Multi-Modal Emotion Recognition from Video Using Audio and Visual Fusion

Satvik V
Computer Science, Engineering
Vellore Institute of Technology
Chennai, Tamilnadu, India

Rohit M
Computer Science, Engineering
Vellore Institute of Technology
Chennai, Tamilnadu, India

Deepesh Raj A Y
Computer Science, Engineering
Vellore Institute of Technology
Chennai, Tamilnadu, India

Jothi R
Computer Science, Engineering
Vellore Institute of Technology
Chennai, Tamilnadu, India

***Abstract*—Emotion recognition plays a crucial role in enhancing human-computer interaction and applications like mental health monitoring and personalized experiences. While previous studies have explored emotion recognition using either visual or audio data independently, this research integrates both visual (facial expressions) and audio (speech) modalities to improve classification accuracy. The study involves two models: a CNN-based model for visual emotion classification and a CNN + LSTM model for audio emotion classification. The visual model achieved an accuracy of 96.88%, demonstrating the power of CNNs in facial expression analysis, while the audio model attained an accuracy of 96.82 %, showing the effectiveness of combining CNN and LSTM for capturing speech-related features. The findings highlight the benefits of multi-modal emotion recognition, suggesting that combining visual and auditory cues can lead to more accurate emotion detection systems with wide applications in human-computer interaction, mental health, and adaptive learning.**

***Index Terms*—Emotion Recognition, Visual Classification, Audio Classification, CNN, LSTM, MFCC, Multi-modal, Deep Learning.**

I. Introduction

Emotion recognition plays a critical role in enhancing human-computer interaction and has the potential to revolutionize a variety of applications, including mental health monitoring, adaptive learning systems, and assistive technologies. The ability to accurately recognize emotions allows machines to respond more empathetically to human needs, enabling personalized experiences. While previous research has focused on emotion recognition from either visual cues (facial expressions) or audio cues (speech), integrating these modalities for emotion detection remains an ongoing challenge. Most systems to date have treated visual and audio emotion recognition as separate tasks, leading to limitations in the overall accuracy and understanding of emotional states.

The research gap lies in the fact that current emotion recognition models typically focus on one modality at a time. While facial expressions provide a wealth of emotional information, they are often incomplete without the context provided by speech. Similarly, emotional speech, captured through features like MFCCs, often lacks the detailed emotional cues conveyed by facial expressions. Thus, effective emotion recognition requires better exploration of each modality separately.

In this study, we aim to address this gap by building and evaluating two separate models: one for visual emotion recognition using CNNs to classify emotions based on facial expressions, and another for audio emotion recognition using MFCC features processed through a combination of CNN + LSTM networks. The models are tasked with classifying eight emotional states: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. The RAVDESS dataset, which contains audio-visual recordings of actors performing various emotional expressions, is used for training and evaluating both models. This approach highlights the importance of both visual and audio cues in emotion recognition, as shown in Figure 1, where various facial expressions represent different emotions. These visual cues are vital in understanding how emotions are communicated beyond just speech.

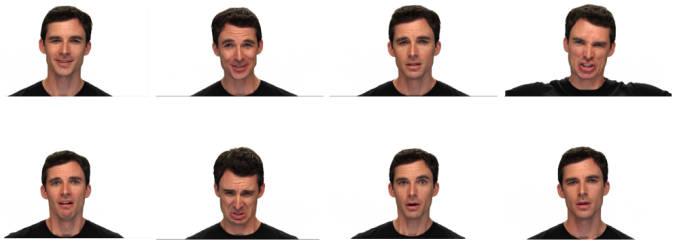

Fig. 1. Multiple emotions

By addressing the challenge of recognizing emotions from visual and audio data separately, this study aims to improve the accuracy of emotion detection systems. These advancements will lead to more reliable models for emotion recognition, capable of adapting to real-world applications across various industries. This paper is organized as follows: Related Works discusses previous research in emotion recognition, focusing on the use of visual and audio modalities. The Methodology section outlines the architecture of the visual and audio mod-

els, describing how each modality is processed. The Results section presents the evaluation of both models, followed by a discussion of the performance and potential applications.

II. Related works

Multimodal Emotion Recognition (MER) is a growing field in affective computing that seeks to identify human emotions by combining data from multiple sources such as speech, facial expressions, and text. By integrating complementary signals across modalities, MER systems offer improved accuracy and robustness compared to unimodal approaches. Recent deep learning models, particularly those using attention mechanisms, graph structures, and cross-modal transformers, have advanced the ability to capture inter-modality dependencies and contextual cues. This literature survey reviews recent MER approaches, categorizing them by methodology and highlighting their techniques, datasets, and performance outcomes.

