# Multimodel Analytics of Collaborative Learning Activities


IIT Gandhinagar
Indian Institute of Technology Gandhinagar

Makkena Lakshmi Manasa
makkena.manasa@iitgn.ac.in

Siju Niyati Samji
niyati.siju@iitgn.ac.in

Rushitha Thipparapu
rushitha.thipparapu@iitgn.ac.in

Under the Supervision of
Prof. Aditi Kothiyal
aditi.kothiyal@iitgn.ac.in


## Introduction

- Collaborative learning activities are rich, multimodal interactions that demonstrate how partners coordinate and communicate.
- Understanding these behaviours helps identify productive teamwork.
- We analysed audio-video recordings of 17 pairs in a hands-on open-ended task.
- Four modalities are used: emotions, gaze, hand movements, and speech.
- Together, they capture the quality of engagement and interaction.
- Features are normalised and reduced with PCA to highlight key patterns.
- Clustering uncovers distinct collaboration profiles among pairs.
- This offers a data-driven approach to characterise effective collaboration.

## Implementation

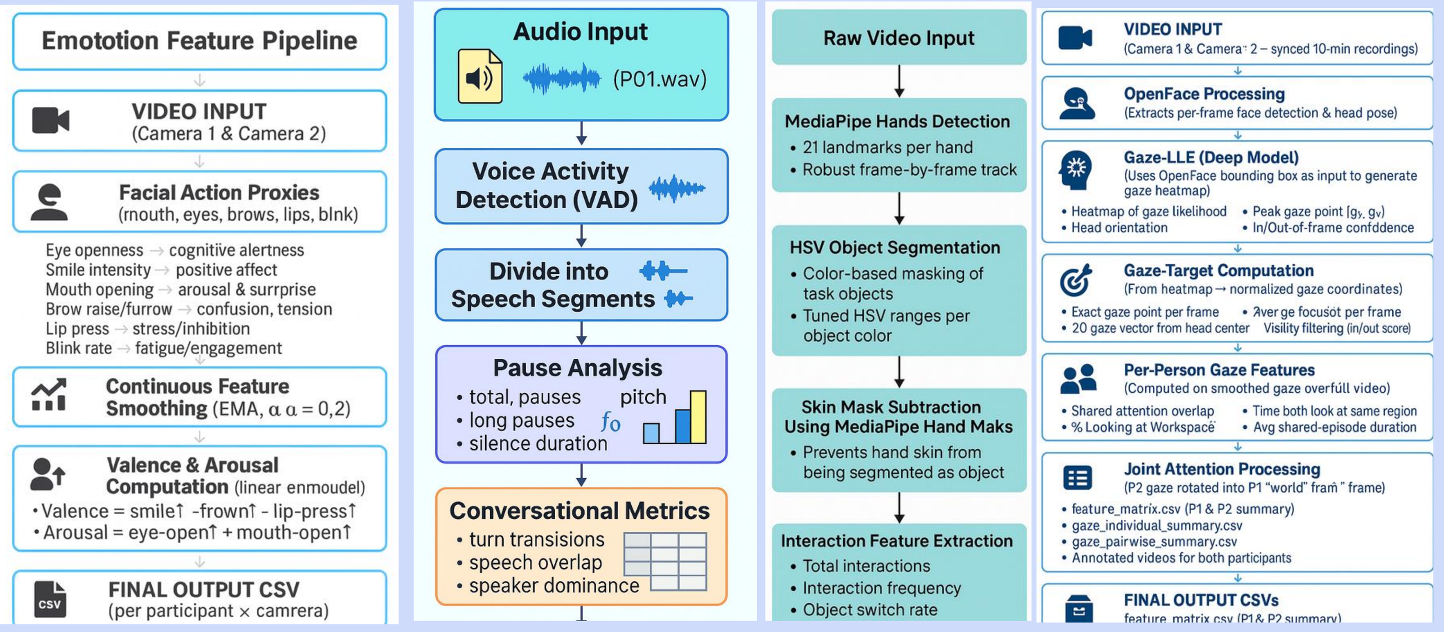



Figure 4: Methodology used for Emotions, Audio, Hand moment and Eye gaze respectively

## Feature Matrix Calulation

**Arousal Calculation** is a weighted combination produces the score:

- arousal = 0.55·eye_open + 0.45·mouth_open.

