

Innovative Stress Detection with Facial Expression Extraction using multimodal data fusion approaches

¹Reshu Gupta, ²Aniket Kumar, ³Dr. Anil Rajput

¹Research Scholar, Shobhit University, Meerut

²School of Electronics, Electrical & Mechanical Engineering (SEE & ME)

Shobhit Institute of Engineering & Technology (Deemed-to-be University)

NH-58, Modipuram, Meerut, Uttar Pradesh, India.

³Co-Supervisor: OSD, HED, Govt. of MP, Directorate Bhopal, MP

Abstract

In recent years, there has been a surge in stress-related problems due to the fast-paced nature of modern life and work, which has greatly raised psychological stress levels. As a result, medical research is now heavily focused on developing efficient methods for diagnosing and preventing depression. In order to create a model for predicting depression using multimodal data fusion approaches, this study investigates the ways in which stress appears in voice, writing, and facial expressions. In order to assist in weeding out extraneous information, the model employs parallel attention to extract important characteristics from each type of input. Fusion Attention also assesses each modality's contribution to the whole image. We use a Gated Recurrent Unit (GRU) to capture the subtleties of stress at various points in time in order to comprehend how these aspects vary over time.

Keywords: Emotional Intelligence, Psychological Stress Indicators, Stress Detection Algorithms, Mental Health Diagnostics

Introduction

Stress is a persistent, long-term mood disorder with symptoms that include, but are not limited to, low mood, pessimism, concentration and memory problems, self-neglect, decreased appetite, and decreased activity levels. In extreme situations, it can also result in delusions, hallucinations, and other mental health conditions (Yung et al., 2004). The number of people suffering from depression has been continuously increasing in recent years, and it is currently the fourth most common condition in the world. However, due to a number of social and personal reasons, the current detection rate for depression is still below 10%. Thus, it is essential to look at ways to increase the prevalence of stress-related diseases. The main resource for diagnosing depression is the Diagnostic and Statistical Manual of Mental Disorders (DSM), which is published by the American Psychiatric Association (Wilhelm et al., 2004; Dingle et al., 2010). The DSM states that five or more of the following nine symptoms—persistently depressed mood, loss of interest or

anhedonia, insomnia or hypersomnia, significant weight loss or gain, and fatigue—must be present almost every day for at least two weeks in order to be diagnosed with major depressive disorder.

It is important to use computer algorithms for the diagnosis and treatment of depression because this evaluation approach is subjective and time-consuming by nature (Graham et al., 2019). Deep learning has made significant strides recently in a number of fields, such as natural language processing, computer vision, and speech recognition. Because of this, more and more studies are concentrating on using deep learning techniques to improve the efficacious treatment of depression.

Nowadays, a number of techniques are used to diagnose depression, such as video-based facial expression analysis (Baltrusaitis et al., 2016; Valstar et al., 2017), speech-based acoustic feature analysis (Cummins et al., 2015; Williamson et al., 2019; Yalamanchili et al., 2020), and text semantic analysis (Hassan et al., 2017; Jia, 2018; Havigerova et al. in 2019). These unimodal techniques are constrained by the complexity and uniqueness of depressed manifestations, even though they have benefits like lesser data dimensionality, fewer computational costs, and simpler algorithmic model design.

The way that people communicate their emotions varies greatly; some people are more comfortable expressing their emotions verbally, while others may be less inclined to do so. In these situations, reading a person's facial expressions or examining their auditory characteristics can reveal important information about their emotional condition. Complementary information that increases robustness and recognition accuracy in diagnosing depression can be obtained by using a multimodal emotional analysis technique. Among the main algorithms used for depression prediction are Support Vector Machine (SVM), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Attention processes. Because SVM uses kernel functions to efficiently translate low-dimensional data to high-dimensional space, it is frequently used for classification problems. Nevertheless, SVM training is prone to both underfitting and overfitting, necessitating careful kernel function selection and parameter modification (Peng et al., 2019).

