MEmoR: A Multimodal Emotion Recognition using Affective Biomarkers for Smart Prediction of Emotional Health for People Analytics in Smart Industries

Akshi Kumar[1], Kapil Sharma[2], Aditi Sharma [3 *]
[1] Dept. of Computing & Mathematics, Manchester Metropolitan University, Manchester, UK
[2] Dept. of Information Technology, Delhi Technological University, Delhi, India
[3] Dept. of Computer Science & Engineering, Delhi Technological University, Delhi, India
{***Corresponding author:*** *aditisharma9420@gmail.com}

Abstract—The intersection of people, data and intelligent machines has a far-reaching impact on the productivity, efficiency and operations of a smart industry. Internet-of-things (IoT) offers a great potential for workplace gains using the "quantified self" and the computer vision strategies. Their goal is to focus on productivity, fitness, wellness, and improvement of the work environment. Recognizing and regulating human emotion is vital to people analytics as it plays an important role in the workplace productivity. Within the smart industry setting, various non-invasive IoT devices can be used to recognize emotions and study the behavioural outcomes in various situations. This research puts forward a deep learning model for detection of human emotional state in real-time using multimodal data from the Emotional Internet-of-things (E-IoT). The proposed multimodal emotion recognition model, MEmoR makes use of two data modalities: visual and psychophysiological. The video signals are sampled to obtain image frames and a ResNet50 model pre-trained for face recognition is fine-tuned for emotion classification. Simultaneously, a CNN is trained on the psychophysiological signals and the results of the two modality networks are combined using decision-level weighted fusion. The model is tested on the benchmark BioVid Emo DB multimodal dataset and compared to the state-of-the-art.

Index Terms— Affective Computing; Visual Analysis; Multi-model; Emotion Recognition; E-IoT; Facial Expression Analysis

1 INTRODUCTION

Smart industry (also called Industry 4.0) is the confluence of trends and technologies of automation and data exchange in manufacturing and related industries. In order to work efficiently, smart industries must support interconnection of wireless devices, sensors, and people through the Internet of Things (IoT) [1]. This intelligent networking of machines and processes enables new ways of production, value creation and real-time optimization. The benefits of Industry 4.0 include providing novel business models, services & products and increasing the operational efficiency (asset utilization, operational cost reduction & employee productivity).

IoT makes the human-machine interaction open to interminable virtual opportunities and connections that facilitate attaining collective goals in the manufacturing process. Typically, the components required to facilitate an efficacious & productive smart factory include data, technology, process, people, and security. A smart industry does not transform into a “dark” industry without workforce. Instead, the transitioning disrupts the status quo of human resources as people are expected to be the key to the radical knowledge-intensive operations. Essentially, IIoT solutions pave the way forward for more efficient regulation of safety standards to secure better working space that ensure physical well-being of employees. Simultaneously, IoT-enabled devices can alert management to worker fatigue, strain, or risk-taking behavior and assist monitoring & tracking the employee's mental well-being. Analyzing worker performance is a not new industry trend but is customarily done by the HR departments and the industrial psychologists [2]. But IoT technology is fundamentally changing this practice by implementing 'people analytics' that is a data-driven approach to manage people at work, identify conduct risk, quantify performance and enhance the workforce output.

Pertinent studies have shown that the employees’ on-job emotional and psychological conditions affect their workplace productivity [3]. Positive, healthy workplace environment can boost employee commitment, engagement, and performance to ultimately sustain the productivity of the business. Anger, frustration, fear, and other "negative emotions" can lead to crippling levels of stress, anxiety or depression [4]. These negative emotions can hamper employees' commitment, creativity, decision making, work quality, and workplace retention. It is therefore imperative to comprehend the on-job emotional state of employees automatically to improve the productivity, performance and working environment eventually benefitting the smart industry.

Automatic emotion recognition an essential requisite in smart industries so as to warrant and support the on-job well-being of employees. However, even though research in affective computing has gained momentum, emotion recognition in practical, real-time non-simulated environments is still challenging. With the advancements in healthcare and information technology, these affective biomarkers (which quantify the affective state of a user by measuring different changes in the body) can now be measured by having sensors in wearable devices like fitness watch or a chest wearable or with the help of implants such as pacemakers [5]. Fig.1 depicts the various affective biomarkers for real-time emotion recognition.

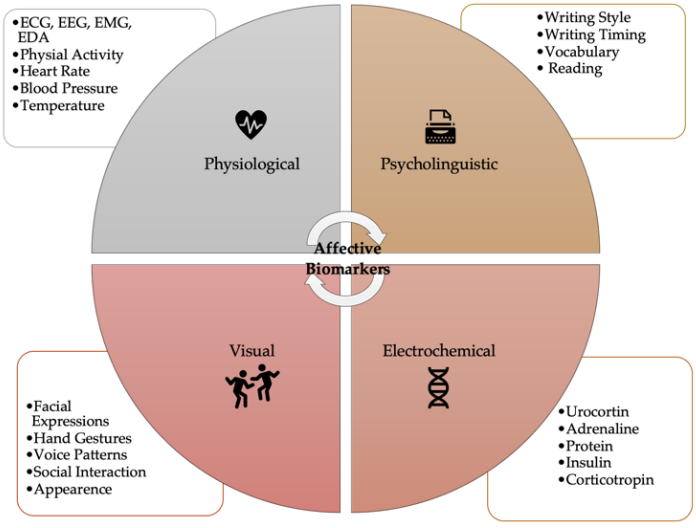


Fig.1. Affective Biomarkers for Emotion Recognition

Within the smart industry setting, non-invasive IoT devices and wearable devices (cameras, microphones, smart watch, smartphones, etc.) can be used to recognize emotions using multimodal (aural, visual, gestural, linguistic and bio-sensing) biomarkers and study behavioural outcomes in various situations. Collectively, referred to as Emotional Internet-of-Things (E-IOT), these sensors can facilitate detection of emotional state in real-time for early diagnosis of negative emotions and poor personal behavior [5]. It can be integrated into the manufacturing tasks or wearable device of employees as an application on smartphones. For example, with the help of AI, E-IOT can detect patterns indicative of depression and other mental illnesses in employees. Cameras can record images of employees at different times of the day and computer vision can extract behavioral patterns to compare them with those of depressed people to determine whether an employee is suffering from anxiety or depression [6]. If computer vision finds that an employee is not emotionally happy, then it can send signals to IoT devices that can alert HR personnel. HR personnel can intervene and help eliminate stress and negativity by organizing counseling sessions for that employee to improve his comfort at work thus creating a positive, smart workplace environment. Similarly, the wearable technology-enabled "quantified self" enormously empowers and allows employers to trade out unproductive and unhelpful behaviors.

