IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 December 2022, accepted 12 January 2023, date of publication 16 January 2023, date of current version 24 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3237545

RESEARCH ARTICLE

Multimodal Hierarchical CNN Feature Fusion for Stress Detection

RADHIKA KUTTALA[1], RAMANATHAN SUBRAMANIAN[2], (Senior Member, IEEE), AND VENKATA RAMANA MURTHY ORUGANTI[1], (Senior Member, IEEE)

[1]Department of Electrical and Electronics Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore 641112, India
[2]Faculty of Science and Technology, University of Canberra, Bruce, Canberra, ACT 2617, Australia

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

ABSTRACT Stress is one of the most severe concerns in modern life. High-level stress can create various diseases or loss of focus and productivity at work. Being under stress prevents people from recognizing their stress levels, so early stress detection is essential. Recently, multimodal fusion has enhanced the performance of stress detection models using Deep Learning (DL) techniques. The low, mid, and high-level features of a Convolutional Neural Network (CNN) are discriminative. A comprehensive feature representation can be obtained by fusing all three levels of CNN's features. This study mainly focuses on detecting stress by exploiting these advantages using a multimodal hierarchical CNN feature fusion. The two multimodal physiological signals used in this study are Electrodermal activity (EDA) and Electrocardiogram (ECG). We develop a hierarchical feature set by concatenating multi-level CNN features for each modality. Multimodal fusion on both hierarchical feature sets is performed using the Multimodal Transfer Module (MMTM). The experiments are carried out with raw frequency domain data and the features from the frequency bands to study the effectiveness of both. The model's performance is compared to the different combinations of hierarchical features from low, mid, and high levels. To verify the generalizability, the proposed approach has been evaluated on four benchmark datasets - ASCERTAIN, CLAS, MAUS, and WAUC. The proposed method showed its effectiveness by outperforming existing models by 1-2%, respectively, on frequency band features. It is observed that the hierarchical feature set from all three levels performed better than all other combinations by 2-4%. As a result, this strategy can be a useful addition to stress detection.

INDEX TERMS Multimodal, EDA, ECG, CNN, hierarchical feature fusion, stress detection, subject-independent.

I. INTRODUCTION

Stress is a way of responding to overwhelming demands or challenges from a scenario that manifests as emotional, physical, or behavioural changes by the human body [1]. The way an individual views the scenario has a significant impact on how stressed they are. When an individual faces a challenge in achieving their goal, they evaluate the scenario in two stages– (i)the need to achieve the desired goal and (ii) the external and internal resources to meet the challenges [2]. Human stress is classified as positive and negative. Positive or acute stress is the stress that lasts for a short time when an individual's capabilities are sufficient to meet the challenge [3]. Negative or chronic stress is the stress that lasts for a long time when a challenge exceeds an individual's capabilities [4]. At some point in life, every individual is exposed to a stressful scenario and will react accordingly. If an individual can cope with stressful scenarios, the next time a similar scenario arises, the individual won't have as much of a stressful impact [5]. Similarly, if an individual cannot cope with a stressful situation and is repeatedly exposed to a similar situation, the individual will develop chronic stress [6]. Each time the body encounters a stressful scenario, the brain triggers the stress response to visual input from the ears, nose, and eyes. This response is known as "fight-or-flight" [7]. Instantly, the hypothalamus receives a distress signal from the brain. The hypothalamus

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

is the brain's command center. The hypothalamus regulates involuntary body activities through the Autonomic Nervous system (ANS) [8].

During stress, most organs are controlled by ANS without human knowledge [9]. The ANS is divided into two main divisions: the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS). The stress response is controlled by the complementary interaction of SNS and PNS in different physiological conditions [10]. The SNS initiates the fight or flight stress response, which results in a series of changes, including physiological, behavioural, and so on [11]. On the other hand, the PNS plays an essential role in reducing stress responses in individuals by suppressing the SNS [12]. The initial symptoms that emerge from a stressful scenario are called acute stress reactions [13]. These symptoms are visible within minutes during a stressful scenario and settle down quickly. Sweating, difficulty breathing, palpitations, nausea, chest pain, headaches, etc., are the physical symptoms of acute stress reactions [14]. If the symptoms last longer, they will cause chronic stress reactions. Depression, anxiety, memory loss, heart attack, stroke, high blood pressure, cholesterol, ulcer, weight loss, shortness of breath, weak immune system, etc., are the long-term health effects linked to chronic stress [15]. Because of the negative impacts of stress, it is crucial to build an effective stress detection system. A timely and accurate diagnosis of stress can improve an individual's life as productive, healthier, and happier [16].

Stress is detected through physiological, psychological, and behavioural markers [17]. Psychological interactions include increased negative feelings like anger, anxiety, depression, etc [18]. Self-report questionnaires or an examination with a psychologist are used to conduct a psychological assessment of stress. The disadvantage of such assessments is that they are only performed once the affected person or those around them recognize the intensity of the stress, which is usually too late [19]. In as short as 24 hours, people can experience memory lapses regarding the day's emotional mood, which lead to inaccurate stress level measurements using self-reports or questionnaires [20]. An individual's behaviour is affected by stress. Emotions like irritation, anger, sadness, etc., are the resulting changes. But they are hard to measure, as individuals can hide these emotions [21]. Physiological signals can reveal an individual's inner affect's strength and quality without any manipulations [12]. These physiological changes are non-voluntary responses that are difficult to notice externally. Hence, hormone monitoring is widely considered reliable for assessing stress [22].

The physiological aspects have several distinct advantages, like reliability, simplicity, continuous readings, cost-effectiveness, user-friendliness, non-maskability, non-invasiveness, etc., which makes them popular among researchers for stress detection [23]. Common physiological signals for stress detection are EDA, electroencephalography (EEG), ECG, respiration pattern, electromyogram, skin temperature, blood pressure, etc. [24]. For most physiological signal-based stress detection research, ECG and EDA signals are widely used either separately or in combination [25]. The ECG signal determines the electrical activity of the heart. As the ANS directly affects the heart rate, there will be variations in the heart rate during stress [16]. The EDA signal determines the change in the electrical characteristics of the skin. During stress, the body sweats more, which leads to increased skin conductance [26].

An innovation that right away benefits society in healthcare is the growing application of machine learning (ML), deep learning (DL), and wearable technology [27]. For different tasks using physiological signals, ML or DL models are trained using benchmark physiological datasets [28]. Support Vector Machine (SVM), random forest, K-Nearest Neighbour, decision tree, linear discriminant analysis, etc., are common ML methods used for the study [29]. ML approaches are frequently employed and get state-of-the-art for most stress detection studies, whereas DL methods are less extensively used because of the need for large data [30]. CNN, Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), etc., are the DL algorithms commonly used for stress detection [31]. Recently, multimodal fusion using the DL approach was found effective for stress detection [32]. There are three different levels of multimodal fusion: early, late, and intermediate [33]. The early fusion method merges feature representations of each modality at the feature level and starts training. After being trained separately, the different models are integrated at the decision level in the late fusion technique [34]. Intermediate fusion begins training by fusing higher-level feature representations of each modality from independent modality models [35]. The multimodal fusion model learns the highly linked representation across multiple modalities simultaneously, which enhances the model's performance over unimodal approaches [36].

