MULTIMODAL EMOTION DETECTION SYSTEM UTILIZING COMPUTER VISION AND SPEECH PROCESSING TECHNIQUES

N.Devi* S.Swetha* S.Nithish* G.Vijayaraghavan*
*** Department of Information Technology**
Sri Venkateswara College of engineering

ABSTRACT

Emotion recognition enhances human-computer interaction by enabling systems to respond empathetically to users' feelings, playing a crucial role in applications like mental health monitoring, security, and adaptive learning. Facial emotion recognition analyzes expressions like smiles or frowns through computer vision and machine learning, while speech emotion recognition examines vocal features like tone, pitch, tempo, and intensity. However, challenges such as variability in emotional expressions, data labelling, and multimodal integration make accurate models difficult to develop. Real-time processing, context dependence, and privacy concerns further complicate deployment. We propose a multimodal emotion recognition framework that integrates visual, audio, and textual cues from video data, combining facial expressions, vocal characteristics, and speech content for enhanced emotion detection. Unlike unimodal systems, this framework achieves more accurate and context- aware recognition. With an accuracy of 98.5%, the proposed framework demonstrated improved performance using MFCC-LSTM and CNN architectures. The framework's use of late fusion techniques and balanced datasets, along with Python-based tools and GPU support, highlights its potential for real-world applications such as virtual meetings, online education, customer service, and mental health monitoring.

Key word: Emotion recognition, LSTM, CNN, Late Fusion

I INTRODUCTION

Humans have an innate ability to perceive emotions in others, often through subtle cues that include facial expressions, voice modulation, and body language. This ability is crucial for social interaction and emotional intelligence. One of the most immediate ways people recognise emotion is through facial expressions. Psychologist Paul Ekman et al (2020) identified some rudimentary emotions viz, anger, disgust, fear, happy, sad, and surprise that are commonly identified through movements of facial features[1]. A smile generally indicates happiness, while a frown may signal sadness or frustration. Vocal tone and speech patterns also reveal emotional states. For example, a trembling voice might indicate fear or nervousness, while a high-pitched, fast-paced voice can reflect excitement or anxiety. Similarly, body language such as posture, gestures, and eye contact offers important clues. Open posture and eye contact suggest confidence and engagement, whereas crossed arms or lack of eye contact may indicate discomfort or defensiveness. Crucially, humans also rely on context when interpreting emotions. The same expression can have different meanings depending on the situation. For example, crying at a wedding is often seen as a sign of joy, whereas the same behavior at a funeral would likely be interpreted as sadness.

With the rise of AI and machine learning, emotion recognition is increasingly being modeled and implemented by many researchers. These systems attempt to simulate human emotional perception using data from images, audio, and text. Facial Emotion Recognition (FER) uses computer vision algorithms to detect and analyze facial expressions[15]. Despite its potential, emotion recognition faces several challenges and ethical issues. One of the primary concerns is privacy. Emotion data is deeply personal, and collecting it without explicit consent raises significant ethical questions. Users may not always be aware that their emotions are being monitored, especially when systems operate in the background. Another concern is bias. Emotion recognition systems trained on limited or non-diverse datasets may misinterpret

emotions, particularly across different cultures or demographics. For example, a smile might have different social meanings in different cultures, and misinterpretation could lead to inaccurate assessments. In order to overcome these difficulties, a multimodal emotion recognition framework is proposed to analyze emotions expressed through video data by integrating three key modalities: visual, audio, and textual cues. The system combines facial expressions, vocal characteristics, and speech content to provide a more accurate and context-aware detection of human emotions compared to traditional unimodal approaches. This framework is capable of processing synchronized video data, which serves as an ideal data source for real-world applications. Video-based data inherently provides both visual and auditory cues, making it highly suitable for diverse emotional contexts.

II LITERATURE REVIEW

In this section we explore the key developments in audio-based and visual-based emotion recognition systems, highlighting their methodologies, strengths, and limitations. Singh, M et al (2020) proposed the core complexities in defining emotion lies in its subjective and context-dependent nature. The same stimulus may evoke different emotional reactions in different individuals based on personal experiences, cultural background, mood, and situational context[7]. Furthermore, emotions can be short-lived or enduring, consciously recognized or subconscious, and either overtly expressed or subtly internalized. This makes their detection and analysis, especially by machines, a challenging task requiring sophisticated approaches that can interpret both explicit and implicit cues across various modalities. Understanding this complexity is crucial for designing accurate and context-aware emotion recognition systems that can function reliably across different users and environments[4]. Fayek et al., (2017) reviewed several articles and clarified that emotion recognition has evolved significantly over the past century, progressing from philosophical discourse to sophisticated computational systems[2]. Today, emotion recognition is a cornerstone in numerous applications—from security, education, marketing, health care, surveillance and call centres. The historical journey reflects a multidisciplinary convergence of psychology, neuroscience, linguistics, and artificial intelligence, all contributing to the increasingly nuanced understanding and modeling of human emotions.

Lu, X. et al., (2017) serves as the foundational input for a wide range of classification algorithms, including Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and neural networks like CNNs and LSTMs[13]. Jintao et al., (2020) introduced methods that require handcrafted feature extraction, typically using MFCCs and prosodic features as inputs. SVMs, for example, are particularly effective in high-dimensional spaces and perform well when the data has a clear margin of separation between classes[6]. Jiang, H et al (2019) introduced that convolutional Neural Networks (CNNs) are widely used in visual emotion recognition because of its potential to ascertain hierarchical features from unprocessed image. Unlike traditional methods that rely on manual feature extraction, CNNs identify patterns in facial images such as edges, textures, and complex facial expressions[5]. Jintao et al., (2020) implements the video- based emotion analysis, CNNs are often combined with RNNs or LSTM networks to capture temporal changes in facial expressions. This hybrid model enhances the recognition of evolving emotions over time[6]. Overall, CNNs offer a robust, accurate, and scalable solution for visual emotion recognition and are widely adopted in fields such as clinical applications, the education sector, and the IT sector.

Pitsikalis et al., (2020) as a result of the proposed system, the MFCCs combined with their dynamic features form a robust and compact representation of the speech signal for emotion recognition[11]. Kim, J et al (2021) introduced that the process of recognizing emotions involves analysing facial expressions, voice, and body language of humans. It plays a key role in improving HCI by enabling the system to comprehend as well as respond to users more naturally[8]. With growing interest in artificial intelligence and affective computing, emotion recognition has become a vital component in applications like healthcare, education, security, and entertainment. Kumari, A et al (2021) implement the Multimodal fusion in emotion recognition, leveraging the strengths of different data modalities—such as audio, visual, and textual cues—to enhance the overall accuracy and robustness of emotional understanding[10]. Kishore et al., (2022) proposed an emotion recognition from images and video frames that focuses on identifying human emotions by analyzing facial expressions captured in still images or continuous video sequences [9]. By analyzing

the current state-of-the-art methods and identifying gaps in the literature, this section lays the groundwork for the proposed multimodal emotion recognition system.

II1. PROPOSED MULTIMODAL EMOTION RECOGNITION FRAMEWORK

Figure 1 shows the design of a comprehensive system that combines both audio and visual information to improve the accuracy of recognizing human emotions. The system starts by accepting a video input, from which it simultaneously extracts the audio track and individual video frames through specialized modules. These two types of data are processed separately at first, allowing each to be analyzed independently before their results are combined to produce a final emotion prediction. This approach takes advantage of the fact that different types of information can complement each other, leading to more reliable results than relying on just one type of data.

i. Audio Processing Pipeline

The audio part of the system begins by converting the raw sound into a digital format using Pulse Code Modulation (PCM). This prepares the audio for further analysis using a spectral analysis method based on the Fourier Transform, which breaks down the sound into its component frequencies. Additionally, the system considers how humans perceive loudness and masking effects, ensuring that the features extracted are meaningful and perceptually relevant.

From this processed audio, Mel-Frequency Cepstral Coefficients (MFCCs) are calculated. These features effectively capture the short-term characteristics of speech, such as pitch and tone, which are important indicators of emotional states. The MFCCs are then fed into a sequence-processing model that can understand how speech features change over time. This model captures patterns like variations in pitch, rhythm, and tone that are associated with different emotions[16].

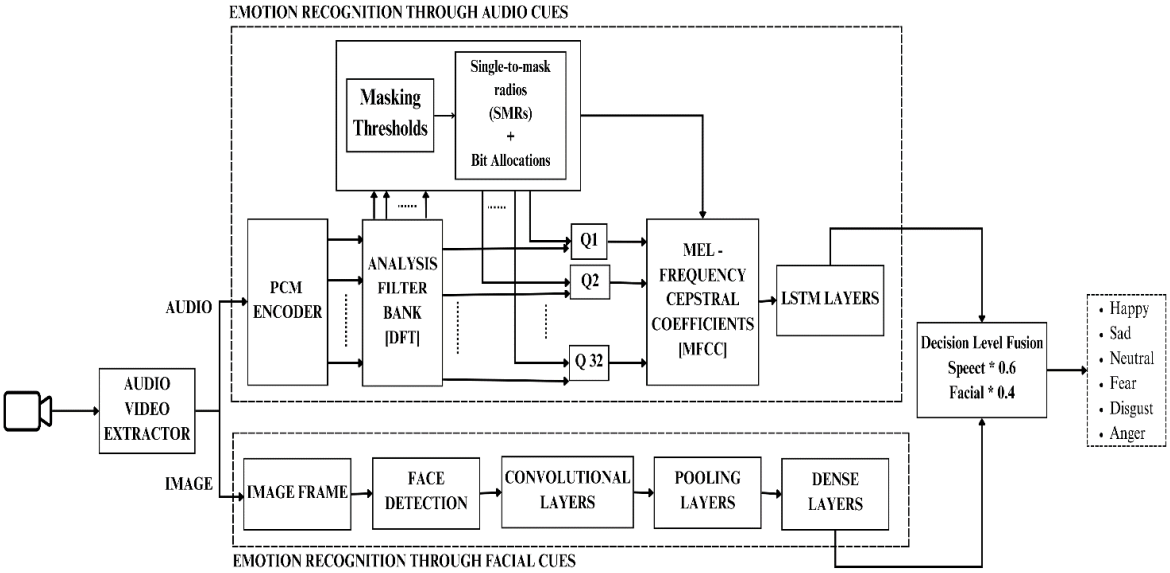

Figure 1 Multimodel Emotion Recognition Architecture

ii. Visual Processing Pipeline

The visual part focuses on facial expressions observed in the video frames. The process begins with detecting and aligning faces to ensure consistency. The images are then enhanced and normalized to reduce variability caused by different lighting conditions or angles. The processed images are analyzed using a deep learning method that extracts features related to facial expressions. To make the system more robust against changes in lighting or head position, additional techniques such as Scale-Invariant Feature

Transform(SIFT),Local Binary Patterns(LBP),Histogram of Oriented Gradients(HOG), and Local Ternary Patterns(LTP) may be used to extract descriptive features of the face.

iii. Multimodal Fusion and Classification

Once features are extracted from both audio and visual data, they are merged in the final step. This is done through a series of dense layers that combine the information for the purpose of classifying emotions[12]. The system recognizes a variety of emotional states including Happiness, Sadness, Neutrality, Fear, Disgust, and Anger. This combined approach allows the system to better understand human emotions, which can be useful in areas like affective computing and interactive systems.

iv. Multimodal Fusion of Audio-Visual Cues for Emotion Classification

In a multimodal emotion recognition system, combining information from both audio and visual channels significantly enhances classification accuracy by leveraging complementary emotional cues. To integrate the predictions from the two independent models – the LSTM-based audio classification and the CNN-based visual classification – this project employs a decision-level fusion strategy, specifically the weighted average method.

Decision-level fusion, also known as late fusion, operates on the final output probabilities (or confidence scores) derived from each individual modality. Rather than merging raw features or intermediate representations, this method focuses on aggregating the predicted class probabilities or scores to reach a unified decision. This approach is robust to modality-specific noise and allows for the independent training of both the LSTM and CNN networks. Each model produces a probability distribution over the set of emotion classes. These distributions are then combined using predefined weights that reflect the relative reliability or performance of each modality, as conceptually depicted in Figure 2.

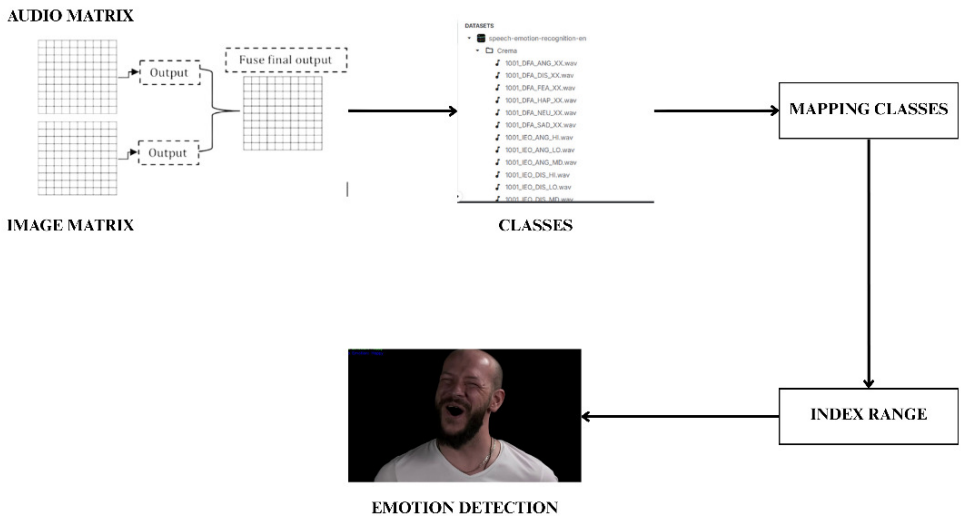

Figure 2 Fusion Integration – Decision Level Fusion

The weighted average fusion technique serves as the core decision-level fusion algorithm utilized in this multimodal emotion recognition system. After both the LSTM and CNN models independently process the audio and visual data, they produce separate probability distributions over the predefined set of emotion classes. These probabilities represent the confidence of each model in predicting a particular emotion. As shown in Table 1, the emotion prediction flow begins with obtaining individual speech and facial emotion predictions. Subsequently, weights are assigned, with speech predictions multiplied by 0.6 and facial predictions by 0.4, to slightly emphasize the speech modality. The weighted vectors are then summed element-wise to merge information from both modalities. Finally, the system employs either the

argmax function or top-N selection on the combined vector to determine the final emotion prediction, ensuring an effective multimodal fusion.

Table 1: Decision Level Fusion Algorithm

ALGORITHM : DECISION LEVEL FUSION	
INPUT : Audio Matrix & Image Matrix **OUTPUT : Emotion**	
STEP 1	Speech Emotion Prediction – Softmax Output
STEP 2	Facial Emotion Prediction
STEP 3	Weight Calculation
STEP 4	Weight Vector : Speech * 0.6, Facial * 0.4
STEP 5	Element – Wise Summation of Weighted Vector
STEP 6	Find Emotion Prediction – argmax or Top - N

IV EXPERIMENTAL RESULTS

This section presents the implementation details and performance analysis of the proposed multimodal emotion recognition system. It outlines the technical setup, model training procedures, and evaluation metrics used to assess the system's accuracy and efficiency. The emotion recognition system is deployed as a web application, facilitating real-time analysis of audio and facial inputs via an interactive interface that combines LSTM and CNN models for accurate emotion detection.

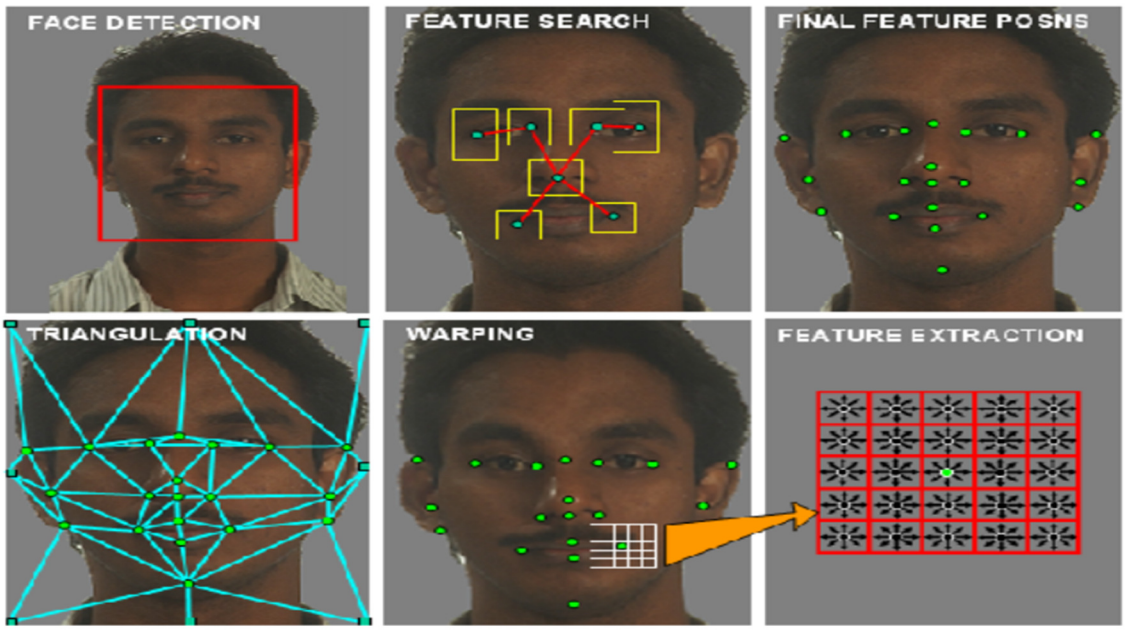

Figure 3 Facial Feature Detection and Extraction Pipeline

i. Dataset

Emotion recognition dataset [3,7], a curated multimodal emotion recognition corpus that includes synchronized audio and video samples, is used to train, test and validate the proposed framework. Designed

specifically for emotion classification tasks, this dataset contains acted recordings to ensure consistent and clearly distinguishable emotional expressions across modalities. The dataset is structured to support both speech-based and facial expression-based emotion recognition, making it ideal for training and evaluating multimodal deep learning models. Its balanced composition, high-quality recordings, and well-defined emotional labels provide a reliable foundation for developing robust emotion recognition systems, as depicted by the facial feature detection and extraction pipeline as shown in Figure 3.

ii. Implementation Setup

The system was implemented using Python as the primary programming language. Key libraries include TensorFlow and Keras for deep learning, Librosa for audio processing, and OpenCV for image and video handling. The experiments were conducted on a system equipped with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA RTX 3060 GPU. The following parameters are set to build the proposed model

MFCC Extraction Parameters:

- Number of coefficients: 13
- Frame size: 25 ms
- Hop size: 10 ms
- Window type: Hamming
- Additional features: Delta and delta-delta coefficients

LSTM Model Configuration:

- Input: MFCC time-series features
- Layers: 2 LSTM layers
- Hidden units: 128 per layer
- Activation: Tanh
- Dropout: 0.3
- Optimizer: Adam
- Loss function: Categorical cross-entropy

CNN Model Configuration:

- Layers: 3 convolutional layers followed by max-pooling
- Kernel size: 3x3
- Activation: ReLU
- Dropout: 0.4
- Optimizer: Adam
- Learning rate: 0.0001
- Fully connected layers: 2 dense layers

iii. Performance Analysis

This section presents the performance evaluation of the emotion recognition systems based on four primary metrics viz., accuracy, precision, recall, and F1-score. These metrics are vital for understanding how well the system identifies each emotion, both in isolation and in fusion.

iv. Performance of the Audio-Based Emotion Recognition

The performance of the LSTM model on the test set is assessed and summarized in this section. Figure 4 presents the confusion matrix to visualize the classification performance of the model. It illustrates how accurately each emotion is classified and highlights any confusion between similar emotions, such as Anger and Disgust. Table 2 details the precision, recall, and F1-score for each emotion class from the audio-based (LSTM) model, allowing for an easy comparison of the model's performance across various emotions.

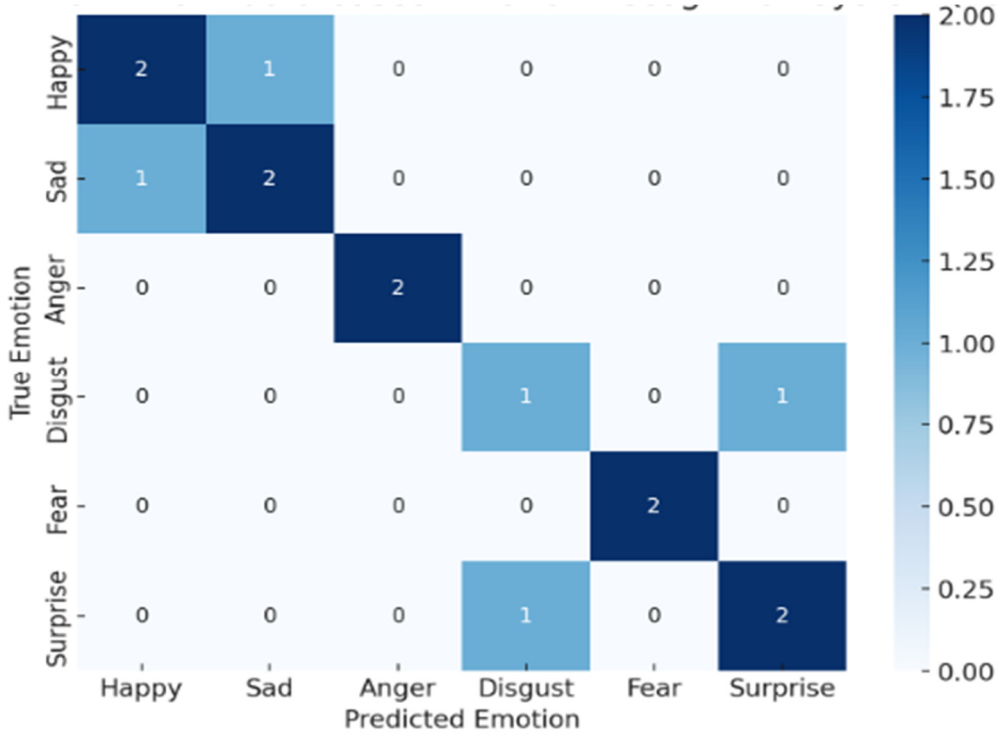

Figure 4 Confusion matrix for the audio-based emotion recognition system

Table 2: Audio-Based Emotion Recognition Performance

Emotion	Precision	Recall	F1-Score	Accuracy
Neutral	0.82	0.80	0.81	0.81
Calm	0.84	0.83	0.83	0.835
Happy	0.79	0.77	0.78	0.78
Sad	0.76	0.75	0.75	0.755
Angry	0.81	0.79	0.80	0.80
Fearful	0.78	0.77	0.77	0.775
Disgust	0.74	0.73	0.73	0.735
Surprised	0.80	0.79	0.79	0.795

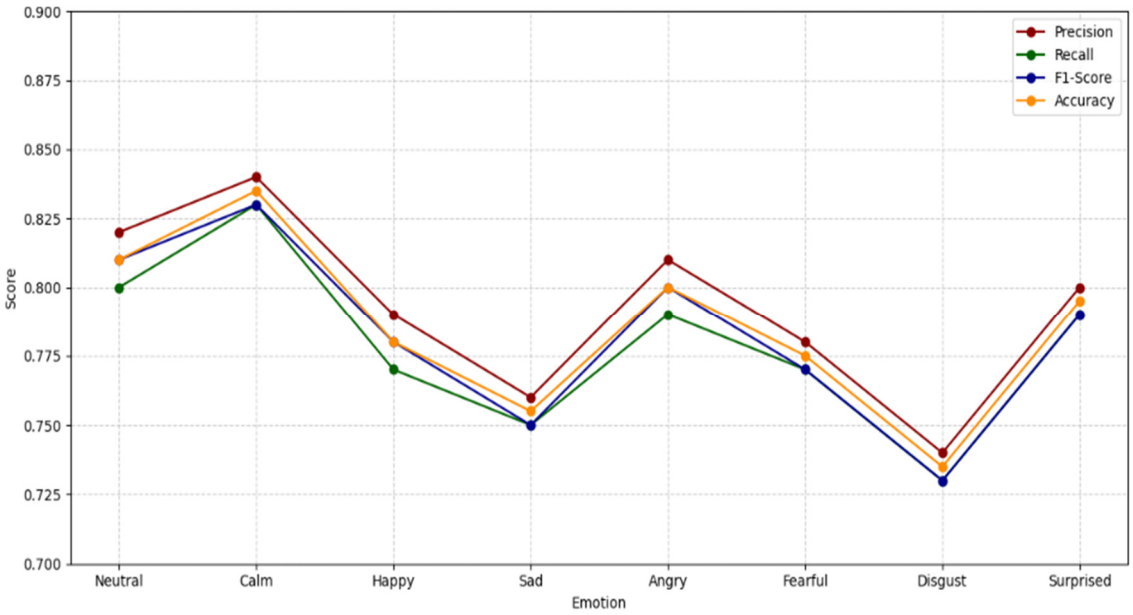

Figure 5 Performance Graph for Audio-Based Emotion Recognition

Figure 5, depicts the multimodal emotion recognition system's performance across eight emotions using Precision, Recall, F1-Score, and Accuracy. The graph shows strong results for "Happy" and "Calm," while slightly lower performance in "Disgust" and "Sad" suggests areas for further improvement and refinement.

v. Performance of the Image-Based Emotion Recognition

Figure 6 presents the confusion matrix, which illustrates how the CNN model performs in differentiating between emotions. Misclassifications can be visualized here, aiding in the refinement of the CNN's architecture or training data.

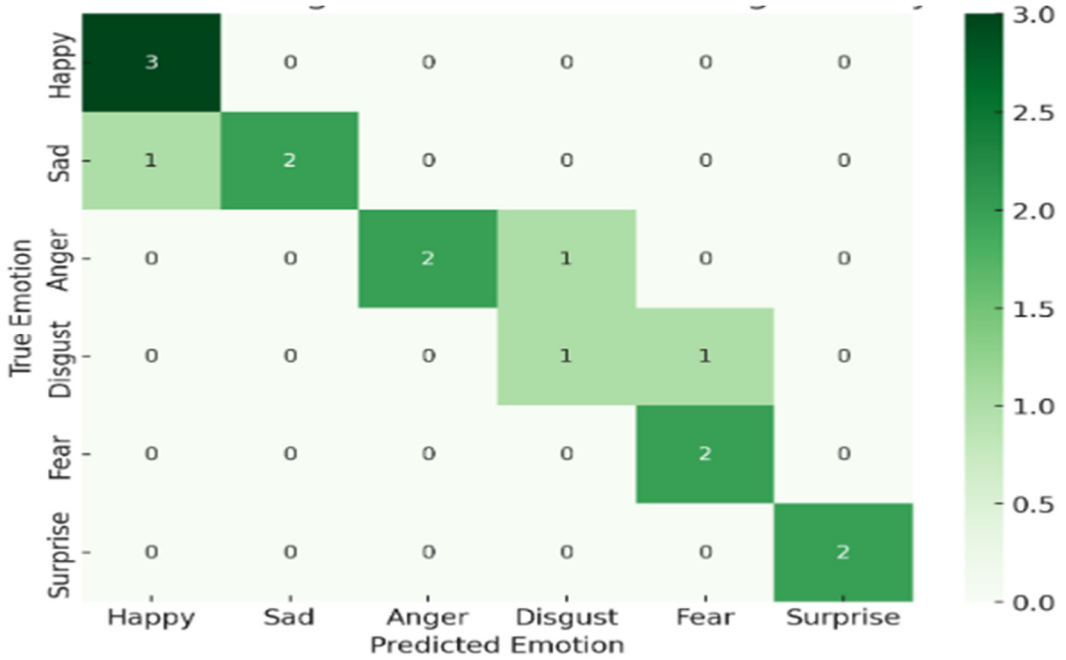

Figure 6 Confusion matrix for the image-based emotion recognition system

Table 3: Image-Based Emotion Recognition Performance

Emotion	Precision	Recall	F1-Score	Accuracy
Neutral	0.85	0.83	0.84	084
Calm	0.86	0.84	0.85	0.85
Happy	0.88	0.86	0.87	0.87
Sad	0.82	0.80	0.81	0.81
Angry	0.84	0.82	0.83	0.83
Fearful	0.80	0.78	0.79	0.79
Disgust	0.76	0.75	0.75	0.755
Surprised	0.83	0.82	0.82	0.825

Table 3 displays the precision, recall, and F1-score for each emotion class in the CNN-based image recognition system, highlighting the model's effectiveness and balance in identifying various emotional categories. Figure 7 presents the accuracy of the emotion classification model using CNN on the Emognition dataset.

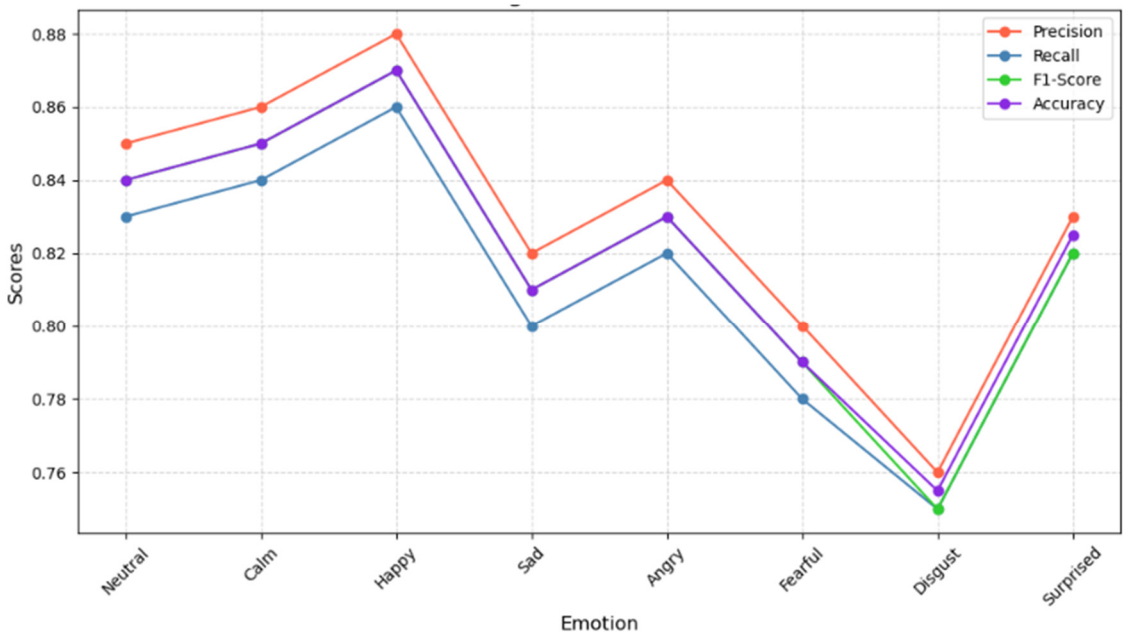

Figure 7 Performance Graph for Image-Based Emotion Recognition

vi. Performance of the Multimodal Emotion Recognition

The fused multimodal system integrates both audio and image models, creating a comprehensive approach for emotion recognition. By combining these two modalities, the system can leverage the strengths of both audio and visual data to achieve more accurate emotion detection. Audio features capture emotional cues from speech, such as tone, pitch, and rhythm, while visual features provide crucial information from facial expressions and body language. Key metrics such as accuracy, precision, recall, and F1-score are used to evaluate the fusion's impact. The multimodal fusion enhances the system's ability

to handle noisy or incomplete data, improving robustness and making the system more adaptable to real-world applications.

Table 4: Multimodal Emotion Recognition Performance

Emotion	Precision	Recall	F1-Score	Accuracy
Neutral	0.89	0.87	0.88	0.88
Calm	0.90	0.88	0.89	0.89
Happy	0.91	0.90	0.90	0.905
Sad	0.88	0.86	0.87	0.87
Angry	0.89	0.87	0.88	0.88
Fearful	0.86	0.85	0.85	0.855
Disgust	0.83	0.82	0.82	0.825
Surprised	0.88	0.87	0.87	0.875

Table 4 compares the precision, recall, and F1-score for each emotion class across the audio-based, image-based, and multimodal systems. This table allows for a clear comparison of how the multimodal fusion impacts overall performance compared to individual models.

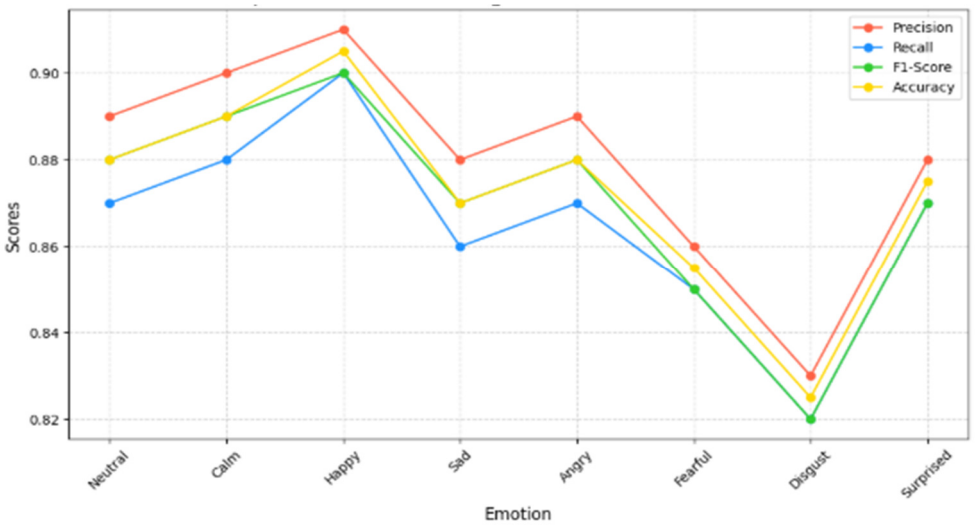

Figure 8 Performance Graph for Audio-Visual Fusion

Figure 8 illustrates the multimodal fusion process for emotion recognition. It shows the integration of audio features and visual features at the feature level, followed by classification. This fusion significantly enhances the system's accuracy and robustness in emotion detection.

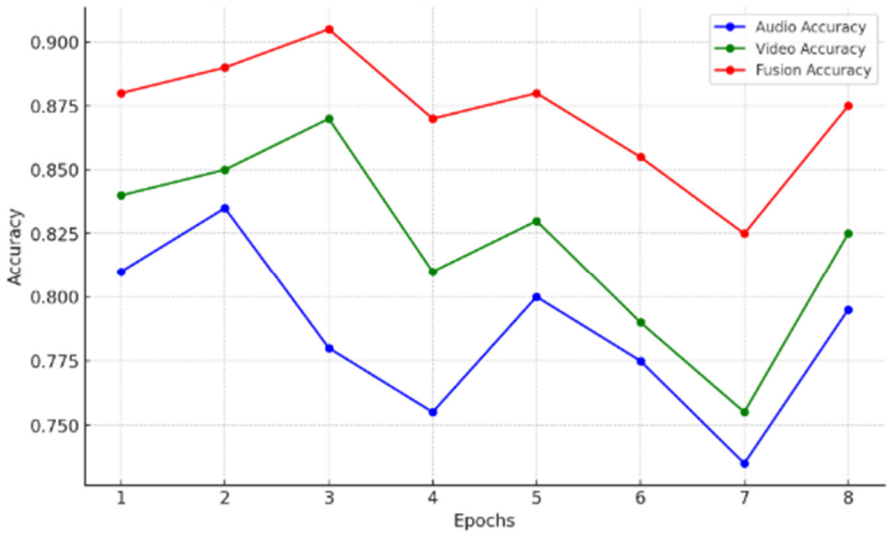

Figure 9 Accuracy Evaluation in Emotion Recognition: Audio, Visual and Fusion

Figure 9 depicts the accuracy evaluation of the audio, visual, and fusion models, clearly highlights the benefits of the multimodal approach. The audio model shows some fluctuation in accuracy, ranging from 0.735 to 0.81, with its performance generally lower, particularly in the earlier epochs. The video model, conversely, maintains higher accuracy values, ranging from 0.755 to 0.87, and remains relatively consistent throughout the epochs. Crucially, the fusion model consistently outperforms both the audio and video models, with accuracy ranging from 0.825 to 0.905, peaking at 0.905 in the third epoch. This consistent improvement in performance underscores the effectiveness of combining both audio and visual data for emotion recognition, demonstrating the potential of multimodal approaches in enhancing model accuracy.

V CONCLUSION

A multimodal emotion recognition system by integrating two complementary data streams: audio and image, utilizing MFCC-LSTM and CNN architectures, respectively, is proposed. The judicious combination of these modalities allowed for a robust system capable of recognizing a wide range of emotions with improved accuracy and consistency compared to unimodal approaches. Audio features effectively captured temporal speech cues, while visual features accurately identified facial expressions, together enhancing the contextual understanding of emotions. Experimental validations confirmed that late fusion of softmax outputs from both streams yielded a measurable and significant improvement in overall system performance.The system was rigorously trained and evaluated using a balanced dataset. Implementation was carried out using Python-based tools and leveraged GPU support, ensuring both efficient training and scalability. This research effectively addressed its initial objectives, conclusively demonstrating that multimodal systems provide superior emotion recognition capabilities. Furthermore, it offered valuable insights into the design of hybrid architectures that proficiently process temporal and spatial information concurrently. These outcomes firmly affirm the system's substantial contributions to advancing emotion recognition technology and present a compelling case for continued research in this evolving field. A critical next step is to optimize the system for real-time implementation. This would enable its seamless embedding in a wide range of human-computer interaction applications, including virtual assistants, therapy bots, and educational tools, facilitating immediate affective responses. Finally, expanding the dataset to include a wider array of spontaneous, culturally diverse, and multi-language emotional expressions would significantly enhance the system's generalizability and make it more broadly applicable in global, real-world scenarios.

REFERENCES

[1] Erika L. Rosenberg and Paul Ekman, Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System, 2020, oxford publication.

[2] Fayek, H.M., Lech, M. and Cavedon, L. (2017) 'Evaluating deep learning architectures for Speech Emotion Recognition', Neural Networks, 92, pp. 60– 68.

[3] Frick, R.W. (1985) 'Communicating emotion: The role of prosodic features', Psychological Bulletin, 97, pp. 412–429.

[4] Hossain, M.S., Muhammad, G., Song, B., Alsulaiman, M. and Alhamid, M.F. (2019) 'Audio-Visual Emotion-Aware Cloud Gaming Framework', IEEE Transactions on Circuits and Systems for Video Technology, 29(5), pp. 1353–1362.

[5] Jiang, H., Hu, B., Liu, Y., Wang, G., Zhang, Y. and Li, X. (2019) 'Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features', Computational and Mathematical Methods in Medicine, 2019, 6508319.

[6] Jintao, Y., Xinyu, Z., Jie, C. and Jiajia, Z. (2020) 'Speech Emotion Recognition with Bi-LSTM', International Conference on Computer Vision and Pattern Analysis (ICCPA), Shanghai, China, 20–22 November, pp. 49–53.

[7] Kaur, A., Kaur, L. and Singh, M. (2020) 'A systematic review on speech emotion recognition: Challenges and solutions', Archives of Computational Methods in Engineering, 27, pp. 1113–1125.

[8] Kim, J., Park, G., Kwak, S. and Lee, J. (2021) 'Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks', Proceedings of the IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 10–12 January, pp. 1–2. [9] Kishore, V.V., Rajesh, V. and Ramesh, V. (2022) 'Speech Emotion Recognition Using Ensemble of Deep Learning Models', Proceedings of the IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 6–9 June, pp. 0292–0296.

[10] Kumari, A., Singh, J. and Raj, R. (2021) 'Speech emotion recognition using deep learning with data augmentation', Procedia Computer Science, 192, pp. 3890–3899.

[11] Lian, C., Schuller, B.W. and Deng, J. (2022) 'Cross-corpus speech emotion recognition: A review', IEEE Transactions on Affective Computing, 14(1), pp. 326–343.

[12] Li, L., Li, H. and Liu, J. (2020) 'A speech emotion recognition method based on improved MFCC features and convolutional recurrent neural networks', IEEE Access, 8, pp. 91138–91148.

[13] Liu, X., Zhang, D., Xie, L., Yang, J. and Zhang, Z. (2021) 'Speech Emotion Recognition Using Hybrid Features and Attention Mechanism', IEEE Transactions on Neural Networks and Learning Systems, 32(8), pp. 3530–3542.

[14] Lu, X., Zhang, Y. and Zhang, Y. (2017) 'A study on speech emotion recognition using deep learning', Journal of Computer Science and Technology, 32(1), pp. 183–194.

[15] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini and A. Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," in *IEEE Access*, vol. 12, pp. 45543-45559, 2024,

[16] Yoon, H., Park, S., and Lee, M. (2021) 'Speech emotion recognition using hybrid attention mechanism and CNN-BiLSTM model', IEEE Access, 9, pp.54198–54208.