Received January 14, 2021, accepted January 27, 2021, date of publication February 1, 2021, date of current version February 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055933

Multi-Modal Physiological Data Fusion for Affect Estimation Using Deep Learning

MURTADHA D. HSSAYENI, (Graduate Student Member, IEEE), AND BEHNAZ GHORAANI, (Senior Member, IEEE)

Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

Corresponding author: Behnaz Ghoraani (bghoraani@fau.edu)

This work was supported by the National Science Foundation under Grant 1936586 and Grant 1942669.

ABSTRACT Automated momentary estimation of positive and negative affects (PA and NA), the basic sense of feeling, can play an essential role in detecting the early signs of mood disorders. Physiological wearable sensors and machine learning have a potential to make such automated and continuous measurements. However, the physiological signals' features that are associated with the subject-reported PA or NA may not be known. In this work, we use data-driven feature extraction based on deep learning to investigate the application of raw physiological signals for estimating PA and NA. Specifically, we propose two multi-modal data fusion methods with deep Convolutional Neural Networks. We use the proposed architecture to estimate PA and NA and also classify baseline, stress, and amusement emotions. The training and evaluation of the methods are performed using four physiological and one chest motion signal modalities collected using a chest sensing unit from 15 subjects. Overall, our proposed model performed better than traditional machine learning on hand-crafted features. Utilizing only two modalities, our proposed model estimated PA with a correlation of 0.69 ($p<0.05$) vs. 0.59 ($p<0.05$) with traditional machine learning. These correlations were 0.79 ($p<0.05$) vs. 0.73 ($p<0.05$) for NA estimation. The best emotion classification was achieved by the traditional method with 79% F1-score and 80% accuracy when all the four physiological modalities are used. This is while with only two modalities, the deep learning achieved 78% F1-score and 79% accuracy.

INDEX TERMS Affect estimation, convolutional neural networks, emotion recognition, multi-modal sensor data fusion, regression, stress detection.

I. INTRODUCTION

Affects are observed dynamic reactions caused by change in our emotions and are directly linked to our mental and physical health [1]. Specifically, the positive affects (PA) and negative affect (NA) give physicians valuable information about mood and anxiety disorders [1]. PA and NA are commonly measured using a self-report questionnaire known as the Positive and Negative Affect Schedule (PANAS) [2]. PANAS can provide a range of measurement timeframes. It can show people's momentary emotional states in addition to their general traits [3]. Automated momentary PA and NA using non-invasive and wearable devices equipped with biosensors [4] could provide a significant benefit over burdensome, self-report questionnaires and a possibility for tracking change in emotion or moods over time. Such information could have numerous applications. For example, in case of patients with Parkinson's disease, it can enable tracking anxiety fluctuation, which could help clinicians to administer individualized and effective medication management [5], [6]. It can enable tracking stress in caregivers, especially in case of caregivers managing stress with asthma, and identifying when a caregiver is required to see a psychiatrist to prevent further health complications [7].

Automated emotion detection using wearable devices have been explored [8]–[10]. However, the existing works are mostly focused on developing machine learning algorithms that use the physiological signals collected from wearable devices to classify emotions. These methods classify emotions into discrete categories, which are not able to quantify the intensity or range of emotions. As a result of missing such quantitative measurements, these emotion detection methods may not be used for tracking the amount of change in the affect scores or provide the inter-subject changes over time. Some other works estimate the severity levels of specific emotions such as stress [10], [11] and anxiety [12]. In this work, we estimated the PA and NA scores that show diverse emotions. Hence, there is a need for developing algorithms

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

for estimating the PA and NA scores from wearable devices that can provide quantitative measurements of the momentary changes in emotion. The contributions of this paper are twofold. (i) We develop a data-driven feature extraction and fusion machine learning method based on deep learning to explore multi-modal physiological signals for estimating PA and NA score. The rationale behind this approach is that a thorough investigation of data is needed as the characteristics of the physiological signals that are associated with the PA and NA scores are unknown and may be different from those that have been used for emotion detection. (ii) To our knowledge, we are the first to develop automated momentary PA and NA estimation algorithms based on physiological signals collected using wearable devices. Such a combined machine learning and sensing system could contribute to the development of a home-based system for tracking mental health.

II. RELATED WORK AND OBJECTIVES

Affects are usually reflected physiologically and physically on different parts of the body, such as the brain, heart, face, and skin [13]. Recent technologies for recording audio, video, and physiological signals can capture affect-related changes, such as heart or respiratory rate variability and muscle tension. Such signals were used in a machine leaning algorithm to objectively detect human emotions such as stress [13], [14]. In this paper, our focus is on estimating the PA and NA scores using wearable sensors with motion and physiological signals such as respiration (RESP), electrocardiogram (ECG), electromyogram (EMG), and electrodermal activity (EDA). Using the wearable sensors over videos provides an opportunity for continuous monitoring of the scores as the person does not require to engage with the system actively. More importantly, the physiological signals reflect the real emotion as controlled by the autonomic nervous system and are less likely to be faked (e.g., fake facial expressions or voice tones) [13], [14].

Emotion detection using physiological signals and machine learning models have been explored extensively. In this scenario, a model is trained to classify emotion into discrete categories such as angry, sad, happy, disgusted, afraid, surprised, and stressed [13], [15]. The majority of the detection models are based on traditional machine learning with hand-crafted features [15]–[17]. The fusion of multi-modal physiological signals [18] has shown to improve performance over ensembling multiple traditional models on a single modality [19]–[21].

Three main types of feature fusion include: early, intermediate, and late fusion [13]. The early fusion is performed before applying the classification models by concatenating the extracted features. This scheme allows for exploitation of the dependencies between different modalities as performed. The intermediate fusion considers the time dependencies between asynchronous modalities. In the late fusion, the outputs from the multiple models that are pretrained separately are combined to get one prediction by training a new model, majority voting or by averaging the probabilities for each class across the models then selecting the class with the highest probability. Early fusion has been adopted by Schmidt *et al.* [17], Oh *et al.* [14], and Bota *et al.* [18]. In Ref. [18], Bota *et al.* shows that early fusion is the most appropriate modality fusion for traditional emotion classifiers in terms of accuracy and computational complexity.