In the last few years, the popularity of CNN has significantly increased. CNN extracts the most discriminating characteristics from the data while learning from it. Recent studies have proven that CNN can generate statistically relevant results for various applications [37]. To link input layers to the output layer, a CNN model consists of several layers, including convolution, pooling, dense, etc. Deep CNN's several layers can encode various low, mid, and high-level features. The deep layers are used to learn high-level features, and the shallower layers are used to determine low-level features [38]. The availability of the most discriminating features is one of the most important factors for increased classification accuracy [39]. Furthermore, the probability of getting a high classification accuracy with just one conventional feature extraction method is relatively low. The model's performance could be better with more information. Researchers also point out that information loss in the network may increase as the layers increase [40]. Due to these reasons, in recent studies, feature fusion methodologies like hierarchical features at each level

are integrated and used for training. The relevant information can be retained, and information loss is minimized by this hierarchical feature fusion [41]. In most end-to-end CNN networks, the last convolution layer's feature maps, mainly global features without hierarchy features, serve as discriminative features. However, low and mid-level features from the initial layers have discriminative features. The model can learn more efficient quality-aware feature representation with the help of hierarchical features (low, mid, and high-level features) [42].

Recent research shows that integrating features are often more efficient than independent features. This motivates us to apply the concept of feature fusion to enhance the efficiency of CNN-based stress detection models. We propose a multimodal hierarchical CNN feature fusion model that uses complementary features from various layers to enhance the performance rate of stress detection models. To the best of our knowledge, the proposed methodology has not yet been systematically addressed for stress detection. Hence, this paper presents a multimodal hierarchical CNN feature fusion model for stress detection using EDA and ECG signals. Initially, frequency domain features from EDA and ECG frequency bands or the raw frequency domain data are given as input to the CNN model. Inspired by the effectiveness of hierarchical feature fusion on CNN from the literature's [38], [39], [40], [41], [42], we concatenate the high, mid, and low-level features from the convolutional layers of EDA and ECG separately to form a hierarchical feature set. Unlike single-level fusion, gradual fusion has shown better performance [43]. So, each hierarchical feature set is used for multimodal fusion using MMTM (gradual fusion). Finally, we perform late fusion on the classification probabilities of each modality. This study also explores the performance of the distinct combination of hierarchical features concatenation from the low, mid, and high-level features. The proposed method is examined on four standard datasets- CLAS [44], ASCERTAIN [45], MAUS [46], and WAUC [47]

The following four folds provide a summary of the major contributions of this study:

1) **Multimodal hierarchical CNN feature fusion:** The low, mid, and high-level features from the initial convolutional layers are concatenated separately for each modality, and multimodal fusion is performed on the hierarchical feature set using MMTM.
2) **Combinations of hierarchical features:** Examine the performance of the concatenated distinct combination of hierarchical features from the low, mid, and high-level.
3) **Raw data and frequency band feature:** Compare the effectiveness of the raw frequency domain data, and the features from the frequency bands.
4) **Generalization ability:** To ensure generalizability, the proposed stress detection model has been evaluated on four benchmark datasets- CLAS, ASCERTAIN, WAUC and MAUS.

Organization: The remainder of this paper is organized as follows. Section II examines recent works on hierarchical feature fusion and the identified research gap. Details of the proposed framework are provided in Section III. The experiment results are presented in Section IV and compared with existing works. The paper is concluded in Section V. We have defined the key terms used in this paper for better understanding and clarity. The list of abbreviations used in the paper is shown in Table 1.

TABLE 1. List of abbrevations.

Abbreviation	Explanation
Deep Learning	DL
Convolutional Neural Network	CNN
Electrodermal activity	EDA
Electrocardiogram	ECG
Multimodal Transfer Module	MMTM
Autonomic Nervous system	ANS
Sympathetic Nervous System	SNS
Parasympathetic Nervous System	PNS
Electroencephalography	EEG
Machine Learning	ML
Support Vector Machine	SVM
Recurrent Neural Networks	RNN
Long Short Term Memory	LSTM
Hierarchical Feature Aggregation	HFA
Multiscale Feature Aggregation	MFA
Synthetic Minority over-sampling Technique	SMOTE
Discrete Cosine Transform	DCT
Heart Rate Variability	HRV
Convolution	Conv
Batch Normalisation	BN
Max Pool	MP
Fully Connected	FC
t-distributed Stochastic Neighbour Embedding	tSNE

II. RELATED WORKS

Recently, hierarchical CNN feature fusion methods were frequently used in image classification tasks. An overview of such works and it's effectiveness is briefly discussed in this section.

In order to classify fruit diseases, Akram et al. [38] proposed a hierarchical pipeline for deep feature fusion and selection.Pre-trained models were used to extract deep features, which were then fine-tuned via transfer learning. Multi-level fusion was performed before feature selection. Fruit diseases were classified with Multi-SVM using the selected features from the plant village dataset [48]. The proposed method's efficiency was revealed in the classification results in terms of accuracy as 97.8%, sensitivity, G-measure and precision as 97.6%.

A face recognition algorithm with hierarchical feature fusion was proposed by Zhang et al. [41]. The proposed framework learned shallow and deep facial aspects using supervisory information. The features are combined to enhance face recognition efficiency in the face of occlusion and illumination. The visual geometry group network and lightened CNN are both altered using this method. The proposed approach provided significant recognition results in

both the AR face database [49] and the labelled faces in the wild [50] database.

In the wild images, blind quality assessment using hierarchical feature fusion was proposed by Sun et al. [42]. The features from the intermediate layers to the final feature representation were hierarchically integrated using a staircase structure. The proposed method allowed the model to fully use visual data at all levels, from low to high. An iterative mixed database training approach was proposed to train the model simultaneously on multiple datasets. The proposed model benefited from the additional training samples and the capacity to learn a more generic feature representation. Experiments were conducted on six real-world image quality assessment datasets, and the results revealed that the proposed model performed significantly better than other state-of-the-art models.

A multiple hierarchical feature fusion for an end-to-end steel surface flaw detection is presented by He et al. [51]. The developed method uses a baseline CNN to produce feature maps to attain good classification abilities at each level. A feature fusion network with multiple levels merges several hierarchical features into a single feature with more details. A region proposal network creates regions of interest based on these multilayer properties. The final detection results are generated for each ROI by a detector composed of a bounding box regressor and a classifier. A defect detection dataset called NEU-DET [52] is compiled to evaluate the proposed method. Using baseline networks, the proposed technique yields 74.8/82.3 mean average precision on the NEU-DET dataset.

A selective feature connection mechanism for concatenating CNN features from multiple layers is proposed by Du et al. [53]. A feature selector created by high-level features links low-level features to high-level features. The proposed method shows universal acceptance, superiority, and efficacy on various challenging computer vision tasks. Ma et al. [54] proposed a multi-layer feature fusion on CNN to classify satellite image scenes. Since combining feature maps of various scales is not practical, the proposed method first transforms each feature map to fit its dimensions. Instead of just the final convolution layer, two methods for fusion were created to combine feature maps of various layers, and these features were given to the next layer or a classifier. Empirical findings showed that the proposed methods perform efficiently on public datasets.

