408 (CIRA 2021-051)] from Khalifa University. *(Corresponding author: M. Sami Zitouni.)*

M. Sami Zitouni is with the College of Engineering and IT, University of Dubai, 14143 Dubai, United Arab Emirates (e-mail: mzitouni@ud.ac.ae).

Cheul Young Park and Uichin Lee are with the Graduate School of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea (e-mail: cheuly@kse.kaist.ac.kr; uclee@kaist.edu).

Leontios J. Hadjileontiadis is with the Department of Biomedical Engineering, Khalifa University of Science and Technology, Abu Dhabi 127788, United Arab Emirates, and also with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (e-mail: leontios.hadjileontiadis@ku.ac.ae).

Ahsan Khandoker is with the Department of Biomedical Engineering, Khalifa University of Science and Technology, Abu Dhabi 127788, United Arab Emirates (e-mail: ahsan.khandoker@ku.ac.ae).

Digital Object Identifier 10.1109/JBHI.2022.3225330

Alternatively, emotions can be identified based on the physiological signals (biosignals), which are electrical, thermal, or mechanical signals measured from the human body over time. The most commonly used types of physiological signals include electrocardiogram (ECG), electroencephalogram (EEG), heart rate (HR), electrodermal activity (EDA), temperature (T), galvanic skin response (GSR), mechanomyogram (MMG), blood volume pressure (BVP), respiration rate (RR), Electrooculogram (EOG), electromyography (EMG) and Photoplethysmogram (PPG). Traditionally, the diagnostic information of these signals is used in clinical practice for disease diagnosis and symptoms progression. However, as these biosignals are captured as outputs from the human body status and functionality, they also are affected by the current human emotional state and convey valuable information about it [7]. The main advantage of using physiological signals for ER is that they are involuntary reactions, caused by changes in the nervous system and endocrine system, making them difficult to be controlled and masked through subjective consciousness. Thus, a more objective and reliable ER can be achieved [8].

In this context, this work aims at investigating the challenge of classifying emotions based on the arousal-valence space using multi-modal peripheral signals that can be collected in non-invasive manner and monitored in the daily life. To this end, we present a framework that uses a convolutional LSTM network at its core, to classify people's emotions into affective levels of arousal and valence. It was proven in the literature, methods using LSTM networks were able to achieve good and robust performances when used in classification tasks of sequences from physiological signals [8], [9], [10]. Further, We consider the naturalistic conversation scenarios, since there are many social activities where multiple people take turns and they interact with one another. A naturalistic conversation is a conversation that people would actually conduct in real life settings, which is neither scripted nor affected by external influences. Thus, we argue that ER in such scenario must carefully consider two aspects: 1) emotion label diversity, 2) inter-person variations, and 3) temporal variations of emotion labels. The main contributions of the work are:

- The proposed ER framework archives robust and accurate performance using non-invasive peripheral physiological signals, that can be continuously obtained via wearable devices.
- A cumulative prediction mechanism is introduced to infer for the predicted emotion class by combining both current and past network prediction scores.
- A multi-perspective assessment scheme is involved in the annotations for training, considering self-, partner- and combined-ratings, and the effects on data distribution at the arousal-valence space as in Russell's circumplex model of affect [11], and on the network's classification performance are explored, in addition to introducing dynamic thresholding for affective rating conversion.
- Emotion labels diversity and their temporal variations, as well as inter-person variations are considered simultaneously, and their effect on the recognition is demonstrated and discussed.

The used physiological signals are collected in real-life settings during naturalistic conversations in a form of a debate between participants, capturing their social interactions. We first perform the training and classification based on two levels of arousal and balance. Then, we extend it into a four class problem based on the four quadrants of the arousal-valence space. Extensive experiments are conducted to demonstrate the performance and robustness of the proposed framework in subject dependent as well as independent tests, and show the superiority in comparison to baseline classification.

II. Related Work

A considerable amount of research is being conducted in interdisciplinary fields, including biomedical engineering, computer science, psychology, and AI, to develop emotion classification schemes that enable machines to detect, analyze, and interpret human emotions [12]. Some of such schemes are based on single-modal data, where emotion information is extracted from one type of signal. Other methods are based on multi-modal data, where the cues from multiple physiological signals are used to achieve higher accuracy and robustness in distinguishing between different emotions. Additionally, the emotions can be classified based on how they are conceived, as discrete emotions (e.g., happiness, sadness, surprise, anger, fear, etc.) [13], pleasantness state (pleasant, unpleasant, neutral) [14], or as dimensional emotions (combination of two parameters: arousal and valence) [15], [16], [17]. In classifying discrete emotional state, discrete emotion model is considered to recognize the groups of emotions, whereas the others refer to circumplex ER. This study focuses on dimensional emotions and the recognition of the arousal-valence state.

A. Peripheral Signals-Based Methods

Many studies in the literature demonstrated that ER can be achieved, using single or multi modal peripheral physiological signals. The major advantage here is the easiness of continuously obtaining such signals using daily life wearable devices. This is highly desirable as it allows continuous monitoring of person's emotional states for healthcare applications, and human-machine interactions.

1) Single-Modal Methods: Shukla et al. [18] performed a study on feature selection for ER, where 40 EDA features from time, frequency, and time-frequency domains were implemented on AMIGOS dataset [19]. In this study, feature selection methods including Conditional Mutual Information Maximization, Joint Mutual Information, and Double Input Symmetrical Relevance, as well as machine learning techniques were used in the experiments [18]. It was concluded that similar numbers of features are needed to achieve optimal average accuracy for arousal (85.75%) and valence (83.9%) recognition, in a subject dependent scenario. Additionally, results obtained in subject independent tests were much inferior. Agrafioti et al. [1] proposed a method that used ECG signals to detect dynamically evolving emotion patterns based on the empirical mode decomposition. Instantaneous frequency and the local oscillation were utilized to extract features for classification. This work considered active, as well as passive arousal, while suggesting that active induction method yields more ECG reactivity. The average accuracy achieved for two class arousal classification is 76.19%.

2) Multi-Modal Methods: A method that utilized nonlinear features extracted for the assessment of emotional responses from ECG, EDA, and RR signals was proposed by Valenza et al. [20]. In this work, 35 participants were presented with images from the International Affective Picture Systems (IAPS) [21], and the signals were acquired simultaneously considering five levels of arousal and valence including neutral ones. It was shown that the classification performance based on a quadratic discriminant classifier, was improved when features were extracted from nonlinear dynamic methods compared to

standard features. Wiem and Lachiri [16] presented a framework to classify emotional statements according to arousal and valence model using ECG, RR, T and GSR signals collected in the MAHNOB-HCI multi-modal tagging database [22]. Prepossessing was performed to remove artefacts and noise, and 169 features were extracted. Emotional states were classified by SVM. It was deduced that ECG and RR signals are the key determinants for recognizing human's feelings. Accuracy of 64.23% and 68.75% was achieved for arousal and valance level classification, respectively. Zhang et al. [23] introduced a few-shot learning algorithm to rapidly converge on a small amount of training data for fine-grained valence and arousal recognition. Tests achieved an averaged accuracy of 76.04% (arousal), 76.62% (valence), and 57.62% (four quadrant with neutral). Additionally, Zhang et al. [24] proposed a method based on deep multiple instance learning, to recognize emotions at a finer granularity level when trained with post-stimuli labels, where instances are weakly-supervised in training stage. In Elalamy et al. [25] work, recurrence plots were used to obtain 2D representations of physiological activity, to get less subject dependent and better suited representation for non-stationary signals such as EDA, in conjunction with ECG and PPG.

B. Brain Signals-Based Methods

A major part of the work conducted in the field is based on detecting emotions from brain (EEG) signals, especially with the popularity of datasets, such as DEAP dataset [2], and the strong evidence of them containing determinant emotion information. However, the setup and devices needed for collecting the signals make it less suitable for daily life applications (e.g., health and mental state monitoring).

