IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received May 13, 2022, accepted May 23, 2022, date of publication May 27, 2022, date of current version June 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3178409

Joint Modality Features in Frequency Domain for Stress Detection

K. RADHIKA[1], RAMANATHAN SUBRAMANIAN[2], (Senior Member, IEEE), AND VENKATA RAMANA MURTHY ORUGANTI[1], (Member, IEEE)
[1]Department of Electrical and Electronics Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Tamil Nadu 641112, India
[2]Faculty of Science and Technology, University of Canberra, Bruce, ACT 2617, Australia

Corresponding author: Venkata Ramana Murthy Oruganti (ovr_murthy@cb.amrita.edu)

ABSTRACT Rich feature extraction is essential to train a good machine learning (ML) framework. These features are generally extracted separately from each modality. We hypothesize that richer features can be learned when modalities are jointly explored. These joint modality features can perform better than those extracted from individual modalities. We study two modalities, physiological signals–Electrodermal activity (EDA) and electrocardiogram (ECG) to investigate this hypothesis. We investigate our hypothesis to achieve three objectives for subject-independent stress detection. For the first time in the literature, we apply our proposed framework in the frequency domain. The frequency-domain decomposition of the signal effectively separates it into periodic and aperiodic components. We can correlate their behaviour by focusing on each band of the signal spectrum. Second, we show that our framework outperforms late fusion, early fusion and other notable works in the field. Finally, we validate our approach on four benchmark datasets to show its generalization ability.

INDEX TERMS Stress detection, ECG, EDA, frequency band, auto-encoder, CRNN, SE module.

I. INTRODUCTION

Stress is defined as the nervous system's reaction to a danger or an instruction [1]. Stress has been taken seriously in recent years as it affects many people. This tendency could be due to changing work styles, cultural demands, varying lifestyles, etc. [2]. In some circumstances, stress can be beneficial up to a point in high-pressure situations such as at work, exams, and so kind. Stress is no longer beneficial once it crosses a certain level; it also harms an individual's emotional state, health, quality of life, and productivity [3]. If certain events occur frequently and a person becomes highly concerned, the body will be stressed for the rest of the time, leading to severe health issues [4]. As a result, the importance of stress detection systems has grown compared to the situation that existed a decade ago. Protecting individuals from the growing effects of stress is critical, mainly because stress is unavoidable. As a result, timely stress diagnosis and control are crucial for improving an individual's mental health and overall well-being [5].

Automatic stress detection mainly uses three modalities: psychological, physiological, and behavioral [6]. The Hypothalamic Pituitary Adrenal (HPA) axis and the Autonomic Nervous System (ANS) are the two key components that respond to stress by attempting to restore physiological balance [7]. This is caused by changes in heart activity, sweat gland activity, skin temperature, etc. As effective stress markers, physiological signals can thus provide information on ANS activity. In addition, among the physiological signals, ECG and EDA provide a realistic view of an individual's stress level [8].

The frequency-domain analysis of physiological signals has received less attention than the time-domain analysis. The signal's transitory properties can be used to comprehend the signal's frequency-domain interpretation [9]. Frequency-domain analysis for stress detection has received little attention. When looking for periodic behavior in a signal, frequency-domain analysis comes most in handy. [10]. This paper describes a joint modality feature learning method for stress detection in the frequency domain. The proposed method uses a deep neural network to learn joint-modal mapping. The ECG and EDA frequency bands are identified, and features are extracted from the PSD. These features are used for joint modality feature learning.

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

This study differs from earlier works in the following aspects. Most physiological signal-based stress detection studies used time-domain and time-frequency-domain features. Frequency-domain analysis, despite its importance, receives less attention than time-domain analysis. As a result, we incorporate joint modality feature learning in the frequency domain for stress detection in this study. We use autoencoders to learn joint representation from different modality features. ECG and EDA's frequency bands, which contribute the highest to stress detection, are also evaluated.

The main contributions of this work are summarized as follows:

1) Frequency domain analysis is performed on ECG and EDA signals. The frequency bands of the ECG and EDA have been identified. We analyze the performance of each frequency band of ECG and EDA separately to identify the band that performs best for stress detection.
2) The ECG and EDA signals are divided into fixed duration segments of varying lengths. The above-developed frequency analysis framework is investigated for each segment duration separately to study the influence of segment duration on overall performance.
3) We propose an Auto-encoder-based framework to learn joint modality feature representation from ECG and EDA signals. Results obtained by using all the bands (whole signal) and individually performing the best bands (band-level) are analyzed.
4) We build an optimal CRNN-SE model consisting of convolutional and Long Short Term Memory (LSTM) layers and Squeeze-Excitation modules for use as a classifier in all of our experiments.
5) Finally, we evaluate the developed framework on four benchmark datasets to study the generalization capability.

The remaining paper is structured as follows. Section II reviews recent works on joint modality feature learning and frequency domain analysis of physiological signals. The research gap has been identified, and the objectives of the current proposal have been established. Section III contains details of our proposed frameworks. Section IV presents the results obtained and analysis performed on four benchmark datasets. Section IV-E compares the performance of the proposed method with other appropriate methods from the recent literature, and Section V concludes the paper.

II. RELATED WORKS

This section reviews prior works in joint modality feature learning and frequency domain analysis on physiological signals.

A. JOINT MODALITY FOR NON PHYSIOLOGICAL SIGNAL APPLICATIONS

Zhen *et al.* [11] proposed a CNN based cross-modal learning framework text-image matching.The modalities used were images and text. Two sub-networks (an image CNN and a text CNN) with weight sharing constraints at the fully connected layer were developed to learn the cross-modal correlation between the modalities. Discrimination loss was used for cross-modal learning. A linear classifier was trained using the features obtained from the cross-modal representation space. For text-image matching, a modality invariant framework was proposed by Liu *et al.* [12]. The proposed framework fine-tunes a pre-trained CNN image network and text RNN network with an auxiliary adversarial loss to improve the distribution consistency of the two groups of embeddings (image and text). The distributions of images and text were more similar after adversarial learning, which improved retrieval accuracy.

A cross-modal representation for audio-video retrieval was proposed by Surís *et al.* [13]. Visual audio embeddings were obtained by projecting them into a common feature space with deep neural networks. The joint features were used for a retrieval task that generated a query from either of the two modalities. Cross-entropy was employed as the classification loss function. This loss is optimized with the cosine similarity loss to provide the best results.

A modality-invariant (MI) representations for multimodal sentiment analysis was proposed by Hazarika *et al.* [14]. Text, image and video were used for multiclass classification using Transformer. Joint modality features were obtained by training encoder with text, image and video. In MI learning, all modalities for the task are mapped to a common subspace for distributional alignment. Although multimodal data come from a variety of sources, they are all used to achieve the same goal. Individual modalities are projected into a common subspace and aligned by minimising the loss of Central Moment Discrepancy (CMD). The learned representations are used as joint modality feature representations.