A multiscale and hierarchical feature aggregation network is proposed for segmenting medical images by Yamanakkanavar et al. [55]. Two modules for feature aggregation are used to effectively combine data across end-to-end network layers: Hierarchical Feature Aggregation (HFA) and Multiscale Feature Aggregation (MFA). To learn deeper fusions of the feature hierarchy, the HFA module blends the features iteratively and hierarchically, and the MFA module gradually accumulates features and enriches feature representation. Having a 0.97 average accuracy score on the UFBA-UESC, PH2, and ISIC-2018 datasets [56], [57], [58], it is noted that the suggested model outperforms conventional methods for skin-lesion segmentation in terms of segmentation performance.

Li et al. [59] proposed a hierarchical feature aggregation network for deep image compression. Two approaches—inter and intra-stage feature aggregation—are put forth. Incorporating multiscale data into the inter-stage feature aggregation results in the production of more contextual features. To enhance representations of a single resolution, intra-stage aggregation joins features from the same stage. According to extensive experiments, the proposed method outperformed SOA methods, showing its effectiveness.

For robust cross-resolution face recognition, a representation learning method using a hierarchical deep CNN feature set is proposed by Gao et al. [60]. The proposed approach adaptively fuses the contextual features from different layers to learn more reliable and discriminative features. A feature set-based representation learning technique was developed to collectively describe the hierarchical features for improved recognition to exploit contextual information effectively. The hierarchical recognition outputs from several phases are combined to enhance recognition performance. Experimental results on several face datasets have proved the efficiency of the proposed approach.

In light of the studies above, hierarchical feature fusion and multimodal feature fusion effectively enhance model performance. Recent research on hierarchical CNN feature fusion has also demonstrated the superiority of fusion features over individual features. However, most hierarchical CNN feature fusion-based experiments are conducted only on image-based tasks, and other modalities have received less attention. This inspired us to propose a multimodal hierarchical CNN feature fusion using physiological signals. We aim to take advantage of both hierarchical feature fusion and multimodal feature fusion. We intend to benefit from the complementarity between high-level and low-level features through hierarchical feature fusion. By implementing multimodal fusion at the intermediate level, we intend to enhance the efficiency of the stress detection model. Hence, we propose a hierarchical CNN feature fusion for stress detection using EDA and ECG signals.

III. METHODOLOGY

Figure 1 depicts the novelty of this study on hierarchical feature fusion. Figure 1-(a) shows the traditional end-to-end deep learning approach. Features from the very last layer are only used as the identification feature in end-to-end networks. These features are frequently more general features without using hierarchical features. For this reason, we built a hierarchical feature learning model for stress detection using physiological signals. As shown in Figure 1-(b), we combined deep and shallow features to suit a hierarchical feature set. Later, these hierarchical features are used for multimodal fusion. We first describe the datasets used for this study in the following subsections. In the following subsections, we first give details about the datasets used for this study. Then we

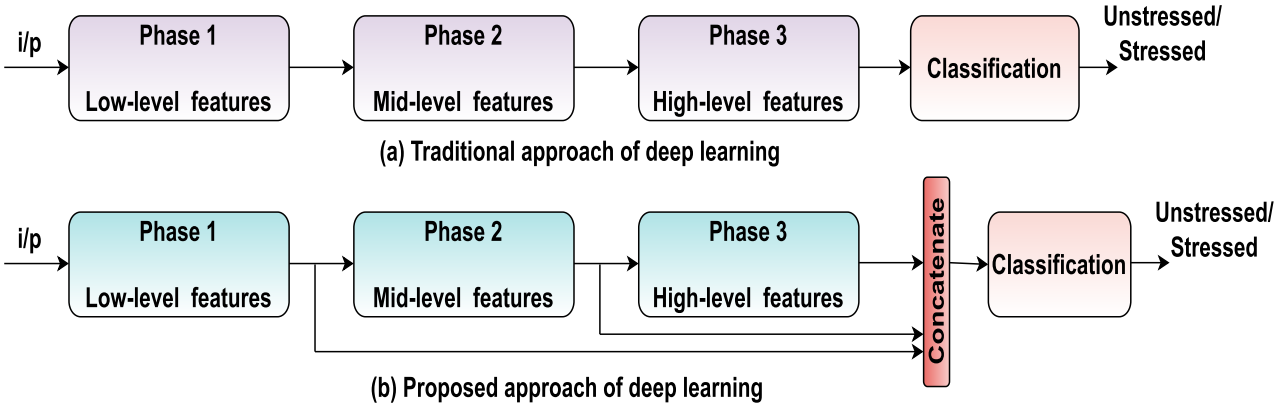


FIGURE 1. Traditional and proposed deep learning techniques are depicted in (a) and (b), respectively.

go into detail about multimodal hierarchical CNN feature fusion's architecture and feature extraction.

A. DATASET DETAILS

This research makes use of the following four benchmark datasets-ASCERTAIN [45], CLAS [44], MAUS [46], WAUC [47], which contain multimodal physiological signals such as ECG and EDA. A detailed explanation of each dataset is given below.

1) ASCERTAIN

The dataset contains 58 subjects physiological signals and face activity recordings. The physiological signals of the subjects were captured while they watched emotional video clips. Emotional video clips of 36 from [61] were used in the study. Based on the previous studies of stress detection using the ASCERTAIN dataset [62], we also used subjective ratings of valence and arousal for stress labelling. In the 2-D valence arousal plane, high arousal values along with low valence values are considered stressed and others as unstressed [63]. The average of the valence and arousal scores are used to decide whether it's high or low [45].

2) CLAS

The dataset contains 62 subjects' physiological data. Emotional video clips were used to evoke the subject's physiological signals. Emotional video clips of 16 from [64] were used in the study. After removing those subjects that didn't have all the data, we were left with 59 subjects. Stress labels have been fixed using the stimulus annotations described in the dataset [44].

3) MAUS

The dataset has recorded physiological data under different cognitive circumstances. The N-back task was used on 22 participants to generate a cognitive load. At the start of the trial, there was a five-minute rest interval. The N-back task required the participant to recall the last N single number from rapidly displayed digits. Whenever a signal matched the N-th number before the stimulation number, the subject was asked to reply by touching the space bar on the computer keypad. After a short rest period, the N-back task with six testing cases was completed. The complexity of the task serves as the ground truth [46].

4) WAUC

The study included 48 subjects who did activities at three different levels of exercise. The speed of a non-rotating cycle or rowing machine was changed to manipulate physiological tasks. During the exercise, sensory signals were captured. The subject's responses to the NASA Task Load Index questionnaire were encoded into binary values. They are classified as high or low cognitive load using the mean score as a cutoff provided in the dataset. After removing those subjects that didn't have all the information, we were left with 45 subjects [47].

To increase the sample count, each signal (EDA/ECG) is split into five-second segments. Subject IDs were established for training and testing to ensure subject independence. The first 36,18, 43 and 42 subject samples from WAUC, MAUS, CLAS and ASCERTAIN datasets are used for training. The remaining 9 WAUC, 4 MAUS and 16 CLAS and ASCERTAIN subject samples are employed for the testing.

