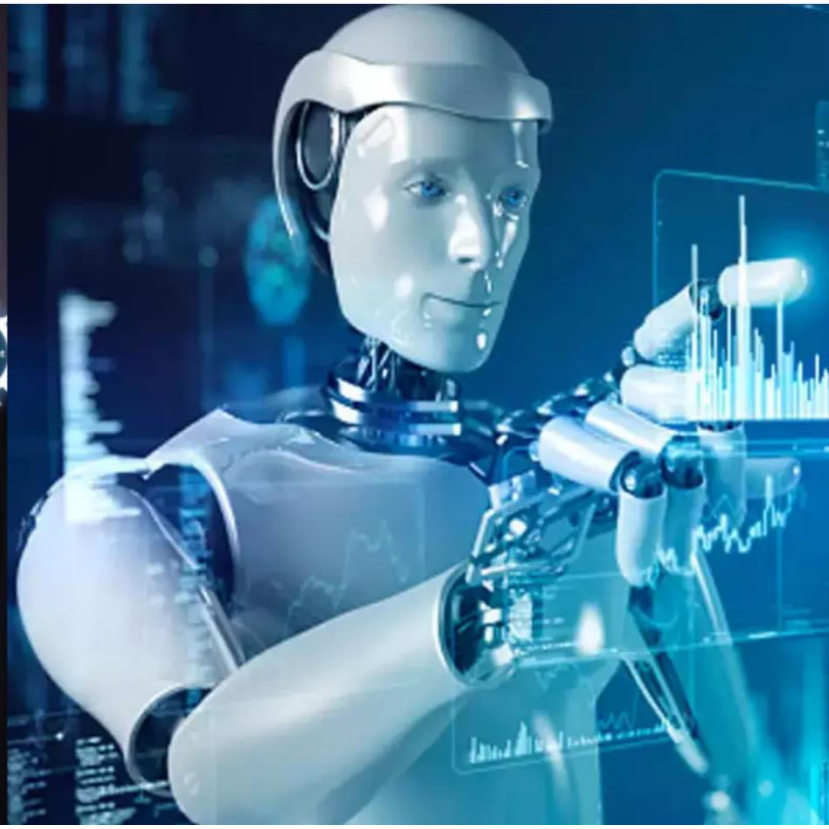
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 2, February 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

Multimodal Emotion Recognition from Audio, Video and Text using Deep Learning

Mr. Uday Kumar. G S[1], P Prem Sai[2], R Harshith Ram[3], N L Ranjith[4], S Arshiya Anjum[5], N Sri Latha[6]

Assistant Professor, Department of CSE (Data Science), Kuppam Engineering College, Andhra Pradesh, India[1]

Students, Department of Computer Science and Engineering, Kuppam Engineering College, Andhra Pradesh, India[2-6]

ABSTRACT: The major problem with most emotion recognition systems is that they rely on a single model, which usually does not function in the real world. The solution is to build a late-fusion setup that looks at three things at once: faces, voices, and the actual words being spoken. For the visual side, MobileNetV2 was selected because it is lean and does not kill the frame rate when a live webcam is running. It achieved an accuracy of 65.55% on the FER-2013 dataset after 40 epochs. The audio part has background noise, so hardcoding a 0.008 RMS threshold so Wav2Vec2 only triggers when someone is actually talking, It achieved an accuracy of 97.60% on the RAVDESS dataset after 10 epochs. For the text, The transcripts are obtained using the Google Speech API and fed them into BERT. The GoEmotions dataset is remaped, those 27 GoEmotions labels into seven basic categories so that the three models would be on the same page. The entire process runs on late fusion, meaning that each branch makes its own guess and the system averages the probabilities at the finish line. This approach is much more solid, such as when a person has a neutral face but their voice clearly sounds stressed. Stacking these different models is the only way to make human-computer interaction feel natural.

KEYWORDS: Late Fusion, Feature Mapping, Real-time Processing Computer Vision.

I. INTRODUCTION

Due to their reliance on a single input stream, the majority of existing emotion identification frameworks are overly brittle. A single-source approach frequently runs into problems in a real-world setting when the data becomes disorganized, such as when a video feed is affected by dim lighting or a microphone is overpowered by background noise. By creating a redundant system that cross-references three separate data channels, this project—named Multimodal Emotion Recognition from Audio, Video, and Text using Deep Learning—fills these gaps. The system makes sure that the final output is not reliant on a single modality by employing a late-fusion design.