B. AUTOENCODER BASED WORKS IN FREQUENCY DOMAIN

A Frequential Stacked Sparse Auto-Encoder (FSSAE) was proposed by Feng *et al.* [15] for detecting Sleep Apnea (SA) using ECG features. The RR intervals are the input to the FSSAE module. This module transforms time-domain RR intervals into frequency-domain RR intervals. Mean Square Error (MSE) was used to calculate the reconstruction loss. Features retrieved from the hidden layer were used to train a separate Time-dependent, cost-sensitive (TDCS) model. An auto-encoder-based system for detecting epilepsy using electroencephalogram (EEG) data was proposed by Sharathappriyaa *et al.* [16]. Harmonic Wavelet Packet Transform (HWPT) and the Katz approach (yielding Fractal Dimension (FD)) are applied to the source EEG signal. The FD and HWPT outcome was supplied into an auto-encoder to map a high-dimensional vector into a lower-dimensional embedding. This lower-embedded feature vector was found to yield higher classification rates. The cost function used to train the autoencoder was MSE. An approach for classifying emotional states in the plane of valence-arousal using a stacked autoencoder was proposed

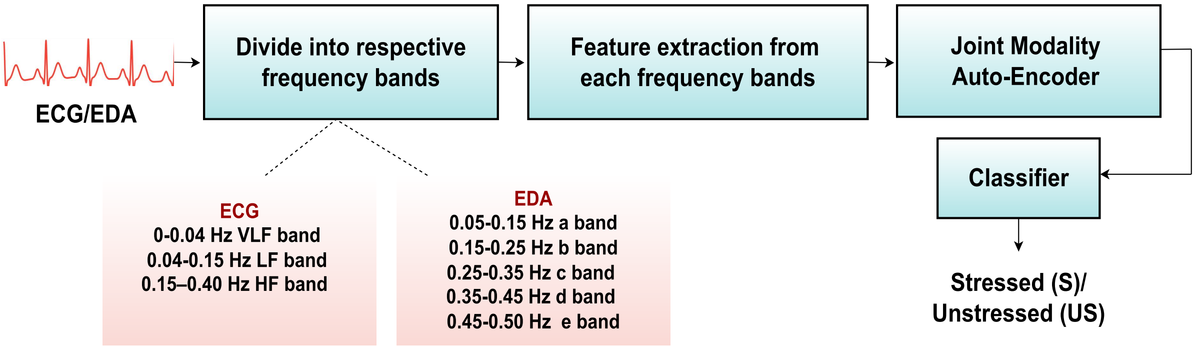

FIGURE 1. An overview of the proposed framework: divided signals into respective frequency bands and extracted features from the bands. Joint feature learning is performed by passing concatenated ECG and EDA features as input to the JMAE module. The joint features from the JMAE module is given as input to the classifier for stress detection.

by Bagherzadeh *et al.* [17]. Physiological signals from the DEAP database, including electromyogram (EMG), electroencephalogram (EEG), and other peripheral signals, were used. Time and spectral features were extracted from these source signals. These features were used to train multiple stacked autoencoders. MSE was used as the reconstruction loss. The majority voting method was used to make the final classification decision. A Supervised Denoising Autoencoder (SDAE) to learn a low-dimensional representation of ECG dynamics to detect false arrhythmia alarms was proposed by Lehman *et al.* [18]. MSE and binary cross-entropy were used to calculate the reconstruction and classification losses.

However, the use of autoencoders for joint modal feature learning in physiological signals, particularly in the frequency domain, has received relatively little attention. Hence, we propose a framework for subject-independent stress detection using features extracted from the ECG and EDA signals.

III. METHODOLOGY

An outline of the proposed framework is given in Figure 1. Frequency bands of EDA and ECG signals are identified. Features are extracted from the PSD. These features are used to learn a joint modality feature representation using an Autoencoder. The obtained joint modality features are used to train a CRNN-SE model to differentiate between stressed and unstressed subjects. Each of the modules is explained in detail below.

A. DATASET DETAILS

The following four benchmark datasets are used in this study.

1) ASCERTAIN

The electroencephalogram (EEG), EDA, ECG physiological signals, and facial activity recordings of 58 subjects are included in this dataset. The average age of the participants was 30. The physiological signals produced by subjects watching the emotional video were recorded. 36 video clips from [19] were used. The length of the videos was 58 to 128 seconds. The sampling rate of EDA and ECG was 128 HZ, and ECG was 256 HZ, respectively. The subjects were asked to give valence arousal ratings on a 7-point scale, expressing their emotional perception after seeing each video clip. Valence rating ranges from −3 to 3, and arousal rating ranges from 0 to 6 [20]. Based on the Valence and Arousal ratings [21], we assigned stress labels as 1 and unstressed as 0 respectively. In the 2-D valence arousal plane, as shown in the Figure 2, HALV is considered as stressed. As a result, those with high arousal and low valence were labeled as stressed, and others as unstressed. The mean value of the ratings is used to determine whether arousal or valence is high or low.

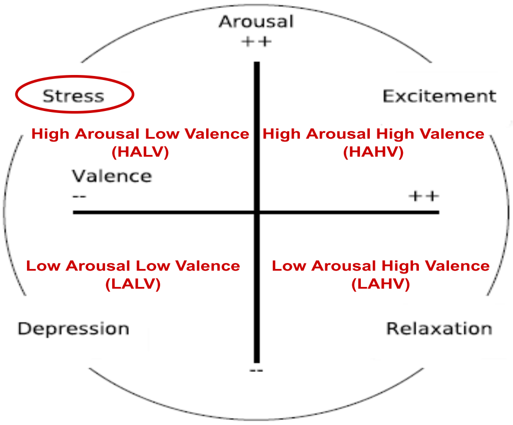
Arousal
++
Stress
Excitement
High Arousal Low Valence (HALV)
High Arousal High Valence (HAHV)
Valence
--
++
Low Arousal Low Valence (LALV)
Low Arousal High Valence (LAHV)
Depression
Relaxation
--

FIGURE 2. Stress is mapped to the upper left quadrant in 2-D circumplex model of valence-arousal proposed by [22], i.e. High Arousal Low Valence (HALV).

2) CLAS

The Plethysmography (PPG), EDA, and ECG physiological data were collected from 62 subjects with a mean age of 20. The sampling rate was 256 Hz. Most of the subjects were students. The subjects are involved in five different activities, including three problem-solving tasks and two perceptive tasks. Image and video-clip stimuli were used for provoking the emotional reactions of subjects in perceptive tasks. 16 emotionally classified 30-second clips from the DEAP database [23] were used as video-clip stimuli. We had 59 subjects after eliminating subjects who didn't have complete information. Stress labels were assigned using pre-defined stimulus tags, which are provided in the dataset [24].

3) MAUS

The dataset captured simple physiological signals under various mental load situations. The N-back task was used to create a mental workload in 22 subjects, 20 of whom were male, and 2 of whom were female. GSR, Wrist-PPG, Fingertip-PPG, and ECG signals were recorded for 35 minutes with a sampling rate of 100 Hz for Wrist-PPG and 256 Hz for others. There was a five-minute rest period at the start of the trial. The N-back task of six trials was performed after a rest interval. The subject had to remember the last N one-digit value in a succession of quickly showing digits in the N-back task. The participant was instructed to reply by pressing the space bar on the keyboard when a stimulus was identical to the N-th number before the stimuli number. The intricacy of the tasks served as ground truth. As the more significant level of N generates a greater level of mental effort, 2 and 3-back tasks were labeled as ''high'' mental workload states, and 0-back tasks were labeled as ''low'' [25].

