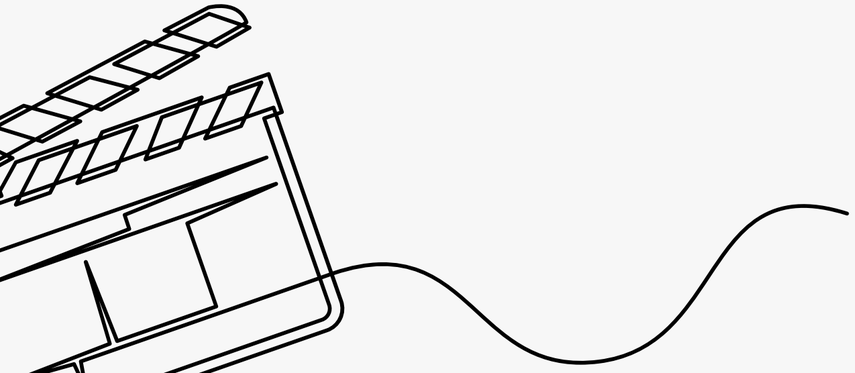
DIRECTORS CUT

IDENTIFYING SALIENT REGIONS OF AUDIO-VISUAL STIMULI



-AVISHI, SHUBHAM, YASH

PROBLEM STATEMENT

Identifying the most **emotionally salient** regions of visual content.

THE CHALLENGE?

- Large amount of time wasted trying to figure out what content resonates the most
- Automated solutions are bulky
- Lot of computational time and power required

WHERE CAN IT BE USED?

Streaming Platforms

Movies and Ads

Improved Mental Health
Monitoring in Therapy

Educational Videos

LITERATURE REVIEW

Multimodal Emotion Recognition via Convolutional Neural Networks: Comparison of different strategies on two multimodal datasets (2023)

- Explores the use of **multimodal approaches** to improve the accuracy of emotion recognition by combining various data types such as facial expressions, motion data (optical flow), and auditory cues (Mel Spectrograms)
- Researchers used **specialised CNN architectures** for each type of input and the information from each was combined into a single system. This produced better results than using just one type of data.
- Used **early fusion** so model can learn patterns across all modalities from the beginning. It helps it **detect relationships** between different modalities more effectively.

LEARNINGS:

- Since we have access to multiple modalities and care about relevance, we should consider using **early fusion** for our project
- The use of **CNNs** for handling each type of input can be applied in our project to process different types of data effectively.

Bilotti, U., Bisogni, C., De Marsico, M., & Tramonte, S. (2023). Multimodal emotion recognition via convolutional neural networks: Comparison of different strategies on two multimodal datasets. Proceedings of the International Joint Conference on Neural Networks (IJCNN), 1–8

Hierarchical Domain-Adapted Feature Learning for Video Saliency Prediction (2021)

- **3D Fully Convolutional Network (FCN):** To capture spatial and temporal features simultaneously across video frames, crucial for learning complex motion patterns.
- **Hierarchical Supervision:** The model generates intermediate conspicuity maps at multiple abstraction levels, enhancing the capture of fine-grained details as well as broader visual context.
- **Domain Adaptatios:** By gradient reversal layers, the model learns domain-invariant features, improving its adaptability to new datasets without specific labeled data.

LEARNINGS:

- Using a **3D FCN** will let our model not only analyze what's happening in a single frame but also understand how things are evolving over multiple frames.

Bellitto, G., Proietto Salanitri, F., Palazzo, S., Rundo, F., Giordano, D., & Spampinato, C. (2021). Hierarchical domain-adapted feature learning for video saliency prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(4), 678–691. https://doi.org/10.1109/TPAMI.2020.3034568

Hybrid time-spatial video saliency detection method to enhance human action recognition systems (2024)

- **Spatial Saliency**: Identifies critical elements within each frame, such as facial features, gestures, or objects, using visual cues like color, edges, and contrast.
- **Temporal Saliency**: Tracks movement and motion across frames using optical flow to detect dynamic changes, such as jumping or gesturing.
- **Nonlinear Combination** (with a Multi-Objective Genetic Algorithm): The process uses a genetic algorithm which optimises a non-linear combination of spatial and temporal saliency map, testing different combinations until it finds the most effective one.