The class imbalance affects the CLAS, ASCERTAIN, WAUC, and MAUS datasets. Real-world datasets frequently have a class imbalance when one class has fewer samples than the other class [65]. For more than two decades, this has been a topic of interest. To solve this problem, continuous enhancement is carried out at the data level, algorithmic level, and through hybrid methods [66]. Sampling techniques have received more attention in the data-level approach to enhance classification performance. Undersampling and oversampling are two categories of sampling methods [67]. Since oversampling creates additional samples from the minority class to compensate for the lack of samples, it is the most effective technique among these [68]. One of the most popular techniques in the literature to generate these new samples is the Synthetic Minority over-sampling

Technique (SMOTE) [69], [70], [71]. It's based on the simple generation of data points on the line segment joining a randomly chosen data point, and one of its K-nearest neighbours was used to sample data from the minority class [72]. This strategy is widely used since it is pretty simple and works incredibly well in reality [73], [74]. We also used SMOTE to train data in line with the literature to address the class imbalance.

B. FEATURE EXTRACTION

The following subsections explain the frequency domain features of EDA and ECG on raw data and in the frequency bands.

1) RAW DATA

The Discrete Cosine Transform (DCT) converts the raw EDA and ECG dataset to the frequency domain. Using the DCT method, a signal can be broken down into essential frequency components [75]. The input signal is more specifically encoded in the DCT as a linear sequence of weighted basis functions connected to its frequency elements. The DCT is given as input to the model.

2) FREQUENCY BAND FEATURES

Based on previous research [76], [77], we have identified three main bands in the frequency spectrum for ECG, as follows:

1) 0.0–0.04 Hz Very-low-frequency band
2) 0.04–0.15 Hz Low-frequency band
3) 0.15–0.40 Hz High-frequency band

According to the literature [78], [79], we also noticed five main bands in the frequency spectrum of EDA

1) 0.05–0.15 Hz–band a
2) 0.15–0.25 Hz–band b
3) 0.25–0.35 Hz–band c
4) 0.35–0.45 Hz–band d
5) 0.45-0.50 Hz–band e

The low-frequency and high-frequency bands are impacted by ANS activities. Therefore, these bands' features will be useful for stress detection [80]. The power spectral density of the Heart Rate Variability (HRV) derived from each band of the ECG is calculated (using Welch's technique). The frequency module pyHRV [81] from the Python library is used for this purpose. We collected 51 frequency-domain measures from these PSDs, including a relative, absolute, peak, and so on. The full list of measures is presented in [81]. Each EDA's power spectral density band is calculated (using Welch's technique). We retrieved 40 statistical characteristics from these PSDs (5 bands with eight features each), including max, min, standard deviation, variance, skewness, kurtosis, median and min.

C. ARCHITECTURE DETAILS

The proposed architecture for stress detection is shown in Figure 2. Phases 1, 2 and 3 are the different levels of features (low, mid and high-level) from the convolutional layers. In each modality, hierarchical features from different levels of convolutional layers are concatenated and given as input to MMTM [43] for multimodal fusion. Multimodal feature information is combined, and the features are recalibrated using MMTM. MMTM makes advantage of the computationally efficient and light-weight squeeze and excitation block [33]. A joint representation is generated in the MMTM module by combining ECG and EDA hierarchical features. The joint representation is used to predict the excitation signals, as explained in [43]. For the excitation, two independent, fully connected layers are used for each modality. One fully connected layer uses ReLU activation, while the other uses sigmoid activation. The excitation output is multiplied by the original features of each modality, which we gave as input to the module.

Four convolution layers consist of filters 32, 64, 128, and 256 with 3×3 as kernel size and ReLu as activation function. Batch Normalisation (BN) and Max Pool (MP) layers are applied after the convolutional layers. The architecture is completed by fully connected FC1 and FC2 and a sigmoid output layer. The Adam optimizer is used for the model training, with the default learning rate and a batch size of 64. As the loss function, Binary Cross-Entropy is used. An early-stopping strategy is used to shorten the training period if after 30 epochs in a sequence the loss does not decrease. The maximum classification probabilities from each model are used to perform the late fusion.

Based on our previous study [82], we perform a multimodal hierarchical feature fusion on the highest performed feature band of ECG ((0.15–0.40 Hz–High-frequency band)) and EDA ((0.15–0.25 Hz–band b). For the experiments, the architecture follows the same as shown in Figure 2 excluding the max-pooling and the kernel size as 2×2.

IV. RESULTS AND DISCUSSION

The experimental findings are presented and discussed in this section. We ran our studies on raw data and frequency band features, since we considered their effects. The proposed model's performance is evaluated using accuracy and F1-score, as shown in equations 1 and 2. In our first set of experiments, we compared the performance of raw data against frequency band features. Results obtained from different concatenation combinations on ASCERTAIN, CLAS, MAUS, and WAUC datasets are shown in Table 2. In our second set of experiments, the performance of the highest-performing band features of the ECG and EDA on the proposed models using the WAUC dataset is shown in Table 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

$$F1score = \frac{TP}{TP + 1/2(FP + FN)} \times 100 \quad (2)$$

TP, FP, TN, and FN are True Positive, False Positive, True Negative and False Negative.