1) Single-Modal Methods: Several single-modal state-of-the-art methods were developed based on the EEG data from DEAP dataset [26], [27], [28]. Other methods based on single-modal brain signals that utilize LSTM networks are [9]. Additionally, Sourina et al. [29] presented a study that performed classification using SVM in a subject dependent manner. Petrantonakis and Hadjileontiadis presented a study for ER that used hybrid adaptive filtering with higher order crossings analysis [30], and a method for evaluating the emotion elicitation procedures using frontal brain asymmetry theory, where classification was performed using SVM in subject dependent and independent manners [31]. Further, Aydin et al. [32] presented a method for detection of emotional dysfunctions through estimating the level of nonlinear inter-hemispheric synchronization using wavelet correlation.

2) Multi-Modal Methods: The following studies used brain signals in conjunction with peripheral signals. A study to detect the arousal level of participants through three types of stimulus was conducted by Anderson et al. [33]. Signals including EEG, ECG, GSR, EOG, and PPG were collected while participants were exposed to exciting and relaxing videos and music before playing Tetris and Minesweeper. Machine learning techniques were used for analysis achieving arousal classification accuracy of 88.9% with SVM. It was concluded that it is possible to determine the type of stimuli used by analysing the biosignals. Liao et al. [8] proposed a method where a Convolutional Neural Network (CNN) was utilized to learn multi-channel EEG spatial representations and an LSTM was used to learn temporal representations of other signals including EOG, GSR, RR, BVP, T, and EMG. For arousal and valence classification, accuracy of 63.06% and 62.41% was achieved, respectively, using peripheral signals representation, and 93.06% and 91.95% by combining both EEG and peripheral representation, accordingly. Saffaryazdi et al. [34] presented a study for recognizing emotions in a face-to-face conversation using hand-crafted features from EEG, GSR, and PPG signals. Classification results achieved F-score of 77.6% and 80% for arousal and valence, respectively, in subject-dependent tests, while 68% and 64% where achieved in subject-independent tests.

Despite considerable attempts in this area, emotions and ER have remained largely unexplored. Specifically, there is a lack of works that focus on detecting emotions with physiological data collected only from wearable devices in naturalistic scenarios. Most of the state-of-the-art studies (as deduced from this section) either use brain signals, which require delicate setups for capturing, or involve specific preparation and conditioning, refraining them from daily life use. This motivate us to develop a framework based on a naturalistic dataset to efficiently provide ER, in terms of high robustness and accuracy in its performance.

III. Dataset and Annotations

A. Dataset Overview

In this work, the K-EmoCom [35] database was adopted, which is a publicly available emotion dataset comprised of multi-modal affective information, including facial expressions, conversation audios, and physiological signals acquired from 32 participants engaging in 10-minute long paired debates on a social issue. Although the dataset involves a social discussion on a single topic to standardize a protocol, it is a dynamic topic for debate since it is about a very controversial subject, in which a conversation can induce wide range of emotions. It enables studying emotions in the context of naturalistic conversations, in particular recognizing continuous emotional states from physiological signals acquired from commercial-grade wearable devices. The dataset consists of annotations of emotions observed during debates at the interval of 5-seconds. Emotions were annotated from three unique perspectives, i.e., of subjects themselves, corresponding debate partners, and external observers. K-EmoCon, according to our knowledge, is the first dataset accommodating multi-perspective assessment of emotions during naturalistic social interactions.

B. Data Selection

Participants were randomly paired to engage in a dyadic face-to-face debate on the Jeju Yemeni refugee issues. Sixteen debates, which sum to 172.92 minutes (M = 10.8 min, SD = 1.04 min), were conducted. All data collection sessions were arranged in a room with controlled temperature and illumination, where two participants sat across a table facing each other with cameras in the middle recording their facial expressions, upper body gestures, and speech audios (see Fig. 1). Throughout a session, participants wore a set of wearable devices for the collection of physiological signals. Table I summarizes the list of devices and respective sampling rates and ranges for signals used in the presented work.

In the proposed framework, we are focusing on signals from the dataset obtained from wearable devices as described in Table I, that have similar low sampling rates (1 or 4 Hz). This will allow us to seamlessly use sensor data in a single, fast, multi-modal detection system. Thus, the signals that are being used in this work include two HR signals (denoted as HR for the signal acquired with the Empatica E4 wristband, and HRp for

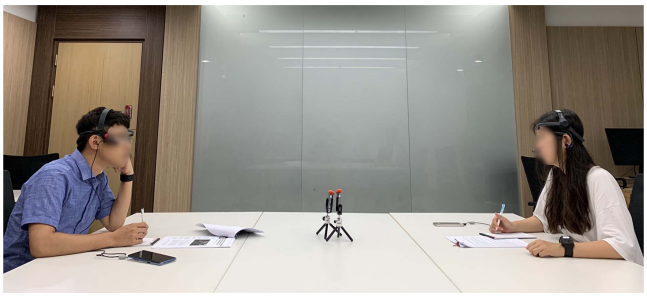

Fig. 1. Two participants in a debate during data collection session for K-EmoCon dataset [35].

TABLE I
DATA COLLECTION DEVICES AND ACQUIRED SIGNALS USED IN THE PRESENTED FRAMEWORK

Device	Signal acquired	*fs*	[min, max]
Empatica E4 Wristband	EDA (or GSR)	4Hz	[0.01 µS, 100 µS]
	HR (from BVP)	1Hz	[n/a]
	Skin temperature	4Hz	[−40 °C, 115 °C]
Polar H7 HR Sensor	HRp (from ECG)	1Hz	[n/a]

the signal acquired with polar H7 sensor, as in Table I), EDA, and T. Data from 21 out of the 32 participants were used in this work, due to unavailability of one or more of these four signals corresponding to 11 participants in the dataset.

IV. METHODOLOGY

A. Physiological Signals

The signals in this work are normalized before being used in the ER system. Based on the data collection protocol, all subjects were introduced to a relaxing video for 2 minutes before starting the debate to measure the baseline of their neutral state. This accounts for individual biases and reduce the effect of previous emotional states. Thus, we use the signals collected within this period as a reference to the subject neutral state. The segmentation of these signals is based on the data collection time stamps. Each participant's signals are normalized based on their personal neutral state, using the last 1.5 minutes of the baseline data. The normalization is done to remove the bias from each participant, leaving mainly the changes in the signal that correspond to the emotional state of the participant during the debate session. This is done by applying the following equation:

$$S_n = S - mean(S_r), \tag{1}$$

where S is the signal in the debate period, and S_r is the signal collected during the relaxing period. Normalization is commonly used in machine learning, where the data is either scaled or transformed, to allow equal contributions and minimize the bias of features whose numerical contribution is higher in discriminating classes [36]. Thus, since this framework is multi-modal, and the data belong to various subjects, the preliminary tests showed that this proposed mean centered based normalization approach enhances the performance of the proposed recognition model.

B. Emotion Classes & Ratings

Emotions are classified based on the level of arousal and valence (affective dimensional emotional model [11]). Both arousal and valence were categorized into two classes, i.e., low (L) and high (H). Further, the emotions are recognized as belonging to one of four affective classes representing the quadrants of the arousal and valence space. These classes include: low arousal low valence (LALV), low arousal high valence (LAHV), high arousal low valence (HALV), and high arousal high valence (HAHV). Taking into consideration label diversity for emotional class generation, we perform the analysis and comparisons by considering the self annotations and partner annotations from the dataset. Further, combined annotations of both ratings are considered, which give equal weights to both self and partner labels so that we can have a unified model for prediction. The category of arousal and valence is determined based on the ratings of the annotators and the level is assigned according to a mid value (threshold). For self annotations and partner annotations, if the rating is below 3 (mid point), then it is considered low, otherwise it is high. On the other hand, for the combined annotations, the rating of both self and partner are added; then, if it is below 6 it is considered as low; otherwise it is high. Table II summarizes the conversion of the annotators' ratings into the corresponding class. Finally, the combinations of the levels of both arousal and valence determine which quadrant the emotion belongs to. Fig. 2 shows the distributions of the labels along the affective classes in each type of annotation.