A. Fusion Strategy-Based Approaches

Qian Ying [1] proposed a multimodal emotion recognition model based on a deep neural network that integrates facial expressions, voice, and text. The model combines CNN for facial emotion recognition, LSTM for speech emotion processing, and BERT for text sentiment analysis. The data was collected from classroom interactions, psychological counseling dialogues, and online learning scenarios. The proposed multimodal fusion model achieved an accuracy of 91.2%, surpassing the individual modalities: 82.3% for facial expressions, 78.5% for voice, and 80.1% for text. The F1 score for the multimodal model was 0.89, demonstrating superior performance, especially in negative emotion recognition, and highlighting the model's robustness in complex emotional contexts.

Varun Vishwanatha Avabratha et al. [2] proposed a multimodal emotion recognition model that combines speech emotion recognition (SER) using MFCC features and facial emotion recognition (FER) using facial coordinates. The speech model uses a network to extract spectral and temporal features, achieving a weighted accuracy (WA) of 71.36%, while the facial model employs a Random Forest classifier after preprocessing facial coordinates, achieving WA of 49.31%. The final multimodal model integrates both SER and FER using decision-level fusion, achieving a WA of 72.9%, demonstrating improved accuracy over unimodal approaches. The model was trained and evaluated on the IEMOCAP dataset. The multimodal approach outperformed the individual models, highlighting the effectiveness of combining both speech and facial expression features for emotion recognition.

Xinxue Du et al. [4] proposed a multimodal emotion recognition model based on feature fusion and residual connection, which utilizes Bi-LSTM and multi-head attention to extract features from voice, text, and video. The model incorporates residual connections to prevent gradient disappearance, improving feature extraction from multiple modalities. The IEMOCAP dataset was used for evaluation, and the proposed model achieved an average accuracy of 61.4% and an F1 score of 61.7%. The results demonstrated better emotion classification performance compared to other single-modal models and baseline methods.

B. Graph-Based and Attention Fusion Approaches

Liang Zhang et al. [7] proposed a multimodal emotion recognition framework called Multi-Modal Graph Attention Network (MM-GAT), which uses graph-based feature representation and attention fusion to capture inter-modality relationships. The model incorporates audio, text, and visual modalities by constructing modality-specific graphs and fusing them via attention mechanisms. The CMU-MOSEI dataset was used for evaluation, achieving an overall accuracy of 83.5% and an F1-score of 82.1%, outperforming conventional fusion strategies like early and late fusion. The graph attention method demonstrated improved interpretability and robustness in capturing emotional cues.

The paper titled Multimodal Emotion Recognition Algorithm Based on Graph Attention Network [10]addresses the challenges and advancements in the field of multimodal emotion recognition, a critical area within affective computing. This domain focuses on interpreting human emotions through various modalities such as language, video, and audio. While prior methods have shown progress, they often struggle with the heterogeneity of multimodal data, which complicates effective feature integration across modalities. A key challenge is determining the relative importance of each modality, as uneven contributions can hinder accurate emotion recognition. To overcome these issues, the authors propose a novel approach called the Graph Attention Distillation Network (GADN), which employs three independent pre-trained models alongside one-dimensional temporal convolutional layers to process different modalities and effectively aggregate temporal and low-level multimodal features. The paper introduces a multimodal interaction gating module to enhance inter- and intra-modal interactions, reducing dependence on a single modality and addressing limitations found in earlier work. Additionally, the graph adaptive distillation module preserves the unique characteristics of each modality during feature fusion by learning distillation weights automatically, thus enabling effective cross-modal knowledge transfer and representing a substantial improvement over conventional fusion methods.

C. Transformer and Cross-Modal Attention Approaches

Minwoo Lee et al. [8] introduced a Hierarchical Cross-Modal Transformer (HCMT) model that performs multi-level fusion of audio, visual, and text data for emotion recognition. The architecture uses cross-modal attention at both feature and decision levels, improving temporal synchronization and contextual understanding across modalities. Evaluated on the IEMOCAP and CMU-MOSEI datasets, HCMT achieved 84.7% accuracy and 83.9% weighted average F1-score, marking a significant improvement over previous benchmarks. The hierarchical fusion and transformer-based attention mechanism helped capture long-range dependencies across modalities effectively.

