International Journal of Advanced Multidisciplinary Research and Educational Development
Volume 2, Issue 1 | January - February 2026 | www.ijamred.com
ISSN: 3107-6513

# Real-Time Multimodal Emotion Recognition

Dr.S.Seetha
*Dept. of Information Science and Engineering*
*CMR Institute of Technology*
*Bengaluru, India*
*seetha.s@cmrit.ac.in*

Manoj S V
*Dept. of Information Science and Engineering*
*CMR Institute of Technology*
*Bengaluru, India*
*masv22ise@cmrit.ac.in*

Pradyumna M P
*Dept. of Information Science and Engineering*
*CMR Institute of Technology*
*Bengaluru, India*
*prmp22ise@cmrit.ac.in*

Samson Anthony
*Dept. of Information Science and Engineering*
*CMR Institute of Technology*
*Bengaluru, India*
*saan22ise@cmrit.ac.in*

***Abstract*—Emotion recognition is an emerging research area that lies at the intersection of artificial intelligence, psychology, and human–computer interaction. Traditional emotion-recognition systems rely on unimodal inputs such as text, audio, or visual cues, which often leads to limited accuracy and reduced robustness in real-world scenarios. To address these limitations, this paper presents a real-time multimodal emotion recognition system that integrates facial expressions from video input, vocal characteristics from audio signals, and semantic information from textual data to predict human emotions more accurately.**

**The proposed system employs deep learning models tailored to each modality. Convolutional Neural Networks (CNNs) are used for facial emotion recognition, Mel-spectrogram-based representations combined with recurrent neural networks are applied for speech emotion analysis, and word-embedding techniques with LSTM-based architectures are utilized for text emotion classification. A unified web-based application developed using the Flask framework enables real-time input processing and emotion prediction.**

**By combining complementary information from multiple human behavioral cues, the multimodal architecture enhances emotional understanding and improves classification consistency compared to unimodal systems. Experimental observations demonstrate that the system performs reliably across different input modalities, highlighting its suitability for real-time applications such as virtual interviews, educational platforms, mental-health monitoring, and human–computer interaction systems. This work emphasizes the importance of multimodal integration and establishes a strong foundation for future advancements in deep learning–based affective computing.**

***Index Terms*—Multimodal emotion recognition, Deep learning, Facial emotion detection, Speech emotion recognition, Text emotion analysis, CNN, LSTM, Human–computer interaction, Affective computing**

## I. Introduction

Emotion plays a fundamental role in human communication, influencing behavior, decision-making, and social interaction. As artificial intelligence systems become increasingly integrated into applications such as virtual assistants, educational platforms, healthcare tools, and intelligent user interfaces, the ability of machines to understand and respond to human emotions has gained significant importance. This need has led to the emergence of affective computing, a research domain focused on enabling computational systems to recognize and interpret human emotional states.

Traditional computational systems process human inputs such as text, audio, and video as purely informational data, often ignoring the emotional cues embedded in speech tone, facial expressions, and linguistic context. This limitation reduces the effectiveness of such systems in applications that require personalized, adaptive, or empathetic interaction. Emotion recognition aims to address this challenge by classifying emotional states such as happiness, sadness, anger, fear, surprise, and neutrality using machine learning and deep learning techniques.

Early emotion-recognition systems primarily relied on unimodal approaches, where emotions were inferred from a single source such as speech, facial expressions, or textual content. Although these systems provided useful insights, they often struggled in real-world scenarios due to ambiguity and incomplete emotional cues. For example, a person may express sadness through facial expressions while maintaining a neutral tone in speech, leading to incorrect predictions in unimodal systems. Such inconsistencies highlight the inherent limitations of relying on a single modality.

To overcome these challenges, multimodal emotion recognition integrates multiple sources of information, typically including text, audio, and visual inputs. Since human emotions are naturally expressed through a combination of linguistic, vocal, and facial cues, multimodal systems provide a more comprehensive and robust understanding of emotional context. By combining complementary information from different modalities, these systems can compensate for the weaknesses of individual channels and improve overall recognition accuracy.

The objective of this project is to design and implement a real-time multimodal emotion recognition system using deep learning techniques. The proposed system analyzes textual input using embedding-based neural networks, processes speech signals through spectrogram-based deep learning models, and detects facial emotions using convolutional neural networks. Each modality is processed through a dedicated pipeline, and the models are integrated into a unified framework accessible through a Flask-based web application.

The scope of this work includes the development of independent preprocessing, feature extraction, and modeling pipelines for text, audio, and video modalities, along with real-time deployment through a user-friendly interface. While the system demonstrates improved reliability compared to unimodal approaches, it also faces limitations related to input

ISSN: 3107-6513

quality, computational requirements, and the absence of advanced multimodal fusion strategies. Despite these constraints, the modular design of the system provides a strong foundation for future enhancements.

The remainder of this paper is organized as follows. Section II reviews related work in emotion recognition across different modalities. Section III describes the system requirements and overall architecture. Section IV presents the proposed methodology for text, audio, and video emotion recognition. Section V discusses implementation details. Section VI analyzes experimental results and system performance. Finally, Section VII concludes the paper and outlines directions for future research.