In the area of depression prediction, CNNs are very good at identifying local information from facial expressions. However, their capacity to depict contextual time-series data is constrained (Mungra et al., 2020). On the other hand, because LSTM algorithms can effectively take into consideration the temporal correlations between various video data segments and their influence on total emotional expression, they are well-suited for depression prediction. However according to Jiang et al. (2020), LSTMs frequently undervalue the significance of the weight connections between high-dimensional characteristics. An LSTM variation with a more compact structure and quicker training times is the Gated Recurrent Unit (GRU). In the meanwhile, high-value data inside large-dimensional datasets are distinguished via attention mechanisms. Although attention algorithms do not automatically take into consideration the temporal

aspects of the data, they can successfully capture the correlations among numerous parameters and reduce the dimensionality of the data when used for depression prediction (Zhao et al., 2019).

This research suggests a depression analysis model based on a BiAttention-GRU framework to efficiently compute the correlation and timing between multimodal features, boosting feature fusion and recognition accuracy. The model makes a number of noteworthy contributions:

Multimodal Feature Fusion: The model is specifically made to fuse depression-related multimodal features. It makes it easier to integrate supplementary information, which improves the model's predictive power and resilience.

Bimodal Attention method: To extract crucial local features from text, audio, and facial expression data, we provide a unique bimodal attention method that makes use of parallel attention. This system makes sure that important data from every modality is efficiently recorded.

Fusion Focus Mode: To allocate weights to features from different modalities according to their respective contributions, the model uses a fusion attention method. This makes it possible to comprehend how many characteristics interact and affect the overall evaluation of depression in a more complex way.

This paper's remaining sections are organized as follows: The pertinent literature and related works are reviewed in Section 2. Section 3 provides specifics on the suggested methodology and its theoretical underpinnings. The materials and experimental setup are described in Section 4, along with a thorough analysis of the findings. Section 5 wraps up the work by summarising the main conclusions and any possible ramifications for further study.

Literature Review

Standardised questionnaires such as the Negative Symptom Assessment Scale (SANS) (Andreasen, 1989), the Hamilton Rating Scale for Depression (HRSD) (Zimmerman et al., 2004), the Baker Depression Scale (BDIII) (Du et al., 2019), and the eight-item Patient Health Questionnaire (PHQ-8) (Kroenke et al., 2009) are the mainstays of traditional methods of diagnosing depression. These methods require subjective assessments by trained and experienced psychologists, which adds time and effort to the process.

The subject of depression prediction is moving more and more towards multimodal, automated assessment techniques as a result of the quick development of technology. In order to enable prompt intervention and therapy, this evolution attempts to improve the efficiency and accuracy of recognising depression symptoms.

The field of stress detection has seen significant advancements through various computational and multimodal approaches.

1. **Text-Based Analysis:** Researchers have utilized natural language processing to identify stress indicators in written communication. For instance, studies by De Choudhury et al. (2013) analyzed Twitter data, revealing that users experiencing stress often exhibit increased negativity and emotional fluctuations in their posts.
2. **Speech Analysis:** Acoustic feature analysis has been pivotal in stress detection. Cummins et al. (2015) found that stress can be inferred from vocal characteristics such as pitch variation and speech rate. Additionally, Krahmer et al. (2019) developed models that successfully classified stressed speech, demonstrating the potential of prosodic features in identifying stress levels.
3. **Facial Expression Recognition:** The analysis of facial expressions has proven effective in stress detection. Baltrusaitis et al. (2016) employed CNNs to recognize stress-related facial cues, while Liu et al. (2020) highlighted how temporal changes in facial expressions can provide insights into an individual's stress response.
4. **Physiological Data Integration:** Some studies have integrated physiological signals, such as heart rate variability and galvanic skin response, with behavioral data to enhance stress detection accuracy. For example, Kwan et al. (2018) demonstrated that combining physiological data with speech analysis significantly improves the prediction of stress levels.
5. **Multimodal Approaches:** Recent work has focused on integrating multiple modalities for more accurate stress detection. Mungra et al. (2020) proposed a multimodal framework that combines text, audio, and visual inputs, achieving improved recognition rates compared to unimodal approaches.