Emotion detection using deep learning models have also been investigated [14], [21]–[25]. These models have achieved higher performance than the traditional models for emotion recognition because of their ability to explore raw signals for data-driven features. The majority of these models were mostly based on a single modality, such as EEG [22], [24], [25] and RESP [23] without feature fusion. Other models used one type of late fusion that was applied on multi deep models using a single modality [21] with deep feature clustering [26], or multi modalities [14], but other types of fusion were not explored. In [14], only one combination of RESP and Heart Rate Variability (HRV) with late fusion was implemented. The different combinations of two or more modalities with deep learning have not been explored.

Our objective in this work was to develop a multi-modal, physiological data fusion framework using deep Convolutional Neural Networks (CNNs) to estimate PA and NA scores using movement and physiological signals collected from wearable devices. This framework was structured as a regression problem to develop two models: one for estimating the PA score and one for estimating the NA score. We investigated two different types of data fusion methods and considered the effect of different types of physiological signals in the estimation. Given that the multi-modal data fusion in deep learning models has not been explored for emotion recognition, we also investigated the application of our proposed framework for emotion detection. This scheme was framed as a classification problem with three classes: baseline, stress, and amusement emotions. For comparison purposes, the well-known gradient tree boosting method [27] was implemented to estimate the PA and NA scores and classify emotion using the features that were recommended previously for emotion recognition.

The paper is organized as follow: Section III describes the dataset used to validate the proposed methods. Section IV explains the preprocessing of the physiological signals, the tradition feature extraction for gradient tree boosting, and the deep learning method with feature fusion. Sections V-VI provides our results and discussion, and the paper is concluded in Section VII.

III. DATASET

The WESAD dataset [17] was used to validate the developed methods. The reason for using WESAD instead of using other datasets such as the DREAMER [28] and CASE [29] is that WESAD includes multi-modal data recorded during different affective states, especially during stress. It also provides PA and NA scores using the PANAS questionnaire, while PANAS was not performed during the other datasets' data

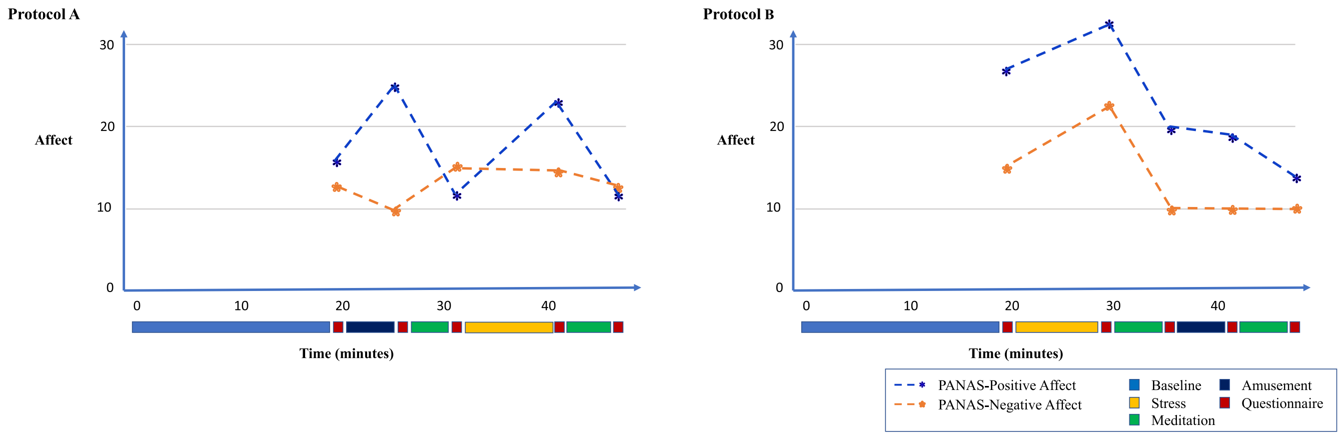


FIGURE 1. Two examples from the two different versions of the protocol showing the PA and NA scores obtained using the PANAS self-report questionnaire after each condition. On average, high scores of PA and NA scores were reported for the stress condition.

collection. The WESAD dataset was collected based on a protocol designed by Schmidt *et al.* for stress estimation [17]. The protocol was approved by the workers' council and the data security officer of their research center. The dataset consists of physiological and motion signals collected from 15 subjects (3 F, 12 M; average age 27.5±2.4 years).

Before starting the experiment, a wearable device was mounted on the subject's chest. The device was equipped with and connected to sensors to measure RESP, ECG, EMG, EDA, skin temperature (TEMP), and acceleration (ACC). The recording sampling rate was 700 Hz.

A brief description of the data collection protocol is as follows, and more details are explained in Ref. [17]. The experiment mainly consisted of three conditions: baseline, stress, and amusement. The stress and amusement conditions were followed by a guided meditation to de-excite the subjects. The order of the stress and amusement conditions was switched between subjects to reduce the order effect, resulting in two protocols, as shown in Figure 1. Half of the subjects performed the experiment while sitting, and the other half were standing to have different postures in each condition. During the **baseline**, the subjects were reading neutral magazines while standing/sitting for 20 minutes. During the **amusement** condition, the subjects were watching 11 funny videos with a 5-second gap between them. The total duration of this condition was about 6 minutes. During the **stress** condition, the Trier Social Stress Test (TST) was applied to induce stress [30]. In this test, the subjects first prepared and gave a personal, 5-minutes public speech to a panel of three people. Second, they were asked to count from 2023 to zero with decrements of 17 and asked to start over if they made a mistake. The total duration of the stress condition was about 10 minutes. To bring the subjects back to the neutral state, they were asked to meditate for about 7 minutes. During this time, they were comfortably sitting and following a breathing exercise with eyes closed.

After each condition, the subjects participated in a self-report PANAS questionnaire [2]. PANAS consists of 20 items to assess the positive (10 items) and negative (10 items) affects as independent dimensions. The ten positive items are active, interested, inspired, strong, excited, proud, enthusiastic, alert, determined, and attentive. The ten negative items are distressed, annoyed, guilty, scared, hostile, irritable, ashamed, nervous, jittery, and afraid. Each of these items has a score of 1 to 5; thus, the positive affect has a range between 10 (sadness and lethargy) to 50 (high energy and full concentration), and the negative affect has a range between 10 (calmness) to 50 (subjective distress).

We used data collected during baseline, stress, and amusement conditions for emotion classification. In addition to these data, for PA and NA estimation, we used the data collected using medication intervals. For the model developments, we did not explore the TEMP modality as it has shown to be irrelevant to emotion or Affect states [17].