## II. Literature Review

Emotion recognition has become a prominent research area within artificial intelligence, driven by the growing demand for emotionally aware systems in human–computer interaction. Over time, research has progressed from simple rule-based and handcrafted feature approaches to advanced deep learning techniques capable of processing large volumes of data. Existing studies highlight the limitations of unimodal emotion-recognition systems and emphasize the advantages of integrating multiple modalities to improve robustness and accuracy.

### *A. Text-Based Emotion Recognition*

Text-based emotion recognition is an important subfield of natural language processing that focuses on identifying emotions expressed through written language. Early approaches relied on handcrafted lexicons and traditional machine learning classifiers such as Na¨ıve Bayes and Support Vector Machines. These methods depended heavily on predefined emotion-related words and syntactic patterns, limiting their ability to capture contextual and implicit emotional cues.

With the advancement of neural networks, deep learning models such as Convolutional Neural Networks and Long Short-Term Memory networks significantly improved text emotion recognition by learning contextual patterns directly from data. Word embedding techniques, including Word2Vec, GloVe, and FastText, enabled richer semantic representation of text, allowing models to detect subtle emotional expressions. More recently, transformer-based architectures such as BERT and RoBERTa have further enhanced performance by capturing long-range dependencies and nuanced sentiment shifts. Despite these advancements, text-only emotion recognition remains limited, as written language does not always reflect vocal tone or facial expressions, which are crucial components of emotional communication.

### *B. Audio-Based Emotion Recognition*

Audio-based emotion recognition focuses on analyzing speech signals to infer emotional states from vocal characteristics. Early research utilized handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients, pitch, energy, and spectral properties, combined with classical classifiers including Hidden Markov Models and Support Vector Machines. While these approaches provided useful insights, they were often sensitive to noise, speaker variation, and environmental conditions.

The introduction of deep learning models transformed audio emotion recognition by enabling automatic feature learning from time–frequency representations such as spectrograms and Mel-spectrograms. CNNs and LSTM-based architectures have proven effective in capturing both spectral and temporal patterns in speech signals. Hybrid CNN–LSTM models, in particular, are capable of learning local frequency-based features while tracking emotional changes over time. However, real-world audio data often contains background noise, overlapping speech, and recording artifacts, which continue to challenge the reliability of audio-only emotion-recognition systems.

### *C. Video and Facial Emotion Recognition*

Facial emotion recognition aims to identify emotional states by analyzing facial expressions in images or video frames. Traditional methods relied on techniques such as Haar cascades, Local Binary Patterns, and geometric feature extraction, followed by classical machine learning classifiers. These approaches were limited by sensitivity to lighting conditions, pose variations, and occlusions.

Deep learning, particularly CNN-based architectures such as VGG, ResNet, Inception, and XCeption, significantly advanced facial emotion recognition by enabling hierarchical feature extraction from raw pixel data. These models can capture subtle facial muscle movements and expression dynamics. More advanced systems incorporate temporal modeling using LSTM networks or three-dimensional CNNs to analyze expression changes across video frames. Despite their effectiveness, facial emotion-recognition systems face challenges related to poor lighting, occlusion, camera angle variations, and cultural differences in emotional expression, which can reduce accuracy in unconstrained environments.

### *D. Multimodal Emotion Recognition Approaches*

Multimodal emotion recognition integrates information from multiple modalities—commonly text, audio, and visual inputs—to achieve a more comprehensive understanding of human emotions. Early multimodal systems used simple fusion techniques that combined the outputs of unimodal classifiers. Over time, research evolved toward more sophisticated fusion strategies, including early fusion at the feature level, late fusion at the decision level, and hybrid approaches.

Recent studies demonstrate that multimodal systems consistently outperform unimodal models, particularly in real-world scenarios where emotional cues may be incomplete or ambiguous. Deep learning architectures designed for multimodal data, including attention-based models and multimodal transformers, enable systems to model interactions between modalities and focus on the most informative emotional cues. These approaches align closely with natural human perception, where emotions are interpreted through multiple sensory channels simultaneously.

ISSN: 3107-6513

### E. Research Gap and Limitations of Existing Systems

Despite significant progress, existing emotion-recognition systems continue to face several limitations. Unimodal models struggle when emotional cues are subtle or contradictory, while multimodal systems require large, diverse, and well-annotated datasets to generalize effectively. Many datasets exhibit biases related to gender, age, accent, or cultural expression, which can affect model fairness and reliability.

Additionally, real-time multimodal processing demands substantial computational resources and efficient synchronization of input streams. Multimodal fusion remains a complex research problem, as determining how different modalities interact and contribute to emotional meaning is still an active area of investigation. These challenges highlight the need for adaptable, scalable, and context-aware multimodal emotion-recognition frameworks, motivating the development of the proposed system.