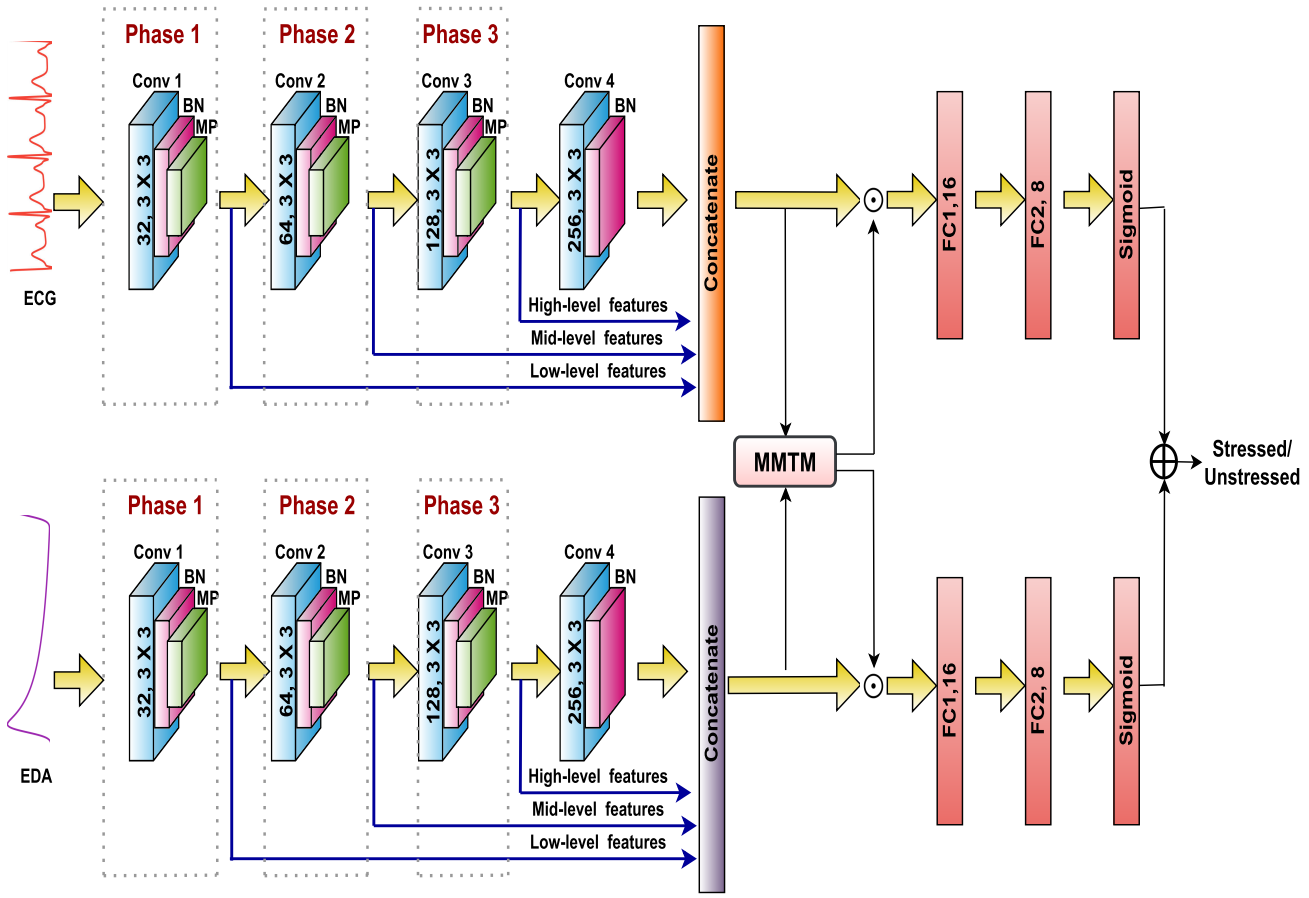


FIGURE 2. Architecture details of the proposed multimodal hierarchical CNN feature fusion framework. A hierarchical feature set made up of low, mid, and high-level features is created and used as input to the MMTM for multimodal fusion. Late fusion is used to categorise subjects as stressed or unstressed after features are recalibrated using MMTM.

A. MULTIMODAL HIERARCHICAL CNN FEATURE FUSION

Hierarchical feature fusion and multimodal fusion are the two fundamental processes that compose the proposed multimodal hierarchical CNN feature fusion model. Convolutional layers encode information at various levels using different layers; for hierarchical feature fusion, we use this concept. As shown in Table 2 and Table 3, the result shows the effectiveness of the proposed methodology because the complementarity between low-level information and high-level information is completely utilized by our efficient hierarchical feature fusion method. It also proves that, besides the hierarchical feature fusion, the multimodal fusion on gradual level and decision level helped to enhance the model's performance by learning across modalities. The promising results suggest that clinical practitioners can use the proposed model for stress detection.

B. DIFFERENT COMBINATIONS OF HIERARCHICAL FEATURES

We performed the proposed multimodal CNN feature fusion on all hierarchical CNN feature fusion combinations. This experimentation phase is essential to show that the proposed model is stable and to identify the best hierarchical CNN feature combination. The features of each convolutional layer are utilized in the proposed architecture to extract the shallow, intermediate, and deep features. As shown in Table 2 and Table 3, first, when comparing the performance from phase1 to phase 1,2 and 3 (concatenation combinations), we observe a consistent increase in performance as the features extracted from phase 1 to phase 3 are added in sequence to the model. Concatenating all level features (phases 1, 2, and 3) enhanced the model's overall performance more than other combinations (phases 1, 2, and 3 alone and its combinations) by 12-15% on raw data, 9-15% on band features and 12.15% on highest band features of ECG and EDA on WAUC dataset. This proves that shallow features are also important to end-to-end networks, along with deep features, and the features extracted from all stages make contributions to enhance the model's performance.

C. RAW DATA AND FREQUENCY BAND FEATURES

We compared the performance of raw data and frequency band features on the proposed model. In the overall study, it is observed that the frequency domain features retrieved from

TABLE 2. Classification results.

S.No	Concatenation combination	Raw data		Band features	
		Accuracy	F1-Score	Accuracy	F1-Score
		ASCERTAIN			
1	Phase 1	79.32%	0.79	84.16%	0.83
2	Phase 2	81.86%	0.82	87.43%	0.87
3	Phase 3	83.14%	0.83	88.75%	0.89
4	Phase 1 and 2	86.25%	0.85	91.38%	0.90
5	Phase 2 and 3	90.63%	0.91	95.23%	0.94
6	Phase 1 and 3	88.76%	0.88	92.85%	0.92
7	**Phase 1,2 and 3**	**93.58%**	**0.93**	**97.61%**	**0.97**
		CLAS			
8	Phase 1	76.94%	0.77	81.83%	0.81
9	Phase 2	79.51%	0.79	83.25%	0.82
10	Phase 3	81.63%	0.82	86.47%	0.85
11	Phase 1 and 2	83.89%	0.84	88.13%	0.87
12	Phase 2 and 3	89.43%	0.89	92.65%	0.93
13	Phase 1 and 3	86.04%	0.86	90.92%	0.91
14	**Phase 1,2 and 3**	**91.32%**	**0.91**	**95.94%**	**0.94**
		MAUS			
15	Phase 1	70.69%	0.71	74.21%	0.74
16	Phase 2	73.22%	0.73	77.73%	0.77
17	Phase 3	75.06%	0.75	79.48%	0.79
18	Phase 1 and 2	78.45%	0.79	80.67%	0.81
19	Phase 2 and 3	82.72%	0.83	85.36%	0.86
20	Phase 1 and 3	80.92%	0.81	83.14%	0.82
21	**Phase 1,2 and 3**	**85.18%**	**0.84**	**88.75%**	**0.88**
		WAUC			
22	Phase 1	69.46%	0.69	75.32%	0.75
23	Phase 2	71.92%	0.72	76.81%	0.77
24	Phase 3	73.53%	0.73	77.64%	0.78
25	Phase 1 and 2	75.39%	0.75	79.18%	0.79
26	Phase 2 and 3	78.60%	0.79	82.52%	0.83
27	Phase 1 and 3	77.16%	0.77	81.49%	0.81
28	**Phase 1,2 and 3**	**81.74%**	**0.81**	**83.86%**	**0.84**

TABLE 3. Classification results WAUC dataset on highest performed band of ECG and EDA.

S.No	Concatenation combination	Accuracy	F1-Score
1	Phase 1	67.16%	0.66
2	Phase 2	68.93%	0.69
3	Phase 3	70.67%	0.70
4	Phase 1 and 2	72.49%	0.72
5	Phase 2 and 3	77.50%	0.76
6	Phase 1 and 3	75.82%	0.74
7	**Phase 1,2 and 3**	**79.31%**	**0.79**

the EDA and ECG frequency bands influenced more for the performance enhancement of the model more than raw data. As shown in Table 2, in all the datasets, we have observed the same shift in the performance of raw data and frequency band features by 2-4%, respectively. We also made a performance comparison on the highest performed band features of EDA and ECG with the latest dataset-WAUC. As shown in Table 3, the results were not encouraging compared to the whole frequency band features. This suggests that features across the entire frequency band influence performance enhancement more than the highest-performing EDA and ECG band features.

D. GENERALIZATION ABILITY

Applying DL techniques in the healthcare industry has several benefits, especially when it comes to predictive modelling. The validity and generalizability of a model are being given more consideration as the development of DL-based models continues to advance. In the healthcare industry, this is particularly important because algorithmic results directly impact patient treatment and clinical judgement. We proposed a subject-independent multimodal hierarchical CNN feature fusion stress detection model. Four benchmark datasets gathered from four separate scenarios are used to validate and examine the generalizability of the proposed methodology. As shown in Table 2, the results prove that the presented framework does not overfit a dataset obtained in a specific setting. In all four datasets, we observed a similar performance shift.