Zhengdao Zhao et al. [5] proposed a transformer-based TDFNet for multimodal emotion recognition, integrating speech and text features to improve accuracy. The model incorporates pretrained embeddings, deep-scale transformer for fine-grained feature extraction, and mutual transformer to capture mutual correlations between speech and text. The IEMOCAP dataset was used for evaluation, and TDFNet achieved an 82.08% WA and 82.57% UA in the RA splitting, outperforming previous methods by 1.78% WA and 1.17% UA. The model's performance was notably improved by the deep-scale fusion and speaker-related mutual correlations.

Xiaoyu Chen et al. [6] introduced a novel Cross-Modal Attention Fusion Network (CMAFN) that focuses on effectively leveraging the interactions between audio, visual, and textual modalities for emotion recognition. Unlike conventional early or late fusion methods, CMAFN dynamically learns inter-modality dependencies by assigning attention weights based on contextual relevance. The model uses CNNs for visual feature extraction, Bi-LSTM with attention for audio sequences, and BERT for textual embeddings. Experiments were conducted on the CMU-MOSEI and IEMOCAP datasets, achieving an overall accuracy of 84.6% and an F1 score of 83.2% on CMU-MOSEI. The model significantly outperformed baseline fusion strategies, demonstrating the effectiveness of cross-modal attention in capturing nuanced emotional cues from heterogeneous data sources.

D. Biometric and Modality Expansion Approaches

Gaoyuan Qin et al. [3] proposed a multimodal emotion recognition model using a 3D-CNN deep learning model for EEG signal recognition, face recognition, and fusion recognition. The model aims to enhance the robustness and accuracy of emotion recognition by incorporating transfer learning and data augmentation. The fusion method combined EEG and facial expression data, leading to a recognition accuracy of 96.79%. The model was evaluated using the DEAP dataset, and the results demonstrated significant improvements in performance through multimodal fusion, achieving higher accuracy compared to single modality models.

The paper Multimodal Biometrics Recognition from Facial Video via Deep Learning [9] explores a biometric recognition system that integrates facial features and ear recognition to improve identification accuracy, particularly under variable conditions. The authors employ supervised denoising auto-encoders to automatically extract robust and non-redundant features from facial video clips, enhancing recognition effectiveness. Gabor filters are used for feature extraction due to their ability to capture scale- and rotation-invariant characteristics. The model addresses challenges in unconstrained environments, such as variations in head pose and facial expressions, by leveraging resilient feature extraction methods. Sparse representation classifiers are used for feature classification, though the authors note their sensitivity to imbalanced training data. To further improve recognition performance, the study implements score-level fusion to combine outputs from different modalities. Experimental evaluations on both constrained and unconstrained datasets demonstrate high recognition rates, validating the practical applicability and effectiveness of the proposed multimodal system.

The reviewed literature reflects significant advancements in multimodal emotion recognition through a variety of fusion strategies, including feature- and decision-level fusion, attention-based graph models, and transformer-based cross-modal frameworks. While fusion approaches have improved performance across datasets like IEMOCAP, CMU-MOSEI, and DEAP, challenges remain in harmonizing heterogeneous modality features, especially in real-time and dynamic environments. Despite notable progress, existing methods often struggle with effective temporal synchronization and inter-modality imbalance, particularly in audio-visual fusion contexts. Our research addresses these limitations by focusing on robust real-time audio and video fusion using CNN and LSTM, enabling adaptable emotion recognition with minimal data and improved temporal coherence—advancing current multimodal systems toward practical, low-data, real-world applications.

III. Methodology

The methodology for this emotion recognition system integrates both visual and audio data to improve accuracy in classifying emotions. As shown in Figure 2, the process starts with extracting frames from the video in the RAVDESS dataset. These frames are resized and normalized to ensure consistency. The resized frames are then passed through a CNN model to classify emotions based on facial expressions. On the audio side, the audio is extracted from the video, and MFCC features are computed and normalized. These features are then processed by a CNN + LSTM model to classify emotions based on speech. Both models independently classify emotions using visual and audio data, and their predictions are evaluated to make the final emotion classification. By combining both types of data, the system can leverage both facial expressions and speech, leading to more reliable and accurate emotion recognition.

A. Dataset Description

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [11] dataset offers a rich collection of audio-visual (AV) files, which are essential for multi-modal emotion recognition research. These files capture both the speech of actors and their corresponding facial expressions. The dataset contains 2880 audio-visual speech files, recorded by 24 actors (12 male and 12 female). Each actor performs 60 emotional trials, with each trial containing one of eight distinct emotions: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. Each emotion is presented at two levels of intensity: normal and strong, except for the neutral emotion, which only has normal intensity.