Motivated by the need to build efficient, productive and safe industries that utilize connected devices to monitor and track employee health, in this research, we propose a deep learning model for multimodal emotion recognition, MEmoR. The proposed MEmoR model considers two modalities, namely, visual and physiological for the classification of emotional state of the subject (employee in case of industry) into one of five discrete categories - anger, sadness, fear, disgust, and amusement. The BioVidEmo multimodal database [7] is used to evaluate and validate the proposed MEmoR model. The dataset includes both bio-signals and video signals for human affective state recognition. The visual input is taken and sampled to obtain image frames, from

which the subject's face is extracted for facial expression recognition. A ResNet50 architecture pre-trained for face recognition is fine-tuned for emotion classification. The physiological signal data is used to train a convolution neural network (CNN) and detect the emotional state of an individual. As the data modalities contain information that are correlated to each other at a higher level and have varied data dimensionality, for combining the results of the two modalities, we have used a late data fusion technique. Thus, the primary contributions of this research are:

- A multimodal approach for emotion classification based on visual-physiological input from E-IOT for real-time industrial setting.
- Application of transfer learning for visual modality component to achieve generalizability and reduce computation costs.
- Multi-stage fine-tuning of pre-trained models in visual modality component using a unimodal emotion dataset followed by our target dataset.
- Identifying discrete emotions as well as valence -arousal affective dimensions of emotion.

The paper is organized into 4 sections. The following section 2 discusses the related work conducted in the domain. Section 3 discusses the details of the proposed MEmoR model Section 4 shows the experimental results obtained whereas the last section gives the conclusion and future work.

2 Related Work

Studies on automatic recognition of emotional expression have been performed by different researchers on various types of data including bio-signals [8, 9], visual-audio signals [10,11], lexicographic data [12,13], and questionnaire based data. The most effective automatic emotion recognition can be captured from Bio-signals, but the most commonly available datasets are the questionnaire based. Kumar has used convolution to resolve the issue of duplicacy in these datasets [14], and to resolve the issue of data overload they proposed optimized summarization on bio-signals [15].

For accurate emotion detection researchers have proposed various model, some of the recent studies has been discussed here. Deep learning has been the most popular approach for affective emotion recognition, due to large size of the datasets [16, 17]. Mittal et. al. has proposed M3ER model for emotion recognition using 3 modalities, facial, textual and speech on IEMOCAP and CMU-MOSEI datasets [18]. Delbrouck detected emotions from the linguistic dataset using the tranfer learning [19]. They evaluated their model on IEMOCAP, MOSI, MOSEI, and MELD datasets. Hagar et. al. proposed a minixception +LSTM model on BioVidEmo dataset to determine the distinction between emotions from videos only [20]. The best distinguishness result was shown between anger and amusement. Iskhakova utilized the BioVid Emo dataset's audio signals only in order to detect negative emotions in individuals using one-dimensional convolution networks [21]. Only two classes of emotions in the dataset, amusement and sadness were evaluated. Xie et al. accessed ECG, EMG and SCL and executed Wavelet transform features and SVM machine learning models on Bio Vid Emo dataset to attain an accuracy of 94.81% with two types of feature selection techniques [22]. Xie and team also explored the impact of different fusion strategies on performance of the model None of the studies considered all the features, bio signals, video signals as well as audio signals for detecting the emotion state of humans.

Few recent studies on multimodal emotion recognition have reported the use of transfer learning. In 2017, a pre-training and fine-tuning (PT/FT) approach for transfer learning with deep neural networks for emotion recognition was used by Gideon et al. [23]. Ouyang et al. [24] proposed a model audio-visual emotion recognition using deep transfer learning and multiple temporal models. Dresvyanskiy et al. [25] reported the use of transfer learning and various fusion techniques for emotion recognition in data with both audio and video modalities. Siriwardhana et al. [26] proposed a "BERT-like" pre-trained self-supervised learning architecture to represent both speech and text modalities for multimodal speech emotion recognition.

3 KEY CONCEPTS

3.1 Emotion and Affect

There are numerous theoretical and practical approaches to probe emotion. The 'discrete emotion models' are often pitted against the 'dimensional emotion models' in literature. Principally, the valence and arousal dimensions are characteristics of affect and not emotion. Therefore, the dimensional theory of emotion hypothesizes the multiple dimensions of affect that hold true for discrete emotion entities. Basically, an affective experience involves affective traits (affectivity) categorized as negative affect and positive affect and affective states are categorized as emotions and mood. Consequently, emotion is a significant observable affective reaction. Further, discrete emotions are organized in a systematic way with regards to valence and arousal. For example, the discrete emotion of 'happiness' implies positive valence and just slight activation/arousal, whereas discrete emotions such as 'fear, anger, disgust' implies very negative valence with much activation/arousal. The levels of emotions are given in table 1.

Table 1. Level of Emotions

AFFECT	**EMOTIONS (Discrete)**	
(Dimensional)	***Primary (Basic)***	***Secondary***
VALENCE (Positive or Negative)	Happy Sad	Proud Relaxed
AROUSAL (Intensity: Agitated or calm)	Angry Disgusted	Confident Embarrassed
ACTION-ORIENTATION (Approach or avoid)	Surprised	Guilty Bored Disappointed

The affect-emotive capabilities are rooted in five basic emotions: joy (amusement), fear, anger, disgust, and sadness [4]. There are various ways to measure emotions as shown in fig. 2.

