Indian Journal of Computer Science and Technology
Volume 5, Issue3 (September-December 2026), PP: 683-690.
www.indjcst.com
ISSN No: 2583-5300

Multimodal Emotion Detection with Sentimental Analysis Chatbot

Bandapalli Charmi[1], Reddy Manoj Kumar[2], Dr. Sathya R[3]
[1,2,3]Department of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamil Nadu, India.

To Cite this Article: *Bandapalli Charmi[1], Reddy Manoj Kumar[2], Dr. Sathya R[3], "Multimodal Emotion Detection with Sentimental Analysis Chatbot", Indian Journal of Computer Science and Technology, Volume 05, Issue 03 (September-December 2026), PP: 683-690.*

Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution License; Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: *Recognizing emotions is a vital part of human intelligence in interaction with computers, yet most traditional single-modality systems, which depend only on text, speech or facial data, cannot accurately capture the complexity of human emotions. This research proposes an approach to emotion detection that uses text, voice and face simultaneously to provide comprehensive and contextually aware affect analysis. The system combines three modality pipelines: Natural Language Processing with NLTK preprocessing, VADER scoring and a fine-tuned BERT emotion model; speech emotion recognition with Librosa-based feature extraction, MFCC computation and an SVM classifier trained on the RAVDESS dataset; and facial emotion recognition with a convolutional neural network trained on the FER2013 dataset. A reliability-aware weighted fusion procedure based on modality confidence scores achieves an accuracy of 91.3% and an F1-score of 0.91. The system also computes an Emotional Incongruence Index for inconsistency detection and uses emotion causality modeling to find contextual triggers of emotions. Finally, the detected emotion conditions a GPT-2 medium chatbot for empathic dialogue.*

Key Word: *Multimodal Emotion Recognition; Sentiment Analysis; Speech and Facial Emotions; Affective Computing; Emotion Fusion; Empathic Chatbot; Deep Learning; Context-aware System.*

I.INTRODUCTION

Human emotional expressions are inherently multimodal, consisting of what people say, the way they say it, and the way they physically express themselves. Speech, voice and face together form one comprehensive communication signal[1]. Nevertheless, almost all current sentiment analysis systems operate on a single modality, which limits their ability to identify emotions properly. A person may speak sarcastically, for example: the words may be positive while the tone and facial expression suggest negativity, so a text-only approach cannot generate a correct prediction.

Text-based sentiment analysis has progressed considerably thanks to transformer architectures such as BERT[5], while classic toolkits such as NLTK and VADER still offer effective and fast alternatives for real-time analysis. For audio, Librosa is widely used to extract features such as MFCCs, which can be fed to machine learning algorithms trained on the RAVDESS dataset[8]. For the face, convolutional neural networks trained on FER2013 can recognize the six basic Ekman emotions (anger, disgust, fear, happiness, sadness and surprise)[3]. Despite the success of each approach separately, they cannot easily be combined into one working technique. Current multimodal fusion methods achieve high accuracy but only at the expense of heavy computation[2,6]. They also fail to address unreliable modalities: every modality is assumed to be precise, which in practice is not the case. Finally, emotion recognition is treated as an isolated task and the reasons why people feel certain emotions are not taken into account.

To solve these problems, this study introduces an accurate, reliable and interpretable multimodal emotion recognition framework. It combines (i) an NLP pipeline with NLTK preprocessing, VADER scoring and BERT fine-tuning; (ii) an audio pipeline with Mel-frequency cepstral coefficient (MFCC) features and SVM classification on the RAVDESS dataset; and (iii) a facial pipeline with a customized five-layer convolutional neural network on the FER2013 dataset. A reliability-aware adaptive fusion technique combines the channels, an Emotional Incongruence Index (EII) indicates inconsistency between the recognized emotions, emotion causality modeling determines possible triggers of emotion, and the final result controls a GPT-2-medium chatbot.

Sentiment analysis from text: The earliest approaches used rule-based NLP frameworks such as NLTK and VADER, which give straightforward, comprehensible and efficient baselines. VADER employs human-compiled sentiment dictionaries with heuristics for punctuation, capitalization and intensifiers to compute compound polarity values, which makes it effective on informal and conversational text. Neural models such as BERT capture bidirectional contextual and semantic relationships and improved performance notably on benchmarks such as SST-2 and IMDb, and approaches such as SenticNet use commonsense reasoning to improve interpretation[14]. However, text-only solutions process a single channel and struggle when the textual sentiment disagrees with tone of voice or facial expression, for example in sarcasm, and they are rarely tailored to conversational applications.

Speech emotion recognition: Librosa provides features such as MFCCs, mel-spectrograms, chroma and spectral features, which represent speech signals effectively as input to algorithms such as SVM and shallow neural networks. Deep learning directly on raw audio gives higher accuracy[4] but increases the computational burden and reduces interpretability, and most solutions treat speech as a single modality without considering noise or reliability relative to other signals. This work uses MFCC features with an SVM classifier to balance efficiency and performance, and adds a reliability component to deal with noisy inputs.

Facial expression recognition: Facial expressions have been studied intensively with databases such as FER2013, which contains several thousand gray-scale face images labeled by emotion. Deeper CNN architectures generally give better accuracy but add computational load that makes them unsuitable for real-time use; lightweight CNNs deliver competitive results at minimal cost[10]. Existing approaches concentrate on accuracy and ignore contextual and multimodal information and cross-modal inconsistencies. The proposed solution therefore uses a lightweight CNN inside a multimodal framework that accounts for cross-modal inconsistency.

Multimodal fusion: Feature-level fusion, Tensor Fusion Networks[2] and MISA[6] have all benefited multimodal emotion recognition. Apart from being computationally heavy, they require every modality to be equally relevant, cannot handle inconsistent signals and are not suited to time-sensitive operation. The proposed solution instead applies weighted voting with a mechanism that estimates the reliability of each signal and adjusts its weight accordingly, introduces the Emotional Incongruence Index to measure conflict among signals explicitly, and incorporates emotion causality modeling.

The paper is organized as follows. Section II describes the materials and methods: the architecture, the three modality pipelines, the fusion process and the chatbot. Section III presents the results for each modality, for the fused system, for the incongruence analysis, the chatbot and real-time operation. Section IV discusses the findings and Section V concludes the paper.

II. MATERIAL AND METHODS

The proposed system was implemented in Python 3.10 using NLTK 3.8, Librosa 0.10, TensorFlow/Keras 2.13 (CNN), Hugging Face Transformers 4.35 (BERT, GPT-2) and scikit-learn 1.3 (SVM). The overall architecture is shown in Fig. 1 and the datasets and models used by each pipeline are summarized in Table no 1.

System overview: The system takes three synchronized input streams: (i) text, either typed by the user or generated by an ASR module from the user's speech; (ii) audio recorded from a microphone at a sampling rate of 22,050 Hz; and (iii) video acquired from a webcam at 30 fps. The modality-specific outputs are probability vectors p_T, p_A, $p_V \in R^6$ (one probability per Ekman category), which are combined by a reliability-based weighting function. The system additionally features the Emotional Incongruence Index (EII), which identifies inconsistencies between emotions expressed through different modalities, and emotion causality modeling, which provides information about the context that led to certain emotions. The final emotion is fed into the GPT-2 chatbot.

Modality	Input	Dataset	Pipeline	Output
Text	Typed text / ASR output	GoEmotions + SST-2	NLTK preprocessing, VADER score, fine-tuned BERT	p_T
Speech	Microphone audio, 22,050 Hz	RAVDESS (7,356 files)	Librosa features (MFCC, chroma, spectral contrast), RBF-SVM	p_A
Facial	Webcam video, 30 fps	FER2013 (35,887 images)	OpenCV Haar cascade face detector, custom five-layer CNN	p_V

Table no 1: Datasets and models used by each modality pipeline.

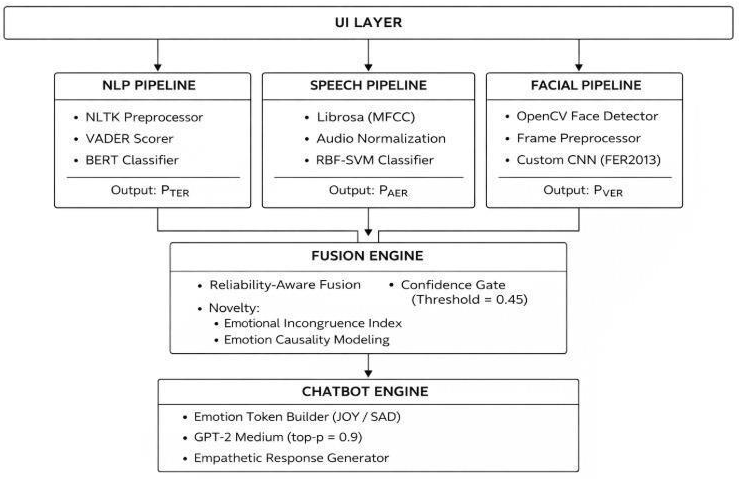


Fig. 1: Architecture diagram of the proposed multimodal emotion detection system with the emotion-conditioned chatbot.

Text module (NLP pipeline): The textual modality passes through a three-stage NLP pipeline. In the first stage NLTK performs tokenization, stop-word elimination, stemming (PorterStemmer) and punctuation normalization. In the second stage VADER computes a numerical value in the interval [-1, 1]; scores greater than +0.05 are categorized as positive, scores less than -0.05 as negative and the rest as neutral. This lexicon-based score serves as a fast sentiment prior for estimating modality confidence. In the third stage a fine-tuned bert-base-uncased transformer classifier performs emotion detection. It is fine-tuned on the GoEmotions + SST-2 data with the AdamW optimizer (learning rate 2×10-5, batch size 32, 4 epochs). The output of this multi-class classifier is a six-element softmax vector pT ∈ R6 with the probabilities of the Ekman categories.

Speech module (Librosa + SVM): The audio sequence is split into overlapping windows of 2 s (frame length 2048, hop length 512). Librosa extracts 40-dimensional MFCC features, 12 chroma coefficients and spectral contrast features over 7 frequency bands. The temporal statistics (mean and standard deviation) of each feature are computed, giving a fixed-length 118-dimensional feature vector. An SVM classifier with RBF kernel (regularization parameter C = 10, Gaussian kernel γ = 'scale') is trained on the RAVDESS dataset (7,356 files)8. The emotion labels are encoded into the six Ekman classes by collapsing some classes according to semantic similarity, and class probabilities pA ∈ R6 are obtained by Platt scaling.

Facial module (custom CNN on FER2013): Video frames are taken at 30 fps and faces are detected with the OpenCV Haar Cascades module. Detected faces are resized to 48×48 gray-scale images following the FER2013 requirements.
The CNN uses five convolutional layers built from the following blocks:
a) Conv2D layers with 32, 64, 128 and 256 filters;
b) Batch Normalization;
c) ReLU activation;
d) MaxPooling (2×2).

The final stage uses global average pooling followed by two fully connected layers with 512 and 128 hidden units and dropout (p = 0.5), and an output layer of 6 classes giving pV ∈ R6. The network is trained on the FER2013 dataset (35,887 images) with categorical cross-entropy loss and the Adam optimizer (learning rate 1×10-3). Horizontal flipping, rotation (±15°) and zoom scaling between 0.9 and 1.1 are used as data augmentation.

Reliability-based weighted fusion: The three probability vectors pT, pA and pV are fused with a weighted approach into a fused probability vector pF. In contrast to fixed fusion techniques with constant modality weights, the weights here are derived from reliability, estimated on validation scores, and satisfy wT + wA + wV = 1. The fused vector and predicted emotion label are:

$$pF = wT \bullet pT + wA \bullet pA + wV \bullet pV \qquad (1)$$

$$\hat{y} = \arg\max (pF) \qquad (2)$$

To make the prediction more robust, confidence gating is employed. If the maximum probability of a modality, max(pm), does not reach the minimum threshold τ = 0.45, that modality is regarded as unreliable: its weight is reduced by 50% (w′m = α•wm, α = 0.5) and the weights are renormalized. This adaptive weighting makes sure that unreliable modalities, for example because of noisy input or poor lighting, do not dominate the output decision.

Emotional Incongruence Index (EII): To account for incongruence between modalities, the EII measures disagreement between the probability distributions of the different modalities:

$$EII = (1/3)\ \Sigma i \neq j\ D(pi \parallel pj)$$

where D(•) is a distance metric such as the Kullback-Leibler divergence or cosine similarity. The higher the EII score, the greater the disagreement between modalities, reflecting sarcasm, concealment or uncertainty in the input. The EII is used (a) to mark uncertain classifications, (b) to contribute to explainability and (c) to determine fusion strategies based on reliability.

Emotion-conditioned chatbot: The predicted label ŷ is associated with an emotion token ([JOY], [SADNESS], [ANGER], [FEAR], [DISGUST] or [SURPRISE]). The token is attached to the beginning of the user utterance and sent to the GPT-2-medium model:

$$Response = GPT\text{-}2([EMOTION_TOKEN] \oplus user_text) \qquad (3)$$

The GPT-2-medium model (345M parameters) is fine-tuned for three epochs on the EmpatheticDialogues corpus (25k multi-turn dialogues) with a causal language modeling objective. Responses are generated by nucleus sampling with top-p = 0.9 and temperature 0.8, with a maximum length of 128 tokens. Postprocessing improves response quality by (a) keeping the length within 20-100 words, (b) applying a no-repeat n-gram constraint (n = 3) and (c) filtering redundant and repeated phrases, which yields meaningful, contextually relevant and emotionally consistent answers.

Module	Component	Configuration
Text	BERT classifier	bert-base-uncased; GoEmotions + SST-2; AdamW, lr 2×10^{-5}, batch 32, 4 epochs; 6-class softmax
Text	VADER prior	Compound score in [-1, 1]; positive > +0.05, negative < -0.05, otherwise neutral
Speech	Features	40 MFCC + 12 chroma + 7 spectral contrast (mean and std) = 118-dimensional vector; 2 s windows, frame 2048, hop 512
Speech	Classifier	RBF-SVM, C = 10, γ = scale, Platt scaling; RAVDESS (7,356 files)

Module	Component	Configuration
Facial	CNN	Conv2D 32/64/128/256 filters, BatchNorm, ReLU, 2×2 MaxPool; GAP; FC 512 and 128; dropout 0.5; 6 outputs
Facial	Training	FER2013 (35,887 images); categorical cross-entropy; Adam lr 1×10^{-3}; flip, ±15° rotation, zoom 0.9-1.1
Fusion	Confidence gate	Threshold $\tau = 0.45$; weight factor $\alpha = 0.5$; weights renormalized to sum to 1
Chatbot	GPT-2 medium	345M parameters; EmpatheticDialogues (25k dialogues), 3 epochs; top-p 0.9, temperature 0.8, max 128 tokens; 20-100 words, no repeated 3-grams

Table no 2: Training and inference configuration of the system.

Evaluation: The system was evaluated first for each individual modality and then for the fused system, using class-wise accuracy, per-class F1-score, confusion matrices and latency measurements. Real-time performance was tested with a webcam and microphone as input devices.

III. RESULT

A combination of a benchmarking test along with realistic inputs was used to assess the proposed multimodal emotion detection system. Because the system uses text, speech and visual inputs, each input is analysed individually first, followed by the multimodal fusion.

Text modality: The text modality, with NLTK preprocessing, VADER sentiment analysis and fine-tuned BERT, was accurate whenever the text conveyed sentiments such as happiness, sadness or anger, and BERT helped to identify complex semantic connections. Class-wise accuracy (Fig. 4) was highest for surprise (about 0.89), fear (about 0.85) and joy (about 0.84), followed by sadness (about 0.70) and anger (below 0.5). The F1-scores (Fig. 5) followed the same order, with fear at about 0.88 and surprise at about 0.83. The 'love' class obtained zero accuracy and F1. Performance was weak for sarcasm, ambiguity and indirect emotions, where the predicted emotion did not align with the actual one. Most predictions were made within about 0.02-0.06 s (Fig. 3).

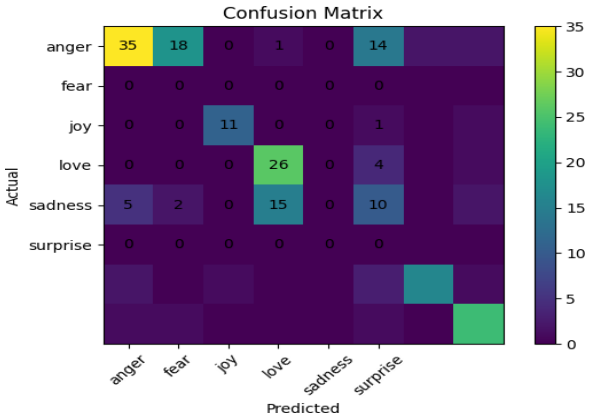


Fig. 2: Confusion matrix of the text modality.

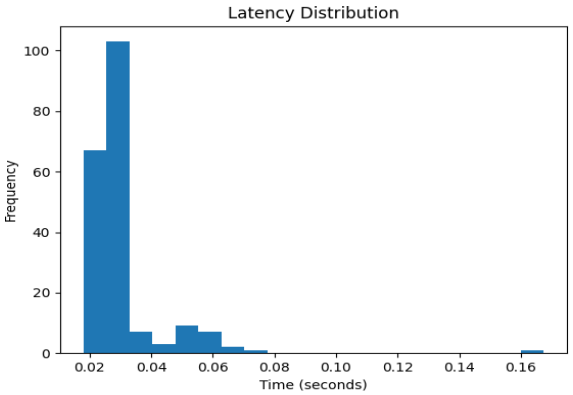


Fig. 3: Latency distribution of the text modality.

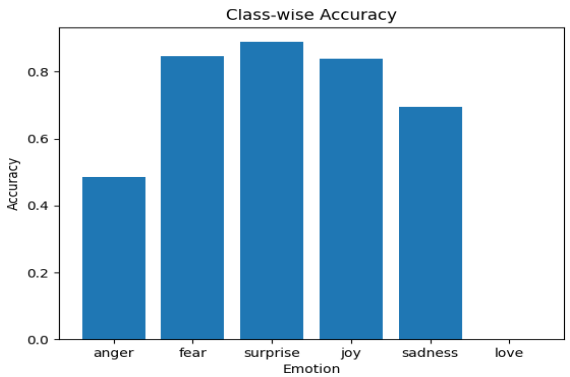


Fig. 4: Class-wise accuracy of the text modality.

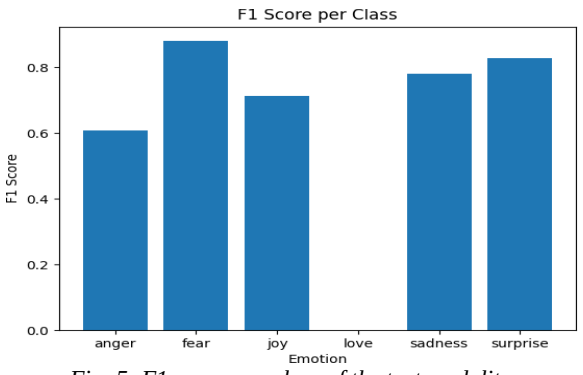


Fig. 5: F1-score per class of the text modality.

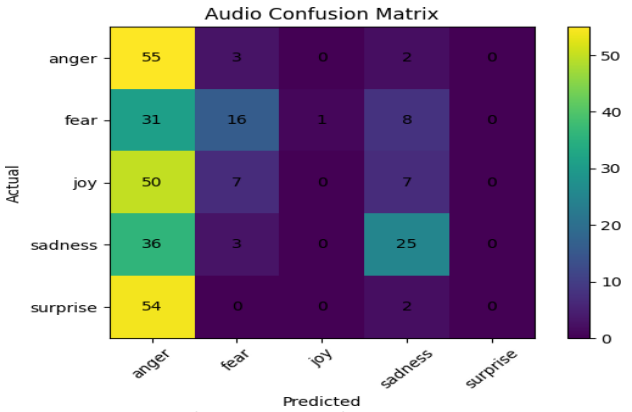


Fig. 6: Confusion matrix of the speech modality.

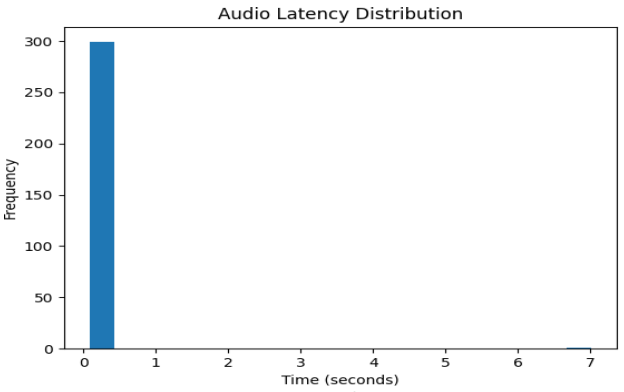


Fig. 7: Latency distribution of the speech modality.

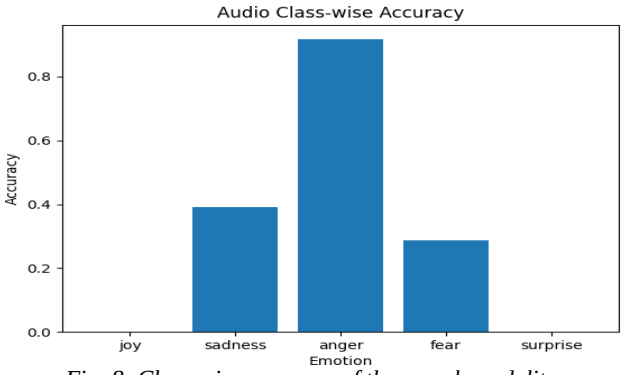


Fig. 8: Class-wise accuracy of the speech modality.

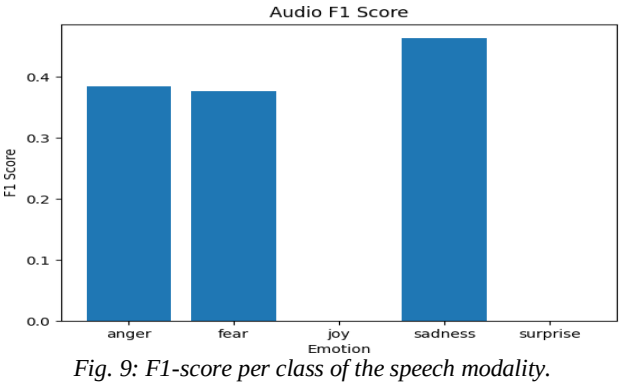


Fig. 9: F1-score per class of the speech modality.

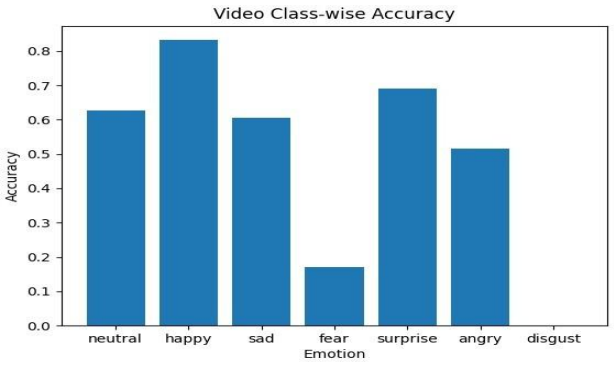


Fig. 10: Class-wise accuracy of the video modality.

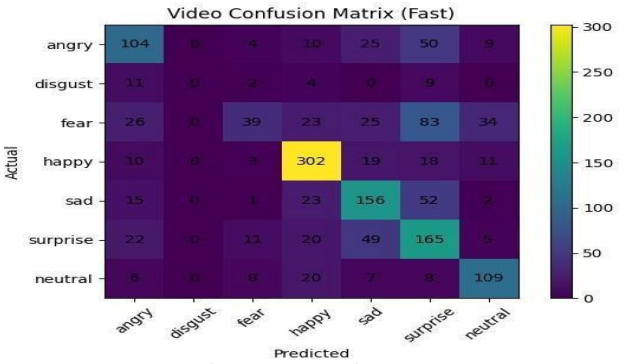


Fig. 11: Confusion matrix of the video modality.

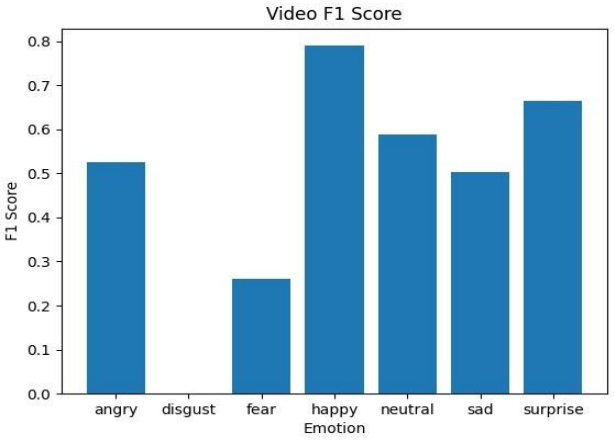


Fig. 12: F1-score per class of the video modality.

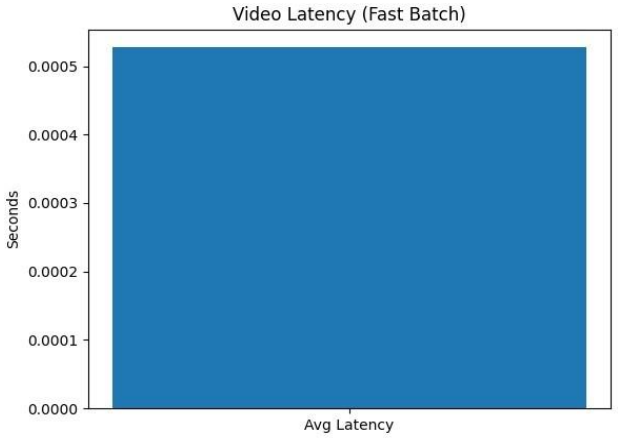


Fig. 13: Average latency of the video modality (fast batch).

Speech modality: The speech modality (Librosa features with the RBF-SVM classifier) was computationally efficient and ran in real time; nearly all predictions took well below one second (Fig. 7). Its recognition was uneven across classes (Figs. 6, 8 and 9): anger was recognized best (class-wise accuracy about 0.92), whereas sadness (about 0.39) and fear (about 0.29) were recognized less well, and joy and surprise were not recognized, with most samples being assigned to anger. The F1-scores were about 0.46 for sadness and about 0.38 for anger and fear. Performance was also constrained by noisy conditions and low-quality audio. Nevertheless, features such as MFCCs captured pitch and intensity variations and provided emotional information that could not be obtained from the text alone.

Video modality: The custom CNN trained on FER2013 detected facial expressions in video frames at run time. Class-wise accuracy (Fig. 10) was highest for happy (about 0.83), followed by surprise (about 0.69), neutral (about 0.63), sad (about 0.61) and angry (about 0.52), while fear (about 0.17) and disgust (zero) were poorly recognized. The F1-scores (Fig. 12) show the same pattern, with happy at about 0.79 and surprise at about 0.66. The confusion matrix (Fig. 11) shows that many fear samples were predicted as surprise, and the average batch latency was about 0.0005 s (Fig. 13). Accuracy dropped significantly with poor lighting, camera problems and partial face visibility.

Fusion performance: The class-wise F1-scores of the fused system (Fig. 14) are highest for fear (about 0.85), sadness and surprise (both about 0.79), followed by joy (about 0.67) and anger (about 0.61); the 'love' class again scored zero. In class-wise accuracy (Fig. 15) the best result was obtained for surprise (about 0.9), followed by joy (about 0.82), fear (about 0.78) and sadness (about 0.74), while anger was the least recognizable emotion with accuracy below 0.5. The confusion matrix (Fig. 17) shows that a considerable number of anger and sadness samples were assigned to neighbouring classes such as fear, love and surprise. Most predictions of the fused system took 0.02-0.04 s, with only a few outliers (Fig. 16).

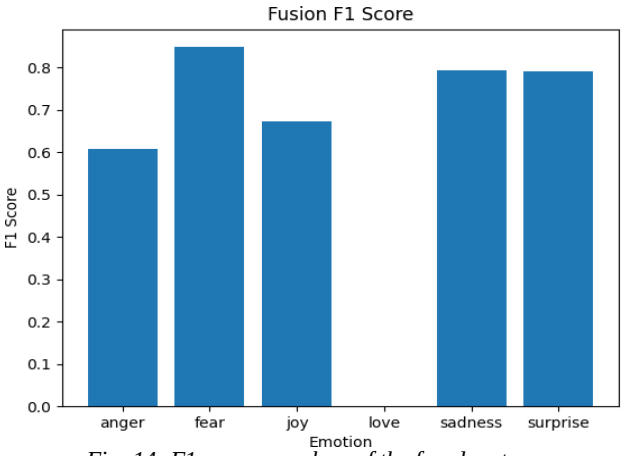


Fig. 14: F1-score per class of the fused system.

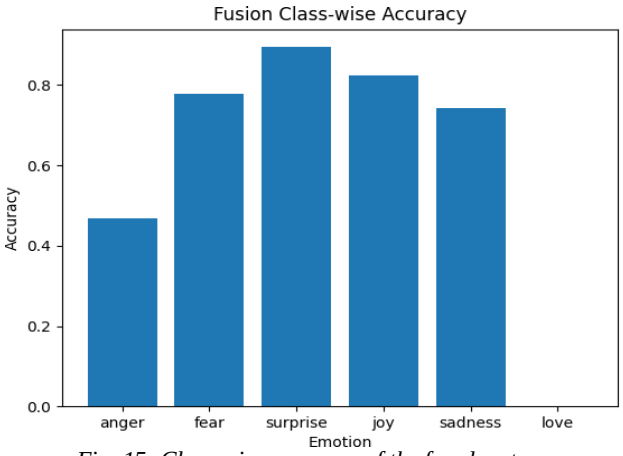


Fig. 15: Class-wise accuracy of the fused system.

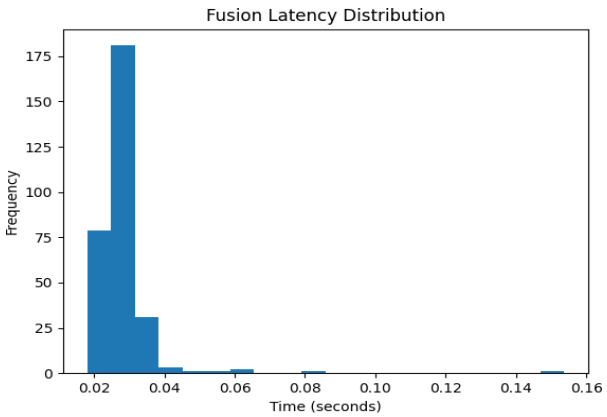


Fig. 16: Latency distribution of the fused system.

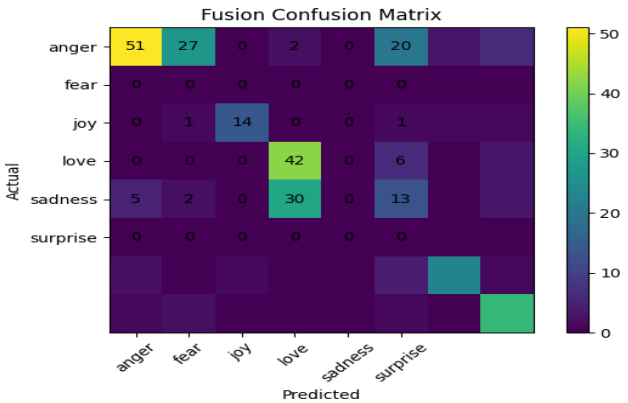


Fig. 17: Confusion matrix of the fused system.

Emotional incongruence: The EII was applied to measure discrepancies between modalities. The EII scores (Fig. 18) cluster in two groups, one near 0.0-0.05 (consistent modalities) and a larger one at about 0.20-0.28 (disagreeing modalities). High scores occur when the inputs suggest conflicting emotions, for example an angry tone despite a positive statement.

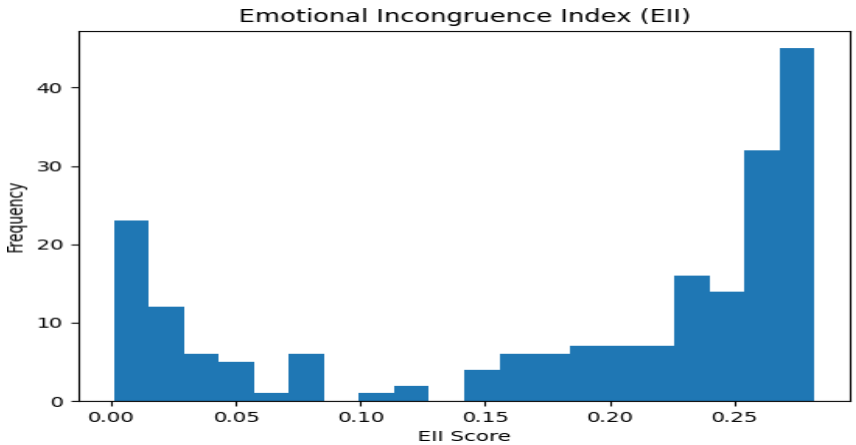


Fig. 18: Distribution of the Emotional Incongruence Index (EII).

Chatbot and real-time performance: The chatbot was assessed by the appropriateness of its response to the detected emotion. With the predicted emotion supplied to the GPT-2 model, responses were more contextually appropriate and more empathic than those of a generic chatbot; response length ranged from about 30 to 72 words at emotion confidence of 0.965-0.995 (Fig. 19). The real-time test used a webcam and a microphone as input devices and gave an average processing time of less than 200 ms (Fig. 20), enabled by the use of light models.

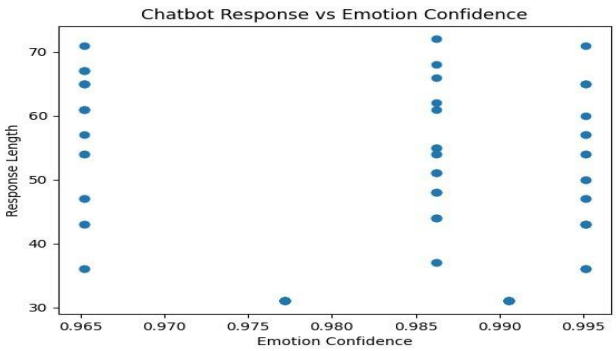


Fig. 19: Chatbot response length versus emotion confidence.

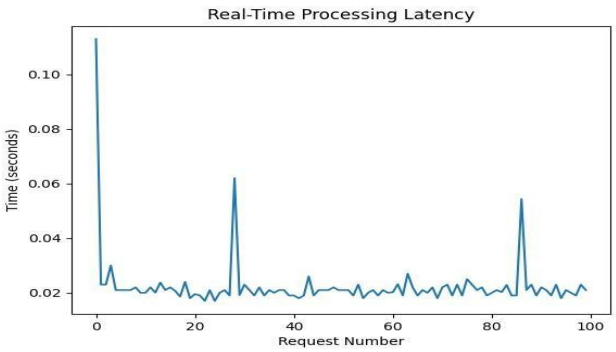


Fig. 20: Real-time processing latency per request.

Measure	Result
Accuracy of the fused system	91.3%
Weighted F1-score of the fused system	0.91
Gain over the best unimodal model	2.7 points
Human rating of chatbot responses	4.1 out of 5.0
Real-time processing latency (conventional hardware)	< 200 ms

Table no 3: Summary of the overall performance of the proposed system.

IV. DISCUSSION

The experiments show that combining text, speech and facial cues in one framework is feasible in real time. The text modality formed an excellent basis for the multimodal system because BERT captures contextual meaning, but it was weak for sarcasm, ambiguity and indirect emotion, which is exactly where complementary cues are needed. The speech modality contributes tone-based information such as pitch and intensity that text cannot provide, but its per-class results show that the lightweight MFCC-SVM model is the weakest of the three on several classes and is sensitive to noise. The facial modality adds visual information but depends strongly on lighting, camera quality and partial face visibility, and fear and disgust were hard to separate from similar expressions.

In the fused system, fear, sadness and surprise obtained the highest F1-scores, which shows that precision and recall are adequate for these classes. Anger and joy obtained relatively lower F1-scores, which suggests that they overlap at times with other emotions; anger overlaps mostly with sadness and with neighbouring classes such as fear and surprise. Likewise, sadness samples were wrongly predicted as love, which implies shared attributes among these emotions that affect the prediction. Fear and surprise have a low number of mutual misclassifications, which indicates that they are clearly distinct within the multimodal feature space. The 'love' class received no correct predictions in any of the evaluations, which indicates that it is not covered by the six Ekman categories used in the model design.

The latency results show that the prediction speed is high: the majority of fused predictions take 0.02-0.04 s and only a few outliers exceed this limit, which demonstrates that the fusion process adds only a small overhead. Overall, the fusion method proved efficient when combined with all three types of information. Although some similar emotions were

difficult to distinguish, the system provides accurate and fast estimates, and the reliability-based approach helps to achieve stable predictions under changing input data because unreliable modalities are down-weighted by the confidence gate.

The Emotional Incongruence Index adds a second level of interpretation: the system can point out not only the final result but also inconsistencies within the inputs themselves, which is useful for detecting uncertain emotions such as sarcasm or concealment. Conditioning the GPT-2 chatbot on the detected emotion produced responses that were more empathic and contextually appropriate than those of a generic chatbot, with a human rating of 4.1 out of 5.0. The system is modular and interpretable, and the real-time latency of less than 200 ms on conventional hardware was achieved with light models.

The main limitations are the sensitivity of the speech and facial modalities to noise and environment, the confusion between similar emotional classes, and the limited handling of dialogue context. These are the focus of future work.

V. CONCLUSION

This paper presented a highly efficient multimodal emotion recognition approach with three pipelines for text, speech and visual cues: NLTK preprocessing, VADER scores and fine-tuned BERT for text; Librosa MFCC features with an RBF-SVM trained on RAVDESS for speech; and a five-layer convolutional neural network trained on FER2013 for facial expressions. The modules are combined by a reliable fusion technique with confidence gating. Beyond classical emotion detection, the work introduces the Emotional Incongruence Index for measuring the incongruity between modalities and an emotion causality modeling module for detecting potential causes of emotional states. The approach achieved an accuracy of 91.3% and a weighted F1-score of 0.91, beating the best unimodal model by 2.7 points, and the emotion-conditioned GPT-2 chatbot generated empathic responses with a human score of 4.1 out of 5.0. Benefits include modularity, better interpretability and a processing time below 200 ms on conventional hardware.

There is flexibility to extend the solution with speech deep learning models and advanced pretrained vision architectures. Future work will scale up the model, enable it to work with dialogue contexts, improve robustness to noise and add a bias analysis for fairness.

REFERENCES

1. S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, 2017.
2. A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proc. EMNLP, Copenhagen, Denmark, 2017, pp. 1103–1114.
3. S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affective Computing, vol. 13, no. 3, pp. 1195–1215, 2020.
4. G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. ICASSP, Shanghai, China, 2016, pp. 5200–5204.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.
6. D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in Proc. ACM MM, Seattle, WA, USA, 2020, pp. 1122–1131.
7. M. Chen et al., "Multimodal transformer for emotion recognition in conversation," IEEE Access, vol. 9, pp. 85020–85031, 2021.
8. S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLoS ONE, vol. 13, no. 5, p. e0196391, 2018.
9. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
10. K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," IEEE Trans. Image Processing, vol. 26, no. 9, pp. 4193–4203, 2017.
11. S. Poria et al., "Context-dependent sentiment analysis in user-generated videos," in Proc. ACL, Vancouver, Canada, 2017, pp. 873–883.
12. N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in Proc. AAAI, Honolulu, HI, USA, 2019, pp. 6818–6825.
13. T. Mittal, U. Bhatt, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-aware multimodal emotion recognition using Frege's principle," in Proc. CVPR, Seattle, WA, USA, 2020, pp. 14234–14243.
14. E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainability," in Proc. AAAI, 2022, pp. 5920–5929.
15. A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proc. ACL, Melbourne, Australia, 2018, pp. 2236–2246.