The actors in the dataset are required to recite one of two predefined sentences in each trial: "Kids are talking by the door" and "Dogs are sitting by the door." These sentences are spoken in both normal and strong emotional intensities,

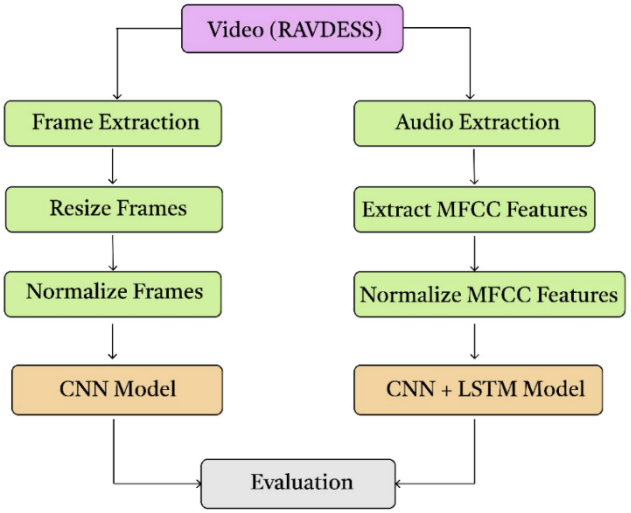

Fig. 2. Workflow

offering variability in how emotions are expressed. Furthermore, each trial is repeated twice to ensure consistency in emotional expression across multiple recordings. These audio-visual (AV) files provide valuable data for emotion detection models by combining both speech and visual cues (facial expressions) of emotions, allowing for more accurate and robust emotion recognition.

By focusing on audio-visual (AV) files, the RAVDESS dataset enables researchers to explore the integration of auditory and visual signals for emotion recognition tasks. This is particularly useful for improving models that detect emotions from both speech and facial expressions, which is crucial for building more comprehensive emotion detection systems. The combination of these modalities helps to capture the full spectrum of emotional expressions, offering a more holistic approach to recognizing emotions in real-world scenarios.

B. Visual Classification

In the data preprocessing step of visual classification, we perform frame extraction from the video files in the RAVDESS dataset to prepare the visual data for further use in emotion recognition. The main goal of this step is to extract key frames from the audio-visual (AV) video files that will be used for image classification tasks based on facial expressions.

For each actor (from Actor 01 to Actor 24), we loop through their respective video folders and extract the video files. The extraction process skips video-only (VO) files, as they do not contain the necessary audio data and are irrelevant for the current analysis. Each remaining audio-visual (AV) video file is processed to extract key frames representing the actor's emotional expressions.

To identify the emotional state in each video, we rely on the filename structure, which encodes the emotion through a specific emotion code. The emotion codes correspond to various emotions such as neutral, happy, angry, fearful, and others. Once the emotion is determined, we create an output folder structure that categorizes the extracted frames into subdirectories based on both the actor and the emotion.

Frames are extracted from the videos at a set interval determined by a frame skipping factor. In this implementation, we extract one frame every 5 frames (i.e., every 5th frame). This helps in reducing the total number of frames and ensures that the frames chosen represent significant points throughout the video. Figure 3 shows different frames extracted from a particular video. Each extracted frame is saved as a JPEG image with a filename that includes both the original video filename and the frame number, ensuring unique and traceable frame identifiers.

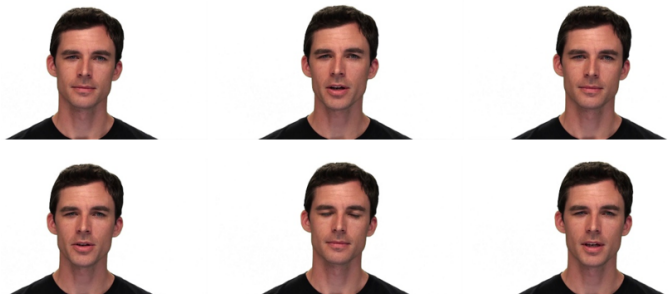

Fig. 3. Frames captured from a video

In the next step of the preprocessing, we reorganize the extracted frames into a new, emotion-based directory structure to facilitate easier access and analysis for emotion recognition tasks. Starting with the original frame dataset, which contains all the frames organized by actor and emotion, we create a new directory structure where each emotion has its own dedicated folder. The emotions, such as angry, calm, disgust, fearful, happy, neutral, sad, and surprised, become the top-level directories within the new dataset. Then, the program iterates over the existing actor folders and checks for the corresponding emotion folders. If the emotion folder exists, it copies the relevant frames from the original dataset into the new, emotion-specific folders, ensuring that each extracted frame is properly categorized according to its emotional expression. By the end of the process, a new, well-structured dataset is created, with each emotion represented in its own folder, making it ready for training and evaluation in image classification models.