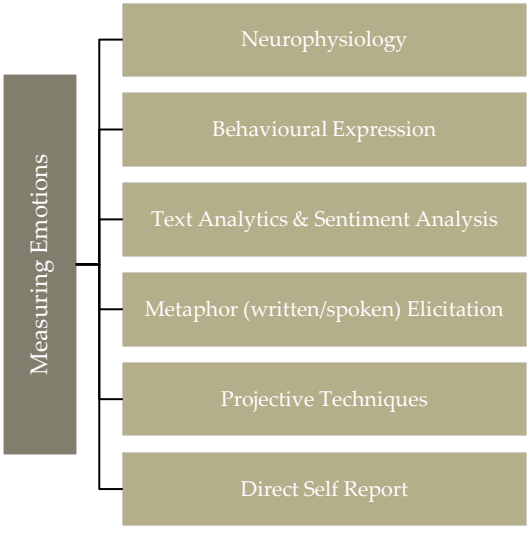


Fig.2. Emotion Measuring Techniques

Nevertheless, the ability to combine and mix emotions and then feel them in a variety of intensities increases the emotional experience exponentially. That is, three key aspects add to the labeling complexity of emotions:

- Firstly, the intensity with which we feel a basic emotion can fluctuate, for example, the highest intensity of anger can lead to destructive or vindictive whereas the lowest level would be furious

- Secondly, syndication of two or more basic emotions can be experienced simultaneously, for example, sadness and disgust results in ashamed
- Thirdly, this syndication of basic emotions can have varied individual intensity, for example, deep sadness with a slight disgust result in regret whereas a lot of disgust with a slight sadness result in guilt.

3.2 Transfer Learning

The use of deep learning models has advanced many application areas as they provide state-of-the-art results. The models are especially useful for feature extraction owing to the ability to learn representations that cannot be modelled manually. But these models require large training datasets to perform well in real-time and building a model from scratch is time consuming and expensive in terms of data collection/labeling, privacy and training time. Transfer learning has emerged as a solution for developing and training deep neural models with less data and compute power. Transfer learning is about "transferring" the learnt representations to another problem. With transfer learning, we basically try to exploit what has been learned in one task to improve generalization in another. It was first popularized in the field of computer vision, particularly for image classification [27]. The essence of transfer learning is extracting knowledge from one or more source tasks and applying it to the target task [26, 27] as it takes representations learned from an extensive training dataset and applies them to a different target dataset or problem which is usually smaller or more limited. The pre-trained models can also be fine-tuned on a comparatively smaller dataset to adapt them to a particular domain.

Transfer learning can be categorized based on the problem and solution or based on the strategy and objective. For the task of emotion recognition, we follow an inductive transfer learning approach. A general representation is pre-trained on a large unlabelled dataset and is then adapted to a supervised target task using the labeled data available [28]. The pre-trained model can also be made available for use to others who can fine-tune it for their target downstream tasks. Thus, while pre-training on a large dataset is computationally expensive, it only needs to be done once. Compared to it, downstream fine-tuning is much cheaper.

3.3 Multimodal Data Fusion

In general, multimodal data fusion can be done in three ways, early fusion, late fusion, and joint fusion. The types of Late Fusion Techniques have been shown in fig. 3.

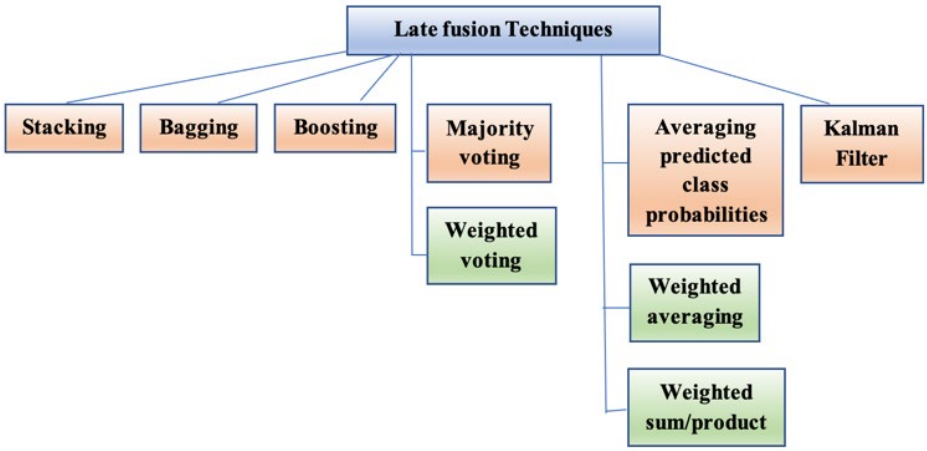


Fig.3. Multimodal late fusion techniques

Early fusion is a customary way of combining features of different modalities prior to training. It is also known as feature-fusion, input- level or data-level fusion. Concatenation of features is the

most used early fusion technique. Early fusion is based on the hypothesis of conditional independence between the different input modalities. However, this hypothesis does not always hold true as different modalities may have highly correlated features, for example video and depth cues. Further, these modalities may hold information that is correlated at a higher level. The main disadvantage of early fusion is that a large amount of data is removed from the modalities to derive common matrices and it requires a common sampling rate for all modalities. Joint fusion or intermediate fusion is a flexible method which allows data fusion at different stages of model training. Alternatively, late fusion uses the prediction scores available for input modalities independently followed by fusion at a decision-making stage [29]. This technique is considerably more straightforward than the early fusion, especially when the inputs vary significantly in terms of sampling rate, data dimensionality, and unit of measurement. It is also known as decision level fusion, a variant of which has been used in the proposed model.

4 MEMOR: MULTIMODAL EMOTION RECOGNITION

The proposed model uses video signals and bio-signals from the BioVidEmo DB and trains four separate deep neural networks, one for video signal and 3 individual bio-signals, named as M1, M_2, M_3 and M_4 respectively for each modality as shown in figure 4.