These studies highlight the diverse methods employed in stress detection, underscoring the importance of multimodal approaches to capture the complex nature of stress responses. This paper seeks to build on these foundations by proposing a novel model that effectively integrates various data modalities for enhanced stress recognition. One approach to assessing a patient's sadness is through automatic stress prediction, which leverages machine learning or deep learning algorithms to analyze their text, audio, or visual expressions. An early method for predicting depression involved text-based emotion analysis, and the rise of social media has provided researchers with a valuable resource for studying emotional expression in text. Eichstaedt et al. (2018) employed the unsupervised machine learning algorithm LDA to develop a prediction model based on electronic medical language gathered from Facebook. However, while the LDA algorithm has shown some success in emotion analysis, it does not account for semantic changes that occur when words are rearranged within a sentence.

With advancements in speech technology, many individuals are opting for voice diaries to document their lives and emotions instead of traditional handwritten journals. By analyzing the acoustic features of a person's speech—such as prosody, tone, loudness, pause duration, and energy—researchers can extract

insights into their mood and mental state, particularly regarding depression. Using a Bayesian algorithm with multi-channel weighted processing, early teenage depression can be predicted based on these vocal characteristics.

While Support Vector Machines (SVM) are effective for standard classification tasks, finding an appropriate kernel function for high-dimensional data classification can be challenging. To address this, Vazquez-Romero and Gallardo-Antolin (2020) developed an ensemble CNN technique. This method involves training multiple CNN models on segmented voice files represented as logarithmic spectrograms, ultimately employing an integrated average approach to derive a final prediction. Single-modal feature recognition for depression, which is characterised by modest feature dimensions that make computation and algorithmic model building easier, is the main focus of the previously described research. Although automated depression recognition has greatly improved with this method, it is important to remember that single-modal features need to meet completeness and intuitiveness requirements. Single modalities usually don't include all the information needed to make precise predictions. Therefore, in order to improve the prediction of depression, the integration of multimodal information had to be given priority.

Motivation

As technology advances, human-computer interaction research undergoes significant changes, resulting in a larger degree of success. All application fields based on human-computer interaction have a constant desire for innovation, thus We choose a method where We can contribute to human-computer interaction in some way. So far, the machine learning and deep learning algorithm has implemented for Facial Expression Detection from the face image. The difference between the machine learning implementation and deep learning implementation for the facial expression detection is about the number of images and its features used to classify the emotions.

METHODOLOGY

A. Collection of Data Set

Primary data : There are many kinds of stressors that may stimulate stress. Playing computer games, answering difficult questions , and solving difficult problems are some example stressors. In this work, we referred to the method of using infrared cameras to record the affective reactions (neutral, relaxed, and stressed) of the participants when they watched three different types of 2-min video clips . The neutral video clips were about scenery or food making. The relaxed ones were highlights of variety show. The stressed ones were science programs with rich knowledge. Each scientific program was followed by a question-answering test. Each test contained ten questions. Half of them were multiple choices and the other half were blank fillings. The total score was 100. We designed the questions in such a way that it was very hard to come up with the correct answers unless the participants could understand the content and grasp the knowledge points well enough in the video. To stimulate the cognitive stress

a bit, before the test, we announced to the participants that they could get some extra rewards if achieving test scores over 50 as incentive.

We have prepared primary data set for machine learning. Data was collected were different age of person. We divide our work in two different parts. In first part, preprocessing of the data is done in different steps and in second part stress analysis.