IV. METHODS

In this section, first we describe the pre-processing methods applied to the raw signals. Next, we provide the hand-crafted feature extraction and machine learning method that were implemented for our comparison purposes. Finally, we provide the proposed deep model architecture for data-driven feature extraction and data fusion of the raw physiological signals. The methods are provided to perform PA and NA estimation and a 3-class emotion classification (baseline, stress, and amusement).

A. PRE-PROCESSING AND SEGMENTATION

Three steps were performed to preprocess and segment the signals. The first step excluded the data recorded during the transitions between the scripted conditions for which the gold-standard labels were not available. The next step filtered the signals using a series of Butterworth filter with an order five similar to the work of Schmidt *et al.* [17]. EDA signals were filtered using a 5Hz lowpass filter, EMG signals using a 0.5Hz highpass filter, ECG signals with a 0.5-45Hz bandpass filter, ACC signals with a 0.1-64Hz bandpass filter, and RESP signals were filtered using a 0.1-1Hz bandpass filter. The third step downsampled the signals by a factor of 2 and then segmented them into 1-minute windows. These selections were based on the work of Kreibig *et al.* and Schmidt *et al.* considering that emotions influence the autonomic nervous system in a period of seconds to few minutes [17], [31]. No overlapping was used between the segments, but instead, the signals in each segment were shifted and rotated as an augmentation method during training.

The 3-class classification methods used the defined conditions (baseline, stress, and amusement) as the gold-standard labels. The segmentation of the data in these conditions resulted in 535 segments ($W \in \mathbb{R}^{2100\times 7}$ where 2100 is the number of samples in each segment and 7 is the number of modalities) for all the subjects. The regression methods to estimate PA and NA used the PANAS scores from all the conditions (baseline, stress, amusement and meditation) as the gold-standard labels. The segmentation of the data in all four conditions resulted in 714 segments for all subjects. The 1-min segments in a condition share the same PANAS scores of that condition.

B. HAND-CRAFTED FEATURE EXTRACTION AND MACHINE LEARNING

In traditional machine learning, hand-crafted features are first extracted to reduce the data dimensionality and provide only the signal characteristics that are important to the application in hand. Next, the extracted features from each modality are fused in case of using multi modalities. Finally, a regression or a classification model is trained depending on the application.

1) FEATURE EXTRACTION

We extracted several features from each of the modalities following the work of Schmidt *et al.* [17] as explained in this section.

- *EDA*: Statistical features were first extracted from the 1-min windows (i.e., mean, standard deviation (STD), minimum, maximum, slope, range, and absolute integral). Next, EDA was separated into skin conductance response (SCR) as a short-term response and skin conductance level (SCL) as a long-term baseline conductivity [32]. Additional features were then extracted from SCR and SCL. These features include: the correlation between SCL and time, integration, duration, and magnitude of SCR segments within each window. A total of 15 features extracted from EDA ($F_{eda} \in \mathbb{R}^{15}$).
- *EMG*: Statistical features were calculated from the time domain signals. Additional features were then extracted from the spectral domain, such as the peak frequency, number, mean and STD of spectral peaks, sum of the peaks' amplitudes, and energy in seven evenly-spaced ranges. The total number of features extracted from EMG was 21 ($F_{emg} \in \mathbb{R}^{21}$).
- *ECG*: Heart rate (HR) and heart rate variability (HRV) were first calculated, and then used to extract the statistical features and HRV triangular index. Additional features were also extracted from the spectral domain, such as the energies in ultra-low, low, high, and ultra-high bands, and the ration between bands' energies. The total number of features extracted from ECG was 16 ($F_{ecg} \in \mathbb{R}^{16}$).
- *ACC*: Statistical features were derived in the time domain of each axis. Additional features were extracted from the spectral domain, such as the peak frequency. The total number of features extracted from ACC was 15 ($F_{acc} \in \mathbb{R}^{15}$).
- *RESP*: First, the inhalation and exhalation were obtained and then used to extract statistical features, stretch range, inhalation volume as the area under the inhalation curve, and breath rate. The total number of features extracted from RESP was 9 ($F_{resp} \in \mathbb{R}^{9}$).

2) FEATURE FUSION

Multi-modal fusion with traditional machine learning has been explored in the literature, early fusion was shown to be the most appropriate [18]. Therefore, in this paper, we used early fusion (i.e. $F_{fusion} \in \mathbb{R}^{T}$, where T is the sum of the number of features in the fused modalities) with gradient tree boosting.

3) GRADIENT TREE BOOSTING

The gradient tree boosting algorithm [27] was used to train two regression and one classification models by assigning an output $\hat{y}^{(w)}$ to a segment w with feature vectors $F^{(w)}_{fusion}$. In the regression models, the output is either the PA and NA score $\hat{y}^{(w)}_r$, and it is the emotion class $\hat{y}^{(w)}_c$ in case of the emotion classification model.

In this algorithm, an ensemble of N_t weak regression/classification trees ($\{f_i\}_{i=1}^{N_t}$) are used to estimate the output $\hat{y}^{(w)}$ according to the function in Eqn. (1).

$$\hat{y}^{(w)} = \sum_{i=1}^{N_t} f_i(F^{(w)}_{fusion}) \tag{1}$$

where $f_i(F^{(w)}_{fusion}) = W_{q(F^{(w)}_{fusion})}$ is the space of regression/classification tree i with L leaves, $q(F^{(w)}_{fusion})$ is the structure of the tree that maps $F^{(w)}_{fusion}$ to an index that represents the corresponding tree leaf, and $W \in \mathbb{R}^{L}$ is the leaf weights. Learning the regression/classification trees is performed using the additive training strategy. In this strategy, one tree is learned at each iteration by minimizing an objective function. This function includes the first and second gradient statistics of a loss function, which in case of regression, is based on the difference between the estimated PA and NA $\hat{y}^{(w)}_r$ and gold-standard $y^{(w)}_r$ score, and difference of the predicted emotion $\hat{y}^{(w)}_c$ and gold-standard $y^{(w)}_c$ labels in case of classification.

C. DEEP LEARNING

First, we describe our proposed CNN architecture and then describe the proposed feature fusion strategy.