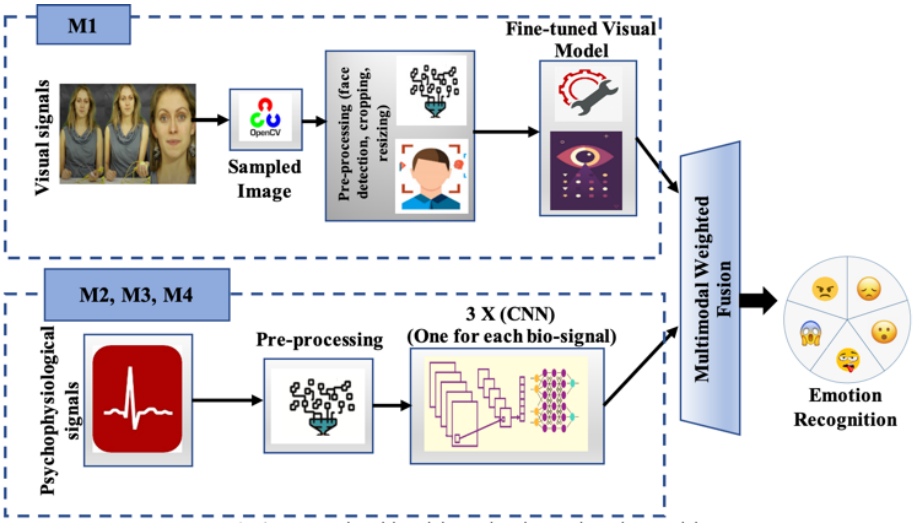


Fig.4. The proposed multimodal emotion recognition model

Subsequently, a weighted fusion of the four networks gives the emotional state of the subject. We select a decision-level fusion as keeping all the four components parallel until the decision will ensure that they are unaffected by the quality of data in another modality, which was our original intent. The steps involved are as follows:

4.1 Visual Modality

A video can be considered as a spatiotemporally connected stack of images. Since our goal is real-time evaluation and our data also consists of short utterances, we choose to focus on facial expressions in sampled images from the video clips. We haven't considered temporal features which would be effective in evaluating longer videos.

4.1.1 Preprocessing

The video clips from BioVidEmo are sampled at 1 frame per three second. Since the utterances are short (3-4 seconds on average), we avoid picking a peak frame for analysis. Instead, we adopt a majority voting strategy after individual emotion classification of the frames tagged to a video clip. Since the sampled images depict scenes, we first need to detect, and extract faces before training our model. For this we use MTCNN in TensorFlow. The extracted face images will be

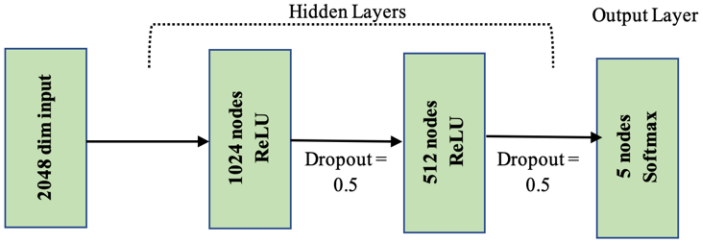
used to fine-tune our pre-trained model.

4.1.2 Fine-tuning

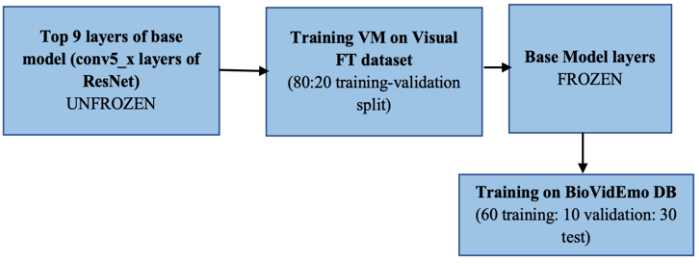
We start with a ResNet50 model loaded with weights pre-trained on the VGGFace2 dataset as our base model. VGGFace2 consists of face images with significant variations in pose, illumination, ethnicity, and age of the subject. This acts as a suitable precursor to the data we will use downstream for emotion classification. ResNet50 is a convolution-based architecture that achieved state-of-the-art results on ImageNet and several other image classification tasks.

ResNet, the winner of ImageNet Large Scale Visual Recognition Challenge, helps to reduce the error rate even on increasing the number of layers [28]. Basic architecture of residual network follows the convolution architecture, containing 1-D convolution layer along with the max pooling layer and ReLU activation layer. The major distinction between two are the skip connections, i.e. in basic deep learning architectures, consecutive hidden layers are connected to each other, but in ResNet50, the ReLU function on every alternative layer is performed after taking the output of ith layer along with the output of $(i+2)^{th}$ layer, and is provided as an input to $(i+3)^{th}$ layer after performing the activation function, known as skip connections or residual connections. ResNet50 preserves the knowledge gained during training and can speed up the new model with more hidden layers.

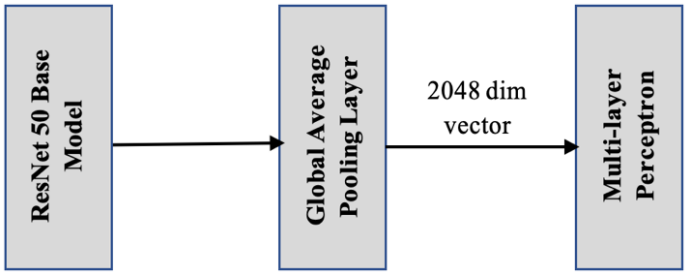
ResNet50 model has been implemented in Keras for emotion detection from face recognition by pre-training the model on VGGface2 dataset for face identification and then using CIFE to map the faces with different emotions, and at last this pre-trained model predicts the affective emotion state of an individual from BioVid Emo's visual signals.



(a) Visual Model (VM)



(b) Multi-layer perceptron Network



(c) Multi-stage fine tuning

Fig.5. The architecture of the visual modality component

CNN architectures learn general features in the lower layers and more task-specific features in the higher layers. We will utilize this property to fine-tune the base model by unfreezing some layers from the top and retraining the network on the CIFE dataset and converging the architecture toward our target task of emotion classification. The CIFE dataset also contains images with varying illumination, age, and ethnicity, making it suitable to act as a bridge between VGGFace2 and BioVid Emo. The fully connected output layer of the base model, which is a SoftMax classifier for 8631 categories is removed. The top 10 layers (average pooling layer and 9 convolutional layers) of the model are unfrozen and a fully connected layer with 5 nodes is added at the top for our target emotion categories for the second round of pre-training on CIFE, to minimize categorical cross-entropy loss using the Adam optimizer. Figure 5 depicts the complete set-up of the visual modality architecture.

4.1.3 Feature extraction and classification

After fine-tuning on CIFE, we freeze all the layers of the model and remove the top layer. The BioVid Emo images are passed through the fine-tuned model to obtain vectors representing the extracted features. We use these feature vectors as input to the MLP network for emotion classification.