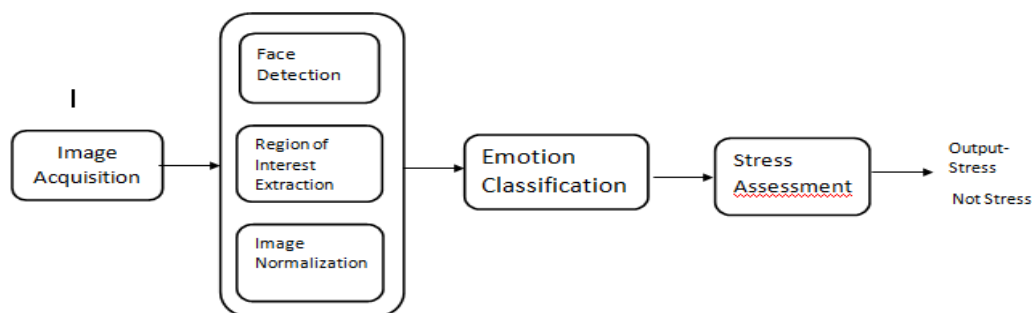
Fig.1. Sample of primary dataset

B. Secondary data

Secondary data set we can get from Kaggle

- Methodology divides in two parts, trained model & second test model.
- First phases - collection of secondary data from Kaggle Exploratory data analysis, preprocessing of images, extract texture.
- Finding the region of interest
- In phase two we get database which we have to used in algorithm.
- Algorithm is based on Support Vector Machine (SVM) and Naive Bayes Algorithm. Though this algorithm we have generated to train the data set.
- Test data set have matched with this training data set and stresses were analysis.
- Stressed data can be sub categorized by 5 different classes to test the level

Proposed Framework: As per the objective of the research work the methodology are as follows:



1. The system's first module will be responsible for capturing images in real-time, from the computer's webcam
2. In the second module will be determined the user's face's location, using a Haar-like feature selection technique and cropped. The face will then be resized to 299x299 pixels and normalized, by dividing the value of all pixels by 255 so that all have values between 0 and 1

3. The Emotion Classification module, consisting of a trained classification model, will classify the face and return a list of seven probability scores, one for each facial expression.
4. The facial expression with the highest probability will then be fed to the fourth module (stress assessment).
5. The various classifications made over time will be recorded and based on the parameters provided to the program, it will then determine whether the user is under stress.