TABLE II
CONVERSION OF THE ANNOTATORS RATINGS (R) INTO THE CORRESPONDING AROUSAL AND VALENCE CLASS

	Class	
Annotations	High	Low
Self	$r \geq 3$	$r < 3$
Partner	$r \geq 3$	$r < 3$
Combined (Self+Partner)	$r \geq 6$	$r < 6$

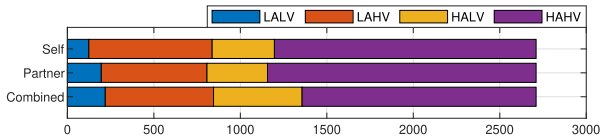

Fig. 2. Distribution of labels in each type annotation.

This way of ratings conversion, however, does not take into account the individual differences between the raters, such as their perspectives and biases, since the same middle value is used to represent the neutrality. Nevertheless, in reality, this is not accurate. For example, in our setup, one rater may provide all his ratings between 3 and 5, and another one may give rates of 1 and 2 only. With the use of the mid value as a threshold, all the ratings of the first one will be converted into high, and all the ratings of the second one into low. Therefore, taking a different conversion threshold for each rater may more accurately reflect their annotations, as one's low may be considered other's high. To that end, we introduced a dynamic threshold for ratings (DTR) to minimize the resultant effect of label diversity, where the conversion is done based on a changing threshold according to the rater's values distribution throughout the session. The DTR is determined as follows,

$$DTR = round(mean(\Sigma_1^m r_i)) - 0.5, \tag{2}$$

where r_i is the corresponding rating value for the ith instance across the total m instances.

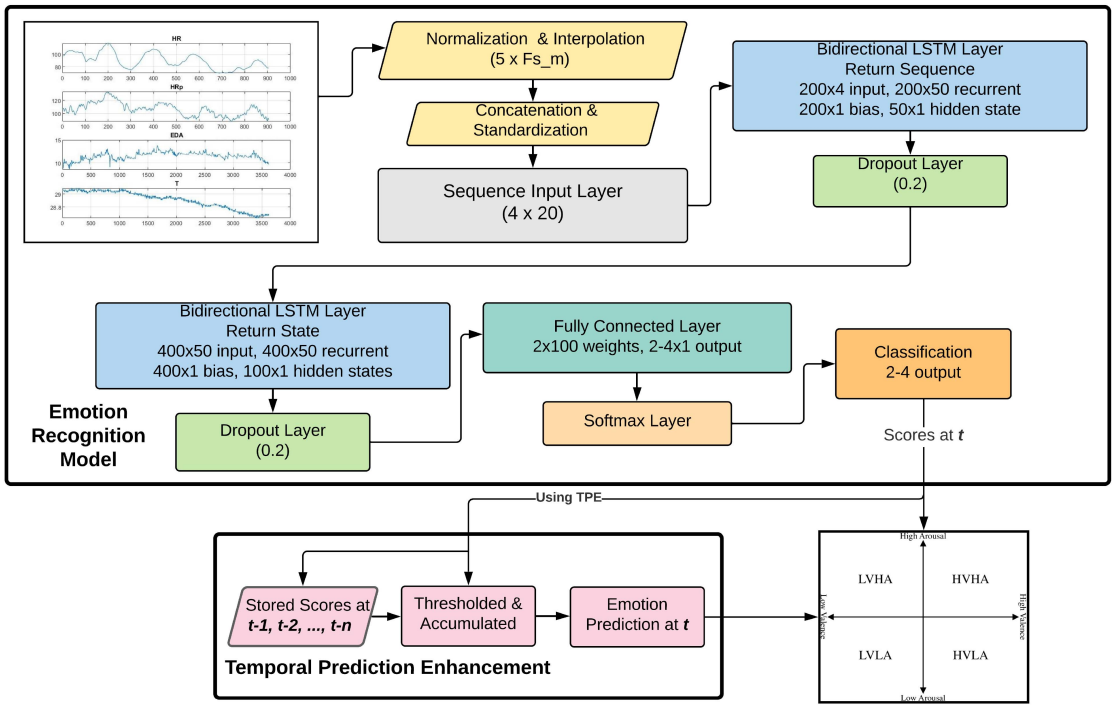

Fig. 3. Block diagram illustrating the proposed ER model with temporal prediction enhancement.

C. Proposed ER Model

Fig. 3 shows a block diagram of the proposed ER system. First, raw physiological signals are used including HR, Hrp, EDA, and T. Before feeding these signals to the network, they are normalized (see Section IV-A) and interpolated. The interpolation is done based on the nearest-neighbor. Then, the signals are segmented into segments of size $w \times Fs_{\max}$, where $Fs_{\max}$ is the highest sampling frequency of the used signals, and w is the sampling period of an emotion annotation. Here, $w = 5$ (i.e., one in every five seconds). All the data obtained from the signals are standardized as follows:

$$S_d = \frac{S_n - mean(S_n)}{SD(S_n)}, \tag{3}$$

where $SD(S_n)$ is the standard deviation of the normalized signal.

Standardization is performed on training and testing data as it can noticeably improve the network performance specially since we are using signals obtained from different sources and have different ranges of values.

Then, the signals are fed to an LSTM network. The LSTM network proposed here consists of a sequence input layer, two bidirectional LSTM (BiLSTM) layers, two dropout layers, a fully connected layer, a softmax layer, and a classification layer. First, the input layer has a size of four (as we are using four signals) and takes the standardized segmented sequences. The first BiLSTM layer contains 50 hidden units and returns a sequence with the same size as the input sequence. A drop layer with a probability of 0.2 is then used to reduce the occurrence of over-fitting [28]. Then, another BiLSTM layer is used that has 20–50 hidden units (different number of hidden units used for various tests), and this BiLSTM layer returns the last state. Another dropout layer with 0.2 probability is used to avoid over-fitting. After that, a fully connected layer is fed with the output, and then a softmax layer is used for activation. Finally,

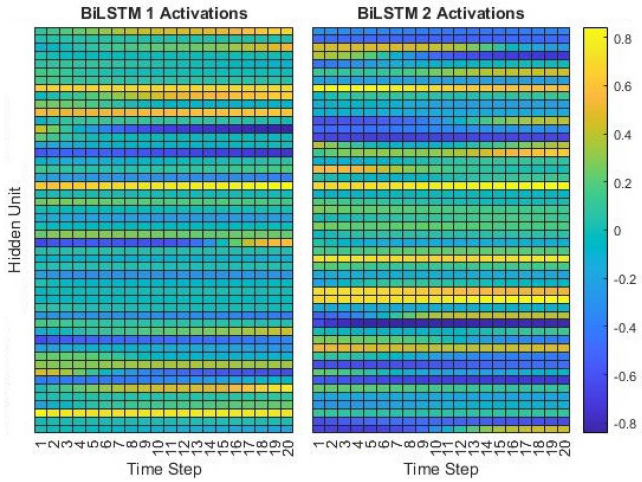

Fig. 4. Visualization of the BiLSTM layers activations.

the classification output has two or four values, based on the usage of the network, whether the system is used for classifying arousal and valence levels into high or low, or it is used to classify the emotion into one of the four quadrants (LALV, LAHV, HALV, and HAHV). Fig. 4 displays visualization of the BiLSTM layers activations' to provide interpretability of the learned features, where the heatmaps show how strongly each hidden unit activates and highlights how the activations change over time.

1) Temporal Prediction Enhancement: For the emotion classification where the annotations were performed at short periods of time (5 seconds in our dataset), we argue that it is not usual for emotions to change dramatically within such a short period. Additionally, this results in increase in the temporal variability of emotion labels. To accommodate this variability, we propose a Temporal Prediction Enhancement (TPE), a mechanism where

a prediction of the emotion at a certain time can take into consideration the predictions at previous periods. This is done by introducing an additional step before deciding the class based on the scores of the classifier, which could significantly improve performance in continuous social interaction scenarios.

First, the current scores of the classification performed using the ER model are obtained for the current period (5 seconds). Then, the scores obtained from instantaneous classification results of the previous periods are fetched. Depending on how many of the past predictions are desired to be considered, the scores needed are determined. Thus, scores at $t-1, t-2, \ldots, t-n$ are used accordingly, where t is the current instantaneous prediction time, and n is the number of past predictions. For example, if it is desired to use predictions at last 10 seconds, scores at $t-1$ and $t-2$ are used, and for last 20 seconds $t-1, t-2, t-3$, and $t-4$. Therefore, for prediction using the past $\alpha \times n$ seconds, scores up till $t-n$ are used, where α is the annotation period.