## III. System Requirements and Architecture

The development of a real-time multimodal emotion recognition system requires careful consideration of functional requirements, performance constraints, and architectural design. Since the system processes text, audio, and visual inputs independently, it must incorporate dedicated preprocessing pipelines, deep learning models, and an integration mechanism that ensures seamless interaction between system components. This section outlines the system requirements, overall architecture, and design justification based on the proposed framework.

### A. Functional Requirements

The multimodal emotion recognition system must support multiple forms of user input, including text, audio recordings, and images or video frames. The system should allow users to submit inputs through a web-based interface and receive emotion predictions in real time. Each input type must be routed to its corresponding preprocessing and emotion-recognition pipeline. The system should process each modality independently, ensuring functionality even when only one type of input is provided.

The system must apply deep learning models to extract relevant emotional features and generate accurate predictions. The output should include the predicted emotion category and may also display confidence values. The architecture should support modular expansion, allowing new models or modalities to be added without disrupting existing functionality. Additionally, the system must handle invalid inputs and processing errors gracefully by providing appropriate feedback to users.

### B. Non-Functional Requirements

In addition to functional capabilities, the system must satisfy several non-functional requirements. Real-time or near real-time performance is essential to ensure responsive interaction. The system should minimize latency while efficiently utilizing computational resources. Usability is a key consideration, requiring a simple and intuitive interface that can be operated by users without technical expertise.

Reliability and stability are critical to ensure consistent predictions across different operating conditions. The system should be modular to facilitate independent updates and maintenance of individual components. Security and data confidentiality must be maintained when handling user inputs. Furthermore, the system should be scalable to accommodate future enhancements, larger datasets, and increased user demand.

### C. Hardware Requirements

The hardware requirements depend on the computational complexity of deep learning models and real-time processing demands. The system requires a machine equipped with a modern multi-core processor to handle simultaneous processing tasks. A minimum of 8 GB RAM is required, although higher memory capacity is recommended for efficient video and model processing. The availability of a GPU significantly improves performance for training and inference, particularly for audio and visual emotion-recognition models. Adequate storage is needed to store datasets, trained models, and application files.

### D. Software Requirements

The software stack is centered around Python due to its extensive ecosystem for machine learning and data processing. Deep learning frameworks such as TensorFlow or PyTorch are used for model development and training. Libraries such as NumPy and Pandas support numerical computations and data manipulation. Librosa is used for audio preprocessing and feature extraction, while OpenCV is employed for image and video processing, particularly face detection.

The Flask framework is used to develop the web application and manage backend routing between user inputs and machine learning models. Additional tools such as Scikit-learn, Matplotlib, and natural language processing libraries support model evaluation and visualization. This software configuration enables efficient development, deployment, and extension of the system.

### E. System Architecture

The system architecture follows a modular design in which each modality is processed through an independent pipeline, as shown in Fig. 1. User inputs are received through the web interface and routed to the appropriate preprocessing module based on the input type. Text inputs undergo tokenization and embedding, audio inputs are converted into Mel-spectrograms, and visual inputs are processed through face detection and normalization.

After preprocessing, modality-specific deep learning models analyze the extracted features and generate emotion predictions, which are then forwarded to the Flask application for formatting and display. This architecture ensures flexibility, scalability, and ease of maintenance, enabling future enhancements such as multimodal fusion or the integration of additional input channels.

ISSN: 3107-6513

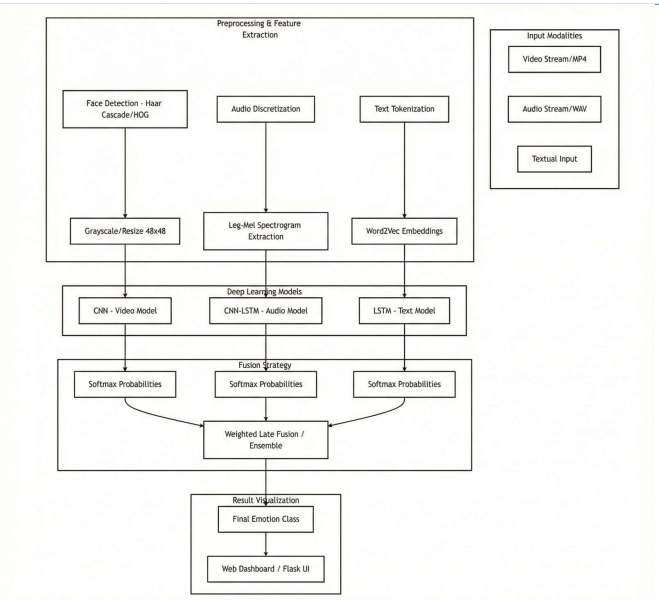


Fig. 1. System Architecture

### F. Module Description

The system is divided into three primary modules: text emotion recognition, audio emotion recognition, and visual emotion recognition. The text module processes written input using embedding techniques and neural network models to identify semantic and contextual emotional cues. The audio module analyzes speech signals through Mel-spectrogram representations and CNN–LSTM architectures to capture tonal and temporal features. The visual module processes facial images or video frames using CNN-based models to detect expression-related patterns.