4.2. Psychophysiological Modality

Psychophysiological signals measure any change a human body experiences, irrespective of any visual symptoms of the change. Digital affective biomarkers can help to track these physiological changes in real-time using wearable devices. The Department of Medical Psychology, University of Ulm, Germany has recorded three different such Psychophysiological signals using 6 Ag/AgCl electrodes fixed on the index and ring fingers, 2 on the upper right and lower left of the body, along with 2 on descending portion of the upper trapezius muscles for recording Skin Conductance Level (SCL), Electrocardiogram (ECG) and Trapezius Electromyogram (tEMG) respectively. Each bio-signal recorded acts as a good indicator of different affective emotion states of the individual. While experiencing sad emotions, the heart rate tends to decline, whereas while in an angry state, the skin temperature rises. To effectively identify the human affective emotional state, three separate deep learning models M_2, M_3 and M_4 were developed to identify the emotion state individually for each signal.

- CNN model for SCL M_2: The SCL signals measure the electrodermal activity of the sympathetic nervous system, it can measure the fluctuations in skin conductance shown due to eccrine sweat and gland activity.
- CNN model for ECG M_3: ECG signals can further help to derive the heart rate, interbeat interval and heart rate variability etc. These signals can help to identify fear, disgust, and amusement in individuals.
- CNN model for tEMG M_4: tEMG measures the muscle activities, an increase in muscle tone relates to increased activity of the sympathetic nervous system indicating anger, amusement, arousal, etc.

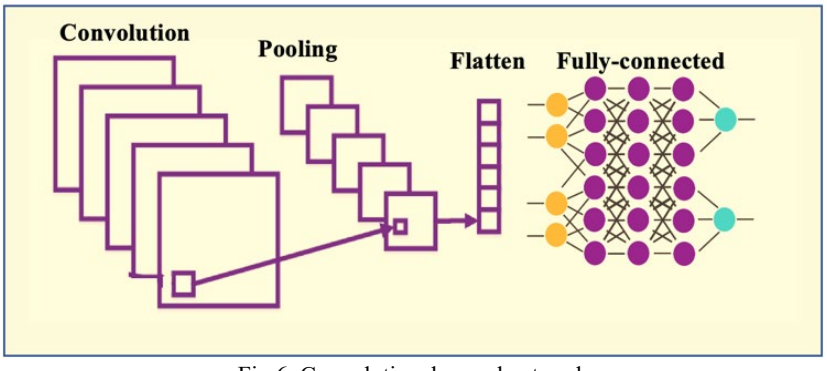


Fig.6. Convolutional neural network

The three models, M_2, M_3 and M_4 use CNN architecture for training. A convolutional Model contains several layers each performing a separate task [24]. The common architectural layers of CNN are convolution layer, ReLU layer, pooling layer and a fully connected layer. CNN is a neural network with a sequence of convolutional layers (often with a pooling step) and then followed by one or more fully connected layers as shown in fig. 6. For Psychophysiological signals, the fully connected layer was trained with 0.001 learning rate, with 256 epochs over a block size of 64.

4.3. Weighted Late Fusion

To process the results of the 4 different models (M_1, M_2, M_3, M_4) the weighted decision fusion technique has been used. Every model has been fitted for varied affective signals, each having results dependent on subjects, not just the cause. To identify the impact of each biomarker the model has been trained separately for the three physiological signals and the video signals. As discussed in section 3.2, fusion can be performed at three different stages, early fusion also known as feature fusion, joint fusion known as classification fusion, and lastly late fusion also known as decision fusion. A thorough review of the impact of different fusion studies has been shown by [30]. In the proposed model, we have used the weighted late fusion framework provided by Tsanousa et al. [31], it is an extension of the weighted averaging late fusion technique. The framework assigns weights on the basis of detection ratio rather than F-scores. The detection Ratio (DR) is shown in equation 1.

$$DR= TP/(TP+TN+FP+FN) \quad (1)$$

where TP represents true positive, TN represents True Negative, FP represents False Positive, and FN is False Negative. DT is calculated for each class, as the model has been executed for both Discrete Emotions as well as the Valence-arousal model, the number of classes is 5 and 2 respectively.
The weight, W of each output class is calculated as:

$$W= 1-DR \quad (2)$$

The weight of each class is then multiplied by the probability vector, P belonging to each model to find the predictive score of the class.

$$S=W*P \quad (3)$$

After calculating the weight of each class for an individual model, the score, S of the model is calculated by adding the scores of each class. The final decision opts through the maximum function, i.e., the model having the highest predictive score for the test case is chosen as the output level. As the proposed architecture contains 4 models for 4 biomarkers, the final output class is provided as:

$$\text{Output Class}= \text{Max } (S_{SCl}, S_{ECG}, S_{tEMG}, S_{video}) \quad (4)$$

This late weighted fusion strategy helps to choose the output class from the model best suitable for the output class. As each bio-signal is measuring the impact of a cause on the physiological state of humans. Each emotion has a varied effect on the body, like when a person is in an angry emotion, the rise in skin temperature is more common than in other emotional states. Therefore, the model trained only on SCL has better class accuracy for anger in comparison to other models, assigning more weight to Model M_1 trained only on SCL for anger class, than other models.

5 MODEL PERFORMANCE

5.1. Datasets

Accurate real-time emotion detection can revolutionize the human-computer interaction industry. But identifying emotion based on one feature will never accumulate all the changes a human feel when going through a change in the affective emotional state. As most of the work conducted in the field is analyzing the human emotion state using bio-signals, we have used multimodal data having 3 bio-signals measuring the physiological changes in the body, along with the visual signals recording the change in expressions a person was feeling. The BioVidEmo is a publicly available dataset for human emotion detection.

BioVid Emo Database: It is a multimodal high-quality physiological database containing discrete fundamental feelings and is created to examine human feelings and to mine emotional affective states. Bio signals like SCL, ECG, EMG are recorded in a controlled environment to identify different emotional states like amusement, sadness, anger, disgust and fear. It contains the data of individuals that were shown various film clips and their evoked feelings and emotional states were recorded. For each individual the most dominant emotion is observed and provided for the dataset. In this dataset, 94 individuals have participated across three diverse age ranges of 18 to 35 years, 36 to 50 years, and 51 to 65 years of age. Of all the participants, no one had any affective or emotional problems. There were 50 female and 44 males participants. But only 86 participants' data is complete and thus available in the dataset and the rest of the data entries are discarded due to incomplete or defiled entries. 15 film clips were selected, and the affective states were documented by instigating emotions through these clips. For instance, to trigger the feeling of sadness, clips like "The Champ" were shown to the participants where "A boxer is seriously injured and dying while his son enters" [7]. BioVid Emo database is now used by researchers for further emotional or affective state mining.