E. T-SNE VISUALIZATION

In DL, we keep seeking data insights; to achieve that, we visualize the data. To visualize the impact of the proposed models, we qualitatively evaluate the proposed hierarchical fusion strategy with the network's feature visualization. This part uses t-distributed stochastic neighbor embedding (tSNE) to assess the network's visual cognition. Features from the FC-16 layers of the frequency band of ECG modality are taken and used for visualization. It is evident from figure 3 that the hierarchical features of the ECG following multimodal fusion are discriminatory enough to classify stressed

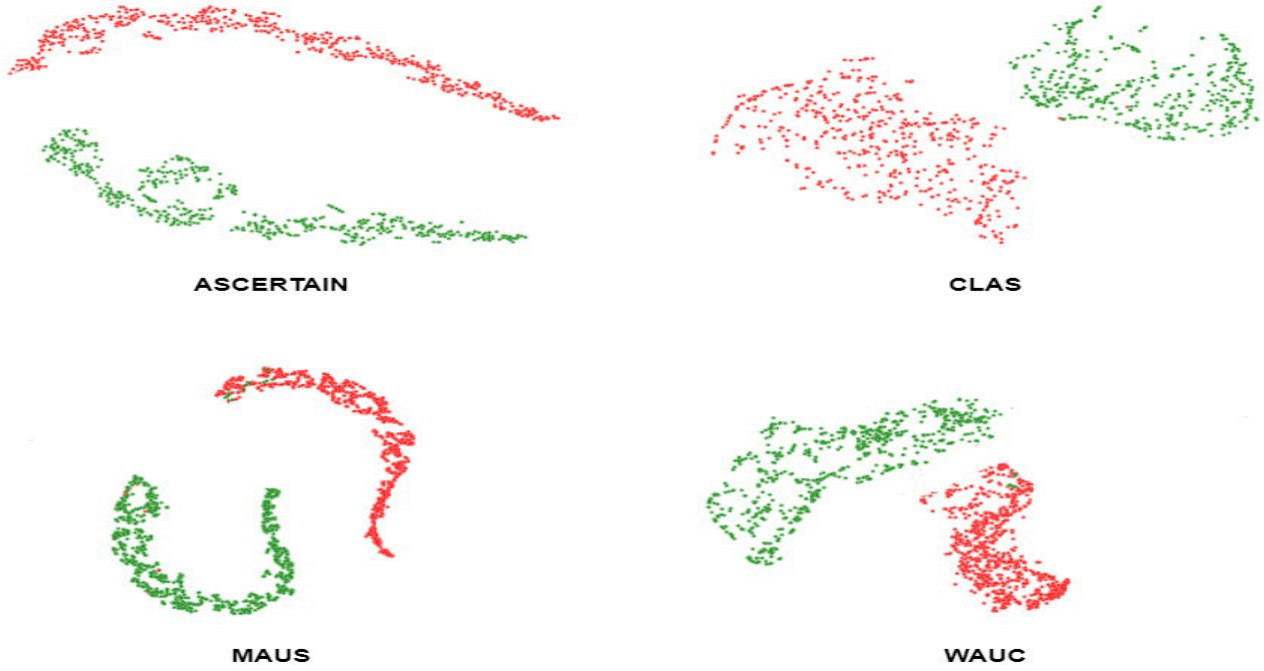


FIGURE 3. tSNE visualization of ECG features from the frequency band of all the datasets. Here, red dots indicate the stress class, and green dots indicate the unstressed class. A clear separation between the two classes is visible in all the datasets.

and unstressed. We can see similar clusters in all the datasets. The plot demonstrates the groups created based on similarity, illustrating the potential capability of the proposed approach for stress detection. The critical t-SNE visualization map's highlighted a distinct separation between stressed and unstressed conditions. We have also noticed similar clusters for EDA frequency band features.

F. COMPARISON STUDY

Nowadays, a critical healthcare challenge is the quick and precise diagnosis of stress. Accurate stress detection is a challenge that has been addressed using various techniques. The traditional DL and ML approaches have shown to be the most successful. This work mainly focused on detecting stress using physiological signals–ECG and EDA using DL. We proposed an efficient multimodal hierarchical CNN feature fusion model for stress detection and compared its performance with several classical ML and DL techniques.

This section analyzes the proposed method's findings with those of existing stress detection research using the four datasets. Table 4 shows a summary of the performance measures. Only a few studies have used the most recent datasets, WAUC and MAUS, in their analyses. Existing works show that the majority of the works are carried on time-frequency domain [44], [62], [83], [84], [85], [86], [87], [88], subject dependent [44], [62], [77], [83], [84], [89], and using machine learning models [44], [62], [83], [84], [89], [90]. Few researchers used traditional deep learning techniques [82], [85], [86], [87], [88]. Compared with the existing works, our work focused on utilizing the full features of an end-to-end network, not only on the last layer features. The proposed approach performs better than all the reported state-of-the-art subject-independent and subject-dependent studies, except for the CLAS dataset. The results of our predictions confirm that our multimodal hierarchical

TABLE 4. Comparison to state-of-art findings.

S.No	Method	Accuracy	F1-Score	AUC
		ASCERTAIN		
1	**Proposed**	**97.6%**	**0.97**	**0.96**
2	[62]	68.7%*	-	-
3	[83]	68%*	-	-
4	[84]	68.7%*	-	-
5	[85]	78.8%	0.79	0.78
6	[86]	75.5%	0.76	0.75
7	[87]	85%	0.85	0.85
8	[88]	89.9%	0.90	0.89
9	[82]	96.2%	0.95	0.96
		CLAS		
10	Proposed	95.9%	0.94	0.94
11	[44]	88.9%*		-
12	[89]	94.8%*		-
13	**[90]**	**97.6%***		**-**
14	[85]	72.6%	0.74	0.73
15	[86]	69.9%	0.70	0.69
16	[87]	80.3%	0.82	0.81
17	[88]	86.1%	0.87	0.86
18	[82]	94.2%	0.93	0.93
		MAUS		
19	**Proposed**	**88.7%**	**0.88**	**0.87**
20	[46]	71.6 ± 11.1%	71.9 ± 10.9%	-
21	[82]	86.5%	0.86	0.85
		WAUC		
22	**Proposed**	**83.8%**	**0.84**	**0.83**
23	[47]	-	-	EDA- 0.66 ± 0.01 ECG- 0.74 ± 0.01
24	**[82]**	82.6%	0.83	0.82

*Subject-dependent

feature fusion model is highly effective for detecting stress in a subject-independent way.