4) WAUC

The study involved 48 participants who performed the NASA Revised Multi-Attribute Task Battery II under three different activity level conditions. The speed of a stationary bike or a treadmill was changed to manipulate physical activity. Six neural and physiological modalities were recorded during the activity: ECG, EDA, breathing rate, electroencephalography, skin temperature, blood volume pulse, and 3-axis accelerometer. After each experimental section, subjects were asked to complete the NASA Task Load Index questionnaire. The NASA Task Load Index questionnaire rating was converted to a binary value and subjects were labeled (low mental workload or high mental workload) using the average rating as a threshold, which is given in the dataset [26]. We had 45 subjects after removing those subjects who lacked the necessary information.

For subject independence, we fixed training and testing subject IDs. The first 42, 43, 18 and 36 subject samples of ASCERTAIN, CLAS, MUAS and the WAUC dataset respectively are used for training. The remaining 16 subject samples of ASCERTAIN, CLAS, 4 subject samples of MUAS and 9 subject samples of WAUC dataset are used for testing. We addressed the class imbalance problem by applying the Synthetic Minority Oversampling Technique (SMOTE) [27] to training data.

B. FREQUENCY BAND AND FEATURE EXTRACTION

Based on prior works in the literature by Kwon *et al.* [28], Rakshit *et al.* [29] and Hsu *et al.* [30] to find the acceptable sampling frequency range of ECG, we observed three major bands in the frequency spectrum – Very Low Frequency (VLF) band 0.0-0.04 Hz, Low Frequency (LF) band 0.04–0.15 Hz, and High Frequency (HF) band 0.15–0.40 Hz. Based on prior works in the literature by Shukla *et al.* [9] and Ghaderyan *et al.* [31], we observed five major bands in the frequency spectrum – a band 0.05-0.15 Hz, b band 0.15-0.25 Hz, c band 0.25-0.35 Hz, d band 0.35-0.45 Hz and e band 0.45-0.50 Hz.

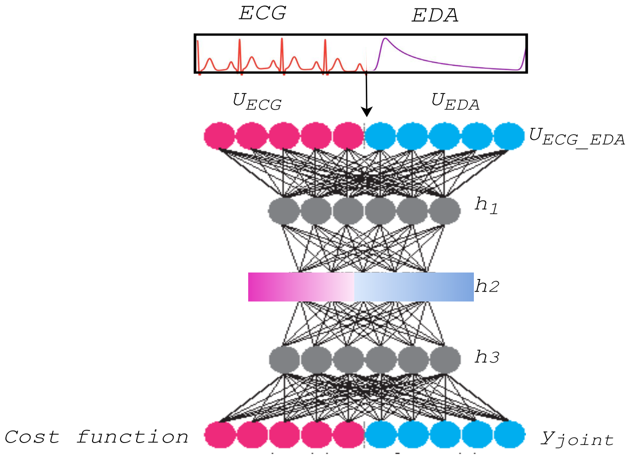

FIGURE 3. An overview of the proposed Auto-encoder to learn the joint modality representation. ECG and EDA features are concatenated (U_{ECG_EDA}) and given as input to the encoder. The embedded layer outcome $h_2(.)$ is taken as joint modality feature representation and used to train a CRNN-SE model.

Power spectral density (using Welch's approach) of the Heart Rate Variability (HRV) extracted from each band of ECG is computed. The python library's frequency module pyHRV [32] is used for this purpose. From these PSDs computed, we extracted a total of 51 frequency-domain measures including Peak, relative powers, logarithmic powers, absolute powers, and so on. Complete list of the 51 measures are available in [32]. Power spectral density (using Welch's approach) of each band of EDA is computed. From these PSDs, we extracted a total of 40 (5 bands with 8 features each) statistical features such as mean, median, min, max, variance, standard deviation, kurtosis and skewness.

C. AUTO-ENCODER BASED JOINT MODALITY LEARNING MODULE

ECG and EDA modalities are simultaneously mapped to a single subspace, and we use adversarial learning to learn this subspace, termed as joint modality. Different from the other works in the literature, we investigate this joint (also referred as shared, cross, common subspace in the literature) modality subspace in the frequency domain for the first time. We propose an auto-encoder based framework to achieve this objective. The architecture of the proposed Joint Modality Auto-encoder (JMAE) is shown in Figure 3.

Firstly, we concatenate the ECG features, U_{ECG} and EDA features U_{EDA} into one single vector input, U_{ECG_EDA}.

The first, second and third fully connected layers are $h_1(.)$, $h_2(.)$ and $h_3(.)$ respectively. The last layer is an output layer, Y_{joint} of the length same as the input vector U_{ECG_EDA}. The first, second and third hidden layers constitute the parameter vector $\theta(.)$ to be learnt by minimizing a cost (reconstruction) function. The cost function is selected such that the distributions of ECG and EDA are aligned in the joint subspace.

Algorithm 1 Pseudo-Code for JMAE

```
procedure INPUT: $U_{ECG\_EDA}$(concatenated ECG and EDA features)
    PARAMETER $W$: Weights of the hidden layers $h_1(.)$, $h_2(.)$ and $h_3(.)$
    $W_{h_1}, W_{h_2}, W_{h_3} \leftarrow$ random // Hidden layer initialization
    $Y_{ECG\_EDA} \leftarrow null$ // Reconstructed input $U_{ECG\_EDA}$
    $l \leftarrow batch\ No$
    $i \leftarrow 0$
    while $i <= l$ do
        // The encoder function converts input $U_{ECG\_EDA}$ into a hidden representation $h_n(.)$:
        $Y_{h1} = f_{h1(U_{ECG\_EDA}, W_{h_1})}$
        $Y_{h2} = f_{h2(U_{ECG\_EDA}, W_{h_2})}$
        $Y_{h3} = f_{h1(U_{ECG\_EDA}, W_{h_3})}$
        /* The decoder function returns a $Y_n$ from a hidden representation $h_n(.)$ */
        $Y_{ECG\_EDA} = f_Y(Y_{h3}, W_Y)$
        Loss = L($U_{ECG\_EDA}$, $Y_{joint}$)
        min$_\theta$ (Loss)
        $i \leftarrow i + 1$
    end while
    return $\theta$
    $\theta \leftarrow Parameters$
    $Loss \leftarrow$ MSE, Cosine similarity and KL divergence
end procedure
```

Based on different works in frequency domain, we investigated the following three cost functions – MSE, cosine similarity, and Kullback-Leibler (KL) divergence. The cost function will represent the differences between the input U_{ECG_EDA} and the reconstructed Y_{joint}. The proposed model was trained with the Adam optimizer using the default learning rate and 64 as the mini-batch size. The pseudo-code for training the JMAE is summarized in Algorithm 1.