To train the transfer learning model for identifying the emotional state of individuals the ResNet based deep neural network is pre-trained on VGGFace2 and then to identify the expression of the face detected the model was trained on Candid-images-for-facial-expression database created by Li et al. [32].

CIFE: Candid-images-for-facial-expression (CIFE) database is a dataset created by Li et al. to construct an improved facial expression model for analyzing real-time facial expression tasks [27]. The CIFE dataset is produced through social media and the Web. Web crawling methods are employed to obtain natural expressions in the seven chosen categories of emotions. These categories are happy, anger, disgust, sad, surprise, fear, and neutral. Utilizing related phrases of these expressions, a huge amount of pictures are accumulated corresponding to the seven classes of expressions. There were 14756 pictures in total for these seven expressions in which pictures of anger, disgust, fear, happiness, neutral, sadness, and surprise are 1785, 266, 781, 3636, 644, 2485, and 997 respectively, and some pictures were added manually to the dataset corresponding to the classes where data was unbalanced. Viola face detector was used to uncover images of faces with these seven expressions. CIFE dataset is a freely available public dataset.

5.2. Results

To evaluate the proposed model, the model is trained using 70% BioVidEmo, and 20% is used for testing and 10 % for validation. The deep learning models were trained with different features and for two types of output classification to observe the effect of each on the proposed model. The various models have been compared based on accuracy and F1-score, since data is not properly balanced, the accuracy alone cannot be considered a reliable performance evaluator, while F1-score provides the weighted average of precision and recall.

The original BioVidEmo dataset is a time series analog data measuring the bio-signals of human using sensors. To process these bio-signals the three second window was applied to split the data and min-max normalization was applied to discretize the data. After preprocessing total of 8905

instances were generated having 1912 instances as Amused, 1776 as Fear, 1689 as Disgust, 1890 as Sad, and 1638 as Anger. The data was also categorized into the Valence-Arousal State, having 4916 cases of Arousal and 3989 instances of Valence. Table 2 shows the classification distribution of discrete emotions.

Table 2. Class Distribution

Class	**Amusement**	**Fear**	**Disgust**	**Sad**	**Anger**
Entities	1912	1776	1689	1890	1638

The models were executed separately on the two types of output categorized dataset. Initially, the model was tested for two output classes Valence and Arousal on 7 different scenarios. First only CNN based M_1 model was used on SCL signals only resulting in 65.54 accuracy, second deep learning model was executed only on ECG signal resulting in 72.89 accuracy. Third the model was trained on only tEMG signals producing 69.81 accuracy, afterwards the three signals were fused at feature level, i.e., early fusion was applied on the three bio-signals and then the CNN model was used resulting in better accuracy of 77.92%.

The face detection was also tested individually for affective emotion state detection using Res-Net50 and CNN, resulting in 74.32 % accurate prediction. The last cases were then combined with late fusion and model accuracy has improved a lot resulting in 79.81% accurate prediction. But the proposed model has surpassed the results with late weighted fusion resulting in 83.79%, implicating that each feature separately works best for a certain class but with class based late weighted fusion, the accuracy of the model can be improved a lot. The results are shown in table 3. The proposed model has performed better than the state of art Biomodal deep belief network (BDBN) by Zhongmin et al. for discrete emotions, as model was tested and trained only for discrete emotions [33].

Table 3. Valence- Arousal Prediction Results

	Accuracy	**F1-Score**
SCL	65.54	0.64
ECG	72.89	0.71
tEMG	69.81	0.70
Feature Fusion of Psychophysiological signals	77.92	0.73
Video Signals	74.32	0.75
Feature Fusion of Psychophysiological & Late Fusion with Video Signals	79.81	0.77
Proposed Model - MEmoR	**83.79**	**0.81**

The same 7 models have also been applied on the discrete emotions and it has been observed there too that weighted late fusion model works best in comparison to early fusion or single features only. These results are highlighted in table 4 along with state of art result.

Table 4. Discrete Emotion Prediction Accuracy

	Accuracy	**F1-Score**
SCL	63.78	0.68
ECG	69.56	0.64
tEMG	70.93	0.61
Feature Fusion of Psychophysiological signals	73.29	0.76
Video Signals	68.51	0.65
Feature Fusion of Psychophysiological & Late Fusion with Video Signals	76.32	0.74
BDBN Model [33]	80.89	-
Proposed Model- MEmoR	**81.54**	**0.79**

Figure 7 shows the variation in the results for two varied output categorized BioVid Emo dataset on all the 7 models.

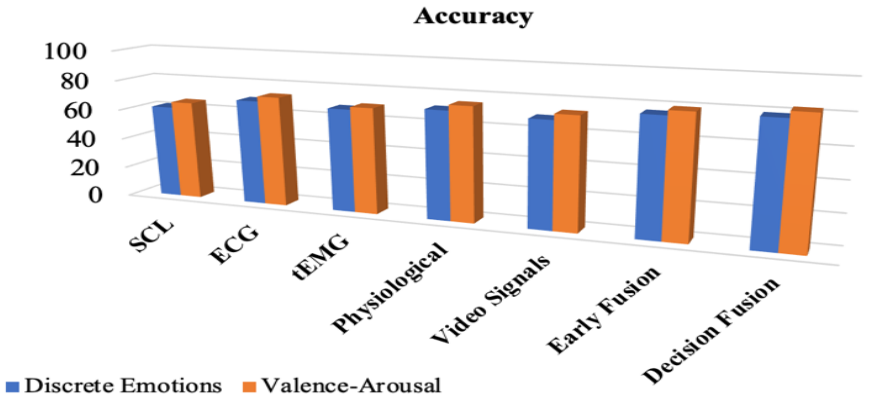


Fig. 7. Accuracy of Discrete Emotion vs Valence-Arousal

The proposed deep learning weighted late fusion model has identified the discrete emotion as well as valence-arousal affective state effectively, figure 8 shows the confusion matrix of the proposed model discrete emotion dataset, with table 5 containing the accuracy of each emotional state for the proposed model with state-of-the-art results, highlighting that Amusement state has been identified better in comparison to other emotional states.