V. CONCLUSION

This paper presents a multimodal hierarchical CNN feature fusion for stress detection. EDA and ECG signals raw data and frequency domain features are used utilized in this study. For identification tasks, the convolutional layer's shallow feature as well as deep feature are useful. We integrate the features on each phase of the end-to-end network to increase efficiency and better utilise the retrieved features in each phase. Low, mid, and high-level features of convolutional layers are concatenated to obtain different combinations, and multimodal fusion is conducted on each hierarchical feature set. Additionally, the combination of features can more effectively convey the characteristics of the physiological signals. The proposed approach is tested on four benchmark datasets - ASCERTAIN, CLAS, MAUS and WAUC. Experimental results show that the proposed approach outperforms previous studies in terms of stress detection in a subject independent manner. Among the different combinations, concatenating all the phases (low, mid and high-level features) yields optimal performance. The proposed approach to feature fusion is a general one that works well in end-to-end networks. To enhance the ability of feature extraction in neural networks, we can use the end-to-end networks deep, medium and shallow features of end-to-end networks and perform feature integration. Thus, in the future, we intend to: (i) expand the studies on hierarchical feature fusion and iidifferent multi-modal fusion techniques on hierarchical features.

REFERENCES

[1] G. Russell and S. Lightman, "The human stress response," *Nature Rev. Endocrinol.*, vol. 15, no. 9, pp. 525–534, 2019.
[2] G. S. Everly and J. M. Lating, "The anatomy and physiology of the human stress response," in *A Clinical Guide to the Treatment of the Human Stress Response*. New York, NY, USA: Springer, 2019, pp. 19–56.
[3] N. Rohleder, "Stress and inflammation—The need to address the gap in the transition between acute and chronic stress effects," *Psychoneuroendocrinology*, vol. 105, pp. 164–171, Jul. 2019.
[4] A. P. Cruz, A. Pradeep, K. R. Sivasankar, and K. S. Krishnaveni, "A decision tree optimised SVM model for stress detection using biosignals," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Jul. 2020, pp. 0841–0845.
[5] S. S. Machiraju, N. Konijeti, A. Batchu, and N. Tata, "Stress detection using adaptive threshold methodology," in *Proc. 5th Int. Conf. Commun. Electron. Syst. (ICCES)*, Jun. 2020, pp. 889–894.
[6] K. A. Demin, A. S. Taranov, N. P. Ilyin, A. M. Lakstygal, A. D. Volgin, M. S. de Abreu, T. Strekalova, and A. V. Kalueff, "Understanding neurobehavioral effects of acute and chronic stress in zebrafish," *Stress*, vol. 24, no. 1, pp. 1–18, Jan. 2021.
[7] R. McCarty, "The fight-or-flight response: A cornerstone of stress research," in *Stress: Concepts, Cognition, Emotion, and Behavior*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 33–37.
[8] A. Agorastos and G. P. Chrousos, "The neuroendocrinology of stress: The stress-related continuum of chronic disease development," *Mol. Psychiatry*, vol. 27, no. 1, pp. 502–513, Jan. 2022.
[9] L. Becker, H. C. Kaltenegger, D. Nowak, N. Rohleder, and M. Weigl, "Differences in stress system (re-)activity between single and dual-or multitasking in healthy adults: A systematic review and meta-analysis: Physiological stress and multitasking," *Health Psychol. Rev.*, pp. 1–45, 2022.
[10] J. Suurland, K. B. van der Heijden, S. C. J. Huijbregts, S. H. M. van Goozen, and H. Swaab, "Infant parasympathetic and sympathetic activity during baseline, stress and recovery: Interactions with prenatal adversity predict physical aggression in toddlerhood," *J. Abnormal Child Psychol.*, vol. 46, no. 4, pp. 755–768, May 2018.
[11] B. Chu, K. Marwaha, T. Sanvictores, and D. Ayers, "Physiology, stress reaction," in *StatPearls [Internet]*. StatPearls Publishing, 2021.
[12] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consum. Electron.*, vol. 64, no. 2, pp. 196–203, May 2018.
[13] C. Regehr and V. R. LeBlanc, "PTSD, acute stress, performance and decision-making in emergency service workers," *J. Amer. Acad. Psychiatry Law*, vol. 45, no. 2, pp. 184–192, 2017.
[14] L. A. Wright, M. Sijbrandij, R. Sinnerton, C. Lewis, N. P. Roberts, and J. I. Bisson, "Pharmacological prevention and early treatment of post-traumatic stress disorder and acute stress disorder: A systematic review and meta-analysis," *Transl. Psychiatry*, vol. 9, no. 1, pp. 1–10, Dec. 2019.
[15] S. J. Lupien, R.-P. Juster, C. Raymond, and M.-F. Marin, "The effects of chronic stress on the human brain: From neurotoxicity, to vulnerability, to opportunity," *Frontiers Neuroendocrinol.*, vol. 49, pp. 91–105, Apr. 2018.
[16] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: In laboratory and real life," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*, 2016, pp. 1185–1193.
[17] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," *Biocybern. Biomed. Eng.*, vol. 39, no. 2, pp. 444–469, Apr. 2019.
[18] E. Turcan, S. Muresan, and K. McKeown, "Emotion-infused models for explainable psychological stress detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 2895–2909.
[19] A. Teilegen, "Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report," in *Anxiety and the Anxiety Disorders*. Evanston, IL, USA: Routledge, 2019, pp. 681–706.
[20] J. Rodríguez-Arce, L. Lara-Flores, O. Portillo-Rodríguez, and R. Martínez-Méndez, "Towards an anxiety and stress recognition system for academic environments based on physiological features," *Comput. Methods Programs Biomed.*, vol. 190, Jul. 2020, Art. no. 105408.
[21] M. Granovetter, "The sociological and economic approaches to labor market analysis: A social structural view," in *Industries, Firms, and Jobs*, 2017, pp. 187–216.
[22] 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.
[23] 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, 2018.
[24] S. Pourmohammadi and A. Maleki, "Stress detection using ECG and EMG signals: A comprehensive study," *Comput. Methods Programs Biomed.*, vol. 193, Sep. 2020, Art. no. 105482.
[25] P. Zontone, A. Affanni, R. Bernardini, A. Piras, and R. Rinaldo, "Stress detection through electrodermal activity (EDA) and electrocardiogram (ECG) analysis in car drivers," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.
[26] R. Sánchez-Reolid, M. T. López, and A. Fernández-Caballero, "Machine learning for stress detection from electrodermal activity: A scoping review," 2020.
[27] 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.
[28] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Comput. Methods Programs Biomed.*, vol. 161, pp. 1–13, Jul. 2018.
[29] S. Elzeiny and M. Qaraqe, "Machine learning approaches to automatic stress detection: A review," in *Proc. IEEE/ACS 15th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2018, pp. 1–6.
[30] C.-Y. Liao, R.-C. Chen, and S.-K. Tai, "Emotion stress detection using EEG signal and deep learning technologies," in *Proc. IEEE Int. Conf. Appl. Syst. Invention (ICASI)*, Apr. 2018, pp. 90–93.
[31] 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.

[32] M. N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, and V. Chandran, "Automatic driver stress level classification using multimodal deep learning," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112793.

[33] F. Liu, J. Chen, W. Tan, and C. Cai, "A multi-modal fusion method based on higher-order orthogonal iteration decomposition," *Entropy*, vol. 23, no. 10, p. 1349, Oct. 2021.

[34] C. I. Patel, S. Garg, T. Zaveri, A. Banerjee, and R. Patel, "Human action recognition using fusion of features for unconstrained video sequences," *Comput. Electr. Eng.*, vol. 70, pp. 284–301, Aug. 2018.

[35] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Mach. Vis. Appl.*, vol. 32, no. 6, pp. 1–18, Nov. 2021.