1) MSE

MSE is calculated, as shown in Eqn. 1, where a_i is the target value - U_{ECG_EDA}. and p_i is the predicted value - Y_{joint}. The cost function value ranges from 0 to ∞. The reconstructed Y_{joint} is more similar to input U_{ECG_EDA} if the MSE value is near to 0 else they are dissimilar.

$$MSE(a, p) = \frac{1}{n}\sum_{i=1}^{n}(a_i - p_i)^2 \tag{1}$$

2) COSINE SIMILARITY

The cosine similarity is computed between the a_i, the target value - U_{ECG_EDA} and p_i, the predicted value -Y_{joint}, as shown in Eqn.2. The cost function has a value between 0 and 1. The value near 0 implies that the Y_{joint} is similar to the U_{ECG_EDA}, while the value near 1 indicates that they are dissimilar.

$$Cos(a, p) = 1 - \frac{\sum_{i=0}^{n-1} a_i \cdot p_i}{\sqrt{\sum_{i=0}^{n-1} a_i^2}\sqrt{\sum_{i=0}^{n-1} p_i^2}} \tag{2}$$

3) KL DIVERGENCE

The KL divergence is the distance metric that computes the similarity between the a_i, the target value - U_{ECG_EDA} and p_i, the predicted value -Y_{joint}, as shown in Eqn.3. The cost function value ranges from 0 to ∞. The two distributions (U_{ECG_EDA} and Y_{joint}) are similar if the value is close to 0, else the distributions are dissimilar.

$$KL(a, p) = \sum_i a_i log\frac{a_i}{p_i} \tag{3}$$

The results of each loss are compared in the result's section Table 3.

D. CLASSIFIER

We selected a CRNN-SE model having 2 convolutional layers, one LSTM layer and two SE modules as our classifier in all our experiments. Details for this choice are given in Appendix A. For frequency domain analysis, each signal is broken into segments of duration 5 sec each. Details for this choice are given in Appendix B.

All the models are trained with the Adam optimizer using the default learning rate and 64 as the mini-batch size. Binary Cross-Entropy (BCE) given by Eqn. 4 is taken as the loss function. Here, y_i is the actual label and $p(y_i)$ is the predicted label for all the N samples.

$$BCE(y_i, p(y_i)) = \frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{4}$$

An early-stopping strategy controls the training duration if the loss does not decrease for 30 epochs in succession. The accuracy and F1-score is used to evaluate the performance of various models.

IV. RESULTS AND DISCUSSION

This sections presents the results obtained by applying our proposed framework on the four benchmark datasets.

A. SELECTION OF FREQUENCY BAND

To study the performance of each of ECG and EDA frequency band, the features obtained from each band used to train separate CRNN-SE classifier. Table 1 shows the frequency band analysis of the ECG dataset, and Table 2 shows the frequency band analysis of the EDA dataset. The results show that the HF band (0.15-0.40 Hz) of the ECG and b band

TABLE 1. Performance of each ECG frequency band.

S.No	Details	Accuracy	F1-score
	ASCERTAIN		
1	0-0.04 Hz VLF band	51.36%	0.51
2	0.04-0.15 Hz LF band	70.83%	0.17
3	**0.15–0.40 Hz HF band**	**72.59%**	**0.73**
4	Late fusion	76.71%	0.77
	CLAS		
5	0-0.04 Hz VLF band	56.50%	0.57
6	0.04-0.15 Hz LF band	61.40%	0.61
7	**0.15–0.40 Hz HF band**	**63.62%**	**0.64**
8	Late fusion	70.59%	0.71
	MAUS		
9	0-0.04 Hz VLF band	54.44%	0.53
10	0.04-0.15 Hz LF band	63.26%	0.64
11	**0.15–0.40 Hz HF band**	**70.89%**	**0.71**
12	Late fusion	73.78%	0.74
	WAUC		
13	0-0.04 Hz VLF band	53.29%	0.53
14	0.04-0.15 Hz LF band	60.43%	0.61
15	**0.15–0.40 Hz HF band**	**63.74%**	**0.63**
16	Late fusion	68.96%	0.68

TABLE 2. Performance of each EDA frequency band.

S.No	Details	Accuracy	F1-score
	ASCERTAIN		
1	0.05-0.15 Hz a band	59.68%	0.60
2	**0.15-0.25 Hz b band**	**66.99%**	**0.70**
3	0.25-0.35 Hz c band	53.41%	0.53
4	0.35-0.45 Hz d band	63.52%	0.64
5	0.45-0.50 Hz e band	60.74%	0.61
6	Late fusion	78.67%	0.79
	CLAS		
7	0.05-0.15 Hz a band	57.54%	0.58
8	**0.15-0.25 Hz b band**	**62.75%**	**0.63**
9	0.25-0.35 Hz c band	51.26%	0.51
10	0.35-0.45 Hz d band	61.87%	0.62
11	0.45-0.50 Hz e band	58.32%	0.59
12	Late fusion	72.57%	0.73
	MAUS		
13	0.05-0.15 Hz a band	60.83%	0.61
14	**0.15-0.25 Hz b band**	**67.39%**	**0.67**
15	0.25-0.35 Hz c band	62.22%	0.62
16	0.35-0.45 Hz d band	61.11%	0.61
17	0.45-0.50 Hz e band	63.57%	0.64
18	Late fusion	74.65%	0.75
	WAUC		
19	0.05-0.15 Hz a band	51.62%	0.51
20	**0.15-0.25 Hz b band**	**66.24%**	**0.66**
21	0.25-0.35 Hz c band	61.96%	0.62
22	0.35-0.45 Hz d band	54.81%	0.55
23	0.45-0.50 Hz e band	63.46%	0.63
24	Late fusion	70.52%	0.71

(0.15-0.25 Hz) of EDA achieved the highest accuracy and F1 score for all the four datasets. It means frequencies from 0.15-0.25 Hz, both ECG and EDA have features with higher discriminative capacity for identifying stress. For a hardware implementation, low pass filter can be used to extract these richer features from the frequency transform on the ECG, EDA signals. It will be interesting to pursue if this band range is valid for other physiological signals e.g. EEG.

TABLE 3. Performance of features obtained from the entire signals and the features obtained from the highest performing frequency bands.