**Speech features**

- Short pause duration < 5 seconds
- Long pause duration >= 10 seconds
- Total Speech = Σ (end – start)
- Total Silence Duration = Σ (NextStart – PrevEnd)
- Dominance = Speech_A / (Speech_A + Speech_B)
- Overlap = Σ intersection(SpeechA, SpeechB)
- Mean Pitch = Average(Pitch across voiced segments)
- Mean Energy = Average(Signal² / Length)

**Eye Gaze features calcation**

- Focus Duration on One Object=mean fixation duration over all fixations
  - Fixation Duration=1/N ∑(endk−startk)
- Average duration of fixation segments (gaze staying within radius 0.03)
- Gaze Shifts per Second=Number of gaze shifts/Video duration (sec)
- % Time Looking at Workspace = (frames looking_workspace)/total frames ×100
- % Time Looking at Partner = (frames looking_to_partner)/total frames ×100
- %Elsewhere=100−(%WS+%Partner)
- Number of Fixations =no of segments where fixation duration is greater than zero.
- Spread = average distance between points in a fixation and its centroid
- Shared Attention Overlap Ratio IoU = Overlapping attention area ÷ Total attention area
  - IoU=∣H1∩H2∣/∣H1∪H2∣

## Results

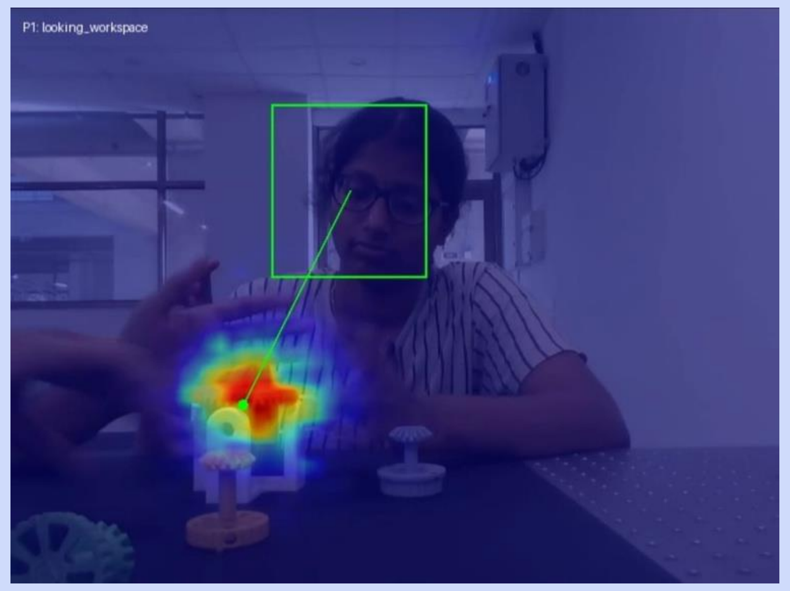



Figure 1: Eye gaze heat map

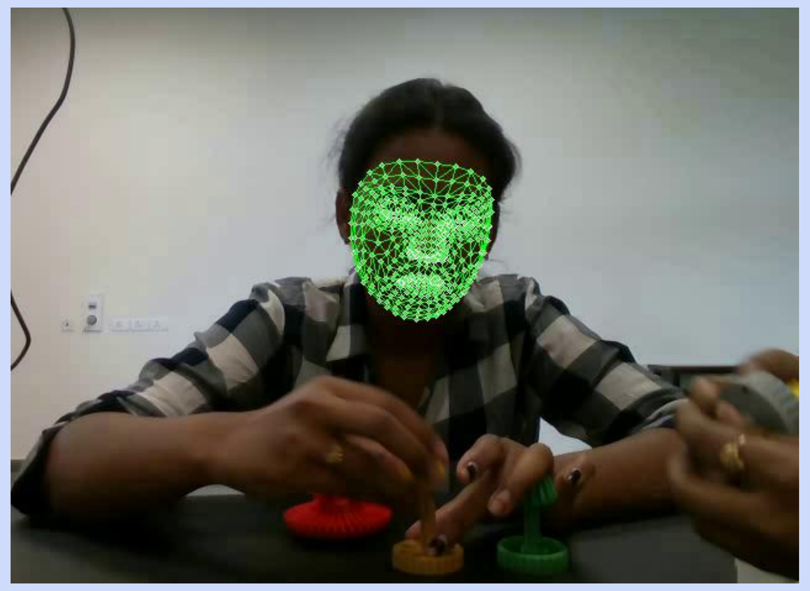
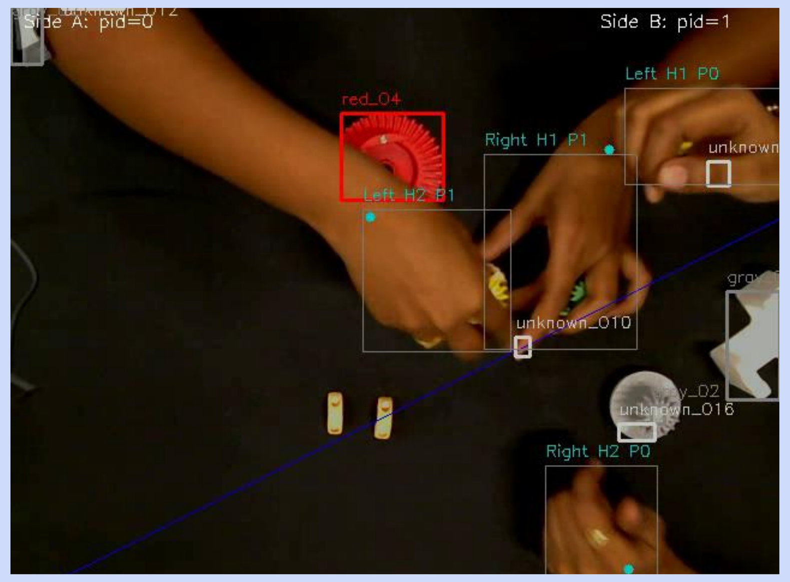