C. Flow Chart

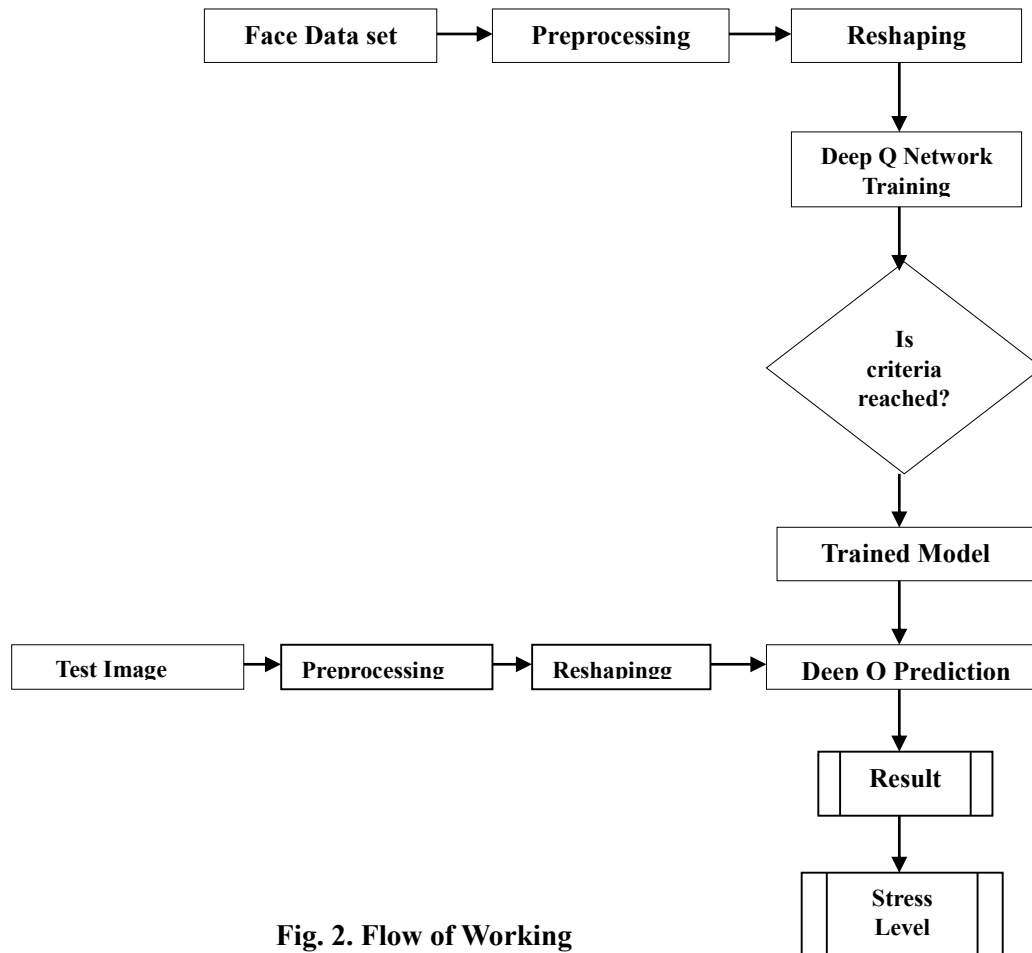


Fig. 2. Flow of Working

Experimental results and Discussions

A. Prepossessing of Data:

▪ Programming

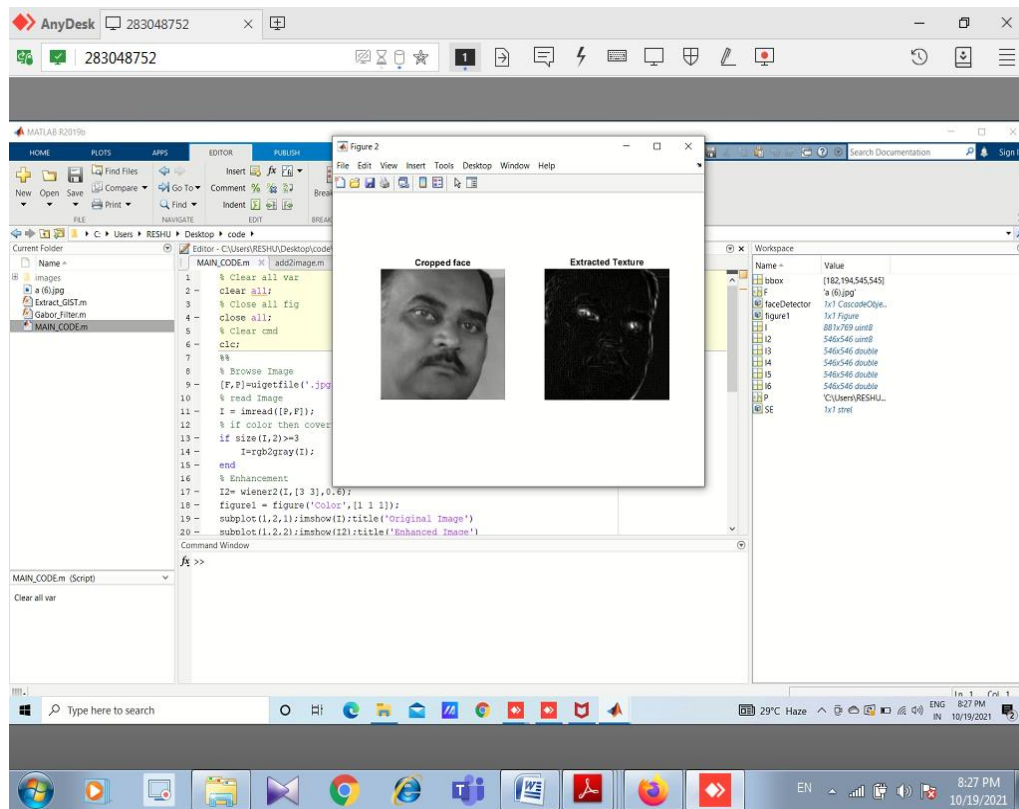
```

% Clear all var
Clear all;
% Close all fig
close all;
% Clear cmd
clc;
  