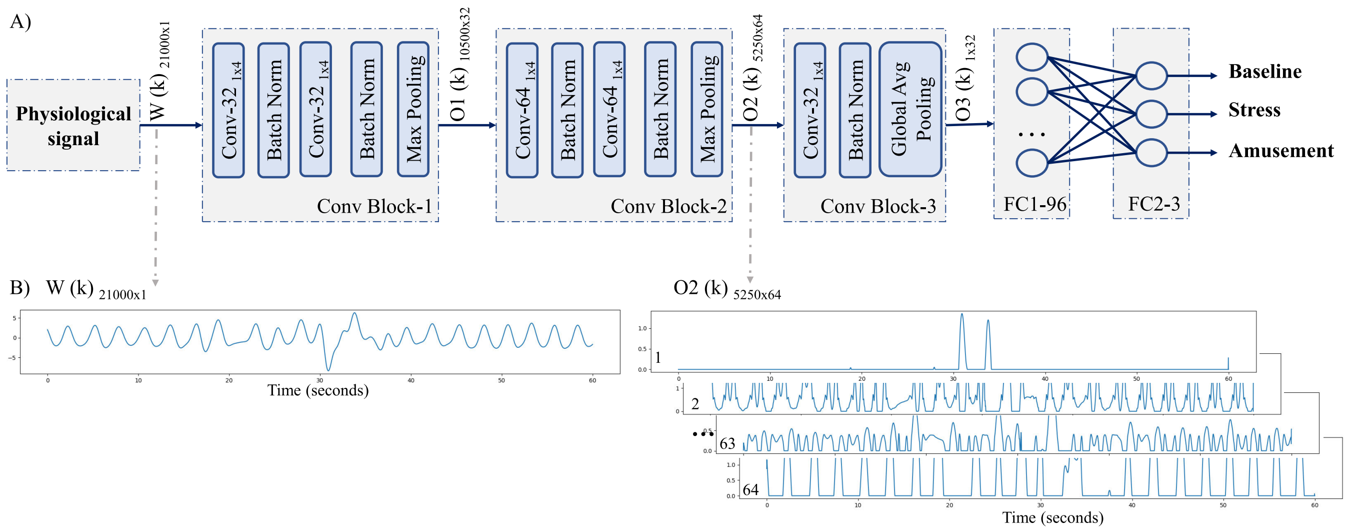


FIGURE 2. Part A shows a CNN architecture with five convolutional layers used for the 3-class emotion classification. For the quantification of PA and NA, the last fully-connected layer is modified to have only one node with a linear activation function. Part B shows an example of the physiological signals (respiration) and the corresponding feature maps from conv block-2 (O2). Note that the first filter in Block-2 has a high activation at around 30 seconds for the abnormal breathing cycle.

1) CONVOLUTIONAL NEURAL NETWORKS

Deep learning models, especially CNNs, have shown to be effective in a wide range of applications, including emotion recognition [14], [21]–[25]. CNNs are applied to the raw signals and can extract features driven by the data structure. Such a data-driven feature extraction is considered the main advantage of CNNs, especially in emotion recognition using physiological signals where the emotion-related signal features may not necessarily be pre-known or need significant priory knowledge. The proposed 1-D CNN on the raw signals is shown in Figure 2A. It consists of three convolutional blocks. The first block consists of two convolutional layers with 32 filters of width four, followed by batch normalization layers and a max-pooling layer. The second block has the same structure but deeper (64 filters). The third block has one convolutional filter, batch normalization layer, and global average pooling layer representing the bottleneck before the two fully-connected layers. Increasing the number of convolutional layers is done by repeating Conv Block-2 multiple times. The output layer has one node for the PA or NA regression and three nodes for the 3-class emotion classification.

Since we are using multi-modal data, extracting the emotion features from each modality may require a different CNN architecture to capture different patterns from these physiological signals. Such a behavior makes the early fusion impractical. Therefore, we choose to use the late fusion in this work as described in the next section. Each 1-D CNN is first trained for each modality, and the architecture is optimized based on the validation data. Next, different feature fusion methods are applied on the feature maps from the global-average pooling layer, or on the pretrained CNN models' outputs to estimate the PA and NA scores or classify the three emotions. Figure 2B shows an example of the input physiological signals to the CNN model and the feature maps from the second convolutional block.

2) FEATURE FUSION

Two types of late fusion methods are investigated. In *late fusion 1*, a separate CNN is trained for each modality. The extracted feature maps from Conv Block-3 (shown in Figure 2A) of the pretrained models are fused to form one feature vector of size M*32 ($F_{fusion} \in \mathbb{R}^{M*32}$), where M is the number of used modalities. Lastly, a fully-connected network is trained with these feature vectors being the input. For the regression problem, the fully-connected network's output has one node representing either the PA or NA score. In the classification problem, the output has three nodes representing the final class of baseline, stress, and amusement. In *late fusion 2*, the average of the three classes' probabilities across the pretrained CNNs is calculated, and the emotion class with the highest average probability is selected. In the case of the PA or NA estimation, the average of the estimated scores across the pretrained CNNs is provided as the estimated affect score.

V. RESULTS

The training and testing of the proposed methods were performed using the WESAD dataset and with subject-based, leave-one-out cross-validation (LOOCV). A validation set with 20% of the training data was used to optimize the models' hyperparameters by minimizing the validation loss. Using such cross-validation strategy, we implemented hundreds of gradient tree boosting and CNN models to design three models: one to estimate the PA score, one to estimate the NA score, and one to classify emotion to baseline vs. stress vs. amusement.

The gradient tree boosting algorithm was implemented using XGboost library [27]. The learning rate was 0.1. A grid search was applied to find the optimal number of regression/classification trees in the range of 10 to 200 with a step of 20. The tree depth was in the range of 3 to 10 with a step of 2. The percentage of used-features per tree was in the

TABLE 1. The performance of the proposed methods in estimating the positive and negative affects using single or multiple modalities with CNN and gradient tree boosting. Chest physio stands for all the physiological signals: ECG, EMG, EDA, and RESP. The best MAE and correlation when using a single or multi modalities are in bold for each method. †indicates a non-significant correlation (p>0.05).