Figure 2: Media pipe facial Landmarks



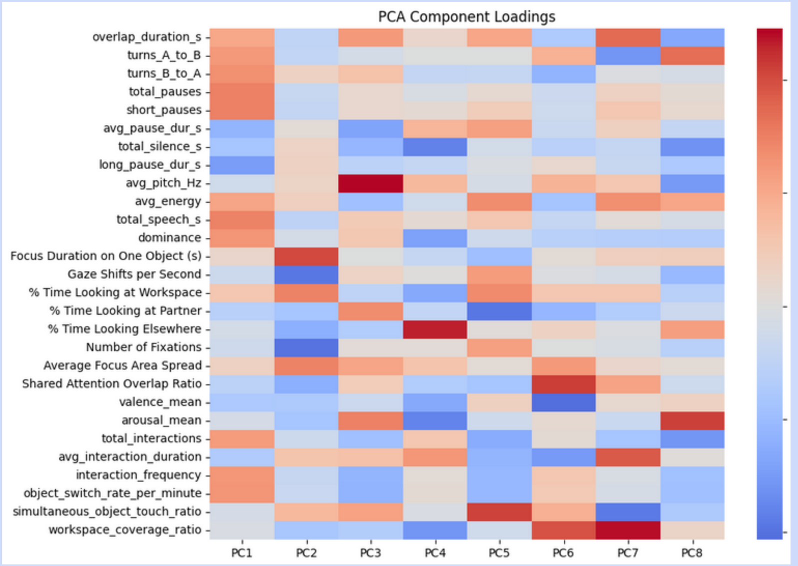



Figure 3: Bounding boxes on objects

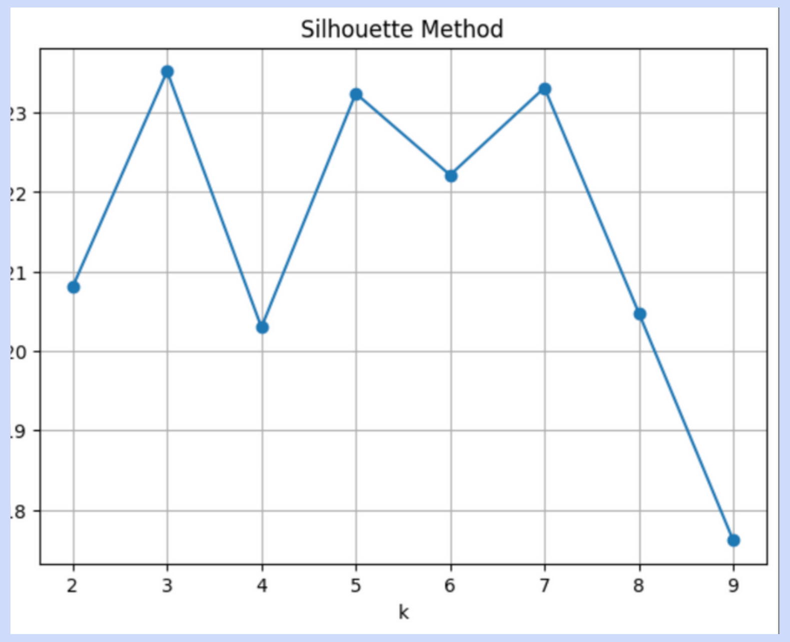



Figure 6: Deciding the K value

**Valence =** = 0.65·smile − 0.25·brow_relax − 0.10·lip_press.

**Hand Movement feature**

- total_interactions = count(hand–object distance < threshold)
- avg_interaction_duration = mean(duration of_continuous_contact_sequences)