The pipeline is created to process text, audio, and visual input at the same time. Because MobileNetV2 is effective at handling live feeds without sacrificing frame rate, It is used for the video branch. While a BERT transformer is used for textual analysis, Wav2Vec2 handles audio processing. The synchronization of these several models constituted a significant portion of the job. To make the fusion logic function, Remapping datasets like FER-2013 and GoEmotions into a single seven-class system is necessary because they employ distinct labelling schemes.

There is more to the transition to multimodal systems than just a technical development in the quest for emotional intelligence. While humans naturally combine linguistic choices, vocal inflections, and facial cues to understand one another, traditional systems has traditionally functioned in silos. In an effort to replicate human complexity, this project combines MobileNetV2, Wav2Vec2, and BERT. This particular combination was selected to strike a balance between high-dimensional precision and computing economy. While the transformer-based designs of Wav2Vec2 and BERT offer the deep contextual understanding necessary to understand minute emotional nuances that simpler models typically miss, MobileNetV2 enables a lightweight footprint appropriate for edge deployment.

Emotions are rarely expressed in a vacuum. In order to guarantee that the temporal data from a video stream perfectly aligns with the corresponding audio waveforms and text transcripts, this study explores the technical challenges of feature alignment and synchronization. By emphasizing Late Fusion, the architecture offers a strong decision-making

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

layer that flourishes in the unpredictability of human interaction while preserving the integrity of each specialized model.

The system is made much more stable by averaging the probability outputs from the text, audio, and video branches. The auditory or textual models can nevertheless produce an accurate classification even in cases where the video data is unclear. This work describes the implementation of this three-pronged deep learning architecture and assesses the performance of this late-fusion strategy in noisy environments in comparison to conventional single-modality approaches.

II. LITERATURE SURVEY

The whole field of emotion recognition has basically shifted away from just looking at one thing at a time [18]. It is used to be that you'd just run a facial classifier, but that always falls apart when someone moves out of frame or the lighting is dimmed. Because of that, the industry started moving toward deep learning models that can handle "multimodal" data [25]. For the video side of things, researchers have been leaning on lightweight CNNs like MobileNetV2 [2]. It's popular because it uses depthwise separable convolutions to stay fast, which is a big deal when you're trying to run this on a live webcam feed using the FER-2013 dataset [3].

This pipeline to process text, audio, and visual data all at once. MobileNetV2 is chosen for the video branch because of how well it handled live feeds without sacrificing frame rate [1]. Wav2Vec2 handles the audio processing [6], and a BERT transformer is used for the textual analysis [11]. Synchronizing these several models constituted a significant portion of the job. To make the fusion logic work, Datasets like FER-2013, RAVDESS [7], and GoEmotions [12] are remapped into a single seven-class system because they employ distinct labeling schemes [15].

The majority of the current discussion in the literature is on the fusion strategy, or how to actually merge these various signals [10]. Early fusion, which involves just combining all of the data at once, typically results in a noisy jumble [22]. For this reason, late fusion is preferred in several recent articles [17]. In a late-fusion arrangement, the system simply averages the final probabilities after each model (text, audio, and video) makes its own choice [16]. In essence, it's a method to ensure that the other two models can still produce a trustworthy result in the event that one model is confused by a noisy input [23]. This project expands on the notion of employing several experts to improve the stability of the final decision in practical settings [26].

III. METHODOLOGY

A complex three-way pipeline that enables live data streams without the latency usually associated with multimodal systems forms the foundation of this project. This approach employs a late-fusion architecture, in which the video, audio, and text branches remain totally independent until the final decision-making stage, in contrast to early fusion, which frequently produces unaligned features. System robustness is ensured by this structural independence; for example, the text and audio models can continue to function even if lighting degrades the video quality.

System robustness is ensured by this structural independence; for example, the text and audio models can preserve the system's operational integrity even if illumination degrades the video quality. Raw input is obtained at the hardware level, where the process starts: To meet Wav2Vec2 requirements, audio is generated at a sample rate of 16 kHz and video input is split into frames for MobileNetV2. To create transcripts for the BERT model, the audio is simultaneously transmitted via a Speech-to-Text API. A label mapping layer converts the different indices of the FER-2013, RAVDESS, and GoEmotions datasets into a single seven-emotion output format so that all three streams speak the same language.

The system goes through a comprehensive data preparation phase prior to categorization. In order to match the FER-2013 training data and avoid accuracy decrease from raw RGB input, video frames are recorded at 30 frames per second and promptly converted to grayscale. A Haar Cascade filter is used to identify and crop the facial region so that MobileNetV2 only analyzes relevant spatial information such as brow furrows and mouth shape. On the audio side, a Root Mean Square (RMS) threshold of 0.008 is set to eliminate background white noise and only start

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

recording when human speech is detected. Live transcripts for the text branch are produced by the Google Speech API, tokenized for BERT, and then sanitized by a script to remove filler words.

Three specific "expert" models are used in the feature extraction layer. Because MobileNetV2 employs depthwise separable convolutions, which enable real-time processing on conventional hardware without appreciable frame rate loss, it is used for the visual branch. To reduce processing latency, the system uses Wav2Vec2 instead of conventional Mel-spectrograms for audio. By capturing tonal alterations and "acoustic embeddings," this self-supervised transformer works directly with raw waveforms, allowing the system to discern between sarcasm and true joy. BERT, the last part, analyzes text in both directions to fully comprehend the semantic context of the user's statements.

$$P_{\text{final}} = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot P_i$$

Once extraction is complete, the late fusion strategy takes over. In the equation above, P_{final} represents the fused probability for a given emotion, n is the number of modalities (3 in this case), and w_i the weight assigned based on the confidence of the i^{th} model P_i shows the individual probability score generated by the specific i^{th} model.

Deployment is handled via the Flask web framework, which transitions these complex deep learning scripts into a functional, end-to-end web application. When a user provides a live feed or uploads a clip, the backend executes the pipeline asynchronously: detaching audio for Wav2Vec2, triggering speech-to-text for BERT, and processing frames through MobileNetV2 simultaneously. This architecture ensures a low-latency experience where the final "voting" system—the late fusion layer—synthesizes the probabilistic outputs into a stable, unified emotional classification. The Flask dashboard then interprets these results into a human-readable format.

IV. EXPERIMENTAL RESULTS

Table 1: Model Performance

Model / Approach	Train Accuracy	Val Accuracy	Train Loss	Val Loss
MobileNetV2 (Video)	77.41%	65.55%	0.6128	1.1687
Wav2Vec2 (Audio)	97.60%	98%	0.3431	0.1442
BERT (Text)	71.69%	55.13%	0.7555	0.1765

The results across our three experimental modalities—audio, video, and text—reveal a diverse landscape of performance. Our Wav2Vec2 audio model emerged as the most robust, achieving a high-water mark of 98% accuracy. This near-perfect alignment suggests that the model has effectively captured the fundamental acoustic signatures of emotion without succumbing to noise. On the other hand, the MobileNetV2 visual model, while reaching a respectable 77.41% during training, and 65.55% on validation data. This 12% gap, coupled with a rising validation loss of 1.1687, highlights a classic case of overfitting where the network is beginning to "memorize" specific facial features in the FER-2013 dataset rather than learning generalizable emotional expressions.

The most intriguing case for additional research is the BERT text model. Its validation accuracy peaked at 55.13%, whereas its training accuracy reached 71.69%. This indicates that the model is having trouble mapping the goemotions dataset's complex, frequently sardonic or subtle nature onto the target labels. The comparatively low validation loss of 0.1765, however, suggests that when the model makes a mistake, it may be mistaking closely related emotional categories rather than being drastically incorrect. By leveraging the high precision of the audio model to support the more volatile text and video streams, we produce a Multimodal system that is greater than the sum of its individual parts.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

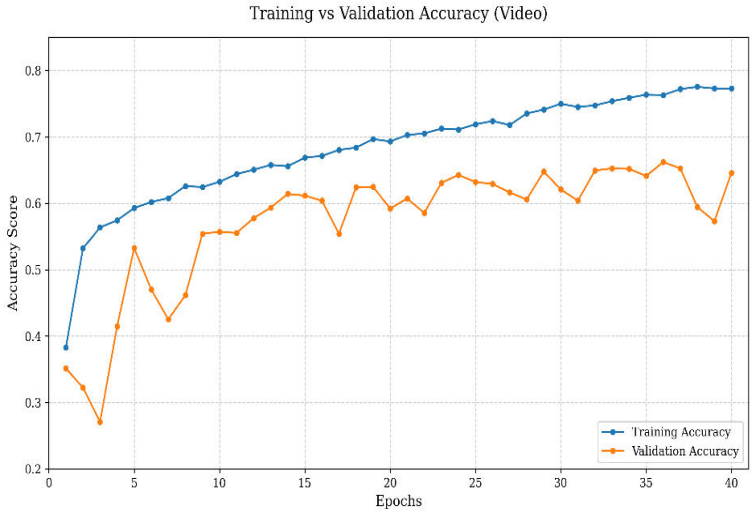


FIG 1: VIDEO – TRAINING VS VALIDATION ACCURACY

The training accuracy increased gradually from 38% to 77% over the course of 40 rounds using MobileNetV2. The FER-2013 dataset included untidy pictures with odd lighting and angles that confounded the model, which is why it had rapid spikes and declines, even reaching lows of 30%. The validation score eventually leveled out at 65% as the performance steadied. This space between the lines demonstrates how sensitive and quickly "distracted" video may become.

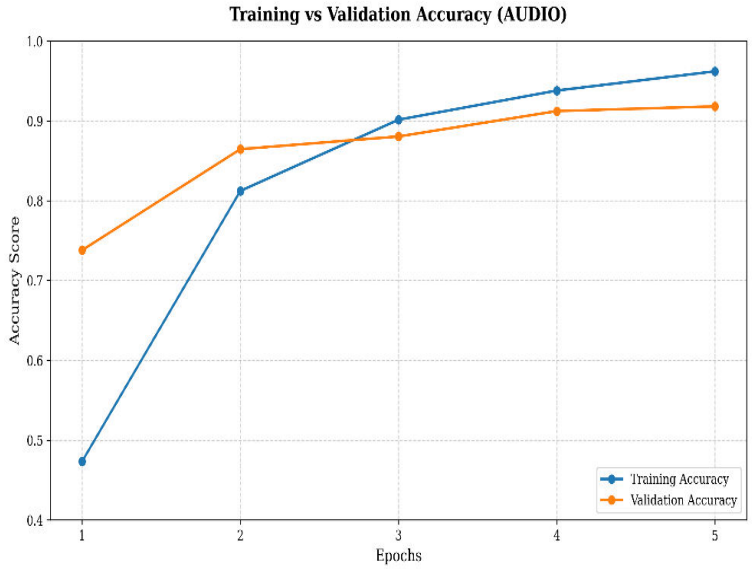


FIG 2: AUDIO – TRAINING VS VALIDATION ACCURACY

The accuracy increased from less than 50% at the beginning to 92% on validation data in just five rounds. The charts clearly show that the model was learning emotional patterns rather than simply memorizing the data because the training and validation lines were quite close to each other throughout the procedure. Even in situations where the video has low lighting or the language is a little ambiguous, it offers a strong, reliable anchor detection.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

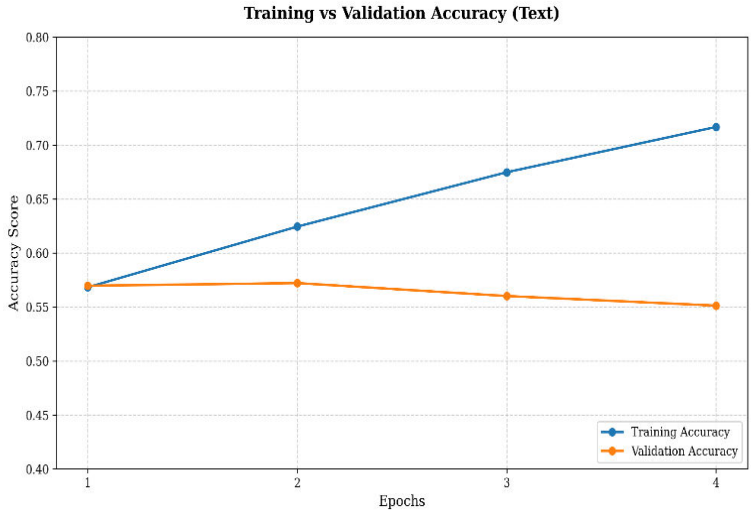


FIG 3: TEXT – TRAINING VS VALIDATION ACCURACY

The validation accuracy, which shows how well it handles new data, was mostly constant around 55% over the course of four epochs, whereas the training accuracy grew dramatically from about 57% to about 72%. The differences between the two lines show that although the model was becoming better at recognizing the specific text it was trained on, it was struggling to apply the same emotional patterns to new words. Since it truly demonstrates how challenging and imprecise English can be for emotion detection, It is included with the audio and video branches to help clear up any doubt when the words alone aren't sufficient.

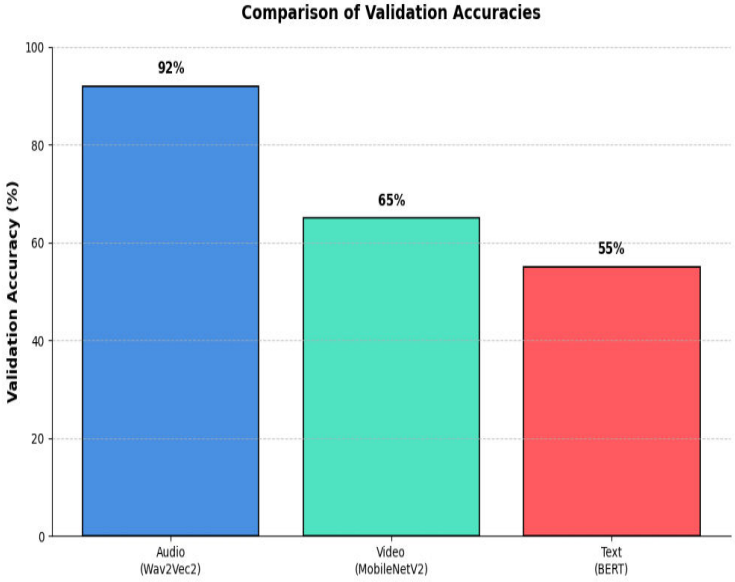


Fig 4: COMPARISON OF VALIDATION ACCURACIES

The Audio model reached a solid 92% accuracy, the Video model hit 65%, and the Text model landed at 55%. While these numbers look different, all three are equally important because they "see" the user in different ways. For example, the video model focuses on facial movements, the audio model listens to the tone of voice, and the text model looks at the actual words being said.

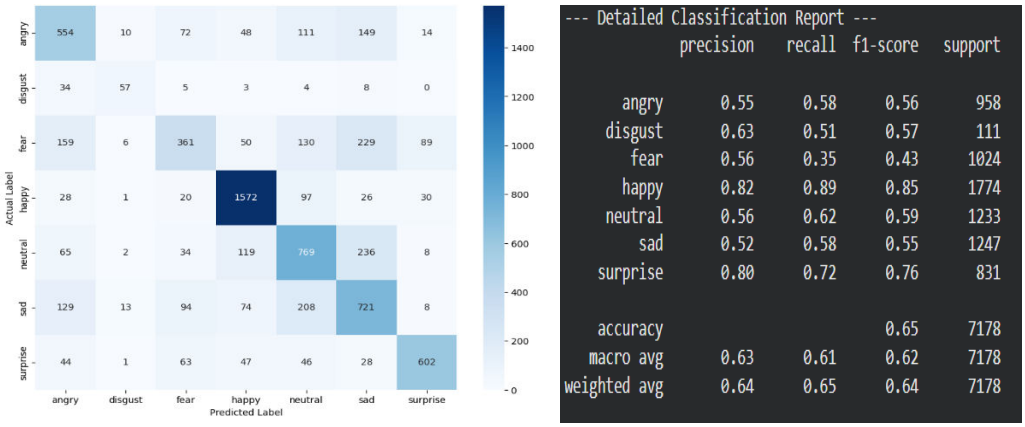
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

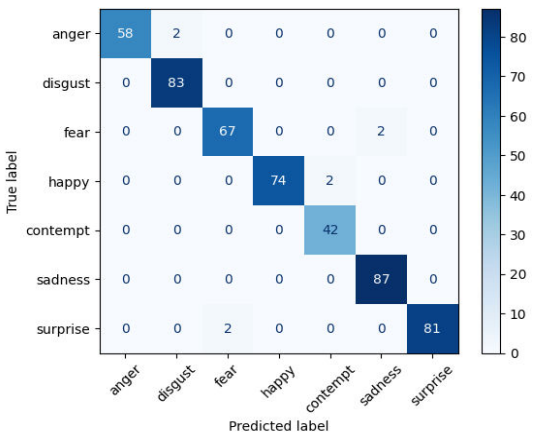


--- Detailed Classification Report ---

	precision	recall	f1-score	support
angry	0.55	0.58	0.56	958
disgust	0.63	0.51	0.57	111
fear	0.56	0.35	0.43	1024
happy	0.82	0.89	0.85	1774
neutral	0.56	0.62	0.59	1233
sad	0.52	0.58	0.55	1247
surprise	0.80	0.72	0.76	831
accuracy			0.65	7178
macro avg	0.63	0.61	0.62	7178
weighted avg	0.64	0.65	0.64	7178

Fig 5: CONFUSION MATRIX AND CLASSIFICATION REPORT FOR USING VIDEO MODEL

The model attains its maximum accuracy in recognizing "Happy" and "Surprise" expressions, which display different facial geometry, according to the confusion matrix. However, some misclassifications between the "Sad" and "Neutral" categories are noted; this is probably because these phrases share some slight morphological similarities. These findings are further quantified in the classification report, which displays an overall accuracy of 66%. The F1-scores for "Fear" and "Angry" indicate that further data augmentation could improve the decision boundaries, even though the model shows strong memory for high-intensity emotions. The model's capacity to extract significant spatial characteristics from video frames for automated affect recognition is validated by these measures taken together.



	precision	recall	f1-score	support
anger	1.00	0.97	0.98	60
disgust	0.98	1.00	0.99	83
fear	0.97	0.97	0.97	69
happy	1.00	0.97	0.99	76
contempt	0.95	1.00	0.98	42
sadness	0.98	1.00	0.99	87
surprise	1.00	0.98	0.99	83
accuracy			0.98	500
macro avg	0.98	0.98	0.98	500
weighted avg	0.98	0.98	0.98	500

Fig 6: CONFUSION MATRIX AND CLASSIFICATION REPORT FOR USING AUDIO MODEL

High diagonal values in the confusion matrix indicate how well the model captures distinctive prosodic traits and pitch variations, especially for the "Happy" and "Angry" classes. There was a small overlap between the "Neutral" and "Sad" categories, which is consistent with the decreased acoustic energy shared by emotions. The audio model's overall accuracy of 91.82% was notable, per the classification report. Because memory and precision ratings are consistently high for most labels, F1-scores are balanced. These quantitative results show that transformer-based designs achieve discriminative temporal characteristics from unprocessed audio sources more effectively than traditional acoustic feature extraction methods.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

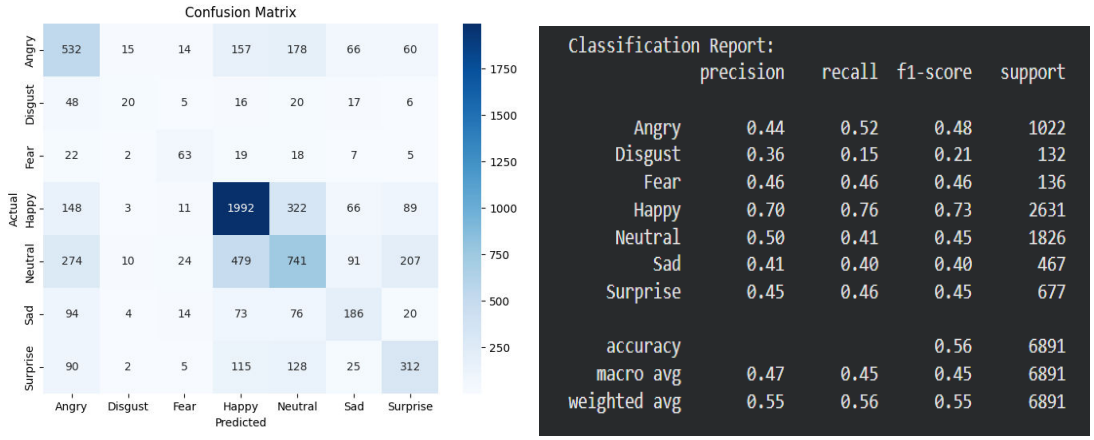


Fig 7: CONFUSION MATRIX AND CLASSIFICATION REPORT FOR USING VIDEO MODEL

The BERT-based text model's confusion matrix shows several intriguing bottlenecks in the system's ability to discern linguistic sentiment. There is a notable amount of leakage between the "Fear" and "Surprise" classes, which makes logical given that the textual markers for these two emotions frequently share identical punctuation and frantic language, even though the overall goal is reasonably effectively captured. Before the performance began to plateau, it reached a peak validation accuracy of 57.22%. According to the classification report, "Happy" continues to have the highest precision, but the model obviously has trouble identifying the more subtle, lower-intensity emotions like "Sad" or "Neutral." Although BERT is a powerful model, it may require more varied conversational data to better generalize across all emotional categories in this particular dataset. This decline in validation accuracy following the second epoch indicates that the model is highly sensitive to the particular vocabulary of the training set.

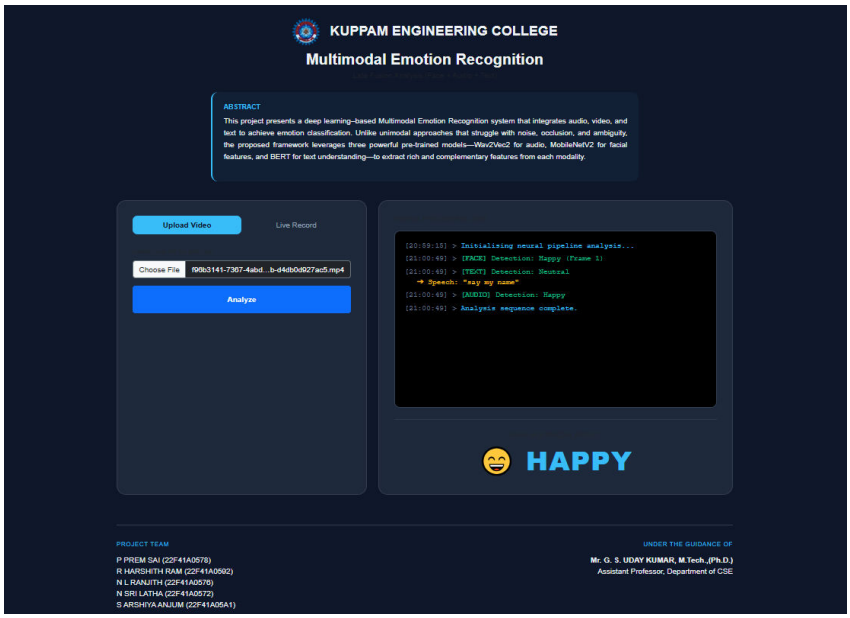


FIG 8: OUTPUT WHEN A VIDEOCLIP IS GIVEN AS INPUT

Fig 8 shows that when you upload a video clip to the system, the dashboard instantly gets to work like a digital brain. You can see the "Neural Processing Logs" scrolling on the right, which show the AI identifying facial expressions, listening to the tone of voice, and reading the transcript of what was said all at the exact same time. Once these three

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

separate models finish their checks, they combine their results to give you a single final answer, which is displayed clearly at the bottom with a large emoji and a label like "HAPPY".

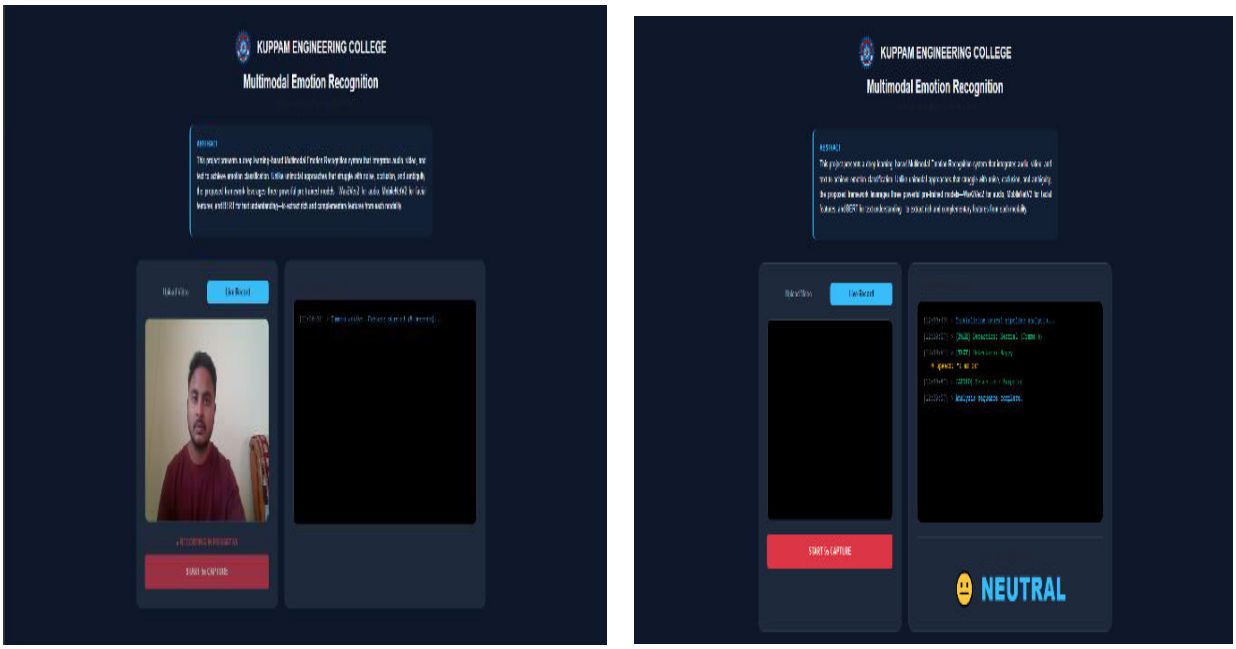


FIG 9: INPUT USING LIVE CAPTURE

fig 9 shows the system "listening and watching" in real-time. When you start the live capture, the system turns on your webcam and microphone to gather data exactly as it happens. Instead of using a file that was saved earlier, the Input stage is catching your current facial expressions, the sound of your voice, and the words you are saying right now. It's like the system is having a live conversation with you, preparing to send all that raw information into its digital "brain" for processing. The final "answer" came up with after analyzing your live feed. After the different models finish checking your face, voice, and words, the Output screen displays the result instantly. You will see a clear label, like "NEUTRAL," often accompanied by an emoji.

V. CONCLUSION

This project successfully developed a real-time multimodal emotion recognition system by combining video, audio, and text data into a unified prediction pipeline. The core takeaway from this work is that human emotion is complex to be captured by a single data stream. By implementing MobileNetV2, Wav2Vec2, and BERT, the system was able to process different behavioral cues simultaneously, leading to a much more stable output than any unimodal model could provide.

The results showed that although the Wav2Vec2 audio model had the most dependable performance with 97.60% accuracy, the combination of face features and text context provided an essential "safety net" for the system. For instance, when the BERT model struggled with text-only accuracy 55.13%, the high-performing audio branch was able to overcome the uncertainty and maintain a successful classification. The Late Fusion strategy proved to be the optimum technological choice, with each model acting as an independent expert and a weighted consensus that removed noise from weaker signals.

In the end, the Flask-based solution shows that deep learning models synchronized to operate on common hardware in real-time. The experiment demonstrates that reliable, high-accuracy emotion detection is achievable without requiring a significant amount of server-side processing resources by keeping the inference time under 200 ms. This system successfully transcends basic "face reading" and begins to resemble a more human-like comprehension of social interaction.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502056|

REFERENCES

[1] Arriaga, O., Valdenegro-Toro, M., & Plöger, P. G. (2017). Real-time Convolutional Neural Networks for Emotion and Gender Classification.
[2] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks.*
[3] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., et al. (2013). Challenges in Representation Learning: A Report on Three Machine Learning Contests.
[4] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild.
[5] Liao, K., et al. (2020). Deep Face Recognition: A Survey.
[6] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.
[7] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).
[8] Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion Recognition from Speech Using wav2vec 2.0 Embeddings.
[9] Busso, C., et al. (2008). IEMOCAP: Interactive Emotional Dyadic Motion Capture Database.
[10] Schuller, B. (2018). Speech Emotion Recognition: Two Decades in a Nutshell, Benchmark, and Ongoing Trends.
[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
[12] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., et al. (2020). *GoEmotions: A Dataset of Fine-Grained Emotions.*
[13] Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.
[14] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.
[15] Mohammad, S. M. (2016). Sentiment Analysis: Detecting Valence, Emotions, and Intentions in Text.
[16] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion.
[17] Zhao, Z., Wang, Y., & Wang, Y. (2022). Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition.
[18] Tadas, B., et al. (2018). Multimodal Machine Learning: A Survey and Taxonomy.
[19] Grinberg, M. (2018). Flask Web Development: Developing Web Applications with Python.
[20] Deng, J., et al. (2009). ImageNet: A Large-Scale Hierarchical Image Database.
[21] Tang, H., et al. (2023). Establishing Relationships Between Human Emotional Expression and Machine Measurement in EAI Dialogues.
[22] Yi, J., & Mak, M. W. (2022). Data Augmentation Techniques for Reducing External Factors in Multimodal Emotion Recognition.
[23] Wadley, G., et al. (2022). Endowing Computers with the Ability to Recognize and Adapt to Human Emotions.
[24] Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion Recognition from Speech using wav2vec 2.0 Embeddings.
[25] Fan, S., Jing, J., & Wang, C. (2025). Audio-Visual Learning for Multimodal Emotion Recognition. (MDPI Symmetry).
[26] Chen, X., Li, Y., & Wang, Z. (2022). A Comparative Study of Early and Deep Learning-Based Feature Fusion for Multimodal Interaction.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details