```

```
%%  
% Browse Image  
[F,P]=uigetfile('.jpg');  
% read Image  
I = imread([P,F]);  
% if color then covert to gray  
if size(I,2)>=3  
    I=rgb2gray(I);  
end  
% Enhancement  
I2= wiener2(I,[3 3],0.6);  
figure1 = figure('Color',[1 1 1]);  
subplot(1,2,1);imshow(I);title('Original Image')  
subplot(1,2,2);imshow(I2);title('Enhanced Image')  
%%  
% Face Detection  
Face Detector = vision.CascadeObjectDetector();  
% Bounding box cordinates  
bbox = step(faceDetector,I2);  
if size(bbox,2)  
    % Face crop  
    I2=I2(bbox(2):(bbox(2)+bbox(4)),bbox(1):(bbox(1)+bbox(3)),:);  
    % Extrcat Texture using 2D Gabor  
    I3=Extract_GIST(I2);  
    SE=strel('disk',5);  
    % Tophat Filtering  
    I4=imtophat(I3,SE);  
    % Tophat+Original  
    I4=imadd(I3,I4);  
    % Bottomhat Filtering  
    I5=imbothat(I3,SE);  
    % Bottomhat Filtering - (Tophat+Original)  
    I6=imsubtract(I4,I5);  
    figure1 = figure('Color',[1 1 1]);  
    subplot(1,2,1);imshow(I2);title('Cropped face')  
    subplot(1,2,2);imshow(I6);title('Extracted Texture')
```

end

B. Results



Model evaluation indicators

The evaluation measures Accuracy, Precision, Recall, and F1 Score are frequently used for classification models. The percentage of accurately predicted samples to all samples is known as accuracy. Recall measures the proportion of true positives to actual positives in the dataset, whereas precision shows the ratio of genuine positive predictions to all projected positives. The F1 Score offers a fair assessment of a model's performance since it is based on both precision and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{1/Precision + 1/Recall}$$

Where:

- TPTPTP (True Positive) indicates the number of correctly predicted positive samples,
- FPFPP (False Positive) denotes the number of incorrectly predicted positive samples,
- TNTNTN (True Negative) indicates the number of correctly predicted negative samples, and
- FNFNFN (False Negative) represents the number of incorrectly predicted negative samples.

Evaluation of Experimental Outcomes

The unimodal features—comprising text, voice, and facial expressions—along with their integrated versions, were assessed using various multimodal data fusion algorithms, including CNN, GRU, BERT, BiAttention, CNN-Attention, GRU-Attention, and BiAttention-GRU.

Both Experiment 1 and Experiment 5 utilized the CNN algorithm. A comparative analysis of predictive performance across different modalities indicates that the CNN algorithm, when applied to image feature extraction, outperforms its efficacy in text or audio sentiment analysis. Similarly, Experiments 2 and 6 employed the GRU algorithm, demonstrating its enhanced effectiveness in text sentiment analysis compared to image and audio modalities.

A comparison of Models 1 and 3 highlights the significant role of the Attention mechanism in enhancing the extraction of critical features from each modality, thereby improving overall prediction accuracy. Furthermore, contrasting Model 3 with Model 7 reveals that integrating a GRU model effectively captures temporal sequence information, which further boosts predictive accuracy.

Moreover, comparing Model 6 to Model 7 emphasizes that the inclusion of the BiAttention model facilitates the extraction of essential intra-class local features and inter-class weighted fusion features, significantly benefiting the performance of multimodal depression prediction.