S.No	Loss	Accuracy	F1-Score	Accuracy	F1-Score
		Multi-Level		**Band-Level**	
		ASCERTAIN			
1	**MSE**	**96.23%**	**0.95**	**94.56%**	**0.94**
2	Cosine similarity	94.54%	0.93	90.47%	0.90
3	KL divergence	91.63%	0.92	87.12%	0.86
		CLAS			
4	**MSE**	**94.15%**	**0.93**	**92.36%**	**0.91**
5	Cosine similarity	90.74%	0.91	88.92%	0.88
6	KL divergence	89.82%	0.90	85.64%	0.86
		MAUS			
7	**MSE**	**86.46%**	**0.86**	**83.12%**	**0.82**
8	Cosine similarity	83.29%	0.82	80.73%	0.79
9	KL divergence	79.57%	0.79	74.38%	0.74
		WAUC			
10	**MSE**	**82.56%**	**0.83**	**80.13%**	**0.80**
11	Cosine similarity	80.38%	0.81	78.42%	0.79
12	KL divergence	77.92%	0.78	74.85%	0.75

B. BAND LEVEL VS WHOLE SIGNAL

We investigated the proposed framework on the whole signal (using all the ECG and EDA frequency bands) as well as on a band level (using the bands with the highest performance, as obtained in Section IV-A). For the whole signal's performance, 51 frequency-domain features from the ECG signal and 40 frequency-domain features from the EDA signal are used to train a $JMAE_{whole}$ module. The first hidden layer $h_1(.)$, second hidden layer $h_2(.)$ and third hidden layer $h_3(.)$ are of length 95, 100 and 95. The joint modality features obtained from the $JMAE_{whole}$ are used to report the results in third and fourth columns of the Table 3. For the band-level performance, 15 frequency-domain features from the ECG signal i.e., HF band (0.15-0.40Hz) features and 8 frequency-domain features from the EDA signal i.e., b band (0.15-0.25Hz) features are used to train a $JMAE_{band}$ module. The first hidden layer $h_1(.)$, second hidden layer $h_2(.)$ and third hidden layer $h_3(.)$ are of length 25, 30 and 25. The joint modality features obtained from the $JMAE_{band}$ are used to report the results in fifth and sixth columns of the Table 3. In all the situation, multi-level features outperformed band-level features performance by 1.7-3.3% (absolute) for MSE loss function, 1.8-4.0% (absolute) for cosine similarity and 4.2-5.2% (absolute) for KL divergence.

We validated the proposed model by performing K-fold cross-validation on the highest performed model (Loss-MSE). The K value is chosen to be 5. The joint features obtained from the JMAE model are split into 5 folds. Classification accuracy and F1-score (mean ± standard deviation) is given in Table 4. In all the datasets cross-validation results

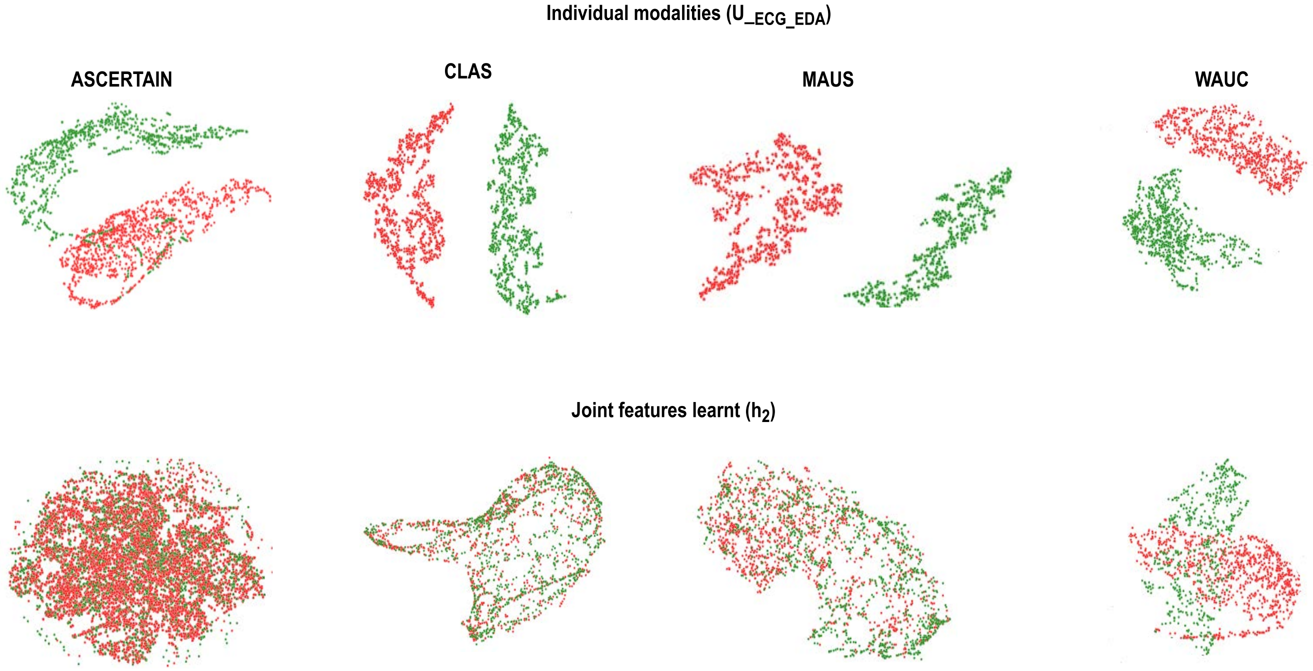

FIGURE 4. t-SNE visualization of ASCERTAIN, CLAS, MAUS and WAUC dataset. When compared to individual modalities (U_{ECG_EDA}) a close alignment between two modalities (ECG and EDA) can be observed in the joint features learnt (h_2).

TABLE 4. Performance of 5-fold cross-validation on highest performing model (loss-MSE).

S.No	Dataset	Multi-Level		Band-Level	
		Accuracy	F1-Score	Accuracy	F1-Score
1	ASCERTAIN	**98.92±0.03%**	**0.98±0.04**	97.34±0.14%	0.97±0.18
2	CLAS	**98.15±0.08%**	**0.97±0.09**	96.57±0.20%	0.95±0.23
3	MAUS	**92.47±0.28%**	**0.91±0.29**	88.42±0.32%	0.88±0.35
4	WAUC	**89.63±0.16%**	**0.88±0.14**	85.81±0.38%	0.86±0.36

outperformed previous results (Table-3) by 2.7-4% (absolute). We infer that this increase is due to subject dependence during cross-validation.

C. T-SNE VISUALISATION

To further investigate the joint feature learning achieved by our model, we plot t-distributed Stochastic Neighbour Embedding (tSNE) before and after joint feature learning. The t-SNE approach projects multi-dimensional points onto two-dimensional or three-dimensional spaces such that if two points have the same distribution, the resulting projection keeps them close. Similarly, in the t-SNE projections, distant points remain far apart. With tSNE, we project the joint features into a 2-D space. The feature visualization of U_{ECG_EDA} (regular features) and $h_2(.)$ (joint features learnt) using MSE cost function on whole signal of all the benchmark datasets are shown in Figure 4. The red dots represent ECG features, and the green dots represent EDA features. Joint feature learning aims to bring different modalities features to a shared space. In the visualization, we observed close overlapping among modalities (ECG and EDA) after joint feature learning. This indicates that the modality gap between the distribution of modalities is significantly reduced.

D. GENERALIZATION CAPABILITIES

The proposed model is tested on four benchmark datasets to assess the proposed framework's generalization capabilities. These tests ensure that our proposed framework is not overfitting to a specific dataset collected in a given environment. We discovered that the performance on all four datasets followed the same patterns. As a result, we ensured that the four benchmark datasets we used were gathered in various scenarios. The CLAS and ASCERTAIN were collected while subjects watched emotional video clips, MAUS and WAUC were collected when subjects undergone physical activity.