[36] J. Aigrain, M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani, "Multimodal stress detection from multiple assessments," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 491–506, Oct. 2018.

[37] L. K. Singh and M. Khanna, "A novel multimodality based dual fusion integrated approach for efficient and early prediction of glaucoma," *Biomed. Signal Process. Control*, vol. 73, Mar. 2022, Art. no. 103468.

[38] M. A. Khan, T. Akram, M. Sharif, and T. Saba, "Fruits diseases classification: Exploiting a hierarchical framework for deep features fusion and selection," *Multimedia Tools Appl.*, vol. 79, nos. 35–36, pp. 25763–25783, Sep. 2020.

[39] M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, S. Hussain, A. Khosravi, U. Rajendra Acharya, V. Makarenkov, and S. Nahavandi, "UncertaintyFuseNet: Robust uncertainty-aware hierarchical feature fusion model with ensemble Monte Carlo dropout for COVID-19 detection," 2021, *arXiv:2105.08590*.

[40] X. Li, D. Song, and Y. Dong, "Hierarchical feature fusion network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 9165–9175, 2020.

[41] J. Zhang, X. Yan, Z. Cheng, and X. Shen, "A face recognition algorithm based on feature fusion," *Concurrency Comput., Pract. Exper.*, vol. 34, no. 14, p. e5748, Jun. 2022.

[42] W. Sun, X. Min, G. Zhai, and S. Ma, "Blind quality assessment for in-the-Wild images via hierarchical feature fusion and iterative mixed database training," 2021, *arXiv:2105.14550*.

[43] H. R. Vaezi Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13289–13299.

[44] 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.

[45] 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.

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

[47] 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, Art. no. 549524.

[48] D. P. Hughes and M. Salathe, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," 2015, *arXiv:1511.08060*.

[49] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in Unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images: Detection, Alignment, Recognit.*, 2008, pp. 1–15.

[50] A. Martinez and R. Benavente, "The ar face database: CVC," Tech. Rep. 24, 1998.

[51] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1493–1504, Apr. 2020.

[52] K. Song and Y. Yan, "A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects," *Appl. Surf. Sci.*, vol. 285, no. 21, pp. 858–864, Nov. 2013.

[53] C. Du, Y. Wang, C. Wang, C. Shi, and B. Xiao, "Selective feature connection mechanism: Concatenating multi-layer CNN features with a feature selector," *Pattern Recognit. Lett.*, vol. 129, pp. 108–114, Jan. 2020.

[54] C. Ma, X. Mu, and D. Sha, "Multi-layers feature fusion of convolutional neural network for scene classification of remote sensing," *IEEE Access*, vol. 7, pp. 121685–121694, 2019.

[55] N. Yamanakkanavar, J. Y. Choi, and B. Lee, "Multiscale and hierarchical feature-aggregation network for segmenting medical images," *Sensors*, vol. 22, no. 9, p. 3440, Apr. 2022.

[56] G. Silva, L. Oliveira, and M. Pithon, "Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives," *Expert Syst. Appl.*, vol. 107, pp. 15–31, Oct. 2018.

[57] T. Mendonca, M. Celebi, T. Mendonca, and J. Marques, "Ph2: A public database for the analysis of dermoscopic images," in *Dermoscopy Image Analysis*, 2015.

[58] N. Codella, V. Rotemberg, P. Tschandl, M. Emre Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[59] W. Li, Z. Du, H. He, J. Tang, and G. Wu, "Hierarchical feature aggregation network for deep image compression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1875–1879.

[60] G. Gao, Y. Yu, J. Yang, G.-J. Qi, and M. Yang, "Hierarchical deep CNN feature set-based representation learning for robust cross-resolution face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2550–2560, May 2022.

[61] 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.

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

[63] 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.*, vol. 13, no. 2, pp. 1044–1056, Apr. 2022.

[64] 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, Jan./Mar. 2011.

[65] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, pp. 1–30, 2018.

[66] S. Sharma, A. Gosain, and S. Jain, "A review of the oversampling techniques in class imbalance problem," in *Proc. Int. Conf. Innov. Comput. Commun.* Singapore: Springer, 2022, pp. 459–472.

[67] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci.*, vol. 505, pp. 32–64, Dec. 2019.

[68] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access*, vol. 4, pp. 7940–7957, 2016.

[69] 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, pp. 321–357, Dec. 2002.

[70] N. Noorhalim, A. Ali, and S. M. Shamsuddin, "Handling imbalanced ratio for class imbalance problem using SMOTE," in *Proc. 3rd Int. Conf. Comput., Math. Statist. (iCMS)*. Singapore: Springer, 2019, pp. 19–30.

[71] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021.

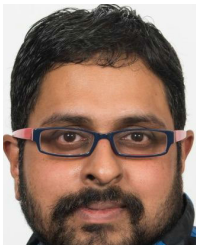
[72] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, Dec. 2021.

[73] A. Özdemir, K. Polat, and A. Alhudhaif, "Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods," *Expert Syst. Appl.*, vol. 178, Sep. 2021, Art. no. 114986.

[74] M. Shuja, S. Mittal, and M. Zaman, "Effective prediction of type II diabetes mellitus using data mining classifiers and SMOTE," in *Advances in Computing And Intelligent Systems*. Singapore: Springer, 2020, pp. 195–211.

[75] S. Banerjee and G. K. Singh, "A new approach of ECG steganography and prediction using deep learning," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102151.
[76] 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.
[77] R. Rakshit, V. R. Reddy, and P. Deshpande, "Emotion detection and recognition using HRV features derived from photoplethysmogram signals," in *Proc. 2nd Workshop Emotion Represent. Modeling Companion Syst.*, Nov. 2016, pp. 1–6.
[78] 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.
[79] 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.
[80] J. Cui, U. A. Leuenberger, F. Aziz, J. C. Luck, J. Stavres, D. J. Kim, Z. Gao, C. Blaha, A. Cauffman, and L. I. Sinoway, "Autonomic responses to acute hyperoxia are impaired in patients with peripheral artery disease," *FASEB J.*, vol. 36, no. S1, May 2022.
[81] 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.*, 2019, pp. 822–828.
[82] K. Radhika, R. Subramanian, and V. R. M. Oruganti, "Joint modality features in frequency domain for stress detection," *IEEE Access*, vol. 10, pp. 57201–57211, 2022.
[83] 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.
[84] V. Markova and T. Ganchev, "Constrained attribute selection for stress detection based on physiological signals," in *Proc. Int. Conf. Sensors, Signal Image Process.*, Oct. 2018, pp. 41–45.
[85] 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.
[86] 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.
[87] K. Radhika and V. R. M. Oruganti, "Stress detection using CNN fusion," in *Proc. IEEE Region 10 Conf. (TENCON)*, Dec. 2021, pp. 492–497.
[88] K. Radhika and V. R. M. Oruganti, "Cross domain features for subject-independent stress detection," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jul. 2022, pp. 1–6.
[89] 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.
[90] 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.

RADHIKA KUTTALA 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. His past affiliations include IHPC, Singapore; University of Glasgow, Singapore; IIIT Hyderabad, India; IIT Ropar, India; and UIUC-ADSC, Singapore. He is currently an Associate Professor with the University of Canberra, Australia. His research interests include human-centered computing, interactive analytics, and explainable machine learning. He is a Senior Member of ACM and AAAC.

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