In the next phase of the project, we proceed with preparing the dataset for training a Convolutional Neural Network (CNN) model to classify emotions based on facial expressions. First, we set a seed value to ensure reproducibility and deterministic behavior throughout the training process, which is crucial for consistent results. The dataset used in this step consists of images organized into emotion-based folders, where each folder contains facial images representing a specific emotion (e.g., angry, happy, neutral, etc.).

The images are loaded and preprocessed to ensure they are in a format suitable for model training. Each image is read using OpenCV, resized to a standard size of 64x64 pixels, and then normalized by scaling the pixel values to a range

between 0 and 1. This normalization helps the model converge faster and prevents issues related to large pixel values. For each image, the corresponding emotion label is assigned based on the folder it was found in, with a dictionary mapping each emotion to a numerical label.

The dataset is then split into training and test sets, with 80% of the data used for training the model and the remaining 20% used for testing its performance. To ensure balanced representation, the data is stratified during the split, maintaining the proportion of each emotion in both the training and test sets. The labels are then converted to one-hot encoding to prepare them for use in the neural network.

With the dataset now prepared, we have a structured set of images and labels, ready for training a CNN model to classify emotions based on facial expressions. This dataset is pivotal for the development of the emotion recognition model and sets the foundation for the subsequent steps, including model training and evaluation.

In this step, we define and train a Convolutional Neural Network (CNN) model for emotion classification based on facial expressions. The model architecture is structured using multiple layers of convolutional and pooling layers, which are designed to automatically extract features from the input images. The CNN starts with a Conv2D layer that applies 32 filters of size 3x3, followed by a MaxPooling2D layer to downsample the feature maps and reduce spatial dimensions. This process is repeated for successive layers with increasing numbers of filters (64, 128, and 256), which helps the model capture progressively more complex features of the images. The working of CNN layers can be better understood by referring to the CNN architecture diagram shown in Figure 4.

After the convolutional and pooling layers, the feature maps are flattened into a one-dimensional vector using a Flatten layer. This vector is passed through two Dense layers, which are fully connected layers responsible for learning the high-level representations from the extracted features. To prevent overfitting, Dropout layers are added after each Dense layer, randomly setting some of the neurons to zero during training to improve generalization.

The final layer is a Dense layer with softmax activation, which outputs a probability distribution across the 8 emotions being classified, corresponding to the number of classes in the dataset. The model uses categorical crossentropy as the loss function, which is suitable for multi-class classification tasks, and the Adam optimizer with a learning rate of 0.0001 to efficiently minimize the loss during training.

The model is trained for 25 epochs with a batch size of 8, allowing the network to learn from the training data while also validating its performance on the test set at the end of each epoch. The training history is captured to monitor the model's accuracy and loss during both training and validation, which will be useful for evaluating the model's performance and identifying potential areas for improvement.

C. Audio Classification

In the data preprocessing step of audio classification, the first task is to extract the audio from the audio-visual (AV) video files in the RAVDESS dataset. The goal is to isolate the audio track from each video and save it as a .wav file, which can later be used for training an audio classification model.

We define a function, extract_audio, which takes the path to a video file, extracts the audio using FFmpeg (a powerful multimedia processing tool), and saves it as a .wav file. The audio is down sampled to mono with a sample rate of 48 kHz, which is a common practice in audio processing to ensure consistency across the dataset. The FFmpeg command is run for each video file to convert the audio from the video into a .wav format, and the audio file is saved with the same name as the original video, but with a .wav extension.

The extraction of audio is performed in parallel using multi-threading with the help of concurrent.futures.ThreadPoolExecutor, which speeds up the process by handling multiple video files simultaneously. This approach significantly reduces the time required to process all the videos in the dataset. Once the audio files are extracted, the program prints a success message, indicating that the multi-threaded audio extraction is complete.

By the end of this preprocessing step, all the audio tracks from the audio-visual (AV) videos have been successfully extracted and saved as .wav files, making them ready for the next steps in training the audio classification model.