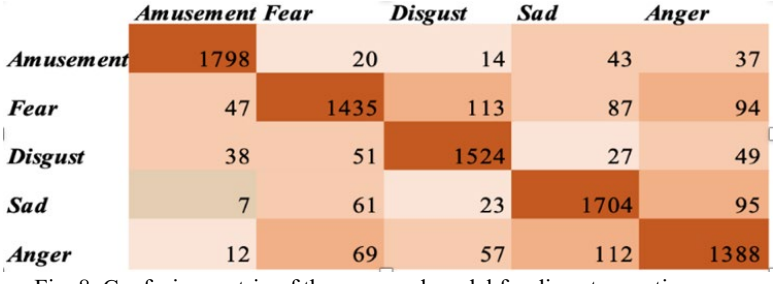


Fig. 8. Confusion matrix of the proposed model for discrete emotions

Table 5. Accuracy of Emotions

	Accuracy	
Emotions	**MEmoR Model**	**BDBM [33]**
Amusement	**91.02**	89.25
Anger	75.31	90.74
Fear	77.64	52.60
Disgust	80.38	86.45
Sad	79.92	85.42
Valence	84.16	-
Arousal	82.71	-

Figure 9 compares the most affective three models for the discrete emotions, that is, early fusion model of all three bio-signals, early fusion of three bio-signals late fused with visual signals, and lastly the proposed model.

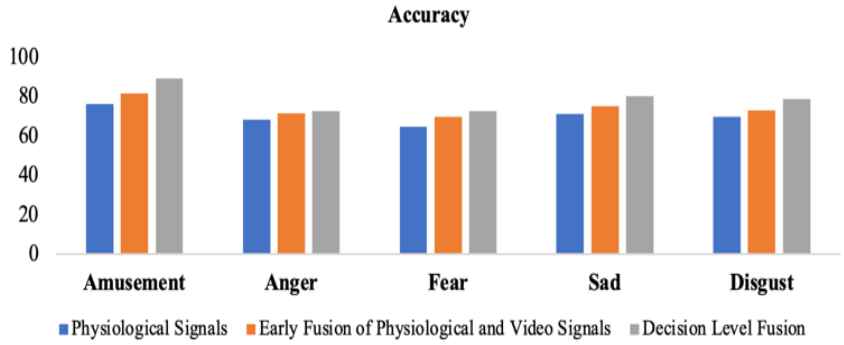


Fig. 9. Comparison of Models for Discrete Emotions

As indicated from results, the weighted late fusion model can act as a promising model for merging the varied modality data for accessing the affective emotional state of an individual in real-time using wearable devices and video surveillance. The superior result of the proposed model confirms the hypothesis that, no one feature(signal) is sufficient for emotion detection. For accurate emotion recognition, both observable and non-observable changes in human body need to be capurted. Visual signals catured the observable changes a human depicts on experiencing a change in mood and emotion, bio signals look into the physiological changes a person feels. As observable changes like frowning, face smirking, hand gestures, tears, change in voice tone indicates the emotional state of a person, but these signs/ changes are controllable by an individual. An actor can display sorrow through facial expressions, while feeling some other emotion, visual signals help in determing the emotional state, but they should never be considered the only source for evaluation. As evident from table 3 and 4 as well, feature fusion of physiological signals performs better than video signals only, but the merger of both observable and non-observable signals are the most appropriate sceneario.

The proposed model- weighted late fusion of all physiological signals and visual signals performs better than early fusion of physiological signals with late fusion of visual signals, as in later scenario, the final output is produced by averaging the probability of the two sub-models, providing more weightage to visual signals in comparison to the proposed model, when the probability weightage of all 4 sub-models are calculated separately, providing more weightage to physiological signals (3 sub-models). For more accurate analysis the proposed work should be tested on a dataset with various observable and non-observable modalities, like facial expression, speech, hand gestures, ECG, EEG, EDA, ACC etc.

6. Conclusion

A perceived situation generates affective responses in the form of emotion reactions and bodily state (physiological sensations, facial, speech or body cues) consequently prompting behavioral actions. Electrophysiological signals, behavioral signals (e.g., posture), speech signals, social interactions and psycholinguistic features in social data can be used as observable traits for detecting human emotional states. These observed emotional states in a working environment can help to monitor the comfort of the employees. A stressed employees produces less output in comparison to a physically and mentally healthy employee. identifying accurate emotions of employees can help make industries and HRs better policies. The proposed model incorporated the capabilities of deep learning, transfer learning and late data fusion to produce state-of-the-art emotion recognition model with an accuracy of 83.79 and 81.54%. The proposed work focused on detecting emotions from reactions captured in a controlled environment while watching videos. For real-time emotion recognition model generation, the study lacks peoples' reactions in a dynamic

environment. For advancements in the field of automated emotion recognition, researchers require a large-sized dataset of audio, video, and bio-signals recorded in a dynamic environment performing various tasks. The proposed model highlights the importance of diverse affective biomarkers in uncovering human emotions.

ABBREVIATIONS

SCL	Skin Conductance Level
ECG	Electrocardiogram
tEMG	Trapezius Electromyogram
CIFE	Candid Images for Facial Expression
CNN	Convolution Neural Network
MTCNN	Multi-task Convoluation network

REFERENCES

[1] Picard, R. W., Vyzas, E., & Healey, J. "Toward machine emotional intelligence: Analysis of affective physiological state". *IEEE transactions on pattern analysis and machine intelligence*, *23*(10), 1175-1191, 2001.

[2] Hodapp, V., Neuser, K. W., & Weyer, G. "Job stress, emotion, and work environment: Toward a causal model". *Personality and Individual Differences*, *9*(5), 851-859, 1988.

[3] Jáuregui, D. A. G. "Toward Emotional Internet of Things for Smart Industry". In *SMART INTERFACES 2017, The Symposium for Empowering and Smart Interfaces in Engineering,* 2017.

