Multimodal Multiclass Emotion Detection using Deep Learning

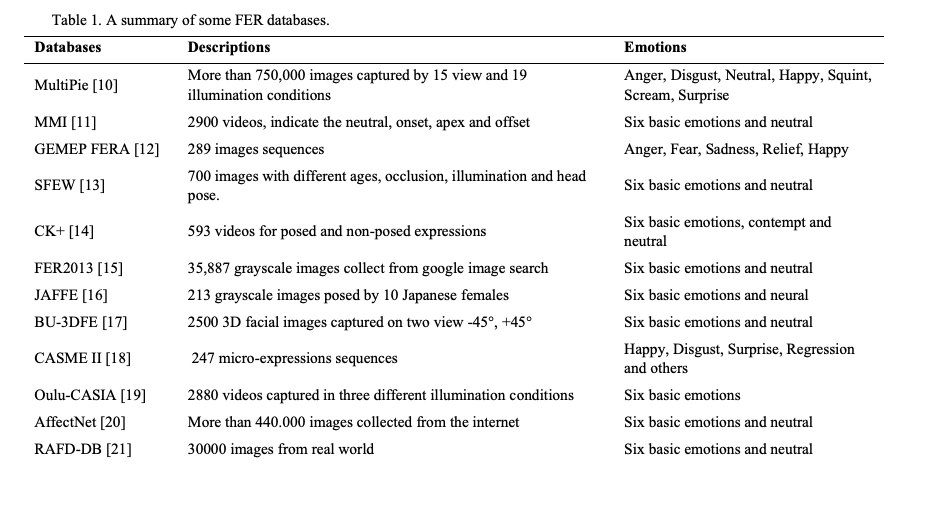
### Team Members :

1. Adithan K – 21011101009
2. Nandhini E-21011101042

### Dataset Availability :

### Dataset Description for the image aspect:

## Facial emotion recognition:



Cite: The 2nd International Workshop on the Future of Internet of Everything (FIoE) August 9-12, 2020, Leuven

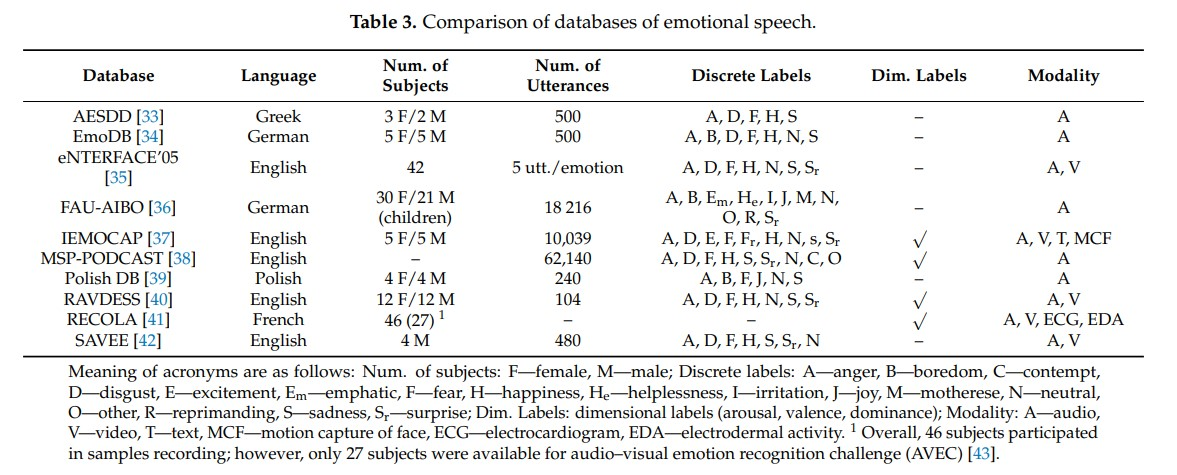
, Belgium Facial emotion recognition using deep learning: review and insights Wafa Mellouka\* , Wahida Handouzi

The best datasets include MultiPie with 750,000 images of 15 views and 19 illuminations, MMI's 2900 videos for six basic emotions, GEMEP FERA's 289 image sequences for five emotions, SFEW's 700 diverse images for six emotions, CK+'s 593 videos for six emotions and contempt, FER2013's 35,887 images, JAFFE's 213 images, BU-3DFE's 2500 3D facial images, CASME II's 247 micro-expressions, Oulu-CASIA's 2880 videos, AffectNet's 440,000 images, and RAFD-DB's 30,000 real-world images for six emotions and neutral.

The ‘IIT Roorkee Speech & Image Emotion Recognition (IIT-R SIER) dataset has been constructed using Balanced Twitter for Sentiment Analysis (B-T4SA) dataset. The recent text-to-speech models generate high-quality audio that can be used as a valid approximation of natural audio signals [21, 22, 25]. A pre-trained state-of-the-art text-to-speech model, DeepSpeech3 [25], has been used to convert the text from the B-T4SA dataset to speech.

Cite : https://arxiv.org/pdf/2208.11868v2.pdf

## Speech emotion recognition:



Cite: A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism

The databases can be divided into three basic categories: 6 of 29

• Simulated (acted): Professional actors express emotions through scripted scenarios.