Classifier scores range from 0 to 1. We decided to remove past predictions that do not have a score corresponding to one class with a value above 0.6. This is to ensure that any uncertainty in the past emotion predictions will not be carried on to the current prediction. The remaining scores are then accumulated and the emotion prediction at the current instant is determined based on the overall score. In the experiments where this mechanism is used, it was applied to the ground truth as well, which was used to measure the performance to have an appropriate assessment. This was done by giving a score of 1 to the correct class, and 0 to the others before the combination.

D. Implementation Setup

The proposed approach was implemented in Matlab 2020a. For training, we tried to balance between the number of epochs and the minimum sequence length, in order to have a decent training speed, and enough iterations to achieve high accuracy. Several tests were performed where the number of epochs was set to be between 200 and 500, while the minimum sequence length is either 20 (equal to the input sequences length) or its multiplication (40, 60, and 80). Additionally, scheduled learning rate was used, where initially the learning rate started at 0.005 and then dropped by a factor of 0.2 at half the way through the epochs.

E. Baseline Setup

Five classifiers, i.e., three heuristic voters (*random*, *majority*, and *class ratio*), *Gaussian Naive Bayes (GNB)*—a simple probabilistic classifier, and *XGBoost* [37], a popular high-performance yet efficient tree boosting system, were trained to establish a comparable baseline to evaluate the performance of proposed ER model. The aforementioned classifiers were chosen in particular to replicate experimental procedures previously employed in well-known works in the field, to evaluate performances of classifiers trained with emotion databases with imbalanced class distributions, against baseline models [2], [22], [38]. For the training of the classifiers, 30 features were extracted from 25-seconds long segments, with each segment containing BVP, EDA, T, and HR measurements. As emotions during debates were annotated every 5-seconds, physiological signals acquired during debates were first divided into 5-second segments such that each segment matches with an (*arousal*, *valence*) tuple. Five such 5 s segments within a 25-second moving window were then concatenated, with a 20-second overlap between windows. Features extracted are as shown in the list below:

BVP: mean BVP; HRV (standard deviation); mean inter-beat interval (IBI); multiscale entropy (MSE) at 5 levels; power spectral density (PSD): spectral power in 0.0–0.1 Hz, 0.1–0.2 Hz, 0.2–0.3 Hz, 0.3–0.4 Hz, and spectral power ratio between 0.0–0.08 Hz and 0.15–0.5 Hz bands; and tachogram power: lower frequency spectral power (LFSP) below 0.08 Hz, medium frequency spectral power (MFSP) in 0.08–0.15 Hz, high frequency spectral power (HFSP) in 0.15–0.5 Hz, and energy ratio between MFSP and (LFSP + HFSP).

EDA: number of peaks exceeding $100\,\Omega$ per second, mean peak amplitude from the saddle point preceding the peak, mean rise time for the signal to reach its peak from the saddle point in seconds, mean GSR, and standard deviation of GSR.

T: statistical moments [*mean*, *stdev.*, *kurtosis*, *skew*], and spectral power in [0.0–0.1 Hz, 0.1–0.2Hz].

HR: mean and standard deviation.

Features used in the baseline classification were selected from the set of features initially proposed by Soleymani et al. in *Toolbox for Emotional feAture extraction from Physiological signals (TEAP)* [38]. A Python script to replicate baseline classification with K-EmoCon is available on *K-EmoCon supplementary codes* GitHub repo, which directly depends on the *Python implementation of TEAP (PyTEAP)* package [39].

TABLE III
CLASSIFICATION ACCURACY (%) BASED ON DIFFERENT BIOSIGNALS COMBINATIONS WITH SELF ANNOTATIONS

Signal	Arousal	Valence
HR	71.10	82.06
T	73.20	82.17
EDA	75.75	82.50
HRp	67.88	81.73
HR, HRp	71.43	82.17
HR, T	76.85	83.06
HR, EDA	79.73	84.39
T, EDA	79.62	85.16
HR, T, EDA	86.27	88.70
HR, HRp, EDA, T	**88.82**	**92.91**

V. RESULTS

A. Subject-Dependent Classification

1) Arousal & Valence Classification: Table III shows the experimental results, in terms of emotion classification accuracy (%) of arousal and valence based on the self-annotations, when each physiological signal is considered alone or combined with the others. The individual signal that provides the highest classification accuracy is EDA (arousal: 75.75%, valence: 82.50%). The use of the four signals together for classification achieves the highest accuracy (arousal: 88.82%, valence: 92.91%). Table IV contains the emotion classification accuracy values for binary levels of arousal and valence. First are the 5 seconds (the annotations period in the used dataset) classification results, where the prediction is applied instantaneously at the output of the ER system's classifier. Additionally, the performance of the proposed approach while using the TPE is also tabulated. Here the results are demonstrated using normal rating conversion, as well as the proposed DTR (see (2)). Various tests' results are shown taking into consideration various numbers of past predictions ($n = \{2, 8\}$). In all experiments, the accuracy increases when TPE is used to deal with temporal variability of the emotions.

TABLE IV
EMOTION CLASSIFICATION ACCURACY (%) FOR TWO LEVELS INSTANTANEOUSLY ($n = 0$) AND USING TPE WITH VARIOUS (n)

Annotations	n	Self	Partner	Combined
Arousal	0	88.82	93.24	87.82
	0 (DTR)	86.60	91.92	85.38
	2	91.56	94.67	89.70
	4	91.68	**94.89**	91.25
	6	**92.67**	93.90	91.81
	8	90.34	94.23	**92.36**
Valence	0	92.91	92.69	89.37
	0 (DTR)	91.81	91.58	86.60
	2	**95.23**	94.34	90.79
	4	94.67	95.56	92.90
	6	95.01	96.45	91.12
	8	94.67	**96.78**	**93.34**

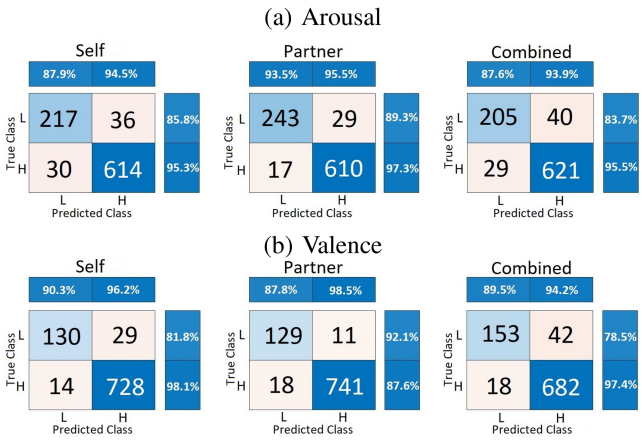

Fig. 5. Confusion matrices of (a) arousal and (b) valence classification according to the tests with bold values in TABLE IV. (a) Arousal (b) Valence.

TABLE V
EMOTION CLASSIFICATION ACCURACY (%) FOR FOUR CLASSES INSTANTANEOUSLY ($n = 0$) AND USING TPE WITH VARIOUS (n)

Annotations	n	Self	Partner	Combined
Four Classes	0	85.16	87.93	81.84
	0 (DTR)	82.83	86.93	80.95
	2	86.35	90.23	83.13
	4	87.24	92.66	84.68
	6	**89.12**	**93.65**	86.35
	8	87.01	92.96	**87.13**

This is also due to the fact that TPE deals with the instability issue of the emotion labeling. Taking into consideration label diversity influence, the highest performance is achieved for both arousal and valence classification when using partner annotations. The highest accuracy for arousal classification is 94.89% (using past 20 s ($n = 4$)), while it is 96.78% for valence classification (using past 40 s ($n = 8$)). When using the self annotations, the best values obtained are 92.67% (using past 30 s ($n = 6$)) and 95.23% (using past 10 s ($n = 2$)) for arousal and valence, respectively. On the other hand, when using the combined annotations, a slightly lower performance is achieved at 92.36% (using past 40 s ($n = 8$)) and 93.34% (using past 40 s ($n = 8$)) for arousal and valence, respectively. Fig. 5 displays the confusion matrices for (a) arousal and (b) valence classification respectively.