[4] Gupta, D., Bhatia, M. P. S., & Kumar, A. "Resolving Data Overload and Latency Issues in Multivariate Time-Series IoMT Data for Mental Health Monitoring". *IEEE Sensors Journal,* 2021.

[5] Kumar, A., Sharma, K., & Sharma, A. "Hierarchical deep neural network for mental stress state detection using IoT based biomarkers". *Pattern Recognition Letters*, *145*, 81-87, 2021.

[6] Luis-Ferreira, F., & Jardim-Goncalves, R. "A behavioral framework for capturing emotional information in an internet of things environment". In *AIP Conference Proceedings*(Vol. 1558, No. 1, pp. 1368-1371). American Institute of Physics, 2013.

[7] Zhang, L., Walter, S., Ma, X., Werner, P., Al-Hamadi, A., Traue, H. C., & Gruss, S. " "BioVid Emo DB": A multimodal database for emotion analyses validated by subjective ratings". In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-6). IEEE, 2016.

[8] Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G., "Human emotion recognition using deep belief network architecture." *Information Fusion* 51, pp: 10-18, 2019.

[9] Sharma, A., Sharma K., and Kumar A.. "Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion." *Neural Computing and Applications*, pp: 1-14, 2022.

[10] Gumaei, A., Hassan, M. M., Alelaiwi, A., & Alsalman, H., "A hybrid deep learning model for human activity recognition using multimodal body sensing data." *IEEE Access* 7, pp:99152-99160, 2019.

[11] Yang, J., Wang, R., Guan, X., Hassan, M. M., Almogren, A., & Alsanad, A., "AI-enabled emotion-aware robot: The fusion of smart clothing, edge clouds and robotics." *Future Generation Computer Systems* 102, pp: 701-709, 2020.

[12] Kumar, A., Jaiswal, A., Garg, S., Verma, S., & Kumar, S., "Sentiment analysis using cuckoo search for optimized feature selection on Kaggle tweets." *International Journal of Information Retrieval Research (IJIRR)*9, no. 1, pp: 1-15, 2019.

[13] Kumar, A., Sharma, A., and Arora, A., "Anxious depression prediction in real-time social data." In *International conference on advances in engineering science management & technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India*. 2019.

[14] Kumar, A. "Using cognition to resolve duplicacy issues in socially connected healthcare for smart cities". *Computer Communications*, *152*, 272-281, 2020.

[15] Kumar, A., Sharma, K., & Sharma, A. "Genetically optimized Fuzzy C-means data clustering of IoMT-based biomarkers for fast affective state recognition in intelligent edge analytics". *Applied Soft Computing,* 2021.

[16] Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. "Learning affective features with a hybrid deep model

for audio–visual emotion recognition". *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 3030-3043, 2017.

[17] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. "End-to-end multimodal emotion recognition using deep neural networks". *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1301-1309, 2017.

[18] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. "M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues". In *AAAI* (pp. 1359-1367), 2020.

[19] Delbrouck, J. B., Tits, N., & Dupont, S. "Modulated Fusion using Transformer for Linguistic-Acoustic Emotion Recognition". *arXiv preprint arXiv:2010.02057*, 2020.

[20] Hagar, A. F., Abbas, H. M., & Khalil, M. I. "Emotion Recognition In Videos For Low-Memory Systems Using Deep-Learning". In *2019 14th International Conference on Computer Engineering and Systems (ICCES)* (pp. 16-21). IEEE, 2019.

[21] Iskhakova, A., Wolf, D., & Meshcheryakov, R. "Automated Destructive Behavior State Detection on the 1D CNN-Based Voice Analysis". In *International Conference on Speech and Computer* (pp. 184-193). Springer, Cham, 2020.

[22] Xie, J., Xu, X., & Shu, L. "WT feature based emotion recognition from multi-channel physiological signals with decision fusion". In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)* (pp. 1-6). IEEE, 2018.

[23] Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., & Provost, E. M. "Progressive neural networks for transfer learning in emotion recognition". *arXiv preprint arXiv:1706.03256*, 2017.

[24] Ouyang, X., Kawaai, S., Goh, E. G. H., Shen, S., Ding, W., Ming, H., & Huang, D. Y. "Audio-visual emotion recognition using deep transfer learning and multiple temporal models". In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 577-582), 2017.

[25] Dresvyanskiy, D., Ryumina, E., Kaya, H., Markitantov, M., Karpov, A., & Minker, W. "An Audio-Video Deep and Transfer Learning Framework for Multimodal Emotion Recognition in the wild". *arXiv preprint arXiv:2010.03692*, 2020.

[26] Siriwardhana, S., Reis, A., Weerasekera, R., & Nanayakkara, S. "Jointly Fine-Tuning" BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition". *arXiv preprint arXiv:2008.06682,* 2020.

[27] Abbas, A., Abdelsamea, M. M., & Gaber, M. M. Detrac: "Transfer learning of class decomposed medical images in convolutional neural networks". *IEEE Access*, *8*, 74901-74913, 2020.

[28] Wu, Z., Shen, C., & Van Den Hengel, A. "Wider or deeper: Revisiting the resnet model for visual recognition". *Pattern Recognition*, *90*, 119-133, 2019.

[29] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. "A review of affective computing: From unimodal analysis to multimodal fusion". *Information Fusion*, *37*, 98-125, 2017.

[30] Hossain, M. S., & Muhammad, G. "Emotion recognition using deep learning approach from audio–visual emotional big data". *Information Fusion*, *49*, 69-78, 2019.

[31] Tsanousa, A., Meditskos, G., Vrochidis, S., & Kompatsiaris, I. "A weighted late fusion framework for recognizing human activity from wearable sensors". In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-8). IEEE, 2019.

[32] Li, W., Tsangouri, C., Abtahi, F., & Zhu, Z. "A recursive framework for expression recognition: from web images to deep models to game dataset". *Machine Vision and Applications*, *29*(3), 489-502, 2018.

[33] Li, W., Abtahi, F., & Zhu, Z. "A deep feature based multi-kernel learning approach for video emotion recognition". In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 483-490), 2015.

[34] Wang, Zhongmin, Xiaoxiao Zhou, Wenlang Wang, and Chen Liang. "Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video." *International Journal of Machine Learning and Cybernetics* 11, no. 4 (2020): 923-934.