• Elicited (induced): Emotions are created via artificially induced situations. With this approach, it is possible to achieve more natural recordings and simultaneously to have control over the emotional and lexical content of recordings.

• Spontaneous (natural): Spontaneous audio recordings are being extracted from various reality shows. The disadvantage of real-world audio samples is that they may be distorted by background noise and reverberation. Another drawback is that the natural or spontaneous databases often contain unbalanced emotional categories.

### COGMEN: COntextualized GNN based Multimodal Emotion recognition:

Dataset:

We experiment for the Emotion Recognition task on the two widely used datasets: IEMOCAP (Busso et al., 2008) and MOSEI (Zadeh et al., 2018b). IEMOCAP is a dyadic multimodal emotion recognition dataset where each utterance in a dialogue is labelled with one of the six emotion categories.

In literature, two IEMOCAP settings are used for testing, one with 4 emotions (anger, sadness, happiness, neutral) and one with 6 emotions. We experiment with both of these settings. MOSEI is a multimodal emotion recognition dataset annotated with 7 sentiments (-3 (highly negative) to +3 (highly positive)) and 6 emotion labels (happiness, sadness, disgust, fear, surprise, and anger).

For IEMOCAP, audio features (size 100) are extracted using OpenSmile (Eyben et al., 2010), video features (size 512) are taken from Baltrusaitis et al. (2018), and text features (size 768) are extracted using sBERT.

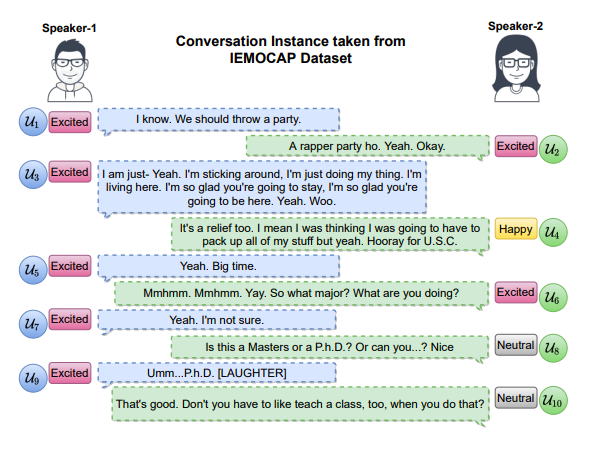
The IEMOCAP benchmark (Busso et al., 2008) consists of a conversation between ten distinct speakers. The dataset contains two-way conversations in videos where every video clip contains a single dyadic English dialogue.

CMU-MOSEI:

The CMU Multimodal Opinion Sentiment and Emotion Intensity (Bagher Zadeh et al., 2018) is an English language dataset containing more than 65 hours of annotated video from more than 10000 speakers and 250 topics.

### Application :

Speech Aspect:



Visual Aspect:

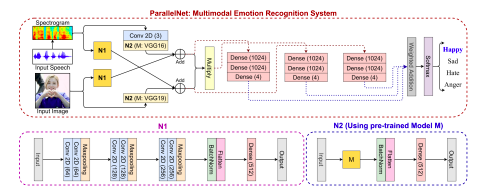


A screenshot of a heat map

Description automatically generated

# Architecture :

Multi-Modal Architecture using speech and series of images



Interpretable Multimodal Emotion Recognition using Hybrid Fusion of Speech and Image Data Puneet Kumar1\*† , Sarthak Malik2† and Balasubramanian Raman1 1Department of Computer Science and Engineering, Indian Institute of Technology Roorkee. 2Department of Electrical Engineering, Indian Institute of Technology Roorkee.

Architecture Cogmen using DL:

A diagram of a graph

Description automatically generated

A diagram of a face

Description automatically generated

Facial emotion recognition using deep learning: review and insights Wafa Mellouka\* , Wahida Handouzia

Speech to emotion classifier:

(speech to text conversion to be done before feeding into this model)

A diagram of a block diagram

Description automatically generated

Speech Emotion Recognition with deep learning Hadhami Aouani1,\* and Yassine Ben Ayed1,\*\*

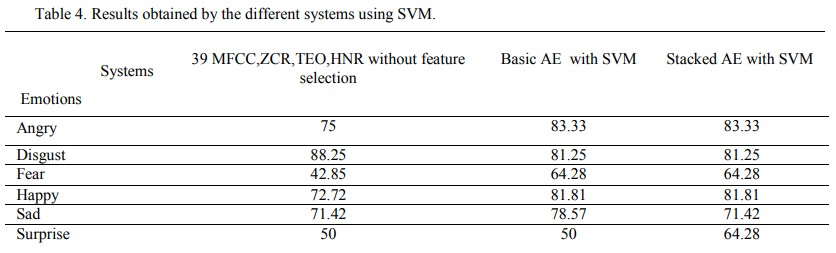
### Result analysis:

## Facial emotions:

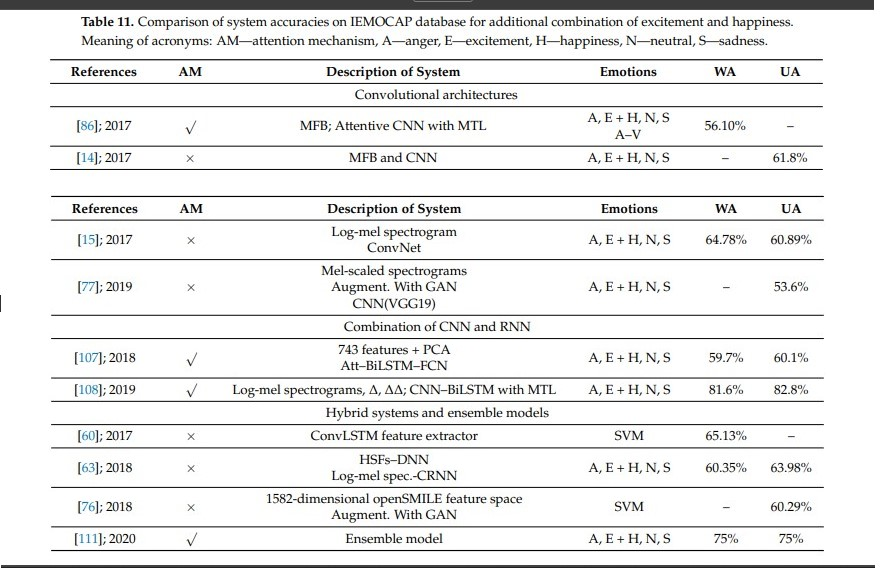
# 

Cite: Facial emotion recognition using deep learning: review and insights Wafa Mellouka\* , Wahida Handouz

## Speech emotions:



Cite: Speech Emotion Recognition with deep learning Hadhami Aouani1,\* and Yassine Ben Ayed



Cite: A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism

## Cogmen:

# 

Cite: COGMEN: COntextualized GNN based Multimodal Emotion recognition

### Proposed architecture:

Domain Selection: specific domains of interest is speech and image processing.

Model Selection: For each domain, curate a set of deep learning models that have shown high performance and state-of-the-art results. This includes models like COGMEN and ParellelNet.

Ensemble Construction: Build an ensemble of models for each domain. An ensemble is a combination of multiple models that work together to improve overall performance. There are several ensemble methods to consider, such as bagging, boosting, and stacking. You'll need to decide which ensemble method is most suitable for your goals.

Ensemble Training: Train the ensemble of models using a diverse set of data from the respective domain. The ensemble's collective predictions will be used to make final decisions. the dataset includes : IEMOCAP and MOSEI, MultiPie.

Integration of audio and video (Hybrid Approach): To further enhance performance, consider integrating audio and image processing methods. This could involve fusing the outputs of image and speech models to leverage the strengths of both modalities. For example, in a multimodal context, you could process both spoken words and corresponding transcriptions to improve accuracy.

Evaluation and Fine-tuning: Assess the performance of the ensemble methods on relevant evaluation metrics, such as accuracy, precision, recall, and F1 score. Fine-tune the ensemble parameters as needed to achieve optimal results.

Comparison with Existing Methods: Benchmark the proposed ensemble approach against pre-existing multimodal methods to demonstrate its superiority in terms of accuracy and loss minimization.

Deployment and Scaling: Once you have a well-performing ensemble model, deploy it in real-world applications. Ensure that the architecture can scale to handle large amounts of data and accommodate updates and improvements over time.

### Code :

<https://github.com/exploration-lab/cogmen>

<https://github.com/declare-lab/conv-emotion>

<https://github.com/exploration-lab/shapes-of-emotion>

<https://paperswithcode.com/paper/interpretable-multimodal-emotion-recognition>

### Citations:

Agarwal, Harsh, et al. *Shapes of Emotions: Multimodal Emotion Recognition in Conversations via Emotion Shifts*.

Aouani, Hadhami, and Yassine Ben Ayed. “Speech Emotion Recognition with Deep Learning.” *Procedia Computer Science*, vol. 176, 2020, pp. 251–260, https://doi.org/10.1016/j.procs.2020.08.027.

Huang, Zi-Yu, et al. *A Study on Computer Vision for Facial Emotion Recognition*. Vol. 13, no. 1, 24 May 2023, https://doi.org/10.1038/s41598-023-35446-4. Accessed 14 July 2023.

Joshi, Abhinav, et al. *COGMEN: COntextualized GNN Based Multimodal Emotion RecognitioN*.

Kumar, Puneet, et al. *Springer Nature 2021 L a T E X Template Interpretable Multimodal Emotion Recognition Using Hybrid Fusion of Speech and Image Data*.

Lieskovská, Eva, et al. “A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism.” *Electronics*, vol. 10, no. 10, 13 May 2021, p. 1163, https://doi.org/10.3390/electronics10101163.

Mellouk, Wafa, and Wahida Handouzi. “Facial Emotion Recognition Using Deep Learning: Review and Insights.” *Procedia Computer Science*, vol. 175, 2020, pp. 689–694, https://doi.org/10.1016/j.procs.2020.07.101.