- interaction_frequency = total_interactions / total_time_minutes
- object_switch_rate = num object ID transitions / total_time_minutes
- simultaneous_touch_ratio = frames both hands touching / total_frames
- workspace_coverage = area(hand centroid path) / total_workspace_area

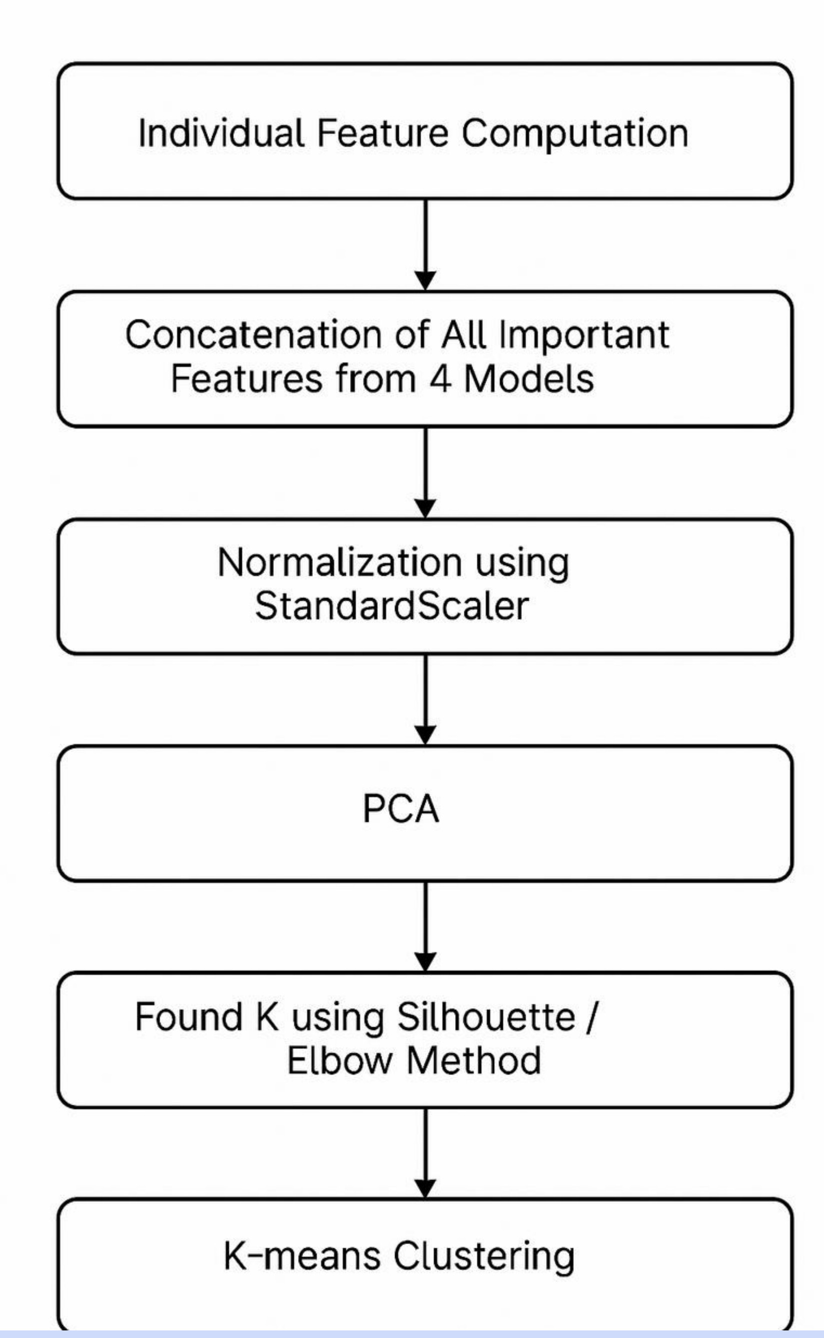
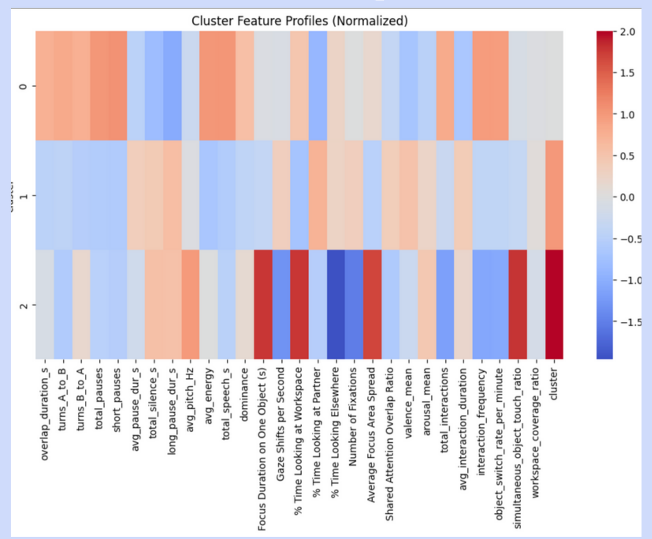
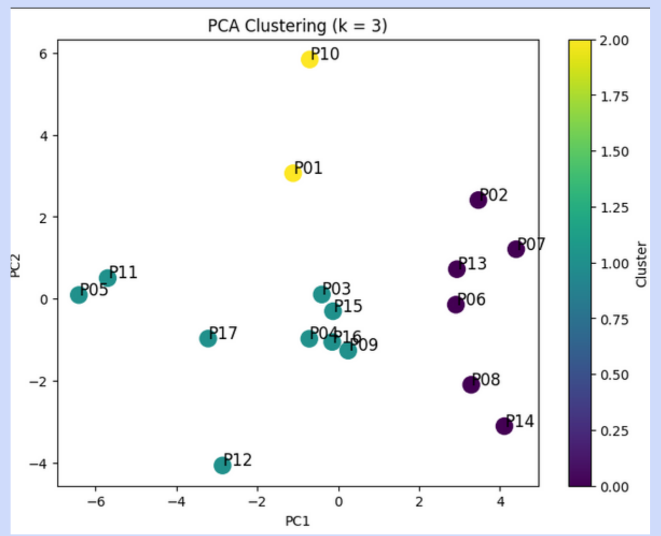



