IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 10, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.04

A Multimodel Approach to Recognize Emotion Using Deep Learning Techniques

Prof.Valarmathi C
Assistant Professor/CSE
Sri Sairam College of Engineering, Bengeluru,India
vinmathi20@gmail.com
https://orcid.org/0000-0002-0127-7410
Jeevan B, Manohar M, Gopinath, Bharat Hanamant Desai
Department of CSE
Sri Sairam College of Engineering, Bengeluru,India
sce22cs120@sairamtap.edu.in, sce22cs090@sairamtap.edu.in
sce22cs107@sairamtap.edu.in, sce22cs101@sairamtap.edu.in

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.11/Issue10/NVISX10083
Research Article Open Access| Double-Blind Peer-Reviewed| Article ID: IJIRIS/RS/Vol.11/Issue10/NVISX10083 Received: 28, October 2025, Revised: 05, November 2025, Accepted: 12, November 2025, Published Online: 21, November 2025.
https://www.ijiris.com/volumes/Vol11/iss-10/04.NVISX10083.pdf
Citation: Prof.Valarmathi,Jeevan,Manohar,Gopinath,Bharat (2025),A Multimodel Approach to Recognize Emotion Using Deep Learning Techniques, IJIRIS: International Journal of Innovative Research in Information Security, Volume 11, Issue 10 of 2025 pages 637-643 **Doi:>** https://doi.org/10.26562/ijiris.2025.v1110.04
BibTeX Key:Prof.Valarmathi@2025Multimodel
IJIRIS papers should be cited as IJIRIS (International Journal of Innovative Research in Information Security, AM Publications, India 2025, ISSN 2349-7017, https://doi.org/10.26562/ijiris.2025.v1110.04 The journal's official abbreviation is IJIRIS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: Emotion recognizing human emotions has become an important research topic as it supports various domains including interactive systems, medical support tools, safety applications, and intelligent services. While traditional systems focus on a single input source, such as text, speech, or facial expressions, relying on one modality often limits accuracy in real-world scenarios. In this study, we propose an integrated approach that merges sentiment identification from text, voice-based emotion interpretation, and visual expression analysis through advanced deep learning models. The proposed system processes user text using an NLP-based model, processes audio signals by extracting key acoustic patterns like MFCC, and determines facial emotions using a CNN trained on expression datasets. By merging outputs generated by the three independent modules, the system produces a more reliable and consistent emotion output. Experimental evaluation shows that the multimodel architecture outperforms individual models, demonstrating improved accuracy and stability, especially in situations involving noise, ambiguous text, or partial visibility. This approach highlights the potential of deep learning driven multimodal emotion detection to build more smart and responsive intelligent systems.

I. INTRODUCTION

Emotion is a fundamental part of human communication and often carries more meaning than spoken words. In recent years, developing systems capable of interpreting and identifying human emotions has become a major area of research. This is mainly due to the rising adoption of virtual assistants, digital learning tools, medical technologies, and interactive applications. Traditional emotion recognition usually depends on a single type of input such as text, speech, or facial expressions. This approach often reduces accuracy when the input is unclear, incomplete, or affected by noise. Advancements in deep learning have improved emotion detection by allowing models to identify patterns from large amounts of data. Approaches like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs) are commonly used for image and audio analysis. Natural Language Processing (NLP) methods help in understanding sentiment, context, and meaning in written text. Even with these developments, relying on a single type of input is insufficient for real-life situations interaction. To overcome this problem, the proposed work uses a multimodel approach that combines text-based sentiment analysis, speech-based emotion detection, and facial expression recognition. Each type of input provides different information. Text helps identify the user's intention and sentiment. Speech reveals tone and pitch variation. Facial expressions show visual cues related to emotional state. When the three components work together, the system produces a more accurate and meaningful result compared to using them individually. This paper explains the architecture, implementation, and evaluation of the proposed multimodel emotion detection system. The findings indicate that merging several inputs improves accuracy and system effective for practical use applications.

II. LITERATURE REVIEW

Poria S et.al [1] discussed how combining text, audio, and visual features improves the performance level of emotion detection systems.

It highlighted the challenges of modal fusion and the importance of deep learning for multimodel processing. Goodfellow I et.al [3]presented various Convolutional Neural Network architectures that significantly improved the recognition of facial expressions in images. Their work demonstrated that Convolutional Neural Network based models outperform traditional feature extraction methods. Tripathi S., Beigi H et al,[4] compared single modality models with multimodal architectures and concluded that combining audio, text, and visual data leads to higher accuracy and robustness, especially in complex real-world situations.Verma A et. al, evaluated conventional machine-learning techniques against deep-learning approaches for speech emotion detection. It found that Convolutional Neural Network and Recurrent Neural Network architectures achieved higher accuracy than Support Vector Machine and K-Nearest Neighbors, especially when using Mel-Frequency Cepstral Coefficients and spectrogram features.Busso C., Bulut M at.al introduced the IEMOCAP dataset, which is commonly applied for identifying emotions from speech. The authors extracted Mel-Frequency Cepstral Coefficients and prosodic features to classify emotions like anger, sadness, and happiness.Zhang's et.al applied Long Short-Term Memory and Gated Recurrent Unit networks for text emotion classification and showed improved performance compared to traditional machine-learning approaches that depend on TF-IDF.Li Xet.al reviewed feature extraction techniques driven by Convolutional Neural Network architectures and emphasized the importance of large labeled datasets to train effective facial recognition models.Yoon S and their colleagues introduced an fusion strategy designed to integrate features from text, speech, and facial expressions to produce highly accurate emotion predictions.Chen J and Wang S., applied transfer learning with pre-trained CNN models such as VGG16 and ResNet to improve facial emotion recognition. The method reduced training time and improved accuracy on smaller datasets. Tzirakis P.and Trigeorgis G., developed an end-to-end deep learning pipeline using Mel-Frequency Cepstral Coefficients and spectrogram inputs. their approach delivered strong accuracy results without manual feature engineering.

III. OBJECTIVES

The primary aim of this project is to design and develop an emotion-detection system capable of accurately identifying human emotions by combining text, speech, and facial expression analysis. The system aims to improve the reliability of emotion detection by utilizing deep-learning methods for each modality and then integrating their outputs.

1. To perform real time text-based emotion analysis
   Analyse user text using Natural Language Processing based models to identify sentiments and emotional categories.
2. To extract emotional features from speech signals
   Use Mel-Frequency Cepstral Coefficients and deep learning classifiers to detect emotions from voice tone, pitch, and energy.
3. To recognize facial expression analysis through deep-learning models
   Implement Convolutional Neural Network based models to classify emotions from facial images captured through a webcam.
4. To combine predictions from multiple modalities
   Fuse text, speech, and facial expression results to generate a final and more accurate emotion output.
5. To improve accuracy and stability of emotion detection
   Reduce errors caused by noisy audio, unclear text, or partial face visibility by using multimodel learning.

IV. METHODOLOGY

The methodology followed in this work describes the complete process involved in designing and implementing the multimodel emotion recognition system. The system is divided into three major components: text emotion detection, speech emotion recognition, and facial expression analysis. Each component is processed separately using deep learning models, and their results are combined to produce a final emotion prediction. The overall workflow consists of data collection, preprocessing, feature extraction, model training, and multimodel fusion.

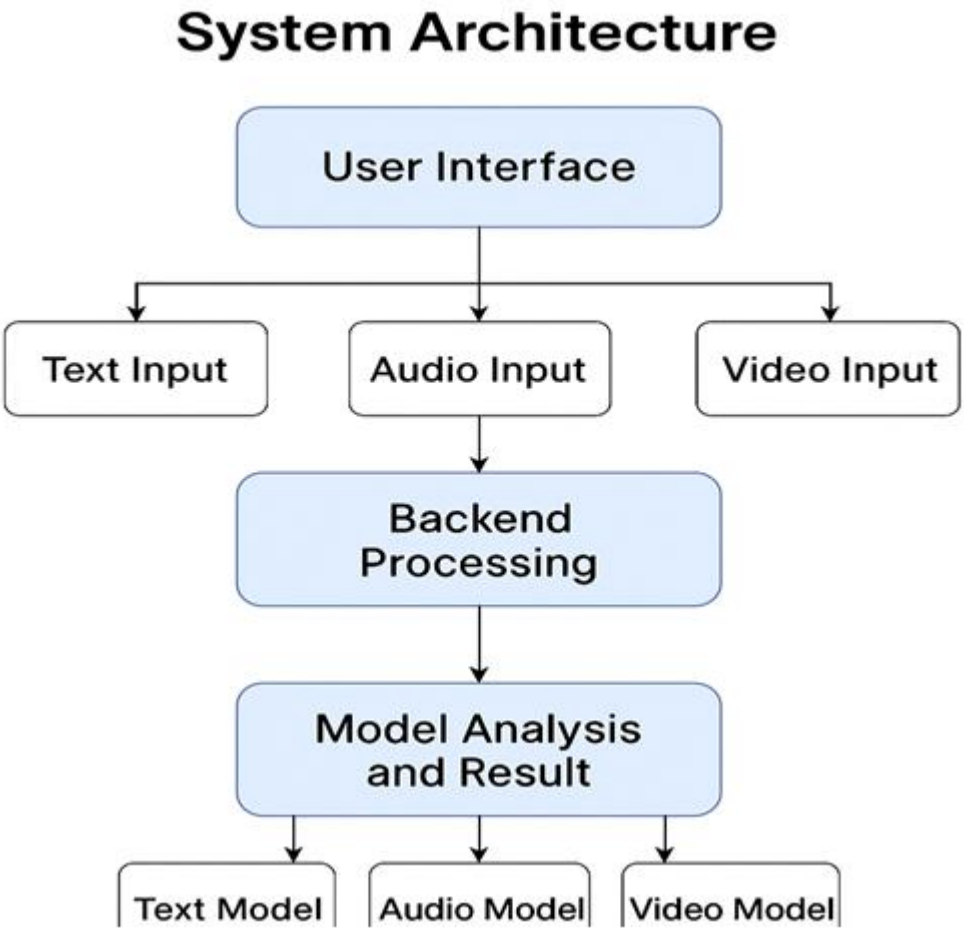
A. System Architecture

The system architecture explains how all three emotion detection modules work together within a structured design. It outlines the flow of data from user inputs to processing and finally to the fusion layer that generates the final emotion.

1. Multimodal Design
   The system is designed using three independent modules for text, speech, and facial emotion detection. Each module processes its input separately and generates a predicted emotion.
2. Input Acquisition
   User input is collected as text through a text field, speech through a microphone, and facial images through a webcam. These inputs are forwarded to the respective processing pipelines.
3. Independent Processing Layers
   Each modality uses a deep learning model suitable for its input type. Text uses Long Short-Term Memory based NLP, speech uses MFCC with CNN/RNN, and facial recognition uses CNN.
4. Fusion Layer
   The final layer combines the outputs of all three modalities to produce a more accurate and stable emotion result compared to using a single modality.

This architecture ensures that each modality contributes to the final output, improving accuracy in real conditions.

IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 10, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.04



B. Text Emotion Analysis

This part explains how the system understands emotional content from user's text. It involves cleaning the text, converting it into meaningful numerical form and classifying it using trained deep learning models.

1. Text Preprocessing
   Raw text is cleaned eliminating stop words, unnecessary symbols, and redundant spaces. This helps in reducing noise and improving model accuracy.
2. Word Embedding Generation
   Cleaned text is converted into numerical form using embedding techniques such as Word2Vec or GloVe. These embeddings capture semantic meaning and word relationships.
3. Deep Learning Model Training
   An LSTM or GRU model is trained with labelled emotion datasets. The model learns emotional by evaluating word order and contextual meaning.
4. Emotion Prediction
   When the user enters a sentence, the trained model analyses it and predicts the emotion class such as happy, angry, sad, fear, or neutral.

C. Speech Emotion Recognition

This part focuses on identifying emotions from the user's speech. The Audio is cleaned and relevant features are extracted and deep learning models classify the emotional tone present in the voice.

1. Audio Preprocessing
   Speech is recorded and converted into a uniform sample rate. Noise removal and silence trimming are performed to enhance signal quality.
2. MFCC Feature Extraction
   Mel Frequency Cepstral Coefficients obtained from the recorded audio. MFCC captures pitch, tone, and energy variations essential for emotion detection.
3. Audio Model Training
   A CNN or RNN model is trained on MFCC feature matrices taken from emotional speech datasets. This helps the system learn patterns in voice tempo and tone.
4. Speech Emotion Output

IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 10, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.04

The trained model classifies the speech input into different emotional categories based on vocal characteristics.

D. Facial Expression Recognition

This module analyzes visual cues from facial images to identify the emotional state. The process includes face detection, preprocessing and classification using deep learning techniques.

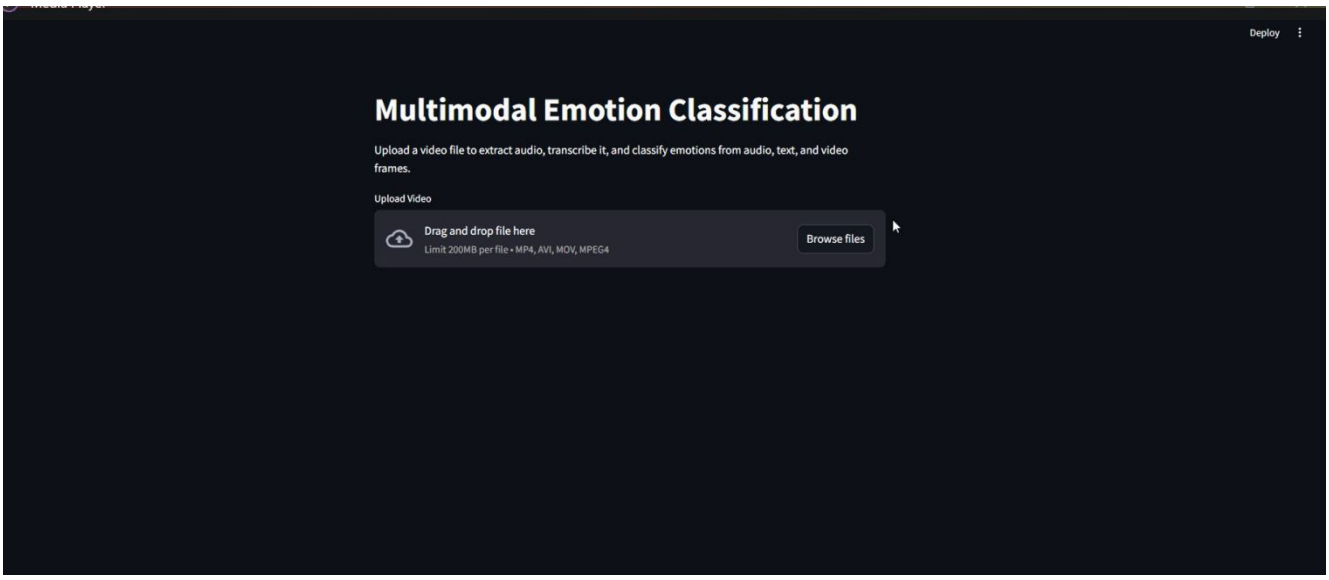
1. Face Image Capture
   The user's facial image is captured using a webcam or sourced from dataset images. Images are resized and normalized for consistent processing.
2. Face Detection
   Algorithms like Haar Cascade or Dlib are used to detect and extract the facial region. This ensures that only the relevant area is sent for emotion analysis.
3. Feature Extraction with CNN
   A CNN model extracts features such as eye movement, eyebrow position, and mouth shape to identify emotional expressions accurately.
4. Facial Emotion Classification
   The CNN predicts emotions like happy, angry, sad, disgust, surprise, or neutral by analysing spatial facial patterns.

E. Multimodal Fusion

This section explains how the system combines predictions from text, speech and facial modules. The fusion process makes the final output more reliable and accurate.

1. **Result Combination**
   Predictions from the text, speech, and facial models are collected and processed together in the fusion layer.
2. **Confidence Score Evaluation**
   Each model provides a confidence value for its prediction. These values are compared to determine the strongest emotion.
3. **Final Emotion**

Output The system generates the final emotion by integrating the strengths of all modalities, improving reliability and accuracy in real time.



F. Implementation Tools

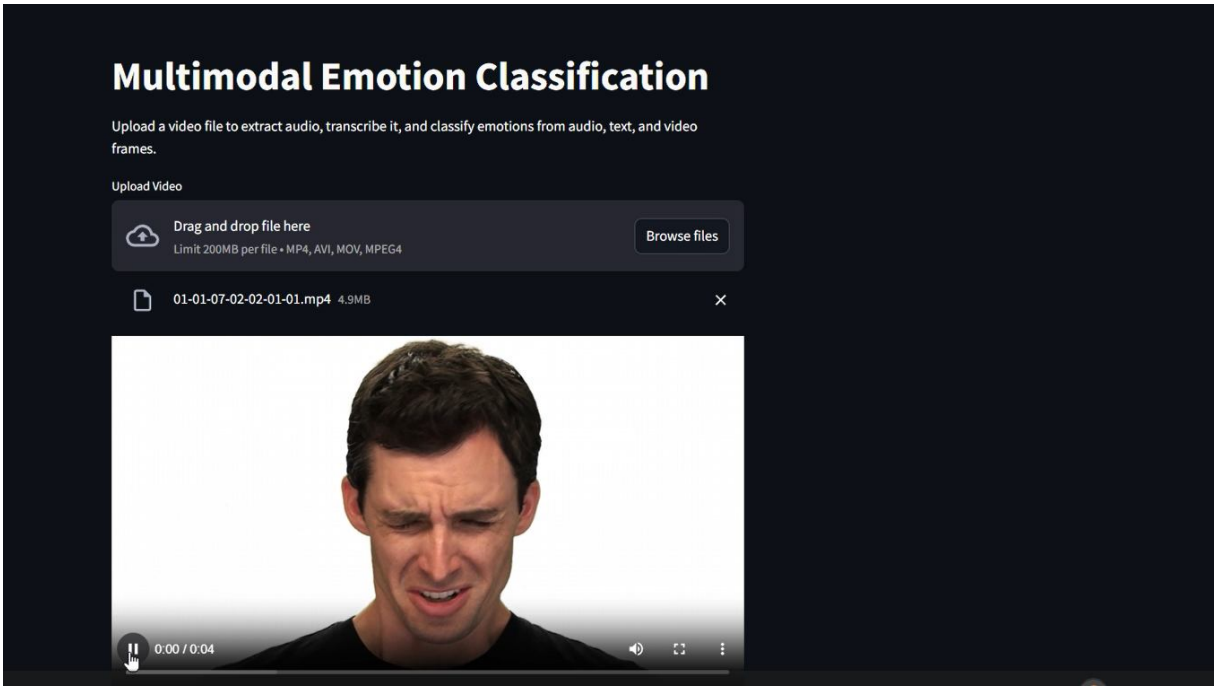
The implementation of the multimodel emotion detection system uses a combination of libraries, frameworks and processing tools. Each tool is selected based on its ability to handle specific tasks such as data processing, model building and real time interaction.

1. **Python Environment**
   Python is used as the primary programming language because it supports machine learning, deep learning and data processing libraries required for text, audio and image analysis.
2. **Deep Learning Frameworks**
   TensorFlow or PyTorch frameworks are used to design, train and evaluate the LSTM, CNN and RNN models. These frameworks provide GPU acceleration and easy model customization.
3. **NLP Libraries**
   Text processing tasks are handled using libraries such as NLTK, spaCy and Gensim. These libraries support tokenization, stop word removal and word embedding generation.
4. **Audio Processing Tools**

IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 10, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.04

Librosa is used for speech processing and MFCC feature extraction. It provides functions for loading audio, trimming silence and generating spectrograms.

5. **Facial Recognition Tools**
   OpenCV is used for image preprocessing, face detection and capturing real time camera input. It helps isolate the facial region before sending it to the CNN model.
6. **GUI Integration**
   A simple interface is developed to collect user input and display the final emotion. This interface connects all three modules and presents the combined output in real time.



G. Testing And Performance Evaluation

Before deploying the system, each module is thoroughly tested to ensure accuracy and consistency. The predictions from all three modalities are evaluated separately and then combined to analyse the overall performance of the multimodel system.

1. **Dataset Based Testing**
   Each model is tested using labelled datasets for text, speech and facial emotions. Accuracy, precision and recall are calculated to validate performance of each module independently.
2. **Real Time Testing**
   The system is tested with real time user inputs for text typing, voice recording and facial expression capture. The response time and correctness of predictions are observed to ensure smooth functioning.
3. **Model Comparison**
   The outputs of individual models are compared with the combined multimodel result. This helps in identifying improvement in accuracy when fusion is applied.
4. **Stability Check**
   Multiple inputs with different emotions are provided to check the stability of the system. The system is observed to see if predictions remain consistent even when one modality is unclear.

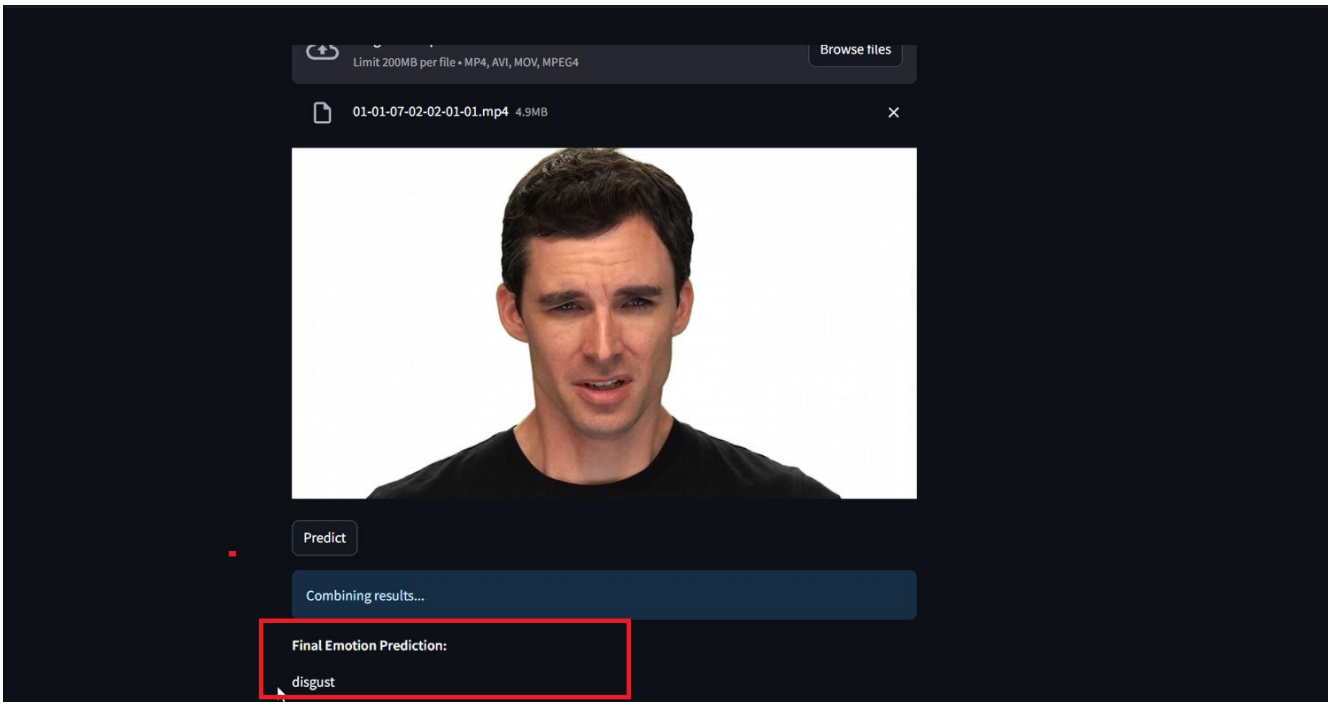
H. Advantages of the Adopted Methodology

The adopted methodology follows a structured approach that improves both accuracy and reliability. The use of deep learning and multimodel fusion helps the system perform better in different real-world conditions.

1. **Improved Accuracy**
   Combining predictions from text, speech and facial modules increases overall accuracy and reduces errors that occur when using a single modality.
2. **Noise Resistance**
   The system remains effective even when audio contains background noise or when the user's facial expression is partially visible.
3. **Better Generalization**
   Deep learning models trained on large datasets help the system understand different emotional patterns and work on input from various users.

IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 10, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.04

4. **Real Time Emotion Detection**
   The methodology supports real time prediction, allowing the system to be used in interactive environments like online learning, customer support or virtual assistants.



V. CONCLUSIONS AND FUTURE SCOPE

The proposed multimodel emotion recognition system successfully combines text, speech, and facial expression inputs to identify human emotions with better accuracy compared to single modality approaches. Each modality contributes unique information, and the use of deep learning models helps in understanding complex patterns present in real world data. Text emotion analysis captures sentiment and context, speech emotion detection identifies variations in tone and pitch, and facial expression recognition provides visual clues about emotional states. By integrating these outputs, the system delivers a more balanced and reliable emotion prediction. The results show that the combined model performs consistently even when one of the inputs is unclear or affected by noise. The approach also demonstrates good scalability and that can be used in applications such as virtual assistants, online learning platforms, healthcare support systems, and human computer interaction. Overall, the developed system proves that multimodel learning can significantly enhance the quality and stability of emotion detection. The system can be improved further in several ways. Additional modalities such as body posture and physiological signals can be included to enhance the depth of emotional understanding. The accuracy can be increased by training the models on larger and more diverse datasets that include different languages, age groups, and lighting conditions. Incorporating transformer-based architectures or attention mechanisms may further improve prediction quality. The interface can be extended for real time monitoring in applications like therapy sessions, online classrooms, and safety monitoring. Cloud based deployment will allow remote access, better storage, and real time processing. The system can also be optimized to run on mobile devices, making it suitable for portable and low power emotion recognition applications. With these improvements, the proposed multimodel system has the potential to evolve into a highly intelligent and adaptive emotion detection solution suitable for future human centred technologies.

REFERENCES

1. Poria S., Cambria E., Hazarika D., "Multimodal Sentiment Analysis: A Review," IEEE Intelligent Systems, 2017. https://ieeexplore.ieee.org/document/7888443
2. Busso C., Bulut M., Lee C. S., "USC IEMOCAP emotional motion-capture dataset," published in the Language Resources and Evaluation journal, 2008.https://sail.usc.edu/iemocap/iemocap_release.html
3. Goodfellow I., Bengio Y., Courville A., "Deep Learning Based Facial Expression Recognition," MIT Press, 2016.https://www.deeplearningbook.org/
4. Valarmathi, C., Velu, A., Prasanth, A., & Dhanaraj, R. K. (2025). NLP-Driven Detection of Cyber-Bullying Comments in Instagram Social Network. In 2025 4th International Conference on Computing and Information Technology (ICCIT) (pp. 383–388). 2025 4th International Conference on Computing and Information Technology (ICCIT). IEEE. https://doi.org/10.1109/iccit63348.2025.10989475.
5. Zhang Z., "Text Emotion Classification Using theDeep Learning Models," International journal of Computer Applications, 2020. https://ijcaonline.org/archives/volume177/number3/31254-312545364.pdf

6. Tripathi S., Beigi H., "Multimodal Emotion Recognition: A Survey and Comparison," Speech Communication Journal, 2018. https://doi.org/10.1016/j.specom.2017.11.003
7. Li X., Zhao G., "Deep Learning Based Facial Expression Recognition," Signal Processing Journal, 2019. https://doi.org/10.1016/j.sigpro.2018.08.027
8. Tzirakis P., Trigeorgis G., Zafeiriou S., Schuller B., "End to End Speech Emotion Recognition Using the Deep Neural Networks," INTERSPEECH, 2017. https://www.isca-speech.org/archive/Interspeech_2017/pdfs/1125.PDF
9. Yoon S., Byun H., "Attention Based Multimodal Fusion for Emotion Recognition," ACM Multimedia, 2020. https://dl.acm.org/doi/10.1145/3394171.3413925
10. Ramamoorthy, R., Velu, A., Valarmathi, C., & Ananthi, M. (2025). Evaluating Mobility Models for IH-VANETs: A Simulation-Based Analysis. In 2025 International Conference on Computing and Communication Technologies (ICCCT) (pp. 1–5). 2025 International Conference on Computing and Communication Technologies (ICCCT). IEEE. https://doi.org/10.1109/iccct63501.2025.11020005.
11. Verma A., Singh R., "A Comparative Study on Emotion Recognition Through Speech Signals," International Journal of Artificial Intelligence Research(IJAIR), 2021. https://ijairjournal.com/uploads/paper/Comparative-Study-Emotion-Recognition-Speech.pdf
12. Chen J., Wang S., "Facial Emotion Recognition Using Transfer Learning," Journal of Image Processing, 2022. https:// jipjournal.org/uploads/papers/facial-emotion-recognition-transfer-learning.pdf.
13. C, V., A, A., A, C., L, H., & L, K. (2024). Enhancing Breast Cancer detection accuracy using U Net architecture. In International Conference on Recent Trends in Computing & Communication Technologies (ICRCCT'2K24). International Conference on Recent Trends in Computing & Communication Technologies (ICRCCT'2K24). International Journal of Advanced Trends in Engineering and Management. https://doi.org/10.59544/zvwa6667/icrcct24p11.
14. C. Valarmathi. (2024). The Combination of Feature Extraction and Classification by Bag of Visual Words to Detect Breast Cancer for Improved Accuracy. Journal of Electrical Systems, 20(3), 2949–2955. https://doi.org/10.52783/jes.4639.