2) Four Quadrants Classification: Table V demonstrates the emotion classification accuracy for the four classes of LALV, LAHV, HALV, and HAHV, using the three annotation types. As in the previous experiments, the classification is performed first without considering the TPE technique, then multiple tests are presented taking into consideration temporal variability. A similar observation to the case of the binary arousal and valence classification can be made here, where for each type of annotation, all the results obtained while using the prediction mechanism are superior. Considering label diversity, the highest accuracy is achieved when using the partner annotations, then comes the self annotations and the combined annotations which is the lowest. For partner and self annotations, the best performances are achieved with the use of past 30 s ($n = 6$), where the resultant accuracy is 93.67% and 89.12%, respectively. For combined annotations, the highest accuracy is obtained using past 40 s ($n = 8$), which is 89.12%. Fig. 6 depicts the confusion matrices of the four quadrants classification results corresponding to the tests which values are in bold in Table V. The amount of data belonging to the HAHV class is the largest, which is expected, as, normally, human beings tend to be in a state of high arousal and valence. Comparing between the confusion matrices, it can be deduced that the one corresponding to the combined annotations has less biased data against the other two. In Fig. 7, plots of ROC curves are shown for arousal, valence, and four quadrants emotion classification. Each plot contains curves that correspond to ER performed using different types of annotations. Area under the ROC Curve (AUC) with 95% confidence interval (CI) values are estimated. For arousal classification, the highest AUC achieved is 0.8454 with CI of 0.7907–0.8838 using partner annotations, while the highest AUC for valence classification is 0.8615 with CI of 0.8017–0.8985 using self annotations. Moreover, the four classes experiments achieved an AUC of 0.8037 with CI of 0.7590–0.8404 with partner annotations.

B. Baseline Classification Results

Table VI shows the baseline classification results for arousal, valence, and four quadrants in comparison with the proposed method, as well as state-of-the-art methods implemented on the K-EmoCon dataset [17], [40]. Accuracy (Acc.) and F1-score averaged across 4-fold cross validation are reported for each classifier as the model's performance for both majority and minority classes is a pertinent issue given the imbalanced dataset. In all test cases, the proposed approach produces the highest performances, followed by *XGBoost* [37] which achieved a classification accuracy of 76.45%–83.18% for arousal, 78.75%–85.19% for valence, and 66.28%–72.88% for four classes. The lowest results achieved in all tests were when using the simplistic *Random* voter classifier.

C. Subject Independent Classification

To further study the impact of inter-personal variations in ER modeling, subject independent experiments were performed for arousal and valence classification. A leave-one-subject-out (LOSO) validation scheme was implemented, where each time training is performed using data from participants except the one on whom emotion prediction is applied. Thus, no prior knowledge of the testing participant exists in the network since none of their data is used in training phase. This experiment was performed using all three types of annotations.

Fig. 8 shows plots of arousal and valence classification accuracy values obtained from the LOSO test for each participant, using different annotations. In the self annotations case, the arousal classification accuracy values range between 38.52%

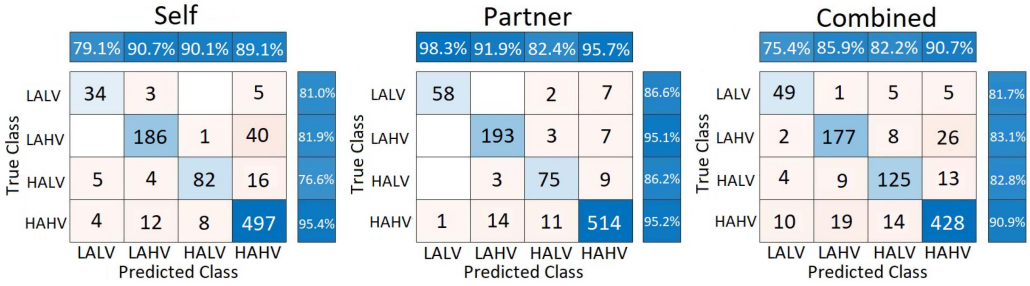

Fig. 6. Confusion matrices of four classes classifications using different types of annotations according to the tests with bold values in Table V.

TABLE VI
BASELINE CLASSIFICATION ACCURACY (%) AND F1 SCORE COMPARISON FOR AROUSAL, VALENCE, AND FOUR QUADRANTS CLASSIFICATION

Method		*Rand.*	*Majority*	*Ratio*	*GNB*	*XGBoost*	*[50]*	*[20]*	*Proposed*
Annotation	Metric				**Arousal**				
Self	*Acc.*	51.19	68.45	56.27	62.55	78.79	-	77.54	**92.67**
	F1	0.5288	0.5564	0.5626	0.5981	0.7739	0.7737	0.7478	**0.9590**
Partner	*Acc.*	50.45	69.77	58.73	67.76	83.18	-	-	**96.11**
	F1	0.5243	0.5735	0.5880	0.6286	0.8209	-	-	**0.9801**
Combined	*Acc.*	50.20	71.32	59.06	65.87	76.45	-	-	**92.36**
	F1	0.5255	0.5939	0.5905	0.6332	0.7353	-	-	**0.9599**
					Valence				
Self	*Acc.*	50.69	80.43	68.70	74.86	85.19	-	81.31	**95.23**
	F1	0.5571	0.7171	0.6890	0.7218	0.8251	0.6911	0.7680	**0.9756**
Partner	*Acc.*	50.65	79.16	66.07	74.12	83.14	-	-	**96.78**
	F1	0.5522	0.6996	0.6622	0.7052	0.8012	-	-	**0.9836**
Combined	*Acc.*	51.06	71.86	60.29	67.88	78.75	-	-	**93.34**
	F1	0.5349	0.6009	0.6032	0.6282	0.7645	-	-	**0.9653**
					Four Quadrants				
Self	*Acc.*	24.60	54.47	38.35	45.93	68.49	-	-	**89.12**
	F1	0.2801	0.3841	0.3850	0.4506	0.6576	0.5635	-	**0.9422**
Partner	*Acc.*	24.93	56.84	38.92	49.17	72.88	-	-	**93.65**
	F1	0.2778	0.4121	0.3901	0.4494	0.7020	-	-	**0.9672**
Combined	*Acc.*	25.75	50.82	33.47	44.09	66.28	-	-	**87.13**
	F1	0.2823	0.3424	0.3352	0.4178	0.6381	-	-	**0.9307**

(P4) and 100% (P21), while for valence they range between 41.03% (P8) and 98.64% (P16), with an overall accuracy of 67.9% (SD: 0.1476) and 75.98% (SD: 0.1521) for arousal and valence classification, respectively. On the other hand, using partner annotations achieves accuracy values between 34.69% (P16) and 93.94% (P11) for arousal classification with overall 64.43% (SD: 0.1884), and accuracy values between 26.9% (P12) and 95.31% (P20) for valence classification with overall 67.68% (SD: 0.1615). Finally, when combined annotations are used, the accuracy of arousal classification ranges from 49.57% (P8) to 93.75% (P20) with an overall of 65.34% (SD: 0.1236), while for valence classification it ranges from 33.79% to 89.80% with an overall of 65.5% (SD: 0.1527).

VI. DISCUSSION

Experiments were conducted to test the proposed framework for classifying binary classes of arousal and valence, as well four emotion classes corresponding to their combination, using the three annotation variations in addition to using DTR. The initial tests performed highlighted the improvement obtained from using multi-modal peripheral signals against single-modal, which is up to 20.94% increase in accuracy (see Table III). Then, in the next experiments for ER, high and satisfactory classification accuracy was achieved in all tests (Arousal 85.38–94.89%, Valence 86.60–96.78%, see Table IV, four classes 80.95–93.65%, see Table V), and noticeable improvements were shown when using the proposed TPE. It can be noticed that when using DTR, the accuracy obtained is slightly lower. Since DTR tends to more accurately represent the annotation of the emotion level, compared to the static thresholding, for several participants' data, using DTR resulted in a lower number of low annotations for arousal and valence. Since detecting the low cases is more critical, having a similar overall accuracy with less number of low annotations is actually an improvement. Furthermore, subject-independent experiments were performed to further challenge the system in a LOSO setup. For many participants/annotations, the results were very promising (up to 100%). Nonetheless, the results on some others were not very satisfactory (27% in P19), due to significant inter-person variation. Finally, a comparison with baseline methods was demonstrated, which showed the superiority of the framework presented in this paper.