In the next phase of the audio classification pipeline, we focus on organizing and processing the audio files extracted from the RAVDESS dataset. The first task is to organize the audio files into a more structured directory format based on both actor and emotion. Each extracted audio file, initially stored in the audio_files folder, is moved into a new directory structure where the audio files are sorted by actor and emotion. This is achieved by splitting the filename to extract the emotion code and the actor ID, which are then used to map the audio file to its appropriate location in the organized_audio_files folder. The emotion mapping is done using a predefined dictionary that converts the emotion codes into human-readable emotion names such as Neutral, Happy, Sad, and so on. Once the files are sorted, they are moved into folders that represent both the actor and the emotion associated with the audio clip.

Once the audio files are properly organized, the next step is to extract audio features that are crucial for training an audio classification model. Here, we focus on Mel-frequency cepstral coefficients (MFCC), which are widely used in speech and emotion recognition tasks. To achieve this, we use the librosa library to load each audio file and compute its MFCC features. The MFCCs are a compact representation of the audio signal that captures the timbral texture, which is important for emotion recognition. For each audio file, the MFCCs are calculated and saved as a NumPy array for later use in model training.

In addition to saving the MFCCs, we also visualize them as spectrograms using matplotlib as shown in Figure 5. These visual representations of the MFCCs help in understanding the

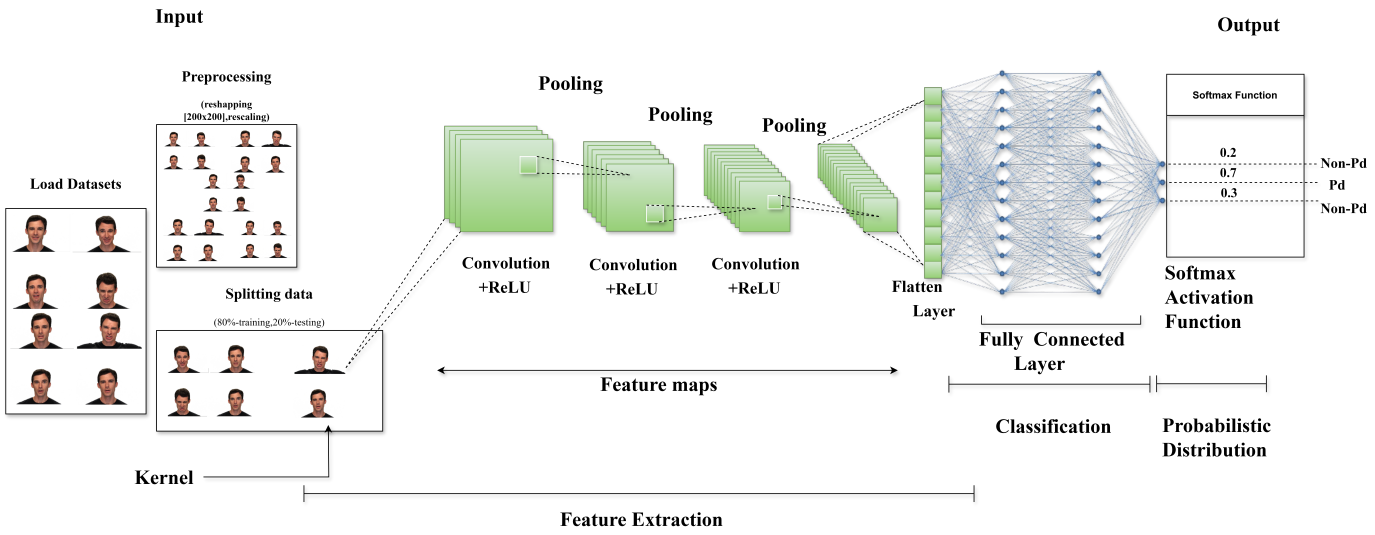

Fig. 4. CNN Architecture

underlying structure of the audio data. Each MFCC spectrogram is saved as a PNG image alongside the corresponding MFCC data. The spectrograms are useful for model training and can also be used for visual inspection of the data.