LEARNINGS:

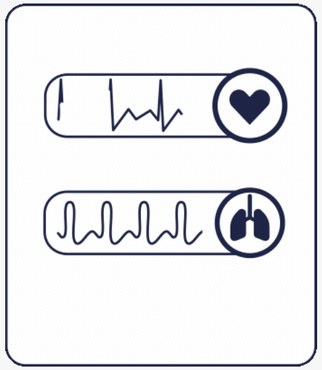
- Use a **genetic algorithm** to find the best way to combine static and motion information, making sure the system identifies both what's happening and where it's happening.

Gharahbagh, A. A., Hajihashemi, V., Ferreira, M. C., Machado, J. J. M., & Tavares, J. (2024). Hybrid time-spatial video saliency detection method to enhance human action recognition systems. IEEE Transactions on Multimedia, 26(1), 123–136. https://doi.org/10.1109/TMM.2023.3267890

DATA SET

DEAP DATASET

The DEAP dataset is a **multimodal dataset** created for **emotion recognition**

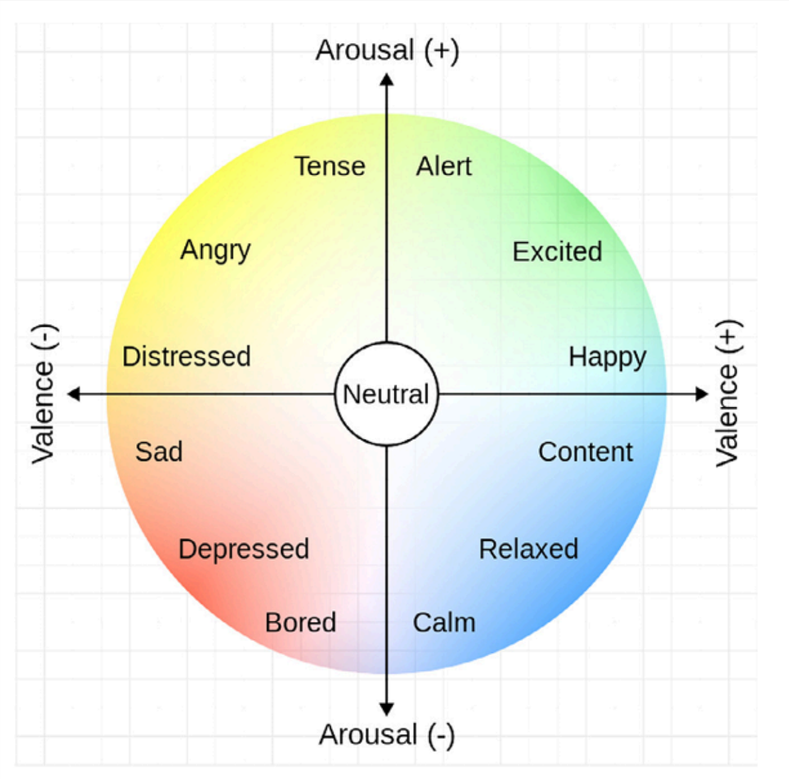


DATA COLLECTION : SELECTING STIMULI

Database content summary

Online subjective annotation	
Number of videos	120
Video duration	1 minute affective highlight (section 2.2)
Selection method	60 via last.fm affective tags, 60 manually selected
No. of ratings per video	14 - 16
Rating scales	Arousal Valence Dominance
Rating values	Discrete scale of 1 - 9

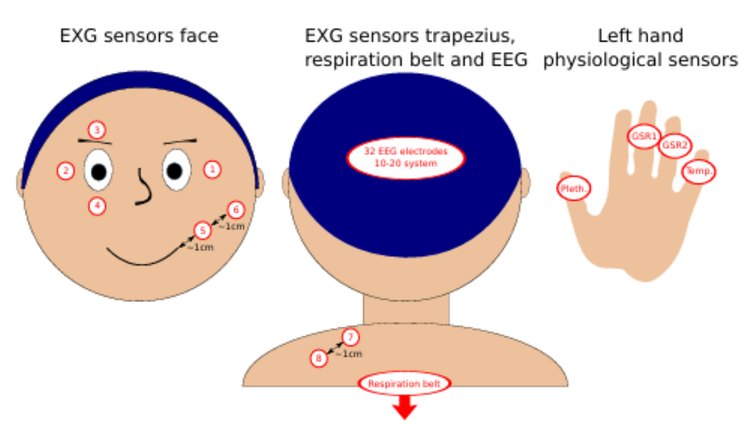
- Source: DEAP (S. Koelstra et al, 2012)