Modalities Used		Positive Affect				Negative Affect			
		CNN		Gradient tree boosting		CNN		Gradient tree boosting	
		MAE	*r*	MAE	*r*	MAE	*r*	MAE	*r*
ECG		6.70	0.59	5.70	0.40	5.00	0.54†	3.77	0.58†
EMG		6.03	0.63	5.53	0.45	3.52	0.64	4.39	0.11†
EDA		6.64	0.29†	5.75	0.28†	3.84	0.48	4.05	0.37†
RESP		**5.45**	**0.64**	**5.31**	**0.58**	3.11	0.70	**3.04**	**0.73**
ACC		5.99	0.53†	5.48	0.47	**3.41**	**0.78**	3.49	0.59
RESP+ECG	**Fusion 1**	7.91	0.63	5.36	0.58	4.29	0.62	3.13	0.72†
	Fusion 2	6.09	0.66			3.38	0.67†		
RESP+EMG	**Fusion 1**	6.32	0.67	**5.15**	**0.59**	3.48	0.74	3.39	0.66
	Fusion 2	**5.15**	**0.69**			**3.04**	0.74		
RESP+EDA	**Fusion 1**	6.57	0.58	5.18	0.58	3.40	0.72	3.43	0.68
	Fusion 2	5.63	0.55			3.27	0.74		
RESP+ACC	**Fusion 1**	5.80	0.57	5.38	0.58	3.08	0.74	**3.14**	**0.73**
	Fusion 2	5.49	0.65			**3.04**	**0.79**		
Chest Physio	**Fusion 1**	8.04	0.61	5.18	0.58	3.92	0.69	3.43	0.68
	Fusion 2	5.38	0.68			3.22	0.76		

range of 10% to 50% with a step of 10%. The proposed CNN architectures (Figure 2A) was implemented in Python using Keras with TensorFlow backend [33]. During the training, the depth of the CNNs was increased by repeating Conv Block-2 up to four times, and the best performing model on the validation data was evaluated on the held-out test data.

A. ESTIMATION OF THE POSITIVE AND NEGATIVE AFFECTS

We used the proposed regression models to estimate the PA and NA scores. The two evaluation metrics were the mean absolute error (MAE) and the Pearson correlation (r) between the estimated PA or NA scores from the models and the gold-standard ones from the PANAS questionnaire. The regression models were trained using the mean squared error loss.

CNNs with different depths were trained in this work for each modality or a combination of modalities to quantify PA and NA. These CNNs had the same architecture in Figure 2A and same number of fully-connected layers but might have different number of blocks (i.e. different depth) based on the optimization performed on the validation data. The output layer had only one node with a linear activation function. The gradient tree boosting algorithm was used in the same fashion described previously to train regression trees to estimate PA and NA.

Table 1 provides the results of the proposed positive and negative affects estimation models using a single modality or a combination of two or all the physiological modalities. When fusing two modalities, RESP was used in all of them as recommended for emotion detection in the work of Schmidt *et al.* [17] as features extracted from RESP have the highest importance. In addition, Supplement Table 1 shows the PA and NA estimation results for the other possible combinations of two modalities without using RESP. Combining the different modalities with CNN was performed using two late fusion methods described in Section IV-C2, and using early fusion with gradient tree boosting.

1) CNN RESULTS

a: SINGLE MODALITY

First, we explored the application of the CNN model with a single modality signal. The proposed CNN achieved the best correlation of 0.64 ($p<0.05$) with 5.45 MAE using RESP to estimate NA, and the best correlation of 0.78 ($p<0.05$) with 3.41 MAE using ACC to estimate PA. The good performance of ACC is mostly due to its ability to capture physical changes such as trembling sensations and shaking caused by anxiety or stress.

b: MULTI MODALITIES

We explored the application of the proposed CNN and fusion methods with different combination of the physiological signals. Regarding the PA estimation, the best performance of both fusion methods was achieved using RESP+EMG, while late fusion 2 resulted in a slightly better performance with r=0.69 ($p<0.05$) and 5.15 MAE when compared to r=0.67 ($p<0.05$) and 6.32 MAE using late fusion 1. Regarding the NA estimation, the best performance of both fusion methods was achieved using RESP+ACC. We observed a similar behavior in terms of the fusion methods.

Late fusion 2 resulted in slightly better performance (r=0.79 (p<0.05) and 3.04 MAE) than fusion 1 with r=0.74 (p<0.05) and 3.08 MAE. A similar MAE but a slightly lower correlation (r=0.74) was achieved using late fusion 2 of RESP+EMG. As shown in the Supplement Table 1, excluding RESP resulted in lower performance for PA and NA estimation.

2) COMPARISON WITH GRADIENT TREE BOOSTING

Based on a *single modality* estimation, the best performance of the implemented tree boosting for PA was r=0.58 (p<0.05) and 5.31 MAE using RESP, and r=0.73 (p<0.05) and 3.04 MAE for estimating NA using RESP, which were both lower than the best performance of the proposed CNN models. As can be seen in Table 1, the proposed CNN models achieved a higher correlation than gradient tree boosting in estimating PA and NA, when the correlation was significant, using all the modalities except for RESP when estimating NA. However, the MAE of CNN was higher than the gradient tree boosting for some modalities. In case of the *multi-modality* estimation, the best performance of the tree boosting for PA was r=0.59 (p<0.05) and 5.15 MAE using RESP+EMG, and r=0.73 (p<0.05) and 3.14 MAE for estimating NA using RESP+ACC. However, their performance was significantly lower than the performance of the proposed CNN models when using the same modalities.

B. EMOTION CLASSIFICATION

We evaluated the proposed classification methods based on two evaluation metrics: the accuracy and the average F1-score of the three emotions of baseline, stress, and amusement. It is essential to consider both of these metrics. For example, in our dataset, baseline class is in the majority, so it is important to increase the accuracy and the average F1-score to ensure that the trained model is not biased towards the accurate classification of the baseline class. The implemented models were trained using categorical cross-entropy loss.

Table 2 provides the results of the three-class classification methods using a single modality or a combination of two or all physiological modalities with CNN and gradient tree boosting. Similar to the PA and NA estimation models, when fusing two modalities, RESP was used in all of them as it shows to have the highest F1-score as a single model in the work of Schmidt *et al.* [17]. In addition, Supplement Table 2 shows the three-class classification results for the other possible combinations of two modalities without using RESP. Combining the different modalities was performed using two late fusion methods with CNN described before in Section IV-C2, and using early fusion with gradient tree boosting. This table also provides the highest performance of hand-crafted features with AdaBoost decision trees reported in [17].

