Early Abnormal Mental Health Detection using Knowledge Distillation Framework on Audio-Visual Data

Team members: Gopagoni sreya, Jyoti Dhiman
Under the guidance of Dr.Mukesh Saini

Problem Statement :

Early Abnormal Mental Health Detection using Knowledge Distillation Framework on Audio-Visual Data.

- The primary objective of this research is to develop an advanced machine learning architecture that facilitates the early detection of abnormal mental health conditions by integrating audio and visual cues through a teacher-student model framework.

Approach:
Develop Teacher-Student Network for Mental Health Assessment

- **Teacher Model (Emotion Mapping from Image/Video Data):** Input: Facial images or video frames.

  Architecture: Use a deep convolutional neural network (CNN) or a combination of CNN and recurrent layers to extract features from the visual data. This model learns to map facial expressions to emotional states.

  Output: Emotion probabilities or embeddings numerical mapping to different emotional categories (e.g., happy, sad, neutral, etc.).

- **Student Model (Emotion Mapping from Speech Data):** Input: Audio samples (speech data).

  Architecture: Employ a recurrent neural network (RNN) or a combination of CNN and recurrent layers to process the audio data. This model learns to map speech patterns to emotional states.

  Output: Emotion probabilities or embeddings numerical mapping to different emotional categories (e.g., happy, sad, neutral, etc.).

- **Abnormal Behaviour Detection:** During inference, compare the emotional predictions made by the student model using speech data with those made by the teacher model using image or video data.

  Detection: If the emotional predictions from audio and visual modalities significantly differ (beyond a predefined threshold), it could indicate abnormal behaviour or abnormal mental health condition.

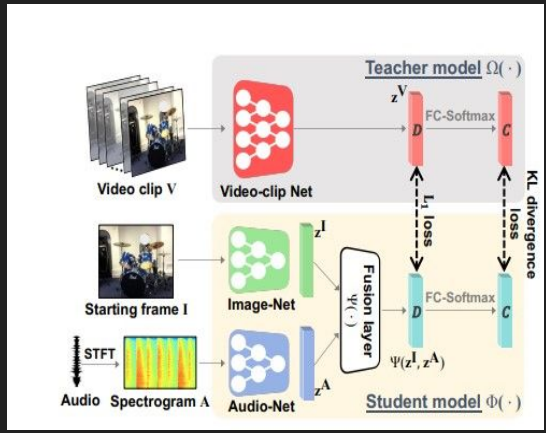
Research before Midsem

1.Teacher-Student Architecture for Knowledge Learning: A Survey

- https://arxiv.org/pdf/2210.17332.pdf
- The paper discusses the concept of a teacher-student architecture exploring its various types, practical applications, and potential future advancements and knowledge distillation.

2. Listen to Look: Action Recognition by Previewing Audio

- https://openaccess.thecvf.com/content_CVPR_2020/papers/Gao_Listen_to_Look_Action_Recognition_by_Previewing_Audio_CVPR_2020_paper.pdf
- This paper talks abouts the optimizing video descriptor model using image and audio instead of untrimmed video by using teacher-student architecture.



Implementation before Midsem

- Successfully implemented the Teacher-student model for numerical data parameters using DNN to detect happy or sad.

  Understood the concept of teacher-student architecture, formulas,loss function, accuracy, predictions.

  https://colab.research.google.com/drive/1HzwaG0zjPik1v1p_K9kDt3DR1_KKHrow?usp=sharing

- Successfully implemented the Deep Image Classifier for Image data using CNN to detect emotions of a person.

  https://colab.research.google.com/drive/1J0xADwUyal5vIwaYCAmvP-lW_nrVpP-6?usp=sharing

Work done after Midsem

Data Collection:

- Attempted to gather 100 video clips featuring only one individual delivering a single dialogue. link for the data is given below.

  Data link: link

- we opted to work with the available dataset of over 1300 videos containing the target dialogue.

  Data link: https://affective-meld.github.io/

- But we encountered challenges. It proved intricate to isolate scenes with only one person, given that films often showcase various expressions and reactions from non-speakers. In the many videos in the dataset there contain other non-speaker faces.
- Hence, our strategy now involves developing algorithms to accurately track and isolate the speaker within the videos containing multiple persons.