A. Temporal Variations of Emotions

There are many factors regarding the data and participants that have an impact on the ER system including the emotion stability and variability of the participants, which indicate how long the emotion state remains unchanged, and how many times an emotion alteration occurs, respectively. This is reflected in the emotion annotations and mainly affected by the rater perspective, the participants emotional expressions, and their personalities. For the 21 participants from whom the data are collected, where emotions were rated each 5 s, the average stability of the arousal

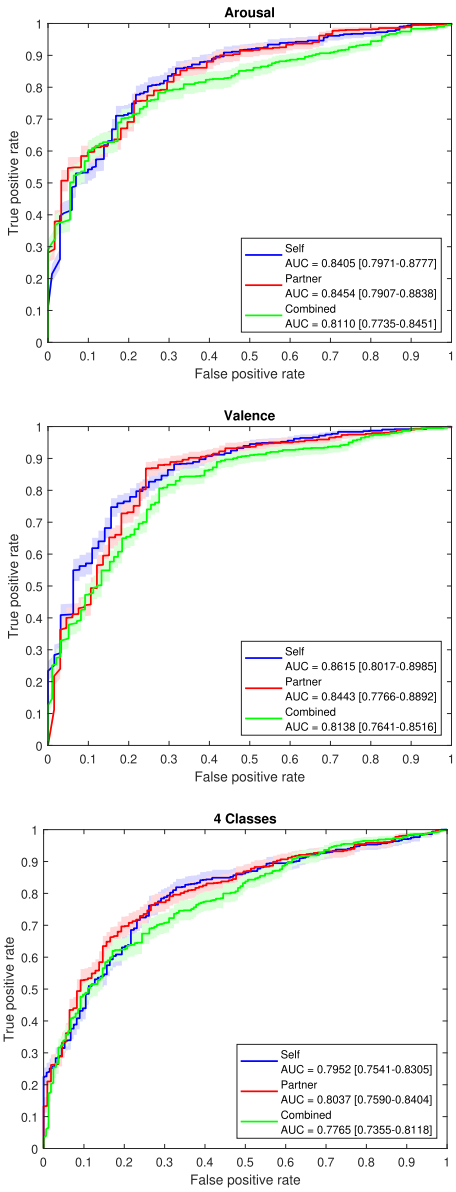

Fig. 7. ROC curves showing classification performance and AUC values with 95% confidence interval.

and valence levels is estimated to be 115 s according to self annotations, and 130 s according to partner annotations. A noticeable observation is that the positive emotion level is sustained for much longer average duration against the negative emotion level. From the self annotations, we estimated that the positive emotion average stability is 175 s against 50 s for the negative emotion, and from the partner annotations, it is 200 s against 70 s. Compared to the annotations frequency (5 s periods), the emotions usually remain unchanged much longer, which explains the enhancement obtained when using the proposed TPE technique, which takes longer time windows into consideration to produce the prediction, as it was demonstrated in Tables IV and V. Nevertheless, over increasing the prediction time window will result in miss-detecting impulsive and short-term emotional alterations, especially in the negative emotional states, which are more critical to detect and have shorter stability periods. Furthermore, to investigate the emotions variability of the participants, we

TABLE VII
STABILITY (BY AVERAGE DURATION) AND VARIABILITY (NUMBER OF ALTERATIONS PER SESSION) OF PARTICIPANTS EMOTIONS

Annotation	Self		Partner	
	Arousal	Valence	Arousal	Valence
Avg. Stability	110s	120s	165s	110s
Pos. Stability	155s	200s	230s	175s
Neg. Stability	65s	40s	100s	45s
Variability	5.9	5.2	3.8	5.7

estimated the average number of alterations in the arousal and valence levels according to self and partner annotations per recording session (approximately 10 minutes for each participant). Accordingly, an average of 5.6 and 4.8 alterations per session were estimated from self and partner perspectives, respectively. However, the variability can drastically vary depending on the rater's perspective, and the participants themselves, as it can be seen in the annotations examples in Fig. 9. More detailed estimations of emotion stability and variability can be found in Table VII.

B. Label Diversity

In fact, annotations used in training and testing the ER model can have great differences when raters with different perspectives are involved, which is investigated in this framework. Participants, by rating their own emotion, may interpret their feelings in a different way than the one observers of their emotional expression follow. Additionally, people express their emotions in different ways and intensity, as well as raters can perceive such expressions in various ways when they are annotating others' emotions. These differences are very noticeable in the naturalistic K-EmoCon dataset, since in each recording session, different couple of people rate (annotate) each other. In Fig. 9, annotations for P19 and P32 are shown, comparing the self and partner perspectives. Therefore, a sequence of data from the same participant can represent a different emotion level when fed into the model, based on the rater. This can affect the recognition performance when such annotation is inconsistent between participants. However, in our framework, this does not seem to be an issue when data from all participants with the respective annotations are used to train the model, as it is shown in the results of Tables IV and V. On the other hand, such inconsistencies can have a major effect on the performance when no data from the testing subject are used in training, as it is seen in some cases some cases of LOSO experiments (see Fig. 8). For example, valence classification test of P19 achieved 27% with partner annotations, against 98% with self annotations. The same observation can be seen in P32 valence classification test which achieved 36% with partner annotations against 100% with self annotations. This concludes that in these cases, the self annotation is more consistent with other participants than the partner. Looking at the exact annotations in Fig. 9, we can see that the valance state is rated high all the time by self, and low almost all the time by partner, which is the same case in the arousal annotations of P32, showing how much differences annotations can have, and, thus, possibly affecting the systems performance when applied on new participants.

C. Inter-Personal Variation

Another important discussion point is the variation of signals distributions and emotional responses between participants