1) CNN RESULTS

Single modality: The proposed CNN using RESP resulted in the best F1-score of 73% and accuracy of 77%. In case of using *multi modalities*, the best performance of 78% F1-score and 79% accuracy was achieved using RESP+EDA and late fusion 1, followed by RESP+ACC with the same F1-score but a slightly lower accuracy, again using late fusion 1. Overall, late fusion 1 resulted in a higher F1-score when compared to fusion 2 using the same signals, while late fusion 2 resulted in a better classification accuracy. When checking the confusion matrix, we noticed that late fusion 1 was less biased towards the baseline class and more sensitive to the amusement class, which could be due to training a new fully-connected layer in this fusion approach. Excluding RESP resulted in a lower performance for emotion classification as shown in the Supplement Table 2.

2) COMPARISON WITH GRADIENT TREE BOOSTING

The best *single modality* detection resulted using RESP with 73% F1-score and 75% accuracy, which was the same F1-score of CNN with RESP but slightly lower accuracy. The worst performance was 49% F1-score and 50% accuracy using EMG, which was much lower that the CNN performance with 61% F1-score and accuracy. The best *multi-modality* performance was achieved using all the modalities with 79% F1-score and 80% accuracy. This is while using only RESP+EDA, CNN achieved 78% F1-score and 79% accuracy.

VI. DISCUSSION

Detection of emotion using physiological signals and machine learning has been commonly investigated in the literature [8]–[10]; however, estimation of PA and NA scores has not been explored yet. This is while automated momentary estimation of PA and NA scores could help with long-term monitoring of mental health and early detection of mood-related diseases. We performed a detailed analysis of CNN and two data fusion strategies for estimating PA and NA scores and emotion detection. We investigated the effect of different combination of physiological signals and compared the results to hand-crafted feature extraction and traditional machine leaning. A summary of the best methods for each signal combination is presented in Table 3.

A. BEST PERFORMANCE

As Table 3 indicates, the proposed CNN and fusion method 2 on RESP+EMG modalities provided the best PA estimation with r=0.69 and MAE=5.15. The same model with RESP+ACC modalities provided the best NA estimation with r=0.79 and MAE=3.04. The hand-crafted features with gradient tree boosting resulted in the best emotion detection of 79% F1-score and 80% accuracy using all the chest physiological modalities. The good performance of gradient tree boosting when using all the physiological signals could be because of its internal feature selection process. However, if only two modalities are considered, the best performance is achieved using CNN and RESP+EMG for PA estimation (r=0.69), CNN and RESP+ACC for NA estimation (r=0.79), and for emotion detection, it is achieved with

TABLE 2. The performance of the proposed methods in classifying baseline, stress, and amusement in comparison with the results of AdaBoost Decision Trees reported in [17] using single or multi modalities. The best F1-score and accuracy when using a single or multi modalities are in bold for each method.

Modalities Used		CNN		Gradient tree boosting		AdaBoost decision trees [17]	
		F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
ECG		55	60	63	66	52	62
EMG		61	61	49	50	38	48
EDA		64	69	59	59	48	54
RESP		**73**	**77**	**73**	**75**	**62**	**72**
ACC		71	77	64	66	44	57
RESP+ECG	Fusion 1	68	68	71	72	-	-
	Fusion 2	65	72				
RESP+EMG	Fusion 1	73	73	78	79	-	-
	Fusion 2	72	77				
RESP+EDA	Fusion 1	**78**	**79**	70	72	-	-
	Fusion 2	75	81				
RESP+ACC	Fusion 1	78	78	73	75	-	-
	Fusion 2	73	81				
Chest Physio	Fusion 1	77	77	**79**	80	**73**	**81**
	Fusion 2	72	79				

TABLE 3. A summary of the best methods used for emotion classification and affect estimation for each of the modalities and their fusion. The best performance of affect estimation and emotion classification is in bold. †indicates a non-significant correlation ($p > 0.05$).

Modalities used	Positive affect estimation			Negative affect estimation			Emotion classificatoin		
	Best method	MAE	r	Best method	MAE	r	Best method	F1-score	Accuracy
ECG	CNN	6.70	0.59	Gradient tree boosting	3.77	0.58†	Gradient tree boosting	63	66
EMG	CNN	6.03	0.63	CNN	3.52	0.64	CNN	61	61
EDA	CNN	6.64	0.29†	CNN	3.84	0.48	CNN	64	69
RESP	CNN	5.45	0.64	Gradient tree boosting	3.04	0.73	CNN	73	77
ACC	CNN	5.99	0.53†	CNN	3.41	0.78	CNN	71	77
RESP+ECG	CNN-Fusion 2	6.09	0.66	Gradient tree boosting	3.13	0.72†	Gradient tree boosting	71	72
RESP+EMG	CNN-Fusion 2	**5.15**	**0.69**	CNN-Fusion 2	3.04	0.74	Gradient tree boosting	78	79
RESP+EDA	Gradient tree boosting CNN-Fusion 1	5.18	0.58	CNN-Fusion 2	3.27	0.74	CNN-Fusion 1	78	79
RESP+ACC	CNN-Fusion 2	5.49	0.65	CNN-Fusion 2	**3.04**	**0.79**	CNN-Fusion 1	78	78
Chest Physio	CNN-Fusion 2	5.38	0.68	CNN-Fusion 2	3.22	0.76	Gradient tree boosting	**79**	**80**

either CNN and RESP+EDA or gradient tree boosting and RESP+EMG (F1-score = 78% and accuracy=79%).

B. DEEP LEARNING VS. TRADITIONAL MACHINE LEARNING

As we expected, CNN provides an ability for the model to investigate raw signals for extracting physiological signals' features that are associated with the subject-reported PA or NA. Regardless of the signal modality, the performance of the proposed CNN models was higher than or equal (in one case) when compared to the traditional machine learning for estimating the PA score. In case of NA estimation, the majority of the modalities provided the best performance when the proposed CNN model was used (seven vs. three cases). A similar behavior was observed in case of emotion detection where in six out of ten modality combinations, the CNN models resulted in the best performance.