- 120 music videos - 60 using tags from last.fm, 60 manually
- 1 minute segments of these 120 videos
- Each video rated by 13-14 volunteers on a 9-point scale for arousal, valence and dominance
- Selected 40 final videos with strongest scores

DATA COLLECTION : EXPERIMENT DATA

Physiological Experiment	
Number of participants	32
Number of videos	40
Selection method	Subset of online annotated videos with clearest responses (see section 2.3)
Rating scales	Arousal Valence Dominance Liking *(how much do you like the video?)* Familiarity *(how well do you know the video?)*
Rating values	Familiarity: discrete scale of 1 - 5 Others: continuous scale of 1 - 9
Recorded signals	32-channel 512Hz EEG Peripheral physiological signals Face video (for 22 participants)



- Source: DEAP (S. Koelstra et al, 2012)

SET UP

- 2 labs - controlled illumination
- minimized eye-movements: 800*600 res
- 1m from screen, high volume

SIGNALS

- EEG: 512 HZ; 32 AgCl electrodes
- 13 peripheral physiological signals
- 22/32 participants: frontal face video

PARTICIPANTS

- 32 healthy participants (50-50 gender ratio)
- unrecorded trial, then 40 recorded trials
- Consent forms, Anonymity and Privacy

DATA COLLECTION : PROCESS

BASELINE

2 min baseline data

VIDEO STIMULI

1 min video display

SELF ASSESMENT

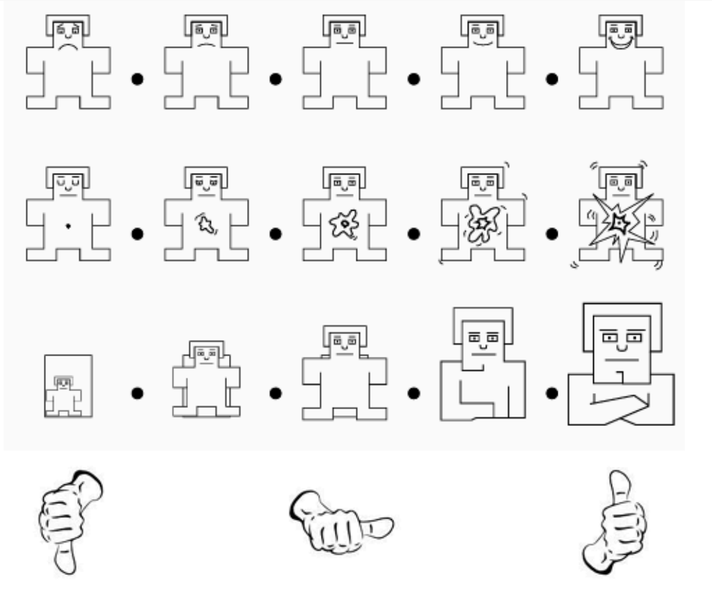
For arousal, valence, dominance

BREAK

Short break after 20 trials

REPEAT

Same process repeated for 20 more trials



The mean values (and standard deviations) of the different ratings of liking (1-9), valence (1-9), arousal (1-9), dominance (1-9), familiarity (1-5) for each affect elicitation condition.

Cond.	Liking	Valence	Arousal	Dom.	Fam.
LALV	5.7 (1.0)	4.2 (0.9)	4.3 (1.1)	4.5 (1.4)	2.4 (0.4)
HALV	3.6 (1.3)	3.7 (1.0)	5.7 (1.5)	5.0 (1.6)	1.4 (0.6)
LAHV	6.4 (0.9)	6.6 (0.8)	4.7 (1.0)	5.7 (1.3)	2.4 (0.4)
HAHV	6.4 (0.9)	6.6 (0.6)	5.9 (0.9)	6.3 (1.0)	3.1 (0.4)

FINAL DATASET

Channel number	Channel name	Channel content
33	EXG1	$hEOG_1$ (to the left of left eye)
34	EXG2	$hEOG_2$ (to the right of right eye)
35	EXG3	$vEOG_1$ (above right eye)
36	EXG4	$vEOG_4$ (below right eye)
37	EXG5	$zEMG_1$ (Zygomaticus Major, +/- 1cm from left corner of mouth)
38	EXG6	$zEMG_2$ (Zygomaticus Major, +/- 1cm from $zEMG_1$)
39	EXG7	$tEMG_1$ (Trapezius, left shoulder blade)
40	EXG8	$tEMG_2$ (Trapezius, +/- 1cm below $tEMG_1$)
41	GSR1	Galvanic skin response, left middle and ring finger
42	GSR2	Unused
43	Erg1	Unused
44	Erg2	Unused
45	Resp	Respiration belt
46	Plet	Plethysmograph, left thumb
47	Temp	Temperature, left pinky
48	Status	Status channel containing markers