E. COMPARISON WITH OTHER WORKS

This section contrasts results obtained by our proposed JMAE framework with recent works on the ASCERTAIN, CLAS, MAUS and WAUC datasets. An overview of the metrics – accuracy, F1-score and AUC are given in Table 5.

It is noted that the majority of the stress detection studies used time and frequency domain features, [24]–[26], [33]–[36], [39]–[41]. Our proposed JMAE based features are learned from the frequency domain measures. Hence, they perform better than the time and frequency domain feature-based frameworks of ASCERTAIN, MAUS and WAUC datasets by *15-17%*.

Most works [24], [25], [33]–[35], [38], [39] reported performance on subject-dependent scenario. The performance of these works are usually higher owing to prior knowledge of

TABLE 5. Comparison with existing stress detection studies.

S.No	Method	Accuracy	F1-score	AUC
ASCERTAIN				
1	[33]	68.7%*	-	-
2	[34]	68.0%*	-	-
3	[35]	68.7%*	-	-
4	[36]	78.8%	0.79	0.78
5	[37]	75.5%	0.76	0.75
6	**Proposed**	**96.2%**	**0.96**	**0.95**
CLAS				
7	[24]	88.9%*	-	-
8	[38]	94.8%*	-	-
9	**[39]**	**97.6%***	-	-
10	[36]	72.6%	0.74	0.73
11	[37]	69.9%	0.70	0.69
12	Proposed	94.2%	0.93	0.93
MAUS				
13	[25]	71.6%*	0.71	-
14	**Proposed**	**86.5%**	**0.86**	**0.85**
WAUC				
15	[26]	-	-	EDA- 0.66 ± 0.01 ECG- 0.74 ± 0.01
16	**Proposed**	**82.6%**	**0.83**	**0.82**

*Subject-dependent

the testing subject during training process itself. However, our proposed framework also outperforms those works in ASCERTAIN and MAUS datasets by *15-27%*. Our JMAE based framework outperformed the subject-independent stress detection works [26], [36] and [37] by *8-22%*.

Few works utilized traditional handcrafted features in conjunction with Machine Learning (ML) models, such as Support Vector Machine and Naive Bayes and Random Forest [24]–[26], [33]–[35], [38], [39]. Few more trained end–to-end deep learning model such as CNN [36] and [37] The proposed framework trains a DL models and uses the outcome of DL models (intermediate layer) to train a DL model. Our framework outperformed existing ML and DL works of ASCERTAIN, MAUS and WAUC datasets by *8-17%*

Using ECG biomarkers several stress related abnormalitiesties can be detected (Coronary Artery Disease (CAD [42], myocardial ischemia [43], stroke, atrial fibrillation, cardiac arrhythmias [44]). Using EDA/GSR biomarker some other set of abnormalities caused by stress can also be detected (brain and heart attack [45], Epilopsy [46], blood pressure [47], Depression [48]). Traditional approaches built separate classifiers using these modalities (biomarkers) and then took the final decision (late fusion techniques) of stressed or not. Our approach concatenates the two biomarkers (feature fusion till here) and then learns joint representation (our contribution) to yield the best feature representation biomarkers for stress detection. This is in line with the clinical practice of diagnosing by simultaneous monitoring physiological signals to take decision. Clinical decisions are rarely made by monitoring only one physiological signal. Our results are performing better than other works in the literature that are based on the single, early, and late fusion of modalities. It is interesting to note that the band-level features based framework ($JMAE_{band}$) performs better than all the other works on ASCERTAIN and MAUS datasets by *11-15%*. This reinforces the richer nature of our proposed JMAE based features.

The results indicate that learning joint features of different modalities from the shared space can enhance the performance of the models. The proposed model is able to perform better than other existing works on ASCERTAIN, MAUS, and WAUS datasets. On the CLAS dataset, the accuracy of [39] is higher due to the ensemble voting on subject dependent model.

V. CONCLUSION

We proposed a joint modality features-based framework in the frequency domain for stress detection. We validated our framework using physiological signal modalities – EDA and ECG. Frequency bands of ECG and EDA are identified. Features extracted from the PSD are used to train CRNN models with SE modules. The proposed framework was tested on four benchmark datasets. The High Frequency (HF) band (0.15-0.40 Hz) of ECG and b frequency band (0.15-0.25 Hz) of EDA were found to have the most impact on the overall performance. Our promising findings encourage us to continue further study into joint modality learning with more than two modalities.

APPENDIX A
SELECTION OF CRNN ARCHITECTURE

The following sections provide information on selecting the number and location of different Convolutional layers, LSTM layers, and SE modules.

A. SELECTION OF CONVOLUTIONAL AND LSTM LAYERS

Referring to Fig 5, we first investigate the CRNN architecture without any SE modules. Model 1 consists of Convolution layer 1, Convolution layer 2, LSTM layer 1, and LSTM layer 2. Model 2 consists of Convolution layer 1, Convolution layer 2, and LSTM layer 2. Model 3 consists of Convolution layer 1, LSTM layer 1, and LSTM layer 2. Model 4 consists of Convolution layer 1 and LSTM layer 1. Each convolutional layer is always followed by Batch normalization and max pool layers. All the models have two fully connected layers (FC1 and FC2) and a sigmoid output layer.

Performance of individual modalities on two datasets for different models is presented in Table 6. Model 2 yielded the highest performance on ASCERTAIN ECG, EDA, and CLAS ECG features. Model 4 yielded the highest performance on the CLAS EDA features. As Model 2 performed best in most cases, we selected Model 2 architecture consisting of two Convolutional layers and one LSTM layer for the rest of the experiments.

B. SELECTION OF SE MODULES

The SE module gained popularity in the ImageNet competition by emphasizing relevant features and suppressing undesirable ones by feature recalibration [49]. The SE module proposed by Hu *et al.* [50] consists of two operations – Squeeze and Excitation. The squeezing block employs global

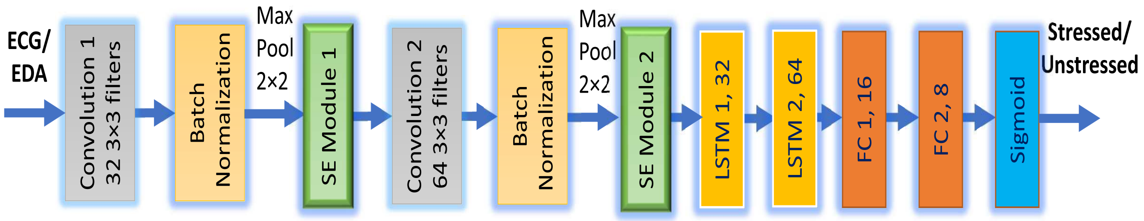

FIGURE 5. Proposed CRNN model by varying convolutional, LSTM layers and SE modules.

TABLE 6. Performance of CRNN model.