C. DATA FUSION

The performance of the two explored fusion methods was different between emotion classification and affect estimation. In PA and NA estimation, late fusion 2 resulted in the better correlation and lower MAE than the late fusion 1 (see Table 1). The reason for the lower performance of the late fusion 1 could be that it increases the dimensionality of the training data when concatenating the feature maps without increasing the number of training windows. As a result, the model leads to the overfitting of the fully-connected network and poor generalization. In case of emotion classification, method 1 of late fusion is better than method 2. The main reason is that the number of windows are imbalanced in the three conditions. Therefore, the trained CNNs using the late fusion 2 are biased towards the majority class, resulting in an overall lower performance. However, in case of affect estimation, the PA and NA scores do not depend on the condition labels, and the data from the meditation intervals are also included in the model development process. Therefore, the distribution of the affect scores is not affected by the imbalanced condition labels. We expect that the late fusion 2 for classification will improve if each model is trained on a balanced training data.

D. MODALITY SELECTION

Minimizing the number of physiological signals reduces the need for collecting and processing redundant signals. As a result, the size of the wearable device required to collect the signals will be reduced in addition to the power consumption during data collection, processing, and affect estimation using deep models. One main observation is that in case of single modality, RESP performs consistently better for PA and NA estimation and emotion detection. In case of multi-modality, the fusion of RESP+EMG or RESP+ACC has a good and consistent performance for all the three applications. Another observation is that some modalities (i.e., EMG, EDA, and ACC) perform well with CNN but not with the traditional methods. One main reason is that the hand-crafted feature extraction may not be optimal for those modalities while a data-driven model such as CNN can improve the performance by offering data-driven feature extraction from the raw signals. Such a rationale can be even further supported by observing that the gradient tree boosting resulted in equal or slightly better performance than the CNN models when using the well-known physiological signals such as ECG and RESP, where the change due to change in emotion has been well-studied.

E. COMPARISON TO THE EXISTING WORK

To the best of our knowledge, this is the first work for automated estimation of PA and NA scores using physiological signals. The reports closest to our proposed work are the continuous estimation of stress [10], [11] and anxiety [12] using wearable sensors. These reports showed a lower correlation when compared to our proposed NA estimation algorithm using CNN when fusing RESP+ACC (r=0.79) and our PA estimation using CNN with RESP+EMG (r=0.69), except for the work by Plarre *et al.* [11] that reported a slightly higher correlation. Plarre *et al.* estimated the stress levels with r=0.72 using a traditional method trained on the hand-crafted features from the ECG and RESP signals recorded using a chest sensor. Siirtola and Röning [10] estimated stress continuously while driving using a wrist sensor containing ACC, EDA, blood volume pulse (BVP), and TEMP. They specified a personalized threshold on stress levels and evaluated the model with balanced classification accuracy. They report an 82% accuracy and a mean r^2 of 0.35. In another work, Nirjhar *et al.* developed a temporal parametric model to estimate public speaking anxiety using speech signals and physiological signals from EDA and photoplethysmography signals [12]. The highest reported correlation was 0.37 ($p<0.05$).

For the three-class classification, AdaBoost decision trees is used by Schmidt *et al.* to detect three emotions on the WESAD dataset that is used in our work [17]. The best performance as reported in [17] is provided in the last column of Table 2. Schmidt *et al.* report the best performance using a single modality with RESP and F1-score=62%. When using all the four physiological signals, the performance improves to F1-score=73%. However, the results reported by Schmidt *et al.* in either the single or multimodality case are lower than the performance of our proposed CNN models (78% F1-score) and the gradient tree boosting algorithm (79% F1-score). In addition, the other methods that are evaluated on the WESAD dataset with LOOCV reported lower F1-scores [34], [35]. Tervonen *et al.* [34] proposed a binary stress classification using the self-organizing map with different personalized levels. F1-score of their semi-personal model evaluated using LOOCV on a field dataset was 62%. The same method achieved an F1-score of 37% on the WESAD dataset when using the chest sensor. F1-score of their personalized model evaluated without LOOCV on a field dataset was 54% and 89% on the WESAD dataset. Bobade and Vani [35] proposed an artificial neural network (ANN) for 3-class stress classifier using the wrist and chest sensors. They reported an F1- score of 78%, which was slightly lower than the performance of Gradient Tree boosting and equal to the F1-score of the CNN model with the advantage in our proposed model when using only a chest sensor with two modalities.

VII. CONCLUSION

We developed algorithms to enable automated momentary estimation of PA and NA scores based on the physiological signals recorded using wearable. Specifically, we developed two late multi-modal data fusion methods with deep CNN models to estimate PA and NA scores and explored the effect of the two late fusion methods and different combination of the physiological signals on the performance. We repeated the experiments for the detection of three emotions of baseline vs. stress vs. amusement emotions. For the evaluation purposes, we also implemented hand-crafted features and gradient tree boosting and compared the results on the physiological and motion signals collected using a chest sensing unit from the WESAD dataset. The best affect estimation was obtained when the estimated scores of the pretrained respiration and chest acceleration CNNs were averaged, which resulted in an average, leave-one-out correlation of 0.65 and 0.79 for PA and NA scores, respectively. The same combination with CNN resulted in the best emotion classification with 78% F1-score. The CNN models provided better performance than gradient tree boosting for affect estimation but performed equally well for emotion detection suggesting that the prior knowledge about the changes in the physiological signals with changes of emotion enhanced the quality of the hand-crafted features and the performance of the gradient tree boosting model. However, as our work indicated, a data-driven feature extraction using CNN is more suitable for affect estimation where the physiological signal characteristics are unknown. Such methods could estimate the momentarily PA and NA over time without the need for frequently filling out burdensome, self-reported questionnaires. Our future work includes collecting additional training data to address the deep models' need for a large-size training set. We expect that increasing the data's size will improve the developed algorithms' performance for the NA and PA estimation and the emotion classification.

REFERENCES

[1] A. Díaz-García, A. González-Robles, S. Mor, A. Mira, S. Quero, A. García-Palacios, R. M. Baños, and C. Botella, "Positive and negative affect schedule (PANAS): Psychometric properties of the online spanish version in a clinical sample with emotional disorders," *BMC Psychiatry*, vol. 20, no. 1, p. 56, Dec. 2020.

[2] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The panas scales," *J. Personality Social Psychol.*, vol. 54, no. 6, p. 1063, 1988.