Research Question: Does the emotion of the person change during a single dialogue delivery in the clips?

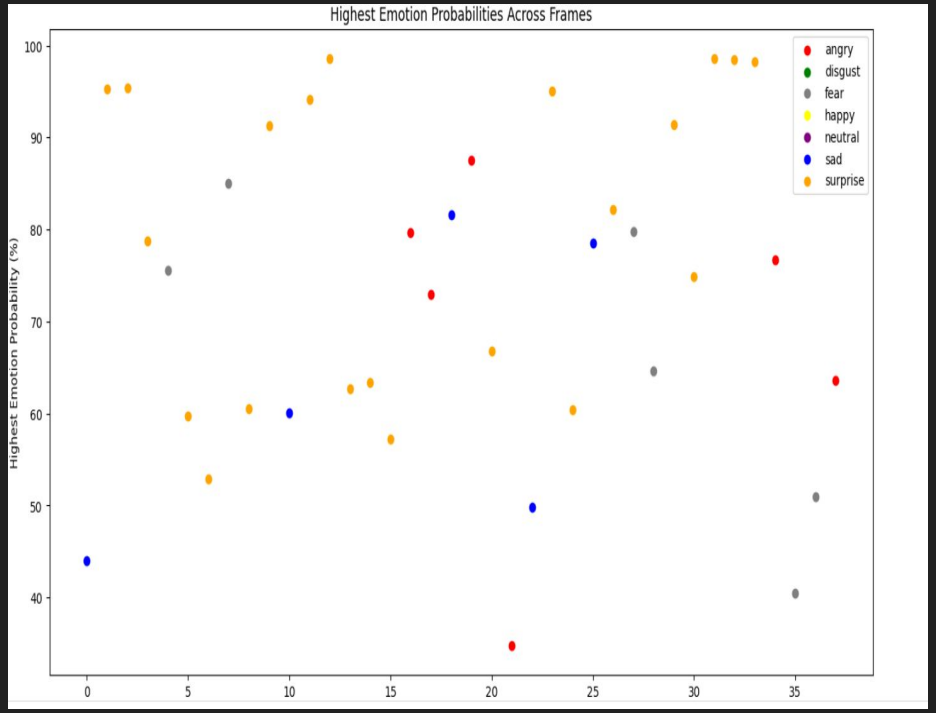
Importance: This question holds significance in shaping the final implementation of the teacher-student architecture framework for abnormal mental health detection. It directly influences the decision on what input to provide to the teacher model.

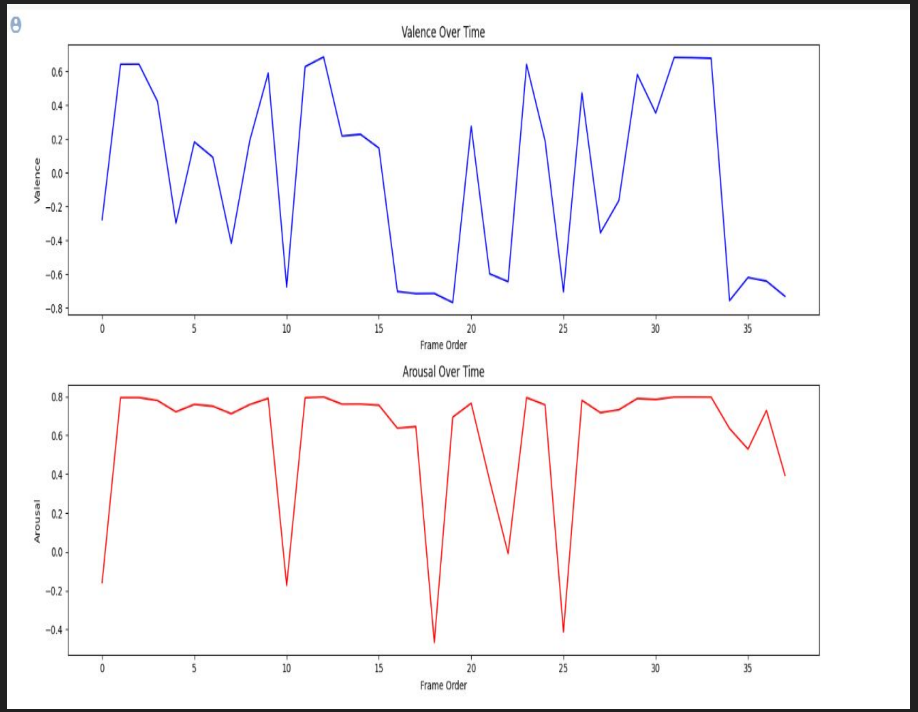
If our hypothesis that emotion does change comes out to be true,we can take a more effective approach like feeding the teacher model single image frames rather than entire videos or exploring alternative method.