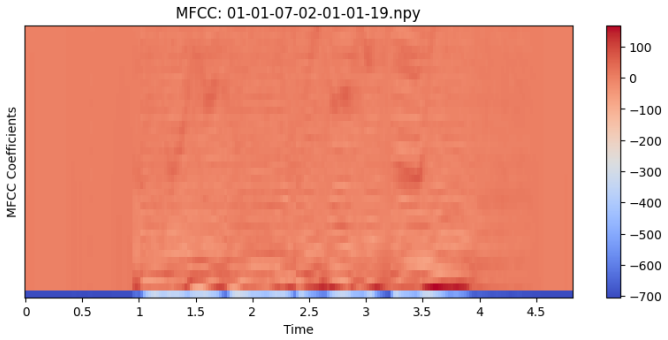

Fig. 5. Mfcc visualization

The entire process is automated through a script that iterates over all the actors and emotions, extracting the MFCCs for each audio file and saving both the MFCC data and the spectrogram images in a structured manner. The feature extraction is done efficiently, ensuring that the dataset is ready for use in training the audio classification model. Once the process is completed, the MFCC features and the spectrogram images are stored in their respective folders, making it easy to access them for further model development and evaluation. This structured organization and feature extraction form the foundation for training an effective audio classification model to recognize emotions from speech.

In the next step of the audio classification pipeline, we perform the data preprocessing required to train an audio classification model using the MFCC features extracted in the previous step. The primary task in this phase is to load the MFCC features stored in NumPy arrays and pair them with their corresponding emotion labels. These MFCC features are stored in a directory structure organized by actor and emotion, and we iterate through this structure to load each file.

For each MFCC file, we transpose the features and expand the dimensions to ensure the correct shape for feeding into the model. This transformation ensures that the MFCC data is in the required format of (time_steps, frequency_bins, channels). After loading the features, we append them to a list for later use in training the model, along with the corresponding emotion labels, which are stored as string values.

The emotion labels are then encoded into numerical values using LabelEncoder from sklearn. After encoding, the labels are converted into one-hot encoding using TensorFlow's to_categorical, as this format is required for multi-class classification tasks.

To ensure that all sequences (MFCC feature arrays) are of equal length, we pad the sequences to the maximum time step length. The padding ensures that all input data has the same shape, which is critical when feeding it into a neural network. The padding is performed in such a way that it preserves the temporal structure of the audio data by adding zero-padding to the shorter sequences.

Finally, the preprocessed data is split into training and testing sets using train_test_split from sklearn, with 15% of the data set aside for testing. This results in two datasets: one for training the model and one for evaluating its performance.

In this step, we define and implement the Audio Classifier (AC) model, which combines both Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) layers to effectively classify emotions from audio features. The model architecture leverages the strengths of CNNs for

spatial feature extraction and LSTMs for temporal sequence modeling, making it well-suited for processing MFCC features extracted from audio signals.

The Audio Classifier (AC) model begins with a Reshape layer to adjust the input dimensions, ensuring that the MFCC features are correctly formatted for the subsequent layers. The reshaped input data has the shape (time_steps, frequency_bins, 1), where time_steps represent the number of frames in the audio, and frequency_bins correspond to the MFCC coefficients.

The first part of the model consists of Convolutional Layers for feature extraction. The initial Conv2D layer uses 64 filters with a 3x3 kernel to capture low-level features from the MFCC spectrograms, followed by Batch Normalization to stabilize the learning process. Another Conv2D layer with 128 filters and a similar setup is added to capture higher-level features. MaxPooling2D is applied to reduce the spatial dimensions, and Dropout is used to prevent overfitting by randomly disabling a fraction of neurons during training.

A third Conv2D layer with 256 filters further refines the feature extraction process, followed by another MaxPooling2D and Dropout for regularization.

Once the CNN layers have extracted meaningful features, the model transitions to LSTM layers for temporal learning. LSTM layers are particularly useful for processing sequential data, such as audio signals, where the temporal dependencies across time steps are crucial for emotion recognition. The model includes three LSTM layers with 256, 128, and 64 units, respectively. The return_sequences=True parameter is used for the first two LSTM layers to pass the sequence of hidden states to the next layer, allowing the model to capture long-range dependencies. The last LSTM layer, however, outputs only the final state, as we are interested in the final emotional prediction. The working of LSTM layers can be better understood by referring to the LSTM architecture diagram shown in Figure 6, where the flow of information through the gates (forget, input, output) and memory cells is depicted.