VIDEO LIST

All the videos used in the online self-assessment and in the experiment, with the Youtube links

FACIAL VIDEOS

Front face video of 22 out of 32 participants

PHYSIOLOGICAL DATA

32 files, with 48 channels at 512Hz. (32 EEG, 12 peripheral, 3 unused and 1 status channel)

- GSR, heart rate, respiration, and skin temperature

Preprocessing

What the Authors Did

Downsampling

EEG data downsampled to 128 Hz

Artefact Removal:

EOG artefacts removed

Bandpass Filtering:

Applied a 4.0–45.0 Hz bandpass filter to EEG data.

Geneva Channel Reordering:

EEG channels reordered according to the Geneva standard.

Segmentation:

Data was segmented into 60-second trials; a 3-second pre-trial baseline was removed.

PREPROCESSING AND FEATURE EXTRACTION

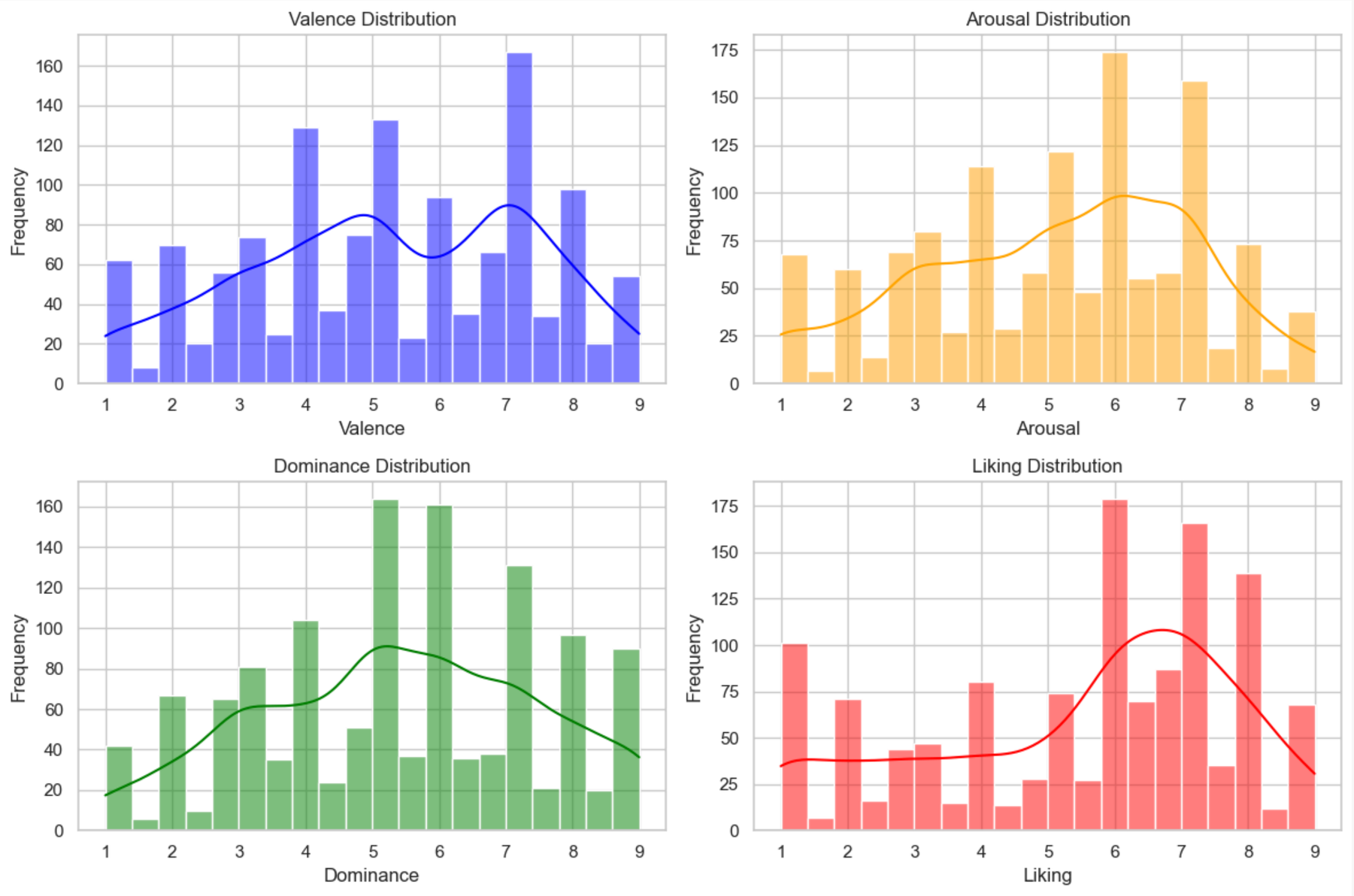
LOADING THE DATA