Approach: For this we tried cropping faces in each video frame, plotting the most probable emotion on a graph, and considering 10 frames per second. The assumption was made that each frame contains either one or zero faces and whole video is of single person.

Plotted the emotions of the person in each frame for whole video/a whole dialogue (Surprise.mp4)

[Link]





Result and Challenges:

Results: The graph shows fluctuations when plotting emotions based on labels and relative stability when using valence arousal methods. This suggests that emotions do change during a video clip in both cases.

Challenges: The accuracy of the results is questionable due to the initial assumption that the entire video contains only one person. The plot contains emotions of non-speakers and frames without faces, causing significant variations in emotions.

Approach: To address this, we are working on algorithms to precisely track and isolate the speaker in videos with multiple persons. We aim to redo the emotion graph plotting and test our hypothesis again.

Final Problem statement for the End Sem

1. Emotion plotting after active speaker detection:

Algorithm: Clustering for Most Frequently Seen Person: Utilizing clustering techniques to identify and track the most frequently seen person in the video. This method aims to focus on the individual who appears most often, assuming that this person is likely the active speaker. After the active speaker was detected, its emotions were plotted.

2. From video clips of an active speaker depicting a particular emotion, we calculated the percentages and corresponding ranks of other emotions occurring with that main emotion.

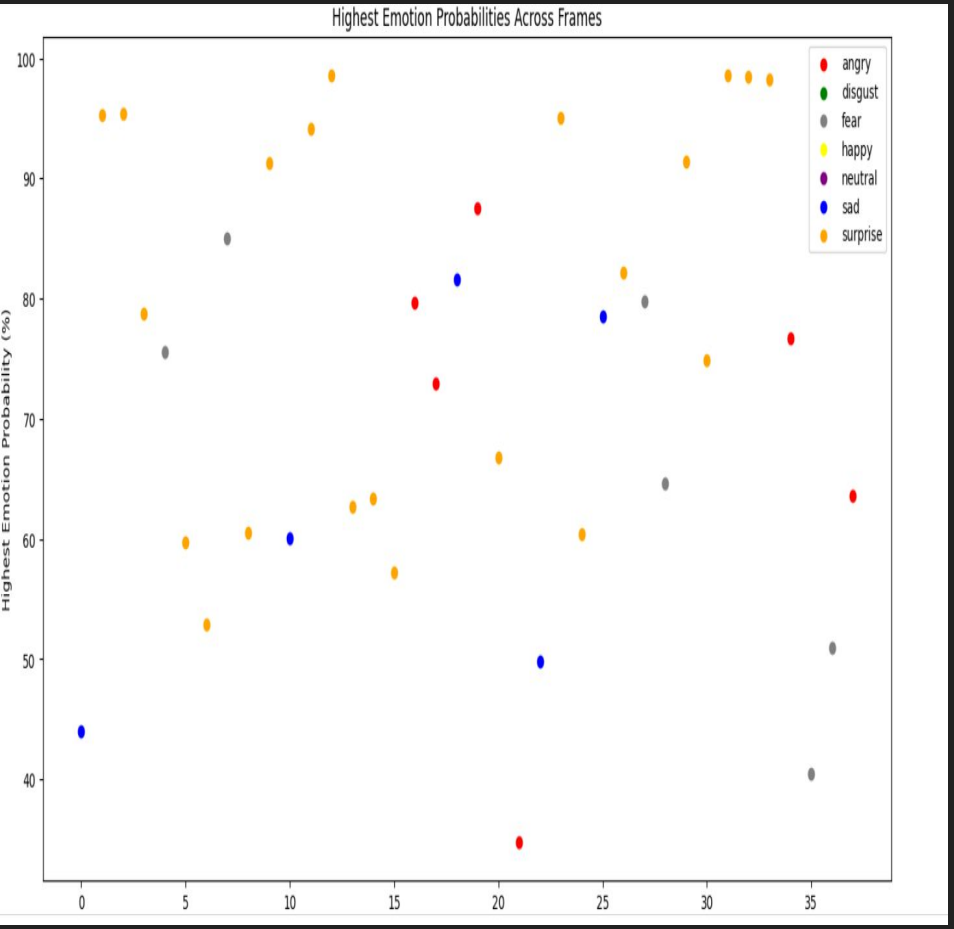
Clustering the most seen person in the video

1. First we have tried to crop the faces in a image frame and store it in a list.
2. The tried to append the list of faces in all the frames in the video.
3. The next step involves feeding this compiled list of faces into a K-means clustering algorithm to group similar images. The underlying assumption is that the largest cluster would represent the person most frequently seen in the video, assumed to be the active speaker.

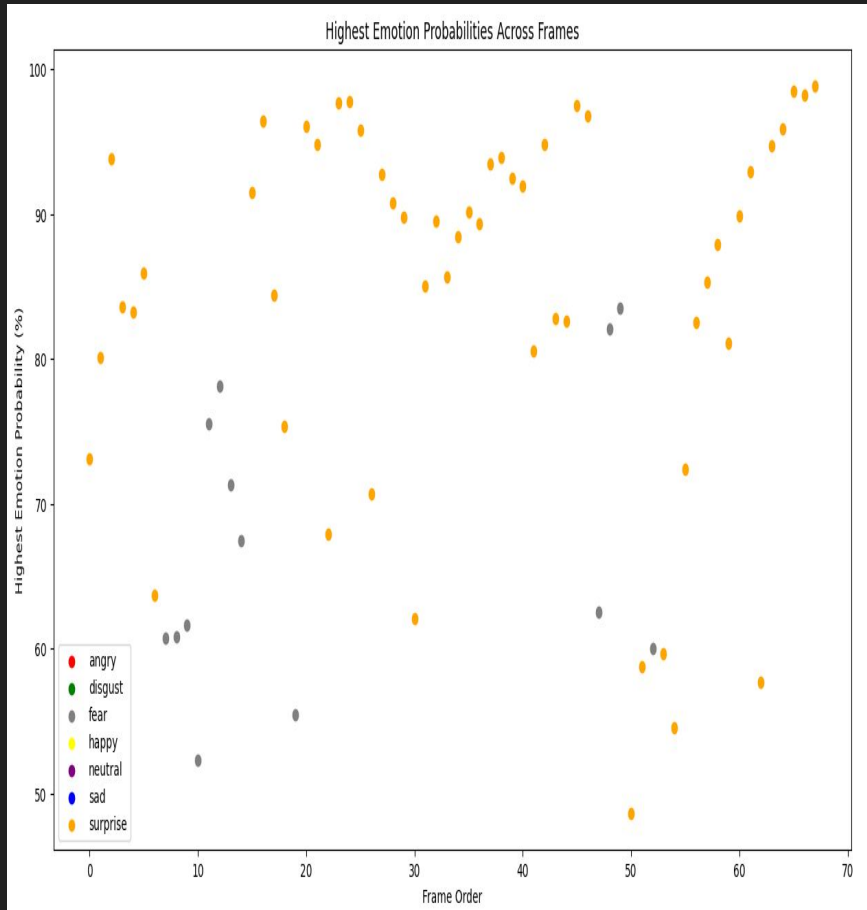
Colab link:
https://colab.research.google.com/drive/1hi_gmQe30GZgu2GGyNndzamZ1P0LgWu8?usp=sharing