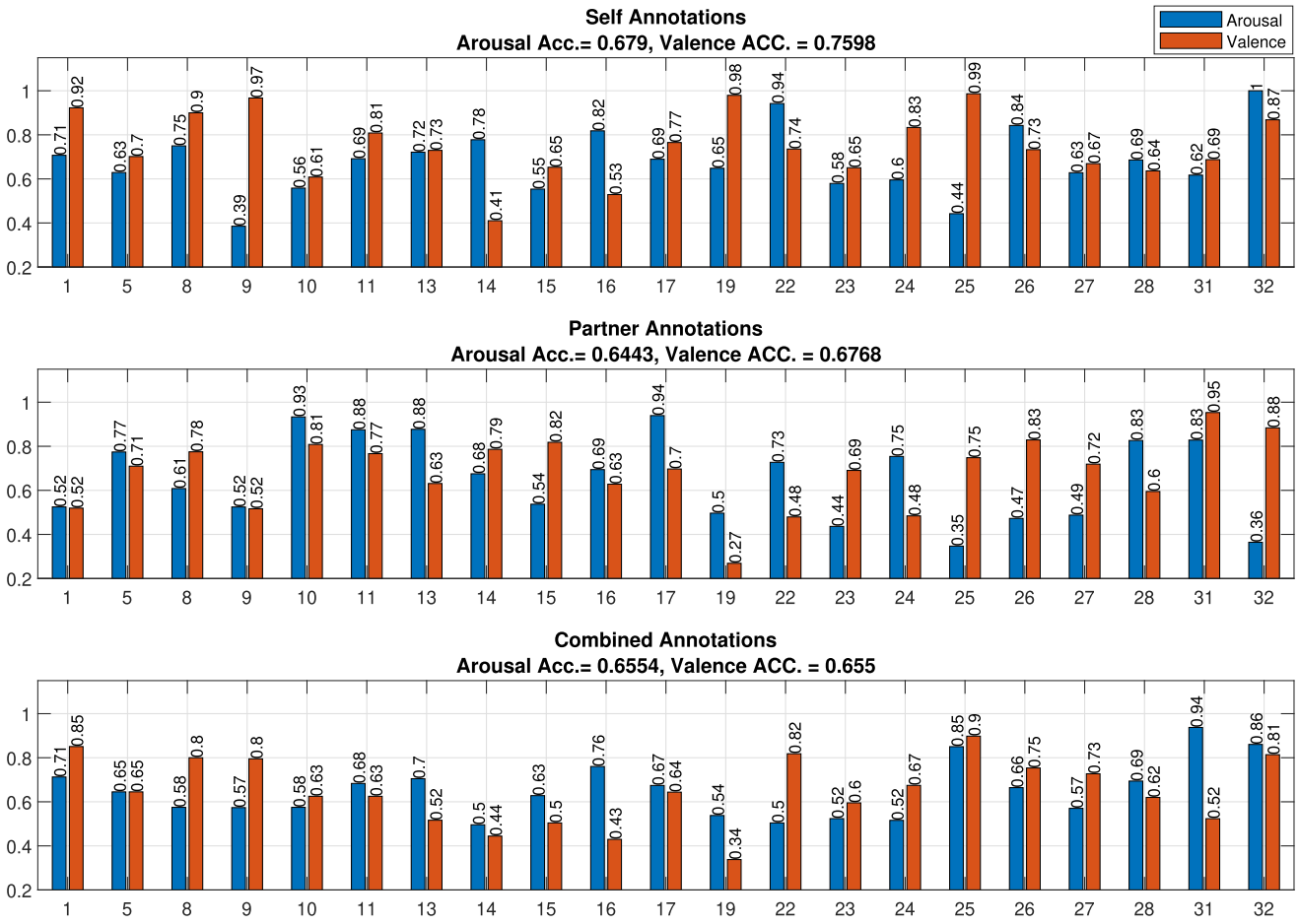

Fig. 8. Subject independent test results showing the accuracy of classification on each participant for arousal and valence levels. The performed is a leave one out using different annotations.

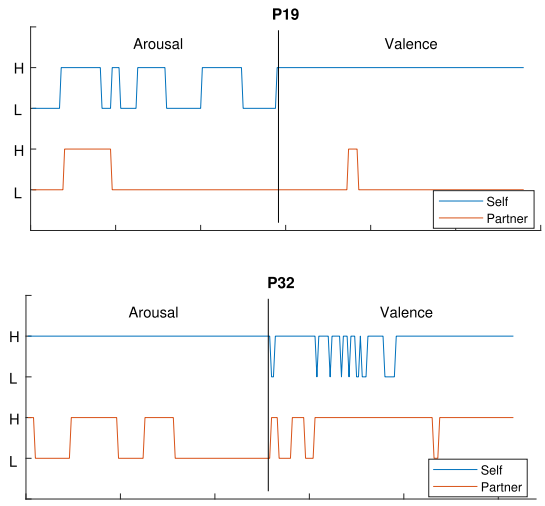

Fig. 9. Arousal and valence level annotation by self and partner raters for participants 19 and 32.

according to elements, such as age and gender, and how it can impact the ER model, especially when transferring it to different environments. The 21 participants followed a gender distribution of 12/9 Male (M)/Female (F), paired with M/M (5), F/F (1) and M/F (8), and their ages were between 19 and 30. In preliminary experiments, we used raw data, without normalization or standardization, from different groups of the participants to verify the system, which resulted in inconsistent and lower performances compared to the reported in Section V. This was mainly due to the differences in signals distributions across the participants, as well as the scales of the signals values. The effect of gender and age on the performance of the proposed approach was examined using a linear regression model. The implementation of the normalization using data from relaxation period of the recording sessions, and the standardization resulted in all p-values > 0.9. This explains the beneficial effect of the data pre-processing that was also reflected in an average improvement of classification accuracy of more than 10%, as well as more consistent performances across data from different groups of participants (based on age and gender), rectifying the impact of such inconsistencies in the data. Preliminary experiments on different groups' only showed around 2% of variations in performance. It is interesting to note that the proposed approach was tested on young adult subjects (19–30 years old). It is expected that older adults express their emotions differently from the young ones. For example, older adults can exhibit higher levels of subjective arousal in negative emotions and tenderness while young adults often show higher levels of physiological arousal in these emotions [41]. Moreover, a variability in the acquired physiological signals is foreseen due to ageing (e.g., for T [42]). Nevertheless, as the proposed ER framework is based on physiological signals

that are normalized and standardized in the resting state, it is anticipated that it could easily be adapted to the characteristics of the physiological signals that are captured from older-adults emotional expression, considering proper annotations and moderate alterations. Additionally, the data collected using Empatica E4 Wristband for P31 and P32 session turned out to be very noisy. Thus, we prelimenarily tested the model by both including and excluding P31 and P32 data, to verify the effect of noisy signals on the model. The resultant performances were similar with no noticeable differences, concluding that the proposed model can tolerate noisy input signals, given that they contain the correct information. This makes it very suitable when used in different environments, where data collected from wearable devices are expected to be noisy to a certain extent.

D. Applications

It is highly desirable for modern human-computer interaction systems to be able to detect emotional cues and synthesize appropriate emotional responses. This can be achieved by equipping machines with robust, accurate, and adaptable ER systems. Having intelligent systems that are emotionally aware can revolutionize various areas including e-health, smart learning, smart homes, online gaming, and neuromarketing. With the recent increase of attention towards the importance of mental health, more work is being dedicated to recognize people emotions in order to monitor the mental health state and be able to intervene when needed [5]. Therefore, an emotionally aware healthcare system should provide real time monitoring of patients' physical and mental states, allowing appropriate therapy and diagnosis to be offered accordingly. Additionally, the presented framework can be used for e-health applications such as using wearable devices and mobile applications for health and fitness self-monitoring [43]. The availability of such technology allows a continuous, uninterrupted collection of data, enabling frequent emotional assessment. Therefore, incorporating ER into health monitoring systems, along with the utilization of cloud computing, can enhance personalized medical assessment and preventive solutions [44].

Cognitive and behavioral therapies aims to help patients to cope with their mental health issues by successive imaginary, mediated, or in-vivo exposures, and the efficiency of such psychotherapy has been recognized for the past years. In this context, affective computing and ER techniques, specifically when using the arousal and valence model which covers a large spectrum of emotions and is widely used in psychology, can be used to improve the therapeutic process since automatic recognition of the arousal and valence components of affective reactions can provide significant information [45]. Further, in psychotherapy, ER offers great help to psychiatrists and patients in supporting early diagnosis of psychiatric diseases. Mid to long term emotional characterization can be used in psychiatry and in conjunction with medical trials. For patients with psychiatric diseases, ER offers the therapists with more insights into the daily variations of the patients mental and emotional state [46].

Learning can be enhanced through an educational system with emotionally aware human-computer interaction that uses the subject response to exercises, its personality, and emotions to adapt the study material and teaching velocity. Additionally, entertainment recommendation systems can utilize ER to obtain emotional responses to adapt music, movie, or TV series recommendation to the user's preferences. This can enhance the user's experience and thus improving the marketing aspect of the service provider. The importance of ER extends beyond human-computer interaction into fields such as psychology, where it could help psychologists identify patients who are unable to express their emotions. For example, ER could be beneficial for patients with autism spectrum disorder or patients diagnosed with the locked-in-syndrome. Moreover, there is increasing interest of utilizing artificial intelligence and its applications in the healthcare system, such as having a virtual avatar that is able read and adjust according to the patients' emotions, can improve their motivation for treatment. Also, this can lead to a faster and higher recovery success in rehabilitation, which enhance the quality of life. Additionally, emotions monitoring can allow establishing an individual profile, which helps to identify causes of depression, anxiety, stress, or chronic diseases, where it can either be shared with professionals or kept for self awareness [47].

Although the proposed framework showed very satisfactory performance in the experiments conducted, the physiological signals were recorded in the laboratory setting which lacks the real life scenarios when people experience emotional changes without even interacting with other people. Further, the results show that the system achieves a consistent performance in subject dependent tests, which is not always the case in subject independent experiments due to inter-personal variations. This issue is still a challenge faced by any of the existing ER methods. To further improve our system, it is required to have a larger sample size of participants in various daily life routines.