- Loaded .dat files and normalized EEG signals with Z-score normalization.
- Segmented data for GSR, PPG, respiration, and EEG.

FEATURE EXTRACTION

- GSR: Extracted average skin resistance, average derivative, and local minima.
- Respiration: Calculated mean and standard deviation of respiration signals i.e breath rate.
- EEG: Extracted spectral power for Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), and Beta (12-30 Hz) bands.

DISTRIBUTION OF LABELS



METHODOLOGY

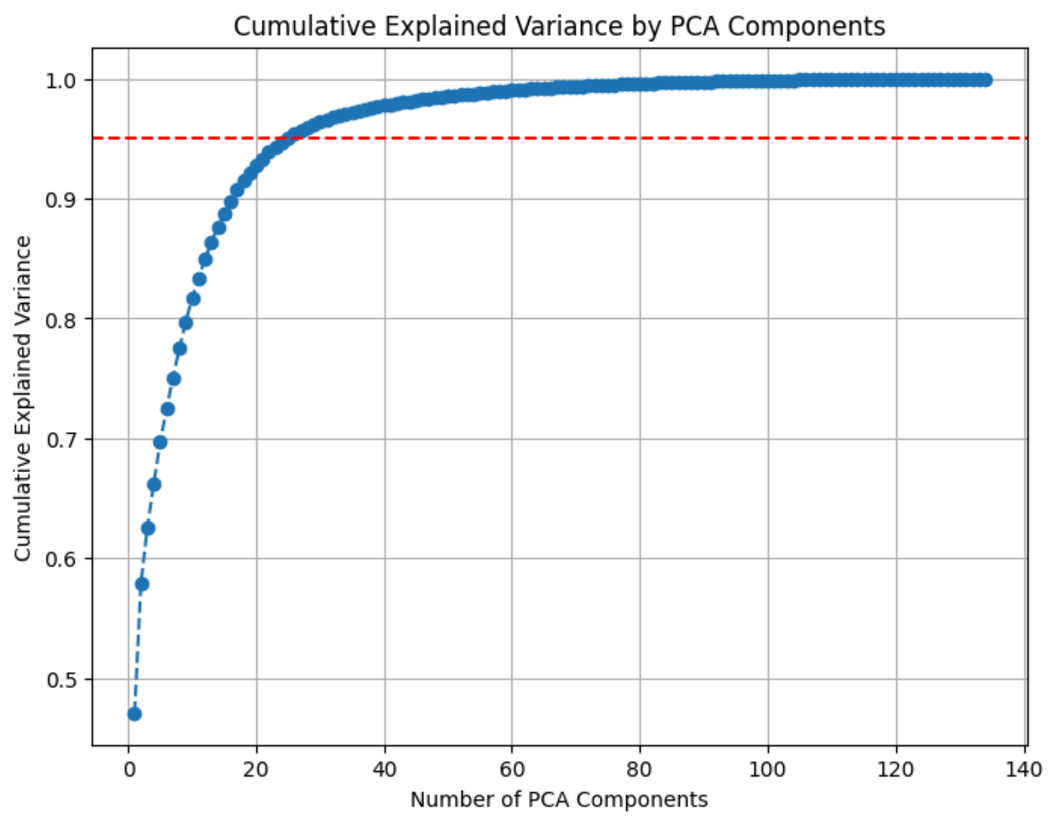
1. Baseline Model

- Data Input:
  - Physiological Signals: Collect EEG, GSR, PPG, respiration.
  - Alignment: Synchronize features with fusion
- Model:
  - Train a model like Random Forest for emotion classification (valence/arousal based emotions.)
- Evaluation:
  - Metrics: Accuracy, F1-score, Confusion Matrix.