[3] G. J. Boyle, E. Helmes, G. Matthews, and C. E. Izard, "Measures of affect dimensions," in *Measures Personality Social Psychology Constructs*. Amsterdam, The Netherlands: Elsevier, 2015, pp. 190–224.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[5] I. H. Richard, S. Frank, M. P. McDermott, H. Wang, A. W. Justus, K. A. LaDonna, and R. and Kurlan, "The ups and downs of parkinson disease: A prospective study of mood and anxiety fluctuations," *Cognit. Behav. Neurol.*, vol. 17, no. 4, pp. 201–207, 2004.

[6] J. Habets, M. Heijmans, C. Herff, C. Simons, A. F. Leentjens, Y. Temel, M. Kuijf, and P. Kubben, "Mobile health daily life monitoring for parkinson disease: Development and validation of ecological momentary assessments," *JMIR mHealth uHealth*, vol. 8, no. 5, May 2020, Art. no. e15628.

[7] M. H. Bellin, J. Kub, K. D. Frick, M. E. Bollinger, M. Tsoukleris, J. Walker, C. Land, and A. M. Butz, "Stress and quality of life in caregivers of inner-city minority children with poorly controlled asthma," *J. Pediatric Health Care*, vol. 27, no. 2, pp. 127–134, Mar. 2013.

[8] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "Emotion recognition involving physiological and speech signals: A comprehensive review," in *Recent Advances in Nonlinear Dynamics and Synchronization*. Cham, Switzerland: Springer, 2018, pp. 287–302, doi: 10.1007/978-3-319-58996-1_13.

[9] P. J. Bota, C. Wang, A. L. N. Fred, and H. Placido Da 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.

[10] P. Siirtola and J. Röning, "Comparison of regression and classification models for user-independent and personal stress detection," *Sensors*, vol. 20, no. 16, p. 4402, Aug. 2020.

[11] K. Plarre, A. Raij, S. Hossain, and A. Ali, "Continuous inference of psychological stress from sensory measurements collected in the natural environment," in *Proc. 10th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, 2011, pp. 97–108.

[12] E. H. Nirjhar, A. Behzadan, and T. Chaspari, "Exploring bio-behavioral signal trajectories of state anxiety during public speaking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1294–1298.

[13] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, Jun. 2018.

[14] S. Oh, J.-Y. Lee, and D. K. Kim, "The design of CNN architectures for optimal six basic emotion classification using multiple physiological signals," *Sensors*, vol. 20, no. 3, p. 866, Feb. 2020.

[15] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, Jan. 2020.

[16] J. Wijsman, B. Grundlehner, H. Liu, H. Hermens, and J. Penders, "Towards mental stress detection using wearable physiological sensors," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 526–532.

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

[18] P. Bota, C. Wang, A. Fred, and H. Silva, "Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet?" *Sensors*, vol. 20, no. 17, p. 4723, Aug. 2020.

[19] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2016, pp. 521–529.

[20] Lin, Lin, Wang, and Wu, "Multiple convolutional neural networks fusion using improved fuzzy integral for facial emotion recognition," *Appl. Sci.*, vol. 9, no. 13, p. 2593, Jun. 2019.

[21] Z. Ahmad and N. M. Khan, "Multi-level stress assessment using multi-domain fusion of ECG signal," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 4518–4521.

[22] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using deep belief networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

[23] Q. Zhang, X. Chen, Q. Zhan, T. Yang, and S. Xia, "Respiration-based emotion recognition with deep learning," *Comput. Ind.*, vols. 92–93, pp. 84–90, Nov. 2017.

[24] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul. 2020.

[25] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Jun. 14, 2019, doi: 10.1109/TAFFC.2019.2922912.

[26] M. A. Asghar, M. J. Khan, M. Rizwan, R. M. Mehmood, and S.-H. Kim, "An innovative multi-model neural network approach for feature selection in emotion recognition using deep feature clustering," *Sensors*, vol. 20, no. 13, p. 3765, Jul. 2020.

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

[28] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

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

[30] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'Trier Social Stress Test'—A tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, nos. 1–2, pp. 76–81, 1993.

[31] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, Jul. 2010.

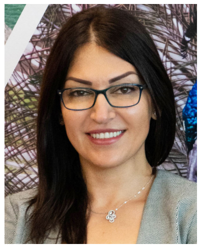
[32] A. Greco, G. Valenza, A. Lanata, E. Pasquale Scilingo, and L. Citi, "CvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2016.

[33] M. Abadi. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: http://tensorflow.org/

[34] J. Tervonen, S. Puttonen, M. J. Sillanpää, L. Hopsu, Z. Homorodi, J. Keränen, J. Pajukanta, A. Tolonen, A. Lämsä, and J. Mäntyjärvi, "Personalized mental stress detection with self-organizing map: From laboratory to the field," *Comput. Biol. Med.*, vol. 124, Sep. 2020, Art. no. 103935.

[35] P. Bobade and M. Vani, "Stress detection with machine learning and deep learning using multimodal physiological data," in *Proc. 2nd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2020, pp. 51–57.

MURTADHA D. HSSAYENI (Graduate Student Member, IEEE) was born in Babil, Iraq, in 1990. He received the B.S. degree in computer engineering from the University of Technology, Baghdad, Iraq, in 2012, and the M.S. degree in computer engineering from the Rochester Institute of Technology, Rochester, NY, USA, in 2017. He is currently pursuing the Ph.D. degree in computer engineering with Florida Atlantic University, Boca Raton, FL, USA. From 2015 to 2017, he was a Research Assistant with the Biomedical Signal and Image Analysis Laboratory. From 2017 to 2018, he was a Lecturer with the University of Technology. Since 2018, he has been a Research Assistant with Florida Atlantic University. His research interests include biomedical signal processing, biomedical image processing, tensor decomposition, machine learning, and deep learning.



BEHNAZ GHORAANI (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Ryerson University, Toronto, Canada, in 2010. From 2010 to 2012, she held a postdoctoral position with the College of Medicine, University of Toronto. She was an Assistant Professor with the Department of Biomedical Engineering, Rochester Institute of Technology, from 2012 to 2016. She is currently an Associate Professor with the Department of Computer and Electrical Engineering, Florida Atlantic University (FAU). Her research interests include generating clinically relevant engineering solutions to tackle significant bottlenecks in data analytics with an emphasis on computer-aided clinical decision making, long-term and continuous health monitoring, remote and personalized therapeutic management, non-stationary and multidimensional signal analysis, adaptive signal feature extraction, traditional and deep learning, and machine learning.