The experimental results indicate that the BERT algorithm outperforms other algorithms in text-based depression prediction, leveraging its self-attention mechanism to extract deeper semantic meanings from the text. Additionally, the CNN-Attention approach demonstrates improved performance in image-based depression prediction. The BiAttention-GRU method excels in multimodal data fusion tasks, effectively integrating critical intra-class local features, inter-class weighted fusion features, and temporal dynamics across modalities, thus enhancing predictive performance in a comprehensive manner.

Conclusion

This study addresses the increasing prevalence of stress-related issues and the need for effective depression diagnosis and prevention methods by exploring multimodal data fusion approaches. It investigates how stress manifests in voice, writing, and facial expressions, employing a BiAttention-GRU framework to

enhance feature fusion and recognition accuracy. The model integrates Parallel Attention for feature extraction and Fusion Attention to assess the contributions of different modalities over time.

Traditional diagnostic methods, such as various psychological questionnaires, rely on subjective assessments, highlighting the need for automated approaches. This research employs infrared cameras to capture participants' emotional reactions to different video clips, with a focus on real-time image processing and classification.

Results indicate that the CNN algorithm is more effective for image feature extraction, while the GRU shows improved performance in text sentiment analysis. The study underscores the significance of attention mechanisms in enhancing feature extraction and predictive accuracy, with the BERT algorithm excelling in text-based predictions. The findings suggest that the proposed multimodal approach can significantly improve stress recognition and depression prediction, demonstrating the potential for automated mental health diagnostics.

References

1. Alankar, B., Ammar, M. A., & Kaur, H. (2021). Facial Emotion Detection Using Deep Learning and Haar Cascade Face Identification Algorithm. In *Springer eBooks* (pp. 163–180). https://doi.org/10.1007/978-981-16-0695-3_17
2. Chandrababha, K., Shwetha, A. N., Kavitha, M., & Sumathi, R. (2021). Real time-Employee Emotion Detection system (RtEED) using Machine Learning. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. <https://doi.org/10.1109/icicv50876.2021.9388510>
3. Darapaneni, N., Singh, D. P., Chandra, S., Paduri, A. R., Kilhore, N., Chopra, S., Srivastav, A., Tomer, R., & Deshmukh, S. S. (2021). Activity & Emotion Detection of Recognized kids in CCTV Video for Day Care Using SlowFast & CNN. <https://doi.org/10.1109/aiiot52608.2021.9454221>
4. Loizou, C. P. (2021). An automated integrated speech and face imageanalysis system for the identification of human emotions. *Speech Communication*, 130, 15–26. <https://doi.org/10.1016/j.specom.2021.04.001>
5. Moolchandani, M., Dwivedi, S., Nigam, S., & Gupta, K. (2021). A survey on: Facial Emotion Recognition and Classification. <https://doi.org/10.1109/iccmc51019.2021.9418349>

6. Sahu, S., Kithani, E., Motwani, M., Motwani, S., & Ahuja, A. (2021). Stress Detection of Office Employees Using Sentiment Analysis. In *Advances in intelligent systems and computing* (pp. 143–153). Springer Nature. https://doi.org/10.1007/978-981-33-4367-2_15
7. Tsuji, S., Sato, N., Ara, K., & Yano, K. (2021). Estimating Group Stress Level by Measuring Body Motion. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.634722>
8. Tagnithammou, T., Monacelli, E., Ferszterowski, A., & Trenoras, L. (2021). Emotional state detection on mobility vehicle using camera: Feasibility and evaluation study. *Biomedical Signal Processing and Control*, 66, 102419. <https://doi.org/10.1016/j.bspc.2021.102419>
9. Xia, Z., Hong, X., Gao, X., Feng, X., & Zhao, G. (2020). Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-Expressions. *IEEE Transactions on Multimedia*, 22(3), 626–640. <https://doi.org/10.1109/tmm.2019.2931351>
10. Zhang, H., Feng, L., Li, N., Jin, Z., & Cao, L. (2020). Video-Based Stress Detection through Deep Learning. *Sensors*, 20(19), 5552. <https://doi.org/10.3390/s20195552>