2. Saliency-based Emotion Recognition (Optimized Model)

- Saliency Detection:
  - Identify key video segments based on viewer attention using 3D-FCN to get regions and averaging saliency scores to get the most relevant segment time frames. (https://github.com/MichiganCOG/TASED-Net)
- Feature Extraction
  - Extract corresponding physiological signals from salient segments only.
- Model:
  - Train a second model (Random Forest) on salient segments.
- Comparison & Evaluation:
  - Compare full video vs. salient segment models using Accuracy, F1-score, and computational metrics.

PCA AND BASELINE MODEL



- Goal: Predict binary valence (high vs. low) using Random Forest.
- Threshold: Valence > 5 = High, Valence ≤ 5 = Low.
- PCA: Reduced dimensions to explain 95% of variance.
- With PCA: Accuracy = 64%, using 26 components.
- Without PCA: Accuracy = 66%, with full feature set for the 20 videos
- PCA simplifies the model with minimal accuracy difference however the accuracy itself is low (Sujata et Patil, 2021)

MODEL TRAINED ON SALIENT REGIONS

	precision	recall	f1-score	support
0	0.91	0.18	0.30	56
1	0.61	0.99	0.75	72
accuracy			0.63	128
macro avg	0.76	0.58	0.52	128
weighted avg	0.74	0.63	0.55	128

- Goal: Predict binary valence (high vs. low) using Random Forest.
- Threshold: Valence > 5 = High, Valence ≤ 5 = Low.
- Shows similar performance to the overall model i.e 63% as compared to the baseline 66%.
- We see a drop in predicting lower valence values but excel on high valence observations.

MODEL TRAINED ON NON- SALIENT REGIONS

	precision	recall	f1-score	support
0	0.22	0.04	0.06	56
1	0.55	0.90	0.68	72
accuracy			0.52	128
macro avg	0.38	0.47	0.37	128
weighted avg	0.40	0.52	0.41	128

- Goal: Predict binary valence (high vs. low) using Random Forest.
- Threshold: Valence > 5 = High, Valence ≤ 5 = Low.
- Shows worse performance to the overall model i.e 52% as compared to the baseline 66%.

FEATURE RELEVANCE

Signal Type	Full Data Accuracy	Salient Parts Accuracy	Non-Salient Parts Accuracy
EEG	63%	61%	54%
GSR	57%	54%	49%
PPG	57%	53%	49%

CHALLENGES

- **Synchronizing Multimodal Features:**

Physiological signals, may have different sampling rates, can be complex. We used resampling to align the data to a common timeline and pre-process each signal for feature consistency.

- **Choosing the Threshold:**

Since we chose the threshold manually, via trial and error, we aren't confident on what the 'correct' threshold should be. There is no existing literature on this field.

- **Data diversity:**

To scale up to multiple applications, we need to ensure that the training data is diverse and generalizes across different contexts and cultures.

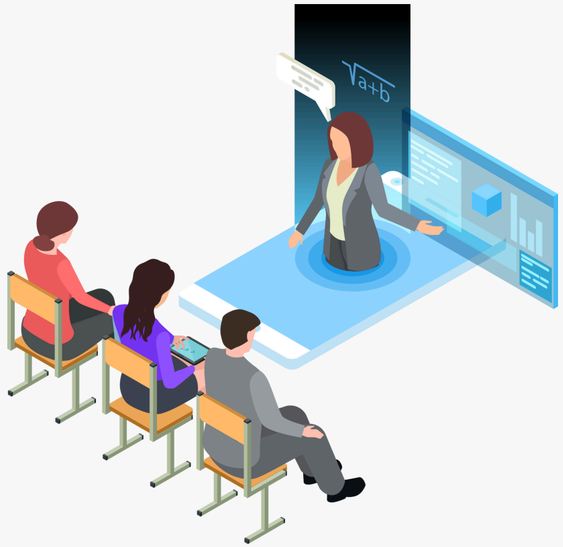
- **The 'suspense' in the films**

If the system identifies and focuses on the key moments—those critical plot twists or suspenseful builds, too early, it could spoil the entire experience for the viewer

IMPACT

Entertainment

Marketing



Education

- Data Efficiency: reduces storage overhead
- Content Accessibility
- Enhances HCI

THANK YOU