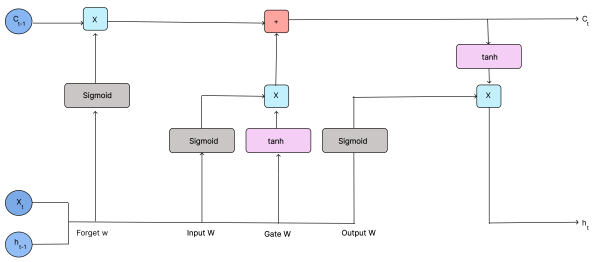

Fig. 6. LSTM Architecture

Following the LSTM layers, the model has a Dense layer with 64 units and Dropout for further regularization. Finally, the model ends with an output layer using the softmax activation function, which outputs a probability distribution across the 8 possible emotions, making it suitable for multi-class classification tasks.

The model is compiled with categorical crossentropy as the loss function, Adam as the optimizer for efficient training, and accuracy as the evaluation metric. The architecture of Audio Classifier (AC) is designed to efficiently learn both the spatial features from the MFCC spectrograms using CNNs and the temporal dependencies in the audio sequences using LSTMs, providing a robust solution for emotion recognition from audio data.

To enhance the training process, we use the ReduceLROnPlateau callback, which adjusts the learning rate when the validation loss stops improving. This callback monitors the validation loss, and if there is no improvement for 3 consecutive epochs, it reduces the learning rate by a factor of 0.5. This dynamic learning rate adjustment helps the model converge more efficiently and avoids overfitting by ensuring that the learning rate is appropriately tuned during training. The model is then trained for 50 epochs with a batch size of 32, and the learning rate schedule is applied to further optimize the training process.

D. Integration of Visual and Audio Emotion Recognition

After training the individual models for visual and audio emotion recognition, the next step is to combine the predictions from both models. The process begins by preprocessing the input video just as we did during training, extracting frames for visual processing and audio for speech processing. For each video, the visual model processes the extracted frames, while the audio model analyzes the corresponding audio. Both models output a probability distribution over the possible emotion classes. These outputs are generated using the Softmax activation function, which transforms the raw predictions into probabilities that sum up to one. This allows us to determine the likelihood of each emotion being present, making it easier to choose the model's most confident prediction.

Once we have the predictions from both the visual and audio models, the next step is to compare their confidence levels. Confidence is simply the probability value assigned to the predicted emotion by each model. For example, if the visual model predicts "Happy" with a confidence of 0.45 and the audio model predicts "Happy" with a higher confidence of 0.65, the final emotion prediction is made by comparing these confidence values. In this case, the audio model, with its higher confidence, would be the deciding factor, and the final predicted emotion would be "Happy."

This approach of combining both models ensures that the system benefits from the strengths of each modality. Visual data, such as facial expressions, offers important emotional clues, while audio data, like tone and pitch in speech, provides deeper emotional context. By integrating the results, the system becomes more reliable, particularly in situations where one model may be less accurate. For example, if the facial expression is subtle and hard to interpret, the audio model may still provide a strong indication of the emotion based on speech patterns. By selecting the emotion prediction with the highest confidence, this multi-modal system significantly

improves the accuracy and robustness of emotion recognition, making it better suited for real-world applications.

IV. Results and Discussion

To evaluate the performance of our multi-modal emotion recognition system, we consistently employed four standard classification metrics throughout this project: Accuracy, Precision, Recall, and F1-Score. These metrics are essential for assessing classification models, especially in the presence of class imbalance and noisy data.

a) Accuracy: Accuracy is a fundamental metric in classification tasks, measuring the proportion of correctly classified instances over the total number of instances. It is computed as shown in Equation (1):

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \tag{1}$$

b) Precision: Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It is defined in Equation (2):

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{2}$$

c) Recall: Recall, also known as Sensitivity or True Positive Rate, evaluates the ability of the model to detect all actual positive instances. It is calculated using Equation (3):

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{3}$$

d) F1-Score: The F1-Score is the harmonic mean of Precision and Recall. It balances the trade-off between the two, especially when class distribution is skewed. It is expressed in Equation (4):

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

These metrics are used repeatedly in evaluating both the audio and visual emotion recognition models, as well as their integrated form, to ensure consistent and comprehensive performance analysis.

A. Model Performance Evaluation-Audio

The performance of the proposed audio emotion recognition model was evaluated using training and validation loss, accuracy curves, a confusion matrix, and a precision-recall-F1 score heatmap.It achieved an accuracy of 96.82% in testing data. These visual tools provide a comprehensive understanding of the model's learning behavior, generalization ability, and class-wise performance.

Figure 10 illustrates the training and validation loss over 50 epochs. Initially, both losses decrease rapidly, indicating effective learning. As training progresses, the model continues to minimize training loss, and validation loss steadily decreