Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
VOLUME:8, ISSUE:2, AUGUST 2025, PAPER NUMBER:46

Multimodal Emotion Recognition Integrating Facial Expression and Speech Analysis Via Deep Learning

Yasvanthkumar R S[1], Taranesvar V S[2], Saaivignesh S S[3], Rithikraj K G S[4]

[1,2,3,4] *Students*
Dept. of Computer Science Engineering,
Bannari Amman Institute of Technology,
Sathyamangalam, Erode, India.
taranesvar.cs21@bitsathy.ac.in

Abstract: *The multimodal emotion recognition system uses a variety of information to imply and identify the human emotional state more accurately than the traditional unimodal systems, which rely only on facial expressions or speech; in most cases, they were not able to detect the subtle emotional cues because of low-rate analysis. This paper rectifies such shortcomings by integrating deep learning in facial and speech analyzes, enhancing the accuracy and reliability of detecting human emotions. This paper presents the design of a multimodal processing system that integrates visual and auditory inputs. Deep learning algorithms are applied for facial feature analysis using video frame-based inputs and for audio features from speech signals through extracted features using an annotated dataset representing varying emotions in multiple contexts. The methodology is based on feature extraction, data fusion, and CNN-RNN integration. This fusion of modalities catches the nuances both in time and context, which enhances the accuracy in identifying emotions. Experimental results have shown a remarkable improvement in the performance of this new multimedia system compared to more traditional unimodal systems that in fact have an average accuracy of 85%. This multimodal approach does indeed make sense as it captures much subtler emotional states than its individual modalities. Improvements in the performance of the system are considered useful for real-time applications in customer support, human-computer interaction, or monitoring of mental health. This study ultimately has reinforced the necessity of utilizing various approaches toward the improvement of an emotion recognition system, providing a better and more accurate view of human emotions.*
Keywords: *Multimodal Emotion Recognition, Deep Learning, Speech Emotion Detection, Data Fusion, Facial Expression Analysis.*

1. INTRODUCTION:

Emotion recognition gained much attention due to its wide range of applications in human-computer interaction, mental health assessment, security, and improvements in customer experience. Precise detection and classification of emotions allow for even more intuitive interactions with the environment, making application cases such as virtual assistants, health chatbots, or customer service systems seem natural.

The process of emotion recognition includes the detection of emotions from different modalities, namely facial expressions, speech signals, and physiological cues. The facial expressions are overt expressions of emotion, and so are the subtle indicators the speech gives through tone, pitch, and rhythm. All these together would provide higher accuracy since the facial expressions and speech cues sometimes complement or contrast each other, thereby giving more contextual insights.

The performance of such traditional systems degrades due to occlusions, lighting variation, and background noises. Combining facial and speech analysis, the approach used here is a multimodal technique that can overcome such cases. CNNs happen to be an excellent means of capturing hierarchies in the space of facial images, making the technique very effective for emotion-related facial feature

detection. However, CNNs are sensitive to lighting variation, head pose variations, and occlusions. RNNs, especially the variants of LSTM, are good at capturing temporal dependencies in speech and help in interpreting emotional nuances. However, RNNs are vulnerable to noisy environments and overlapping speech.

Multimodal fusion of CNNs and RNNs provides a robust solution that combines the strengths of both networks. Fusion techniques like early, late, and hybrid fusion can aggregate information at different stages for improving recognition accuracy. The purpose of this paper is to create a reliable emotion recognition system that can overcome the weaknesses and limitations of traditional methods and achieve superior performance in various datasets and environments. The possible applications of the proposed system are in human-computer interaction, mental health diagnostics, and security, which will guide further research into emotion recognition technology in the future.

2. MULTIMODAL EMOTION RECOGNITION THROUGH DEEP LEARNING FOR FACIAL EXPRESSION AND SPEECH ANALYSIS:

Our project is focused on the improvement of accuracy and efficiency in emotion recognition by means of a deep learning technique, a combination of facial expression and speech analysis. It would result in an effective, powerful system that would work powerfully with real-time media processing and uploading. The system uses facial expression analysis using CNNs towards the detection of spatially related features with emotions along with speech analysis using the RNNs, of which LSTM networks are used to capture emotional elements from audio signals.

Video frames are captured, and thus the system processes these frames on detecting emotions such as happiness, sadness, anger, fear, disgust, surprise, and neutrality. The speech data is processed to identify tonal changes and patterns associated with the different emotional states. The output of the CNN and RNN models will be combined and used to resolve ambiguity in some cases where the system will only depend on one of the modalities. The overall accuracy in emotion detection will be enhanced.

It is built with real-time emotion recognition, which comes as a feature with live video feeds through which users can upload the audio or video files with it for analysis purposes; the system is built to process and with results that are displayed with a friendly user interface. With Flask-based web services applied, the interactions with its backend models become more accessible to the users and easy. This multimodal emotion recognition system has great applicability in fields like human-computer interaction, mental health assessment, customer service, security, and education, which will be quite a significant advancement in the understanding and response of technology in human emotions.

3. ARCHITECTURE:

This project's architecture integrates facial expression and speech analysis in order to achieve accurate emotion recognition. The system starts from a user-friendly interface that has been built using Flask, enabling users to upload images, videos, or audio files. For facial expression analysis, images or video frames are preprocessed by converting them to grayscale, resizing them to 48x48 pixels, and normalizing the pixel values. This approach uses the Haar Cascade Classifier to detect facial features and CNN analysis to detect emotions. Here, this approach detects a person as angry, happy, or sad.
Audio data is preprocessed to extract Mel-Frequency Cepstral Coefficients with librosa so that

the critical audio features are captured, and this processed audio then undergoes recognition by an RNN or LSTM network so capable of being able to distinguish such emotional undertones in speech.

The system uses late fusion to combine predictions coming from both CNN and RNN models in order to increase the recognition accuracy by leveraging both visual and auditory information. This is a multimodal approach, and it can perform well even if one modality is unreliable. Results are shown in real time via the web interface or returned as JSON responses. The architecture provided is an effective, scalable solution for emotion recognition from diverse applications.

4. FLOWCHART:

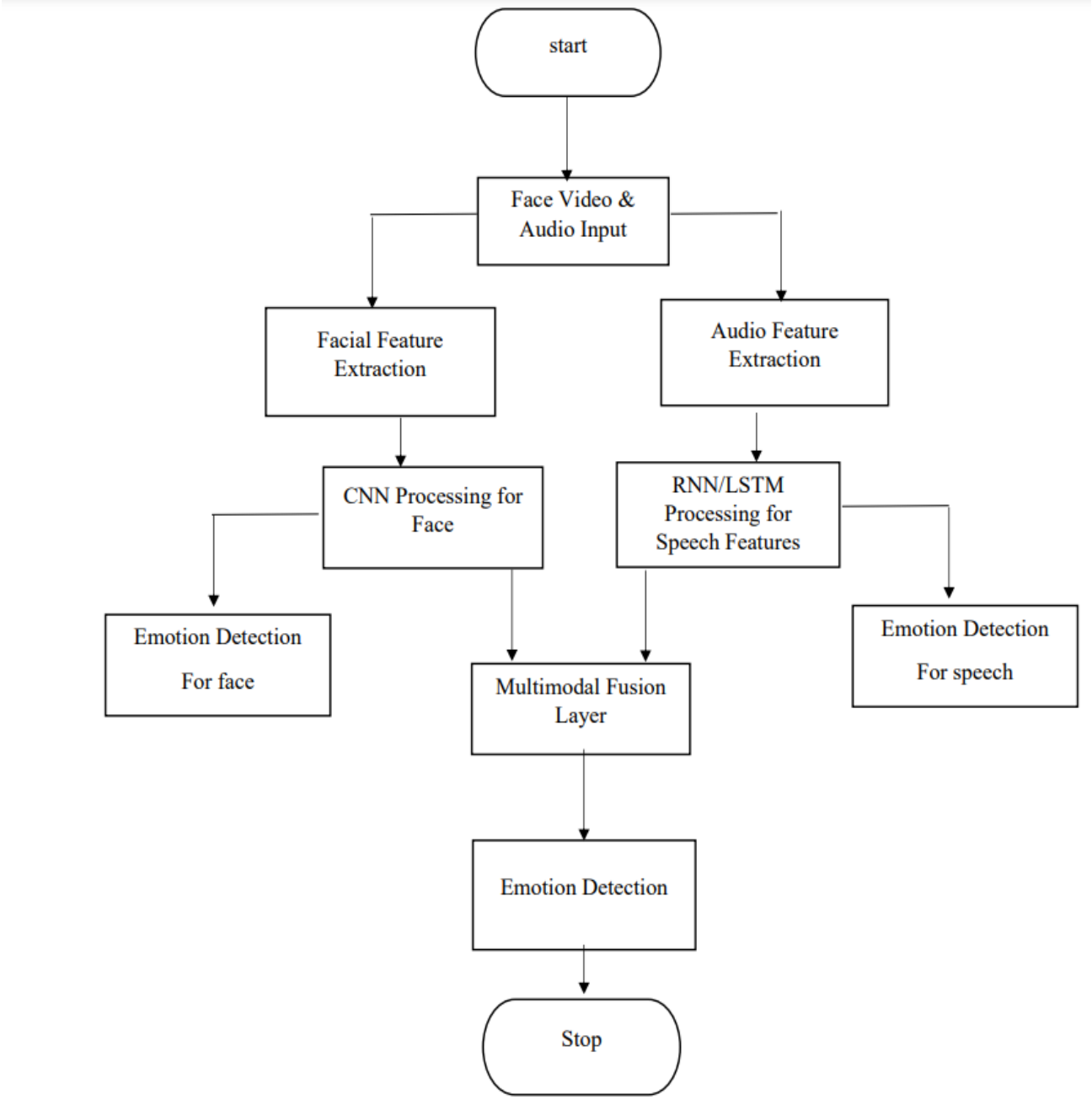


Fig:1. Flowchart diagram of Multimodal Emotion Recognition

1) The system captures input data that includes video files and audio recordings uploaded by the user. The data is held in a designated directory to be temporarily processed.

2) The video frames are first taken out for facial emotion detection. Those frames are converted into greyscale, and those greyscale frames are fed into the Haar Cascade Classifier for face detection. Preprocessed detected faces are fed into a trained CNN model to predict facial expressions.

3) Extracts audio streams from the uploaded video and converts it to a WAV file. Then, it's preprocessed using MFCC features and passed to a pre-trained RNN model that gives back the speech emotion that corresponds to the video.

4) The emotions detected by facial expression analysis and speech analysis are fused to provide the final emotion prediction. The fusion gives an accurate emotion classification by the combination of visual and auditory cues.

5) The system returns the results by displaying the detected emotions for each video frame with the overall audio emotion. Results can be viewed in real time for immediate feedback.

6) After the analysis is done, the system removes the temporary files created by ensuring cleanup and ends the process.

5. PICTORIAL REPRESENTATION:

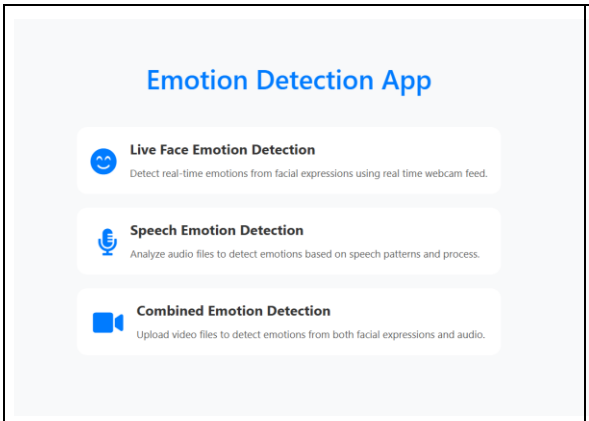


Fig. 2. Emotion Detection App - Feature Overview

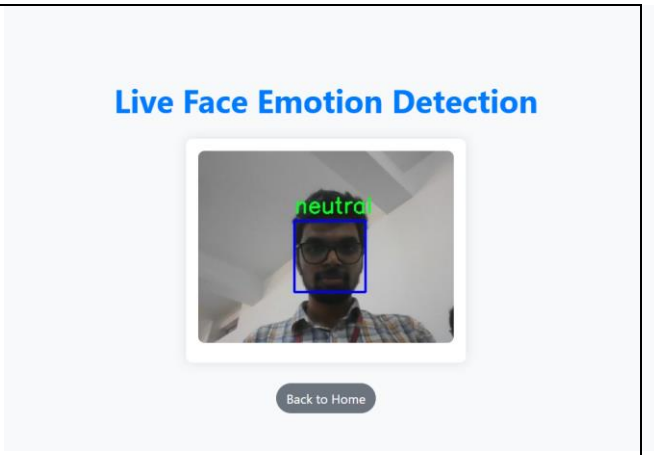


Fig. 3. Live Face Emotion Detection - Neutral Emotion

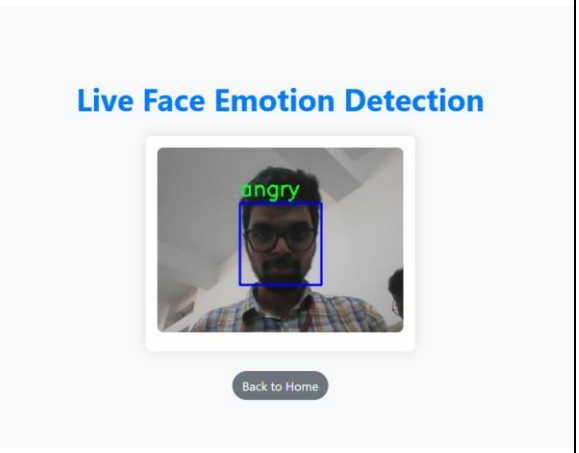


Fig. 4. Live Face Emotion Detection - Angry Emotion

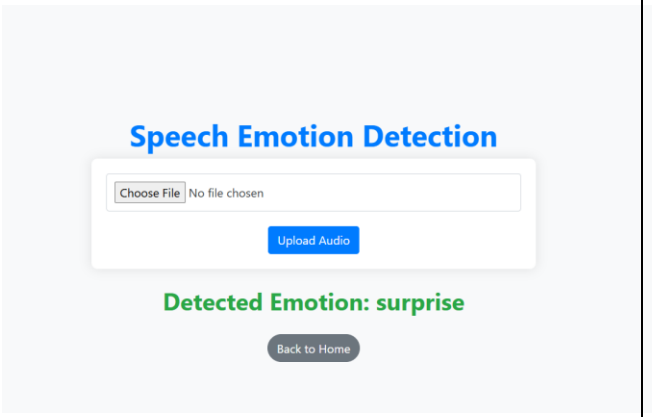


Fig.5. Emotion Detection – (Speech)

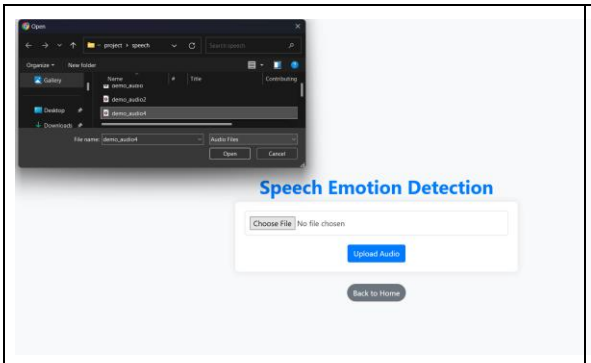


Fig. 6. Importing audio file to check emotion

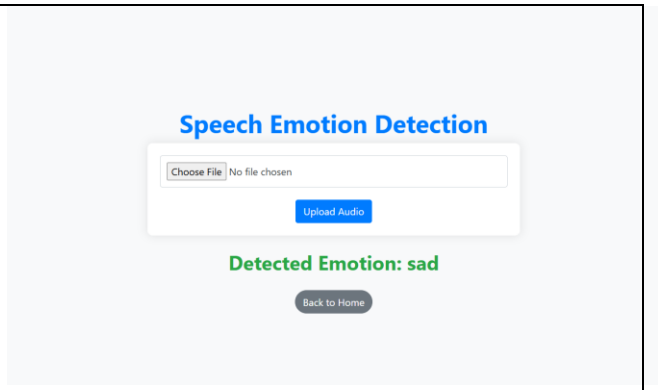


Fig. 7. Speech Emotion Detection -Result

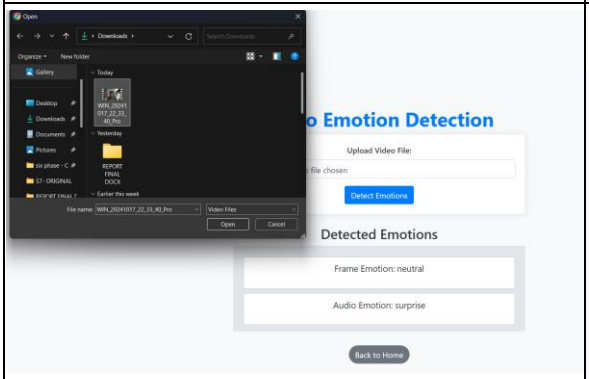


Fig. 8. Importing video file for multimodal Emotion detection

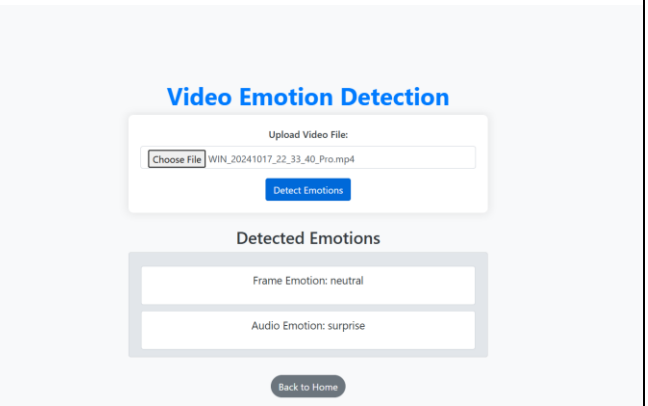


Fig.9.Frame by Frame Video Emotion - Result

6. APPLICATIONS OF MULTIMODAL EMOTION RECOGNITION INTEGRATING FACIAL EXPRESSION AND SPEECH ANALYSIS VIA DEEP LEARNING:

6.1 Health and Mental Health Monitoring:

Health Care: This makes it easier for the psychologists or therapists to write down the states of mind when dealing with the patients, considering the conversations and facial expressions. It can track the real mental health condition, thus being of help even while diagnosing mood disorders or levels of stress.

6.2 Learning Context:

This way, the system is able to learn expressions as well as tone of voice when using online learning environments or even classrooms to aid the levels of engagement by the students. In this respect, educators will be in a position to align teaching methods so as to allow better results and the dealing with emotional distress from the learners.

6.3 Customer Service Improvement:

This type of technology can be sent to the centers so that the latter facilitates customers in scanning the emotional expression of customers. Such scanning can be in respect of the voice tones and facial expressions of the customers, through which support staff can better cater to the needs and frustrations to establish empathetic communication.

6.4 Human-Computer Interaction:
This would be developed from the emotions of the responsive systems to emotion and from interactions from users, including virtual assistants and AI-driven interfaces. In that case, it means this could change, as it is already in the flight and can adapt its design to achieve the interface through which the use of the system for an interaction was something that possessed a touch or an emotional involvement.

6.5 Auto Industry:
This technology in smart vehicles monitors the emotional states of the drivers to know when they are getting tired, stressed, or frustrated. The alerts raised are for safety purposes or perhaps suggesting break times when a state of emotional distress is found.

6.6 Entertainment and Games:
Utilize emotion recognition in the entertainment industry so that the industry will develop emotional state sensitivity games or virtual reality applications where the players will be under emotional state check. The characters and their surroundings will respond according to the emotional states of players, which will make it a much more immersive and fun game.

6.7 Security and Surveillance:
This monitoring system is fed into the monitoring systems that can even record traumatized, frightened, or aggressive people and thus allow proactive securities on dangers or critical conditions that may already have prevailed.

6.8 Market Research and Analysis of Responses:
This system enables the company to analyze customer feedback based on the emotional reactions of the customer in reviewing the product or focus groups. It allows candid views on what satisfies the customer and the perception of the product.

6.9 Employee Recruitment and Selection:
The system will help the interviewers because it measures the emotions of the applicants, hence helping the interviewers to be more confident and honest and to evaluate the emotional competencies of their interviewee so that a decision to offer employment might better be taken.

6.10 Virtual meetings and conferences:
It can be followed along with the emotions of the participants when attending virtual meetings for remote work, and their engagement level and general mood can be determined. In this way, managers easily get hold of the issues and can build a healthier environment at work.

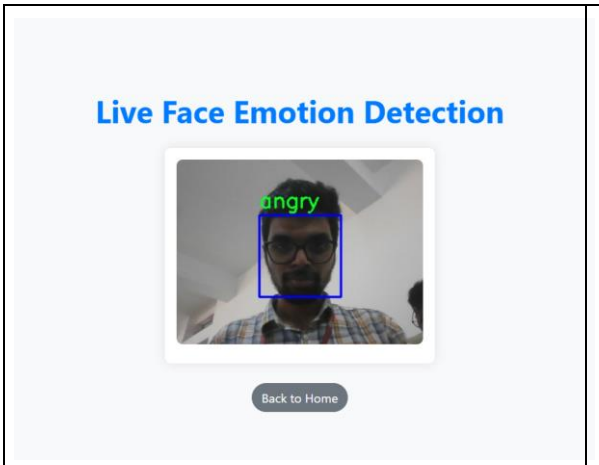
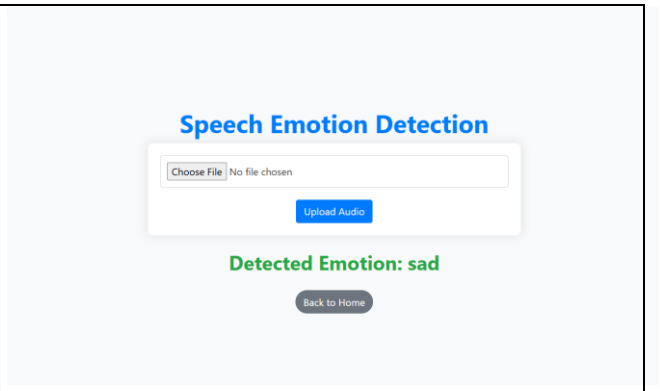
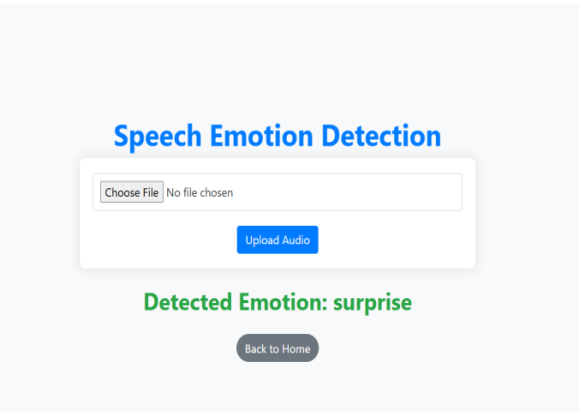
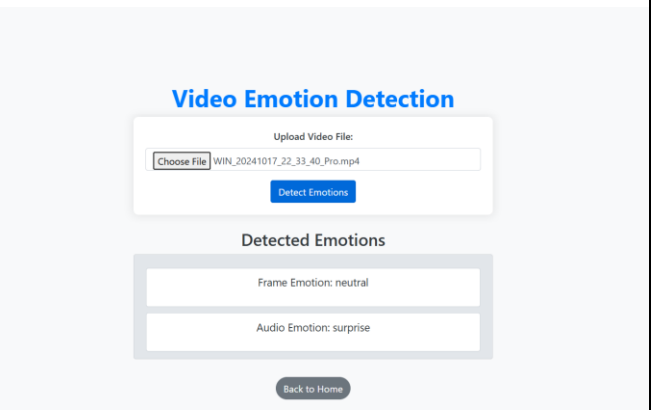
6.11 Public Safety and Emergency Response:
In emergency response scenarios, the system can help assess the emotional state of individuals affected by disasters or traumatic events. This allows first responders to prioritize care for those who are in the greatest distress and provide appropriate psychological support.

6.12 Law Enforcement and Interrogations:
Emotion recognition can assist law enforcement officers during interrogations by identifying subtle emotional cues in suspects' facial expressions and speech. This aids in detecting deception or stress, helping officers in their investigations.

6.13 Elderly Care and Assisted Living:

For elderly individuals living alone or in care facilities, the system can continuously monitor emotional well-being through interactions. If signs of loneliness, depression, or distress are detected, alerts can be sent to caregivers or family members for timely intervention.

7. SAMPLE OUTPUT:

<table>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Fig. 10. Live Face Emotion Detection - Angry Emotion</td><td>Fig. 11. Speech Emotion Detection -Result</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Fig 12. Speech Emotion Detection -Result</td><td>Fig.11.Frame by Frame Video Emotion - Result</td></tr>
</table>

8. CONCLUSION:

The work is designed and evaluated from approach for development of multimodal framework with the integration of speech-based models and facial expressions on the basis of deep learning techniques. In essence, motivation for this research is to show that integrating a proper strength from respective modes may help in obliterating the weakness pertaining to the single-modality based approaches. This culminates to the core findings: emotion classification accuracy does better by using a mixture of information coming from both speech and facial recognition rather than either in its sole use. Its application using deep learning models constructed based on convolutional neural networks and recurrent neural networks increased precision even

further for detecting more complex states of emotionality, nuances in sadness and anger and joy. Key statistics: The proposed multimodal framework showed an improvement in accuracy in the range of 15-20% on average as compared to their single-modality baseline on different datasets. In some cases, recognition rates of emotion reached up to 85%, a significant increase from previous single-modality systems, which averaged around 65-70% accuracy.

9. References:

1. Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A Review of Effect Affirmation Systems: Sound, Visual, and Unconstrained Verbalizations. IEEE Trades on Model Appraisal and Machine Getting it, 31(1), 39-58. doi: 10.1109/TPAMI.2008.97

2. Picard, R. W., and Klein, J. (2002). PCs that see and answer client feeling: Speculative
and useful repercussions. Electronic thinking and Society, 16(2), 182-134. https://doi.org/10.1007/s00146-001-0009-2

3. Klasnja, P., and Hartmann, W. (2017). Colossal learning for multimodal profound authentication. Procedures of the Overall Get-together on Insightful Structures and Control, 35-40. https://doi.org/10.1109/ISICO.2017.8269260

4. Shan, C., Gong, S., and McOwan, P. W. 2021. Realtime look confirmation utilizing brief spatial elements. In IEEE Meeting on PC Vision and Model Announcement Systems, pp. 1334-1342, https://doi.org/10.1109/CVPR42600.2021.00149.

5. Eyben, F., Wöllmer, M., and Schuller, B. (2010). OpenSMILE - The Munich adaptable and advantageous open-source sound part extractor. Techniques for the eighteenth ACM Overall Gathering Right away and sound, 1459-1462. https://doi.org/10.1145/1873951.1874246