Figure 7: Flow of the Project

## Experimental Results




| Cluster ID | Average Learning Gain (LG) | Assigned Name | Hypothesized Profil |
|---|---|---|---|
| Cluster 2 (Yellow points) | 0.75 | Productive Learners | High engagement, variability, balance pairs) |
| Cluster 1 (Green points) | 0.72 | Effective Collaborators | Low initial activity, time for reflection, entrainment (9 pai |
| Cluster 1 (Blue Points) | 0.42 | Passive Learners | Low total talk time, expression, high sp overlap/interruptio |

Figure 5: Final Clusters and their avg. LG

## Challenges Faced

**Hand & Object Actions**

- YOLO struggled with small colored objects.
- Hand interactions (reach/touch/grasp) were hard to classify due to occlusions and varied hand poses.

**Eye Gaze**

- Slow face detection and high GPU load affected tracking.

**Speech**

- Background noise and overlapping speech.
- Emotion/tone cues are hard to capture; audio–video sync is required.
- Getting an accurate transcript without ground truth

**Emotions**

- Struggling with OpenFace Installation on GPU system

## Future Work

- Add more modalities (posture, facial cues, task metrics).
- Use advanced clustering to find deeper learner patterns.
- Train supervised models to predict learning gain from multimodal signals.
- Test on larger, diverse datasets for stronger validation.

## References


[1] Nasir, J., Kothiyal, A., Bruno, B. et al. Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. Intern. J. Comput. Collab. Learn 16, 485–523 (2021). https://doi.org/10.1007/s11412-021-09358-2
[2] https://openface-api.readthedocs.io/
[3] https://mediapipe.readthedocs.io/
[4] https://github.com/fkryan/gazelle
[5] https://github.com/snakers4/silero-vad