S.No	Model	Accuracy	F1-score	Accuracy	F1-score
		ECG		EDA	
		ASCERTAIN			
1	Model 1	65.28%	0.64	78.12%	0.78
2	Model 2	**70.49%**	**0.71**	**79.86%**	**0.80**
3	Model 3	67.71%	0.68	76.39 %	0.76
4	Model 4	67.71%	0.50	78.07%	0.77
		CLAS			
5	Model 1	64.84%	0.65	60.16%	0.61
6	Model 2	**66.80%**	**0.67**	55.08%	0.54
7	Model 3	60.94%	0.61	55.47%	0.56
8	Model 4	56.30%	0.57	**66.71%**	**0.67**

average pooling on the outputs from the previous convolutional block yielding feature maps. To add nonlinearity, these feature maps are passed through a fully connected layer with the ReLU activation function. For smooth gating, the ReLu outcome is passed to the second fully connected layer with a sigmoid activation function. The output of the sigmoid function is weighted by the output of the Convolution layer (used earlier as input to the Squeeze block). The entire process of fully connected layers and weighing is referred to as an excitation operation.

Referring to Fig 5, we now investigate Model 2 with different numbers and location of SE modules. Model 5 consists of Model 2 with SE Module 1 only. Model 6 consists of Model 2 with SE Module 2 only. Model 7 consists of SE Modules 1 and 2. Each convolutional layer is always followed by Batch normalization and max pool layers. All the models have two fully connected layers (FC1 and FC2) and a sigmoid output layer. Performance of individual modalities on two datasets for different models is presented in Table 7. Model 7 yielded the highest performance in ASCERTAIN EDA, and CLAS EDA features. Model 5 yielded the highest performance in ASCERTAIN ECG and CLAS ECG features. We selected Model 7 architecture for the rest of the experiments using ECG features and EDA features.

APPENDIX B
SELECTION OF SEGMENT DURATION

Model 7 is used as framework for EDA features and ECG features. Each signal (ECG/EDA) is divided into segments of fixed duration. Four cases are considered – 2 sec, 5 sec, 10 sec and 15sec duration each. In each case, Model 7 is trained, and the performances obtained are reported in the Table 8. The highest performance is observed for segment duration 5 sec. The overall performance of 5 sec segmented signals (Table 8) rows 2 and 6, for both ECG and EDA features) is higher than the baseline performance of the full signal (Table 7 rows 1 and 5 for ECG features, rows 3 and 6 for EDA features).

TABLE 7. Performance of CRNN model with SE modules.

S.No	Model	Accuracy	F1-score	Accuracy	F1-score
		ECG		EDA	
		ASCERTAIN			
1	Model 5	**70.49%**	**0.69**	68.40%	0.68
2	Model 6	63.19%	0.64	68.75%	0.69
3	Model 7	64.76%	0.65	**80.90%**	**0.81**
		CLAS			
4	Model 5	**66.80%**	**0.67**	54.69%	0.55
5	Model 6	64.84%	0.64	57.42%	0.58
6	Model 7	65.23%	0.65	**59.77%**	**0.60**

TABLE 8. Performance on different segmentation time.

S.No	Segment time	Accuracy	F1-score	Accuracy	F1-score
		ECG		EDA	
		ASCERTAIN			
1	2 Sec	71.36%	0.71	80.72%	0.81
2	5 sec	**77.82%**	**0.78**	**86.13%**	**0.86**
3	10 sec	76.21%	0.75	84.65%	0.85
4	15 sec	72.54%	0.73	81.93%	0.82
		CLAS			
5	2 Sec	68.98%	0.69	63.79%	0.64
6	5 sec	**74.34%**	**0.74**	**69.86%**	**0.70**
7	10 sec	72.96%	0.73	67.53%	0.67
8	15 sec	70.81%	0.69	65.47%	0.65

REFERENCES

[1] S. Uday, C. Jyotsna, and J. Amudha, "Detection of stress using wearable sensors in IoT platform," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 492–498.

[2] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," *Biocybernetics Biomed. Eng.*, vol. 39, no. 2, pp. 444–469, Apr. 2019.

[3] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *J. Biomed. Informat.*, vol. 92, Apr. 2019, Art. no. 103139.

[4] G. Aalbers, R. J. McNally, A. Heeren, S. De Wit, and E. I. Fried, "Social media and depression symptoms: A network perspective," *J. Experim. Psychol., Gen.*, vol. 148, no. 8, p. 1454, 2019.

[5] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, Apr. 2019.

[6] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *J. Biomed. Inform.*, vol. 59, pp. 49–75, Feb. 2016.

[7] K. Kyriakou, B. Resch, G. Sagl, A. Petutschnig, C. Werner, D. Niederseer, M. Liedlgruber, F. Wilhelm, T. Osborne, and J. Pykett, "Detecting moments of stress from measurements of wearable physiological sensors," *Sensors*, vol. 19, no. 17, p. 3805, Sep. 2019.

[8] 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.

[9] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. 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. 2021.

[10] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion recognition from multiband EEG signals using CapsNet," *Sensors*, vol. 19, no. 9, p. 2212, May 2019.

[11] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10394–10403.

[12] R. Liu, Y. Zhao, S. Wei, L. Zheng, and Y. Yang, "Modality-invariant image-text embedding for image-sentence matching," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1, pp. 1–19, Feb. 2019.

[13] D. Surís, A. Duarte, A. Salvador, J. Torres, and X. Giró-i Nieto, "Cross-modal embeddings for video and audio retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 1–6.

[14] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131.

[15] K. Feng, H. Qin, S. Wu, W. Pan, and G. Liu, "A sleep apnea detection method based on unsupervised feature learning and single-lead electrocardiogram," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2020.

[16] V. Sharathappriyaa, S. Gautham, and R. Lavanya, "Auto-encoder based automated epilepsy diagnosis," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2018, pp. 976–982.

[17] S. Bagherzadeh, K. Maghooli, J. Farhadi, and M. Z. Soroush, "Emotion recognition from physiological signals using parallel stacked autoencoders," *Neurophysiology*, vol. 50, no. 6, pp. 428–435, Nov. 2018.

[18] E. P. Lehman, R. G. Krishnan, X. Zhao, R. G. Mark, and H. L. Li-Wei, "Representation learning approaches to detect false arrhythmia alarms from ecg dynamics," in *Proc. Mach. Learn. Healthcare Conf.*, 2018, pp. 571–586.

[19] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Jul. 2015.

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

[21] M. Dahmane, J. Alam, P.-L. St-Charles, M. Lalonde, K. Heffner, and S. Foucher, "A multimodal non-intrusive stress monitoring from the pleasure-arousal emotional dimensions," *IEEE Trans. Affect. Comput.*, early access, Apr. 20, 2020, doi: 10.1109/TAFFC.2020.2988455.

[22] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Develop. Psychopathol.*, vol. 17, no. 3, pp. 715–734, 2005.

[23] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jun. 2011.

[24] V. Markova, T. Ganchev, and K. Kalinkov, "CLAS: A database for cognitive load, affect and stress recognition," in *Proc. Int. Conf. Biomed. Innov. Appl. (BIA)*, Nov. 2019, pp. 1–4.

[25] W.-K. Beh, Y.-H. Wu, An-Yeu, and Wu, "MAUS: A dataset for mental workload assessmenton N-back task using wearable sensor," 2021, *arXiv:2111.02561*.

[26] I. Albuquerque, A. Tiwari, M. Parent, R. Cassani, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "WAUC: A multi-modal database for mental workload assessment under physical activity," *Frontiers Neurosci.*, vol. 14, Dec. 2020.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[28] O. Kwon, J. Jeong, H. B. Kim, I. H. Kwon, S. Y. Park, J. E. Kim, and Y. Choi, "Electrocardiogram sampling frequency range acceptable for heart rate variability analysis," *Healthcare Informat. Res.*, vol. 24, no. 3, pp. 198–206, Jul. 2018.

[29] R. Rakshit, V. R. Reddy, and P. Deshpande, "Emotion detection and recognition using HRV features derived from photoplethysmogram signals," in *Proc. 2nd Workshop Emotion Representations Modeling Companion Syst.*, Nov. 2016, pp. 1–6.

[30] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ECG-based emotion recognition in music listening," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 85–99, Jan. 2017.

[31] P. Ghaderyan and A. Abbasi, "An efficient automatic workload estimation method based on electrodermal activity using pattern classifier combinations," *Int. J. Psychophysiol.*, vol. 110, pp. 91–101, Dec. 2016.

[32] P. Gomes, P. Margaritoff, and H. Silva, "PyHRV: Development and evaluation of an open-source Python toolbox for heart rate variability (HRV)," in *Proc. Int. Conf. Electr., Electron. Comput. Eng. (IcETRAN)*, 2019, pp. 822–828.

[33] V. Markova and T. Ganchev, "Three-step attribute selection for stress detection based on physiological signals," in *Proc. IEEE 27th Int. Scientific Conf. Electron. (ET)*, Sep. 2018, pp. 1–4.

[34] V. Markova and T. Ganchev, "Automated recognition of affect and stress evoked by audio-visual stimuli," in *Proc. 7th Balkan Conf. Lighting (BalkanLight)*, Sep. 2018, pp. 1–4.

[35] V. Markova and T. Ganchev, "Constrained attribute selection for stress detection based on physiological signals," in *Proc. Int. Conf. Sensors, Signal Image Process. (SSIP)*, 2018, pp. 41–45.

[36] K. Radhika and V. R. M. Oruganti, "Transfer learning for subject-independent stress detection using physiological signals," in *Proc. IEEE 17th India Council Int. Conf. (INDICON)*, Dec. 2020, pp. 1–6.

[37] K. Radhika and V. R. M. Oruganti, "Deep multimodal fusion for subject-independent stress detection," in *Proc. 11th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2021, pp. 105–109.

[38] K. Kalinkov, T. Ganchev, and V. Markova, "Adaptive feature selection through Fisher discriminant ratio," in *Proc. Int. Conf. Biomed. Innov. Appl. (BIA)*, Nov. 2019, pp. 1–4.

[39] M. Kang, S. Shin, G. Zhang, J. Jung, and Y. T. Kim, "Mental stress classification based on a support vector machine and naive Bayes using electrocardiogram signals," *Sensors*, vol. 21, no. 23, p. 7916, Nov. 2021.

[40] K. M. Dalmeida and G. L. Masala, "HRV features as viable physiological markers for stress detection using wearable devices," *Sensors*, vol. 21, no. 8, p. 2873, Apr. 2021.

[41] R. Sánchez-Reolid, A. Martínez-Rodrigo, M. T. López, and A. Fernández-Caballero, "Deep support vector machines for the identification of stress condition from electrodermal activity," *Int. J. Neural Syst.*, vol. 30, no. 7, Jul. 2020, Art. no. 2050031.

[42] J. Macleod, G. D. Smith, P. Heslop, C. Metcalfe, D. Carroll, and C. Hart, "Psychological stress and cardiovascular disease: Empirical demonstration of bias in a prospective observational study of Scottish men," *Brit. Med. J.*, vol. 324, no. 7348, p. 1247, 2002.

[43] M. Hammadah, S. Sullivan, B. Pearce, I. Al Mheid, K. Wilmot, R. Ramadan, A. S. Tahhan, W. T. O'Neal, M. Obideen, A. Alkhoder, and N. Abdelhadi, "Inflammatory response to mental stress and mental stress induced myocardial ischemia," *Brain, Behav., immunity*, vol. 68, pp. 90–97, 2018.

[44] M. Kivimäki and A. Steptoe, "Effects of stress on the development and progression of cardiovascular disease," *Nature Rev. Cardiol.*, vol. 15, no. 4, pp. 215–229, Apr. 2018.

[45] A. Fernandes, R. Helawar, R. Lokesh, T. Tari, and A. V. Shahapurkar, "Determination of stress using blood pressure and galvanic skin response," in *Proc. Int. Conf. Commun. Netw. Technol.*, Dec. 2014, pp. 165–168.

[46] D. B. Setyohadi, S. Kusrohmaniah, S. B. Gunawan, P. Pranowo, and A. Prabuwono, "Galvanic skin response data classification for emotion detection," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, pp. 31–41, 2018.

[47] R. F. Navea, P. J. Buenvenida, and C. D. Cruz, "Stress detection using galvanic skin response: An Android application," *J. Phys., Conf.*, vol. 1372, no. 1, Nov. 2019, Art. no. 012001.

[48] A. Joshi and R. Kiran, "Gauging the effectiveness of music and yoga for reducing stress among engineering students: An investigation based on galvanic skin response," *Work*, vol. 65, no. 3, pp. 671–678, Mar. 2020.

[49] A. G. Roy, N. Navab, and C. Wachinger, ''Recalibrating fully convolutional networks with spatial and channel 'squeeze and excitation' blocks,'' *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Feb. 2018.

[50] J. Hu, L. Shen, and G. Sun, ''Squeeze- and-excitation networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

K. RADHIKA received the master's degree in computer science from Amrita Vishwa Vidyapeetham, India, in 2018, where she is currently pursuing the Ph.D. degree with the Department of Electrical and Electronics Engineering. Her research interests include multi-modal interactions and deep learning for applications in affective computing.

RAMANATHAN SUBRAMANIAN (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from NUS, in 2008. He is currently an Associate Professor with the University of Canberra, Australia. His past affiliations include IHPC, Singapore; U Glasgow, Singapore; IIIT Hyderabad, India; IIT Ropar, India; and UIUC-ADSC, Singapore. His research interests include human–centered computing, interactive analytics, and explainable machine learning. He is a member of the ACM and AAAC.

VENKATA RAMANA MURTHY ORUGANTI (Member, IEEE) received the master's and Ph.D. degrees in electrical engineering from IIT Delhi, India. He is currently an Assistant Professor with the Department of Electrical and Electronics Engineering, Amrita Vishwa Vidyapeetham, India. His past affiliations include NUS, Singapore; NTU, Singapore; University of Canberra, Australia; and Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include medical image processing and affective computing. He is a member of the ACM.