VII. Conclusion

A framework for ER on the arousal-valence space was presented. We primarily focused on using peripheral physiological signals as the source of emotion information. In this study, the data used were collected during a debate between pairs of participants where the emotions were rated by both parties mutually. Thus we investigated the use of annotations based on participants themselves, their partners, and combining both ratings for training the recognition system. Additionally, the conversion of the ratings to emotion classes was done based on a mid point following the literature, as well as a proposed dynamic thresholding, which allows us to effectively deal with individual differences in emotion rating tendencies. The proposed LSTM-based ER model was demonstrated. Additionally, TPE, a mechanism based on post processing of the past resultant scores of the classifier, was introduced. Experiments were conducted to demonstrate the performance of the proposed framework, where classification accuracies of up to 96.11% for arousal, 96.78% for valence, and 93.65% for four classes were achieved. Finally, future works should consider validating the findings of this work by using the new datasets possibly collected from a longitudinal, large-scale experience sampling method for an in-the-wild study of ER in the realistic scenario involving many users from different ages, educational level and cultural settings, while employing multi-task learning.

References

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

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

[3] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul.–Sep. 2019.

[4] M. Z. Soroush, K. Maghooli, S. Setarehdan, and A. M. Nasrabadi, "Emotion classification through nonlinear EEG analysis using machine learning methods," *Int. Clin. Neurosci. J.*, vol. 5, pp. 135–149, 2018.

[5] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, 2020.

[6] M. S. Zitouni, P. Lee, U. Lee, L. J. Hadjileontiadis, and A. Khandoker, "Privacy aware affective state recognition from visual data," *IEEE Access*, vol. 10, pp. 40620–40628, 2022.

[7] N. Ganapathy, R. Swaminathan, and T. M. Deserno, "Deep learning on 1-D biosignals: A taxonomy-based survey," *Yearbook Med. Informat.*, vol. 27, no. 1, pp. 98–109, 2018.

[8] J. Liao, Q. Zhong, Y. Zhu, and D. Cai, "Multimodal physiological signal emotion recognition based on convolutional recurrent neural network," *MSE*, vol. 782, no. 3, 2020, Art. no. 032005.

[9] B. H. Kim and S. Jo, "Deep physiological affect network for the recognition of human emotions," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 230–243, Apr.–Jun. 2020.

[10] M. S. Zitouni, C. Y. Park, U. Lee, L. Hadjileontiadis, and A. Khandoker, "Arousal-valence classification from peripheral physiological signals using long short-term memory networks," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, pp. 686–689.

[11] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[12] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 35–55, 2019.

[13] W. Xie and W. Xue, "WB-KNN for emotion recognition from physiological signals," *Optoelectron. Lett.*, vol. 17, no. 7, pp. 444–448, Jul. 2021.

[14] S. Aydin, S. Demirtaş, M. A. Tunga, and K. Ateş, "Comparison of hemispheric asymmetry measurements for emotional recordings from controls," *Neural Comput. Appl.*, vol. 30, no. 4, pp. 1341–1351, Aug. 2018.

[15] C. A. Frantzidis et al., "On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 309–318, Mar. 2010.

[16] M. B. H. Wiem and Z. Lachiri, "Emotion classification in arousal valence model using MAHNOB-HCI database," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, 2017, doi: 10.14569/IJACSA.2017.080344.

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

[18] J. Shukla, M. Barreda-Angeles, J. Oliver, G. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 857–869, Oct.–Dec. 2021.

[19] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr.–Jun. 2021.

[20] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dynamics in affective valence and arousal recognition," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 237–249, Apr.–Jun. 2012.

[21] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center Study Emotion Attention*, vol. 1, pp. 39–58, 1997.

[22] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.

[23] T. Zhang, A. E. Ali, A. Hanjalic, and P. Cesar, "Few-shot learning for fine-grained emotion recognition using physiological signals," *IEEE Trans. Multimedia*, early access, doi: 10.1109/TMM.2022.3165715.

[24] T. Zhang, A. E. Ali, C. Wang, A. Hanjalic, and P. Cesar, "Weakly-supervised learning for fine-grained emotion recognition using physiological signals," *IEEE Trans. Affect. Comput.*, early access, doi: 10.1109/TAFFC.2022.3158234.

[25] R. Elalamy, M. Fanourakis, and G. Chanel, "Multi-modal emotion recognition using recurrence plots and transfer learning on physiological signals," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2021, pp. 1–7.

[26] Y. Luo et al., "EEG-based emotion classification using spiking neural networks," *IEEE Access*, vol. 8, pp. 46007–46016, 2020.

[27] J. Luo, Y. Tian, H. Yu, Y. Chen, and M. Wu, "Semi-supervised cross-subject emotion recognition based on stacked denoising autoencoder architecture using a fusion of multi-modal physiological signals," *Entropy*, vol. 24, no. 5, 2022, Art. no. 577.

[28] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Emotion*, vol. 8, no. 10, pp. 355–358, 2017.

[29] O. Sourina and Y. Liu, "A fractal-based algorithm of emotion recognition from EEG using arousal-valence model," in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process.*, 2011, vol. 2, pp. 209–214.

[30] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 81–97, Jul.–Dec. 2010.

[31] P. C. Petrantonakis and L. J. Hadjileontiadis, "A novel emotion elicitation index using frontal brain asymmetry for enhanced EEG-based emotion recognition," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 5, pp. 737–746, Sep. 2011.

[32] S. Aydin, S. Demirtaş, and S. Yetkin, "Cortical correlations in wavelet domain for estimation of emotional dysfunctions," *Neural Comput. Appl.*, vol. 30, no. 4, pp. 1085–1094, Aug. 2018.

[33] A. Anderson, T. Hsiao, and V. Metsis, "Classification of emotional arousal during multimedia exposure," in *Proc. 10th Int. Conf. PErvasive Technol. Related Assistive Environments*, 2017, pp. 181–184.

[34] N. Saffaryazdi et al., "Emotion recognition in conversations using brain and physiological signals," in *Proc. 27th Int. Conf. Intell. User Interfaces*, Helsinki, Finland, 2022, pp. 229–242.

[35] C. Y. Park et al., "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Sci. Data*, vol. 7, no. 293, pp. 1–16, 2020.

[36] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, 2020, Art. no. 105524.

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[38] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological signals (TEAP)," *Front. ICT*, vol. 4, 2017, Art. no. 1.

[39] C. Y. Park, "PyTEAP, a python implementation of toolbox for emotion analysis using physiological signals (TEAP)." Accessed: Dec. 13, 2022. [Online]. Available: https://pypi.org/project/PyTEAP/

[40] K. Yang et al., "Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer," in *Proc. Int. Conf. Multimedia Retrieval*, 2022, pp. 562–570.

[41] L. Fernández-Aguilar et al., "Differences between young and older adults in physiological and subjective responses to emotion induction using films," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020.

[42] I. H. Gomolin, M. M. Aung, G. Wolf-Klein, and C. Auerbach, "Older is colder: Temperature range and variation in older people," *J. Amer. Geriatrics Soc.*, vol. 53, no. 12, pp. 2170–2172, 2005.

[43] S. A. Nasrat, U. Lee, M. S. Zitouni, A. H. Khandoker, S. Kang, and H. F. Jelinek, "Emotion recognition in the wild from long-term heart rate recording using wearable sensor and deep learning ensemble classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2021, pp. 1676–1678.

[44] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," *J. Med. Biol. Eng.*, vol. 40, pp. 149–157, 2020.

[45] B. Herbelin, P. Benzaki, F. Riquier, O. Renault, H. Grillon, and D. Thalmann, "Using physiological measures for emotional assessment: A computer-aided tool for cognitive and behavioural therapy," *Int. J. Disabil. Hum. Develop.*, vol. 4, no. 4, pp. 269–278, 2005.

[46] D. Tacconi et al., "Activity and emotion recognition to support early diagnosis of psychiatric diseases," in *Proc. 2nd Int. Conf. Pervasive Comput. Technol. Healthcare*, 2008, pp. 100–102.

[47] P. J. Bota, C. Wang, A. L. N. Fred, and H. P. D. Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140990–141020, 2019.