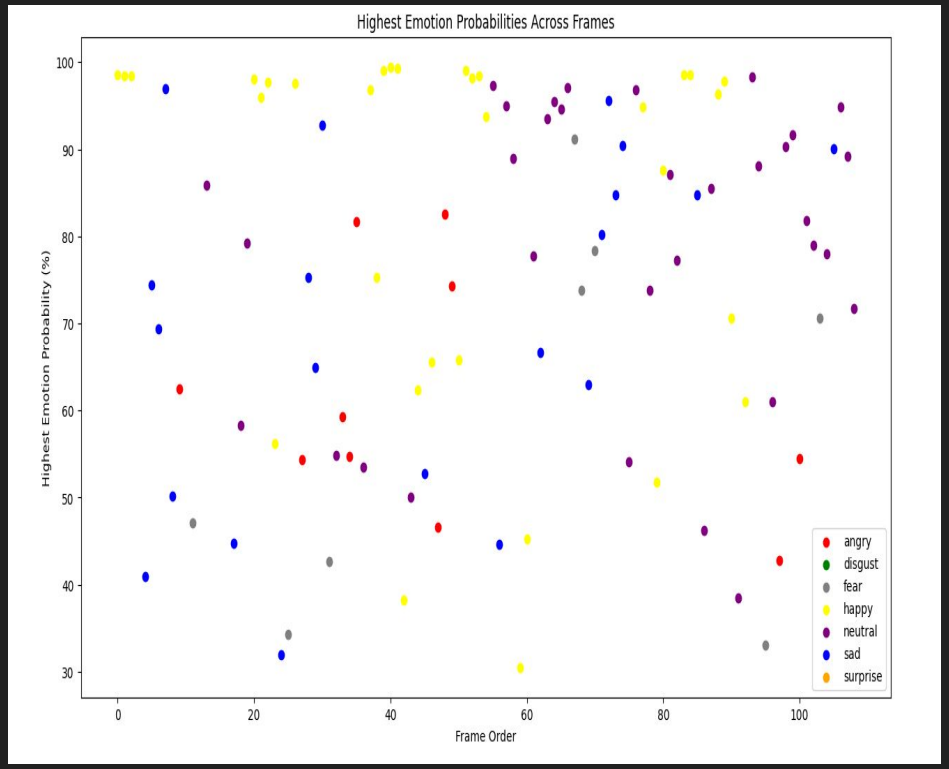
For surprise.mp4

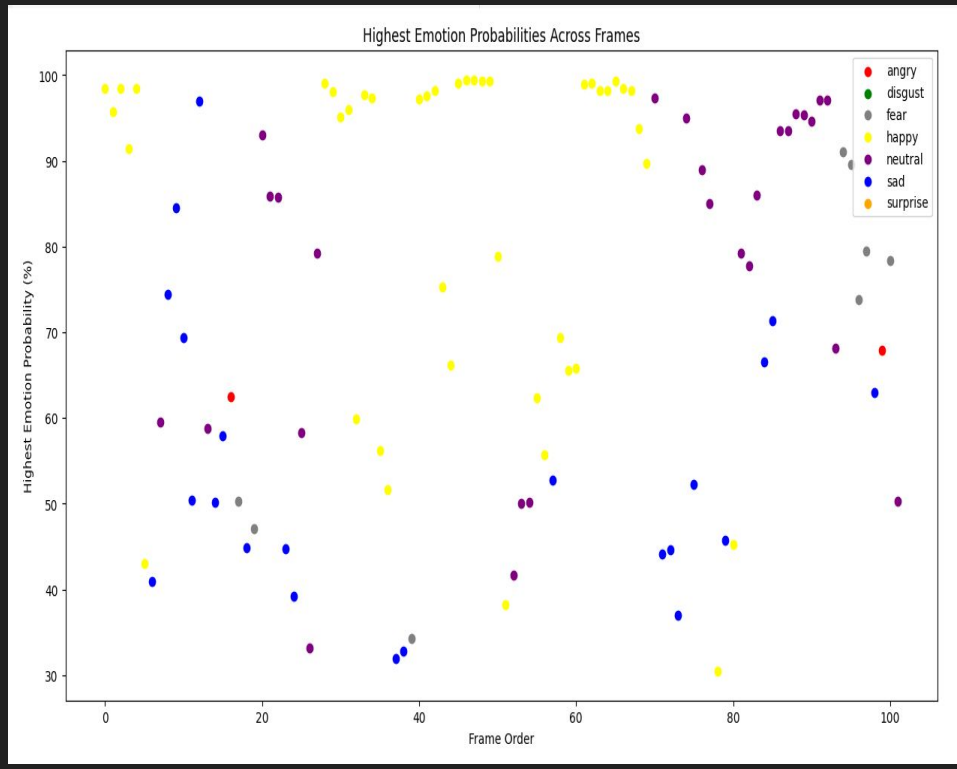
Before



After

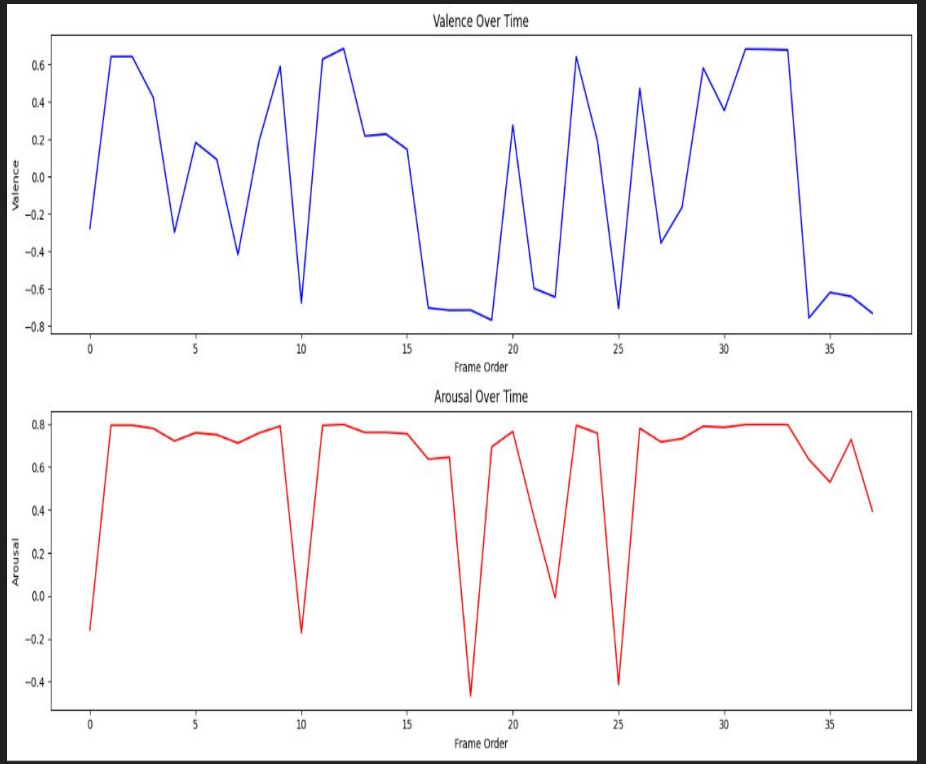



Highest Emotion Probabilities Across Frames
Highest Emotion Probability (%)
Frame Order
angry
disgust
fear
happy
neutral
sad
surprise

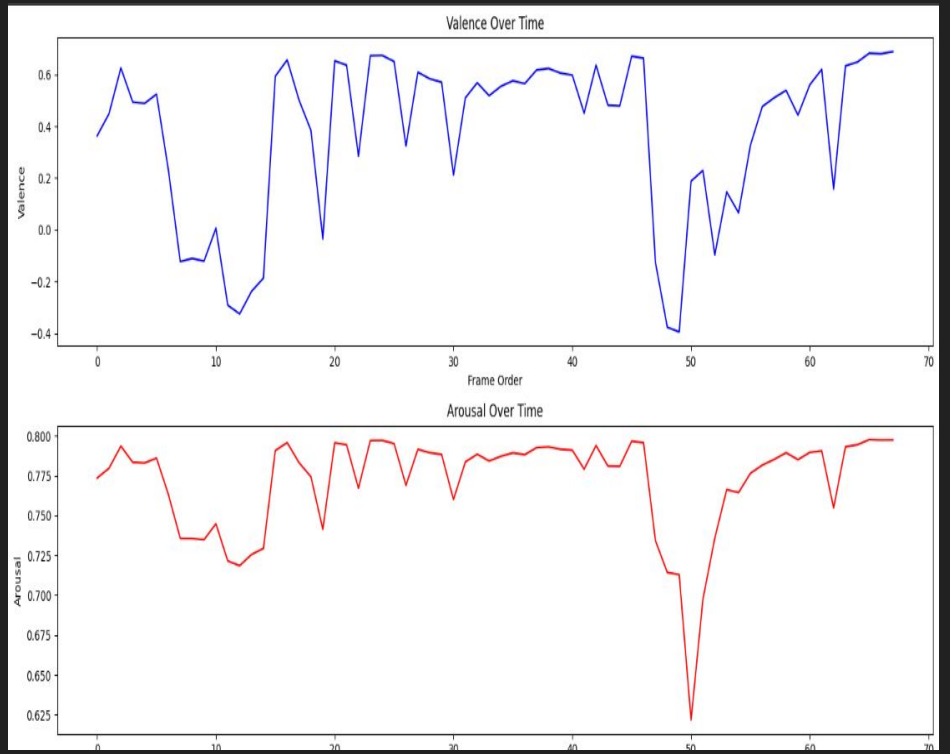

Highest Emotion Probabilities Across Frames
Highest Emotion Probability (%)
Frame Order
angry
disgust
fear
happy
neutral
sad
surprise

Valence arousal comparison:

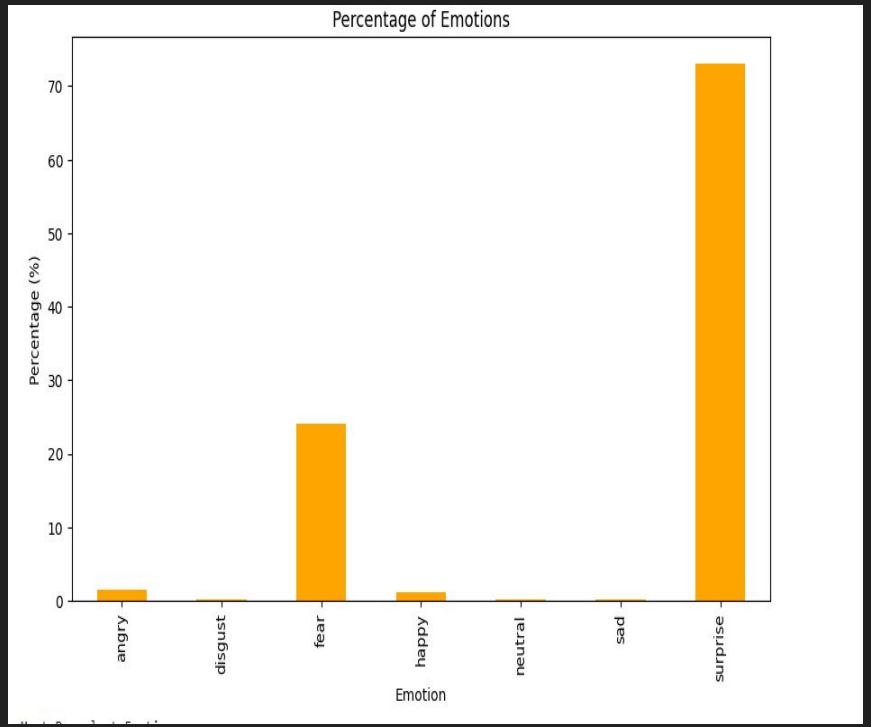
Before (more fluctuations)



After (less fluctuations)



The percentages and corresponding ranks of other emotions occurring with that main emotion (from 4 surprise videos)




```
Most Prevalent Emotion: surprise
Emotion Percentages:
angry          1.409846
disgust        0.185263
fear          24.042019
happy          1.127031
neutral        0.080902
sad            0.090263
surprise      73.064676
dtype: float64
rank:
angry          3.0
disgust        5.0
fear           2.0
happy          4.0
neutral        7.0
sad            6.0
surprise       1.0
dtype: float64
```

Conclusions

We have plotted emotions using image and video frames till now. Our aim in the next semester is to use audio and speech in emotion detection by mapping it to the image model using teacher student architecture.

Develop a mechanism to measure and quantify the alignment or disparity between the emotion predictions of the teacher and student models. Design an alignment loss that guides the student model to predict emotions that are consistent with those predicted by the teacher model, fostering coherence between audio and visual emotion assessments.

If the emotional predictions from audio and visual modalities significantly differ (beyond a predefined threshold), it could indicate abnormal behaviour or abnormal mental health condition.

THANKYOU