Each module operates independently, ensuring that the system remains functional even when only one modality is available. This modularity also simplifies debugging, maintenance, and future expansion.

### G. Design Justification

The chosen system design emphasizes modularity, scalability, and real-time performance. Independent modality pipelines allow the system to adapt to varying input availability and quality. The use of deep learning models ensures effective feature extraction from complex data, while the Flask-based web interface provides a lightweight and efficient integration platform. Overall, the architecture balances computational efficiency, usability, and research-driven design, making it suitable for both academic experimentation and practical deployment.

## IV. Proposed Methodology

The proposed multimodal emotion recognition system is designed to analyze human emotions by processing textual, audio, and visual inputs using deep learning techniques. Each modality captures different aspects of emotional expression, and therefore the methodology treats them independently while ensuring they can be integrated into a unified application framework. This section describes the complete methodology followed in the system, including preprocessing, feature extraction, modeling, training, and integration.

### A. Overview of Methodology

The methodology begins with the acquisition of user input through a web-based interface. Depending on the input type—text, audio, or video—the data is routed to the corresponding preprocessing pipeline. Each pipeline converts raw input into a standardized representation suitable for deep learning models. Feature extraction is then performed to capture meaningful emotional patterns, which are subsequently analyzed by modality-specific neural network architectures. The predicted emotion is finally returned to the application layer and displayed to the user. This modular approach ensures flexibility, robustness, and real-time performance.

### B. Text Input Preprocessing

Text preprocessing transforms raw textual input into a machine-interpretable format. The process includes cleaning operations such as removing special characters, converting text to lowercase, and normalizing repeated symbols. Tokenization is applied to break sentences into individual tokens. These tokens are converted into numerical representations using word-embedding techniques, which capture semantic relationships between words. The resulting sequences are padded or truncated to a fixed length to ensure uniformity across samples before being passed to the text emotion-recognition model.

### C. Text Feature Extraction and Modeling

Text feature extraction is achieved through embedding layers or pretrained word embeddings that represent each token as a dense vector. These embeddings are fed into deep learning models such as Long Short-Term Memory networks or Convolutional Neural Networks. LSTM-based architectures are particularly effective in capturing long-term dependencies and contextual emotional patterns within text sequences, while CNNs identify local n-gram features associated with emotional expression. In some cases, hybrid architectures combining CNN and LSTM layers are employed to leverage both local and contextual information. The model is trained to associate learned patterns with specific emotion categories.

### D. Audio Input Preprocessing

Audio preprocessing begins by loading the speech signal and standardizing the sampling rate. The waveform is segmented into frames to capture temporal variations in speech. Noise reduction techniques may be applied to improve signal quality. The preprocessed audio is then transformed into a time–frequency representation using Mel-spectrograms, which approximate human auditory perception. This representation highlights variations in frequency and energy over time and serves as the primary input for the audio emotion-recognition model.

ISSN: 3107-6513

### *E. Audio Feature Extraction and Modeling*

The Mel-spectrogram generated during preprocessing is used for audio feature extraction, as shown in Fig. 2. Deep learning models such as CNNs and CNN–LSTM hybrids analyze the spectrogram to capture emotional cues present in speech. CNN layers extract spatial features related to frequency patterns, while LSTM layers model temporal dynamics such as changes in pitch, rhythm, and intensity. These combined architectures enable effective recognition of emotions conveyed through vocal tone. The audio model is trained on labeled speech data to map acoustic patterns to corresponding emotional states.

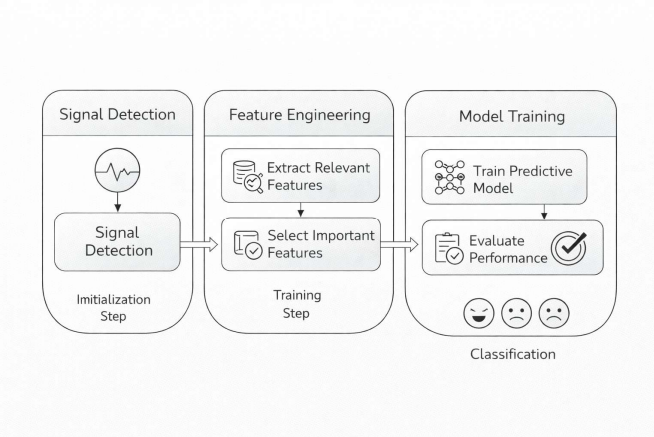


Fig. 2. Audio Processing Pipeline

### *F. Video Input Preprocessing*

Video preprocessing involves extracting frames from video input or processing static images captured through a camera. Since emotions are primarily expressed through facial expressions, face detection techniques are applied to locate and isolate facial regions within each frame. The detected face is cropped, resized, and normalized to match the input requirements of the visual emotion-recognition model. Additional preprocessing steps, such as grayscale conversion or contrast enhancement, may be applied to improve the visibility of facial features under varying lighting conditions.

### *G. Video Feature Extraction and Modeling*

Visual feature extraction is performed using CNN-based architectures that analyze facial images to identify expression-related patterns, as shown in Fig. 3. These models learn hierarchical representations of facial features such as eye movement, eyebrow position, and mouth shape. During training, the network becomes sensitive to subtle variations in facial expressions associated with different emotions. The extracted features are passed through fully connected layers to generate emotion predictions. Although the system primarily processes individual frames, the feature extraction approach generalizes well to real-world facial emotion recognition scenarios.

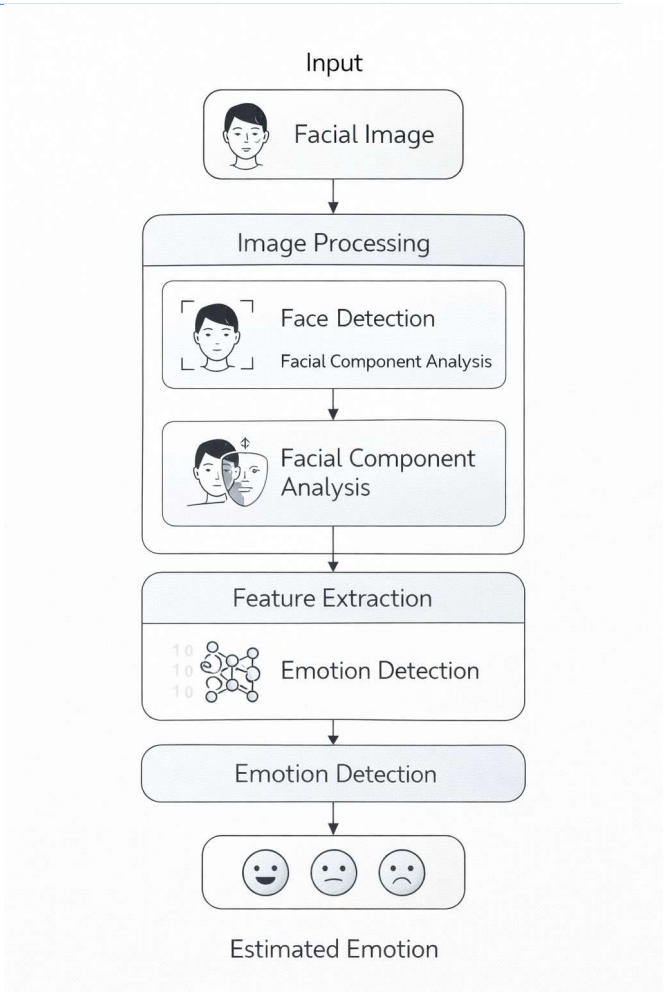


Fig. 3. Video Processing Pipeline

### *H. Model Training and Evaluation*

Each modality-specific model is trained independently using carefully curated and labeled datasets, as shown in Fig. 4. This independent training strategy allows each model to focus on learning modality-specific characteristics without interference from other input types. To ensure reliable learning and prevent overfitting, the dataset is systematically divided into training and validation subsets, enabling continuous performance monitoring throughout the training process.

During training, loss functions such as categorical cross-entropy play a crucial role in guiding the optimization process by quantifying the difference between predicted and actual emotion labels. Optimization algorithms such as Adam are employed to efficiently adjust model parameters, allowing faster convergence and stable learning. Additionally, regularization techniques, including dropout and batch normalization, are incorporated to reduce model complexity, improve training stability, and enhance generalization to unseen data.

Model performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score, which provide a comprehensive understanding of the model's effectiveness

ISSN: 3107-6513

across different emotion classes. Together, these evaluation measures ensure that each modality-specific model achieves balanced and reliable performance, contributing to the overall robustness of the emotion recognition system.

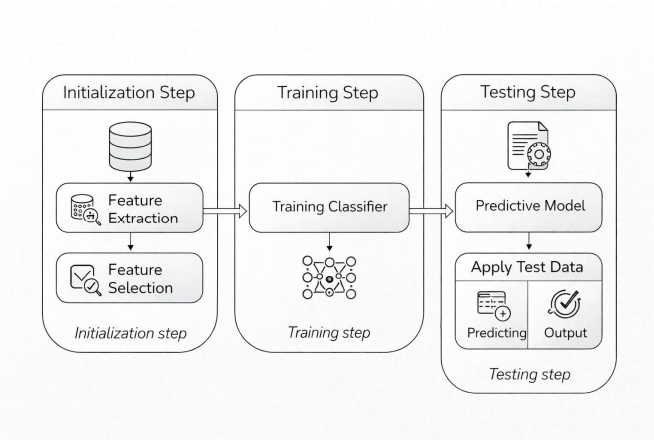


Fig. 4. Training Testing Flow

### *I. Integration of Modalities and Application Layer*

After training, all models are integrated into a unified application layer implemented using the Flask framework. When a user submits input, the system identifies the modality and forwards the data to the appropriate processing pipeline. The predicted emotion label and confidence score are returned to the application layer and displayed through the web interface. Although the current system processes each modality independently, the architecture is designed to support future implementation of multimodal fusion strategies.

## V. IMPLEMENTATION

The implementation phase translates the proposed system design and methodology into a fully functional real-time multimodal emotion recognition application. The system is implemented in a modular manner to ensure maintainability, scalability, and efficient execution. Each modality—text, audio, and visual emotion recognition—is developed as an independent component and integrated through a centralized Flask-based web application.

### *A. Project Structure and Organization*

The project follows a structured directory organization that separates functionality based on modalities. Dedicated folders are maintained for text-based, audio-based, and video-based emotion recognition modules. Each folder contains preprocessing scripts, trained model files, and inference code specific to that modality. The root directory hosts the Flask application, which acts as the central controller connecting the user interface with the backend models. A requirements file is included to list all software dependencies, simplifying environment setup and deployment.

### *B. Text Emotion Recognition Implementation*

The text emotion recognition module processes user-provided textual input through a preprocessing pipeline that includes text cleaning, tokenization, and conversion into numerical representations using embedding techniques. The preprocessed text is passed to a trained deep learning model based on LSTM, CNN, or hybrid architectures. The model analyzes semantic and contextual patterns within the text and predicts the corresponding emotional category. The predicted emotion label is post-processed into a user-readable format before being sent to the web interface.

### *C. Audio Emotion Recognition Implementation*

The audio emotion recognition module begins by loading the audio file submitted by the user. The speech signal is processed using audio processing libraries to extract Mel-spectrograms, which represent the distribution of energy across frequency bands over time. These spectrograms are reshaped and normalized before being fed into a trained CNN–LSTM model. The model captures both spectral and temporal characteristics of speech to predict the emotional state. The module includes error handling mechanisms to manage noisy or corrupted audio inputs and ensures efficient execution to support near real-time inference.

### *D. Video and Facial Emotion Recognition Implementation*

The video emotion recognition module handles static images or video frames captured from user input. Face detection techniques are applied to locate and isolate facial regions within the input. The detected face is cropped, resized, and normalized before being passed to a convolutional neural network trained for facial emotion recognition. The model extracts expression-related features and predicts the corresponding emotion. The implementation accounts for variations in lighting and facial orientation through preprocessing steps and provides meaningful feedback when face detection fails.

### *E. Integration with Flask Application*

The Flask framework serves as the integration layer that connects all modality-specific models with the user interface. When an input is received, the backend identifies the modality and routes the data to the appropriate processing pipeline. To reduce response time, trained models are loaded once during application startup. The Flask routing mechanism ensures clean separation of endpoints for text, audio, and video inputs. The predicted emotion labels and associated confidence information are returned to the frontend for display.

### *F. Execution Flow of the System*

The execution flow begins when a user submits an input through the web interface. The system validates the input and forwards it to the corresponding preprocessing module. After preprocessing, the data is passed to the trained deep learning model, which generates an emotion prediction. The result is then returned to the Flask application and displayed to the user. This structured flow ensures consistent, predictable,

ISSN: 3107-6513

and efficient processing from input acquisition to emotion prediction output.

## VI. Results and Analysis

This section presents the experimental results and performance analysis of the proposed multimodal emotion recognition system. The evaluation focuses on assessing the effectiveness of individual modality-specific models and examining the overall reliability of the system in real-time scenarios. Each modality captures different emotional cues, and the results reflect both their strengths and limitations under varying input conditions.

### *A. Text Model Performance*

The text-based emotion recognition model demonstrated reliable performance in identifying emotions expressed through written language, as shown in Fig. 5. During training and evaluation, the model effectively learned semantic and contextual patterns associated with emotional categories such as happiness, anger, sadness, fear, and surprise. The use of word embeddings enabled the model to capture relationships between words and emotional context.

The model performed well when emotions were clearly articulated through text. However, misclassifications were observed in cases involving ambiguous phrasing, sarcasm, or emotionally neutral statements. These results indicate that while text-based models are effective for structured inputs, their performance is influenced by linguistic complexity and contextual ambiguity.

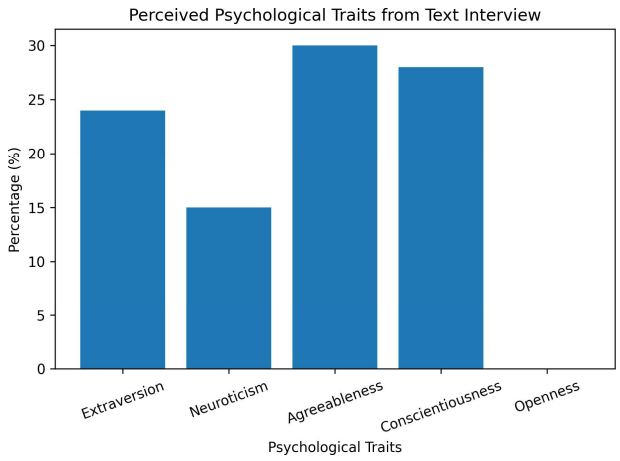


Fig. 5. Text Result

### *B. Audio Model Performance*

The audio-based emotion recognition model exhibited strong performance in classifying emotions conveyed through speech signals, as shown in Fig. 6. By utilizing Mel-spectrogram representations and CNN–LSTM architectures, the model successfully captured both frequency-based and temporal emotional cues. Emotions such as anger, sadness, and calmness were recognized with high consistency in controlled testing environments.

Performance degradation was observed when processing low-quality audio containing background noise, echo, or recording artifacts. Despite these challenges, the audio model demonstrated robustness in clean audio conditions and contributed significantly to the system's overall emotion-recognition capability.

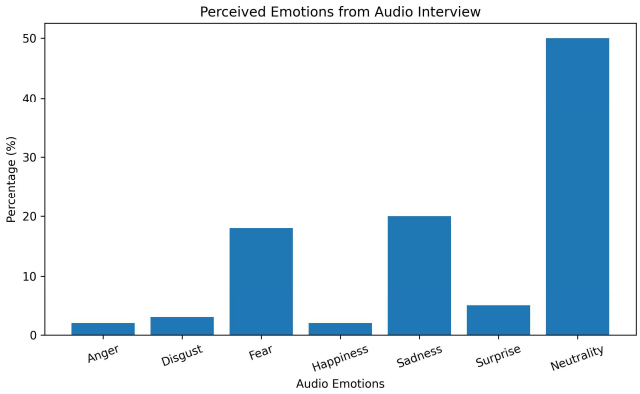


Fig. 6. Audio Result

### *C. Video Model Performance*

The video or facial emotion recognition model achieved effective and consistent results when analyzing clear facial images or video frames captured under adequate lighting conditions, as shown in Fig. 7. The CNN-based architecture demonstrated a strong ability to interpret facial expressions by learning discriminative features such as eye movement, mouth curvature, eyebrow position, and subtle muscle variations across the face. These learned representations enabled the model to reliably identify emotions including happiness, anger, and surprise in the majority of test cases.

During evaluation, the model showed particularly strong performance when facial expressions were well-defined and frontal views of the face were available. This indicates that the network effectively captured spatial patterns associated with expressive facial regions and translated them into meaningful emotional predictions. The results highlight the suitability of CNN-based approaches for real-time facial emotion analysis in controlled or semi-controlled environments.

However, several challenges were observed under less ideal conditions. When faces were partially occluded, rotated at extreme angles, or captured in poor lighting environments, the model's performance declined. In such scenarios, critical facial features were either distorted or insufficiently visible, making emotion discrimination more difficult. Additionally, neutral expressions were occasionally misclassified due to subtle facial movements that resembled emotional cues. These observations emphasize the sensitivity of facial emotion recognition systems to environmental conditions, facial visibility, and input quality, underscoring the need for robust preprocessing and potential

ISSN: 3107-6513

temporal modeling to enhance performance in real-world applications.

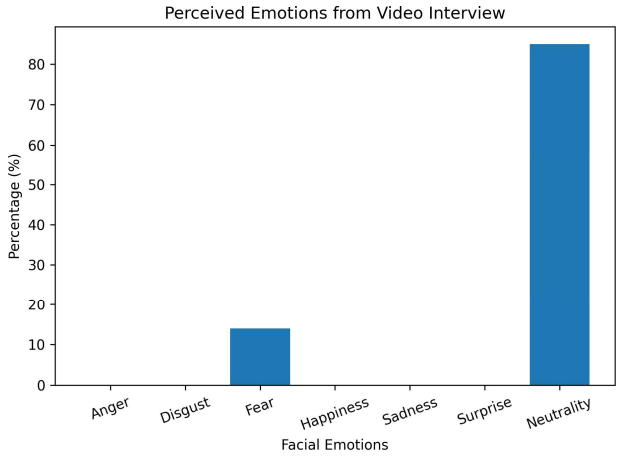


Fig. 7. Video result

### D. Comparison of Modalities

A comparative analysis of the three modalities revealed distinct strengths and weaknesses. The text model performed best when emotional intent was explicitly expressed in words, while the audio model excelled in recognizing emotions through vocal tone. The video model provided strong cues when facial expressions were clearly visible. No single modality consistently outperformed the others across all scenarios.

These findings reinforce the importance of a multimodal approach, as different modalities complement one another. Even though the current system processes each modality independently, their combined analysis offers a more reliable understanding of emotional states.

### E. System Accuracy and Reliability

Overall, the multimodal emotion recognition system demonstrated consistent and reliable performance across all modalities. The system produced stable predictions during repeated evaluations and maintained acceptable accuracy levels consistent with deep learning benchmarks. The integration of models into a unified Flask-based application enabled real-time inference with manageable latency.

The reliability of predictions was influenced by input quality, with clean text, clear speech, and well-lit facial images yielding the most accurate results. Despite these dependencies, the system maintained robustness across diverse testing scenarios.

### F. Real-Time System Evaluation

Real-time testing confirmed that the system operates smoothly and delivers emotion predictions with minimal delay. Facial emotion recognition using live camera input performed effectively, although variations in lighting and camera angles occasionally affected prediction quality. Overall, the system demonstrated practical real-time performance, making it suitable for applications requiring immediate emotional feedback.

### G. Summary of Results

The experimental evaluation indicates that each modality-specific model performs effectively within its respective domain, and their combined use enhances the overall reliability of the system, as shown in Fig. 8. While individual modalities exhibit certain limitations when used in isolation, integrating multiple sources of emotional information allows the system to compensate for these weaknesses and achieve more consistent predictions.

The complementary nature of text, audio, and visual cues contributes to a more comprehensive understanding of emotional states, particularly in real-world scenarios where a single modality may be ambiguous or degraded. By leveraging multimodal inputs, the system demonstrates improved robustness and adaptability across diverse interaction conditions. These results highlight the advantages of multimodal emotion recognition approaches and validate the effectiveness of the proposed real-time system in delivering accurate and reliable emotion predictions.

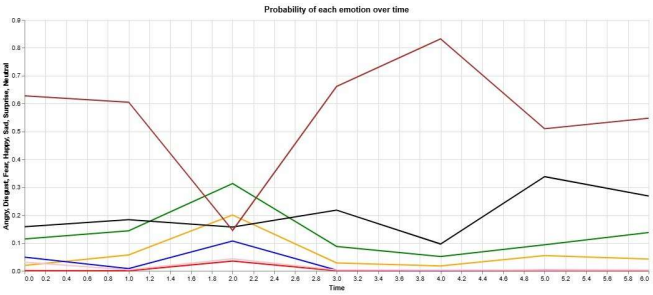


Fig. 8. Overview Result

## VII. Conclusion and Future Work

### A. Conclusion

This paper presented a real-time multimodal emotion recognition system that integrates text, audio, and visual inputs using deep learning techniques. The system was designed to address the limitations of unimodal emotion-recognition approaches by leveraging multiple sources of emotional information. Each modality captures distinct aspects of emotional expression, and the independent models developed for text, speech, and facial emotion recognition demonstrated effective performance within their respective domains.

The experimental results confirm that the system is capable of reliably identifying emotional states when provided with clean and well-structured inputs. The integration of modality-specific deep learning models into a unified Flask-based web application further demonstrates the practical feasibility of deploying multimodal emotion recognition systems in real-time environments. The findings highlight that multimodal approaches offer improved robustness and contextual understanding compared to single-modality systems, even when modalities are processed independently.

Overall, the project successfully meets its objectives of designing and implementing a functional, real-time multimodal emotion recognition framework. The work contributes to the field of affective computing by illustrating the advantages, challenges, and practical considerations involved in building emotion-aware intelligent systems.

### B. Future Work

While the current system establishes a strong foundation, several opportunities exist for future enhancement. One major direction involves implementing multimodal fusion strategies to combine information from different modalities at the feature level or decision level. Such fusion techniques could further improve emotional-context understanding and prediction accuracy.

Future work may also explore the integration of advanced transformer-based architectures for text, audio, and visual emotion recognition. Expanding the training datasets to include greater diversity in terms of cultural background, accents, expressions, and environmental conditions would help improve fairness and generalization. Additionally, improving noise-handling mechanisms for audio input and robustness of face detection under challenging lighting conditions would enhance real-world applicability.

The system can be extended to analyze continuous video sequences rather than individual frames and to incorporate additional physiological signals such as heart rate or body movement. With further refinement and deployment, the proposed framework has strong potential for applications in virtual interviews, mental-health monitoring, e-learning platforms, and intelligent human–computer interaction systems.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the management and faculty of the Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, for their continuous guidance and support throughout the project. Special thanks are extended to the project guide for valuable suggestions, constructive feedback, and encouragement during the development of this work.

## REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.
[2] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," in *Proc. COLING*, 2016, pp. 1601–1612.
[3] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.
[4] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE WACV*, 2016, pp. 1–10.
[5] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
[6] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.
[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. ICLR*, 2013.
[8] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.
[9] F. Eyben, M. Wo¨llmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM MM*, 2010, pp. 1459–1462.
[10] Z. Zhang, J. Han, J. Deng, A. Metallinou, S. Narayanan, and B. Schuller, "Multi-modal speech emotion recognition using recurrent neural networks," in *Proc. ICASSP*, 2014, pp. 6814–6818.
[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[14] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, 2009, pp. 312–315.
[15] P. Ekman and W. V. Friesen, "Facial action coding system," Consulting Psychologists Press, 1978.
[16] D. G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Proc. Affect Computing*, 2008, pp. 92–103.
[17] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Multimodal sentiment intensity analysis in videos," in *Proc. EMNLP*, 2017, pp. 1986–1996.
[18] J. Deng, Z. Zhang, F. Ringeval, E. Marchi, and B. Schuller, "Recurrent neural networks for dimensional emotion recognition from audio," in *Proc. INTERSPEECH*, 2013, pp. 681–685.
[19] S. E. Kahou et al., "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.
[20] Y. Wang, Y. Guan, and L. Wang, "A survey on emotion recognition using speech," *J. China Univ. Posts Telecommun.*, vol. 21, no. 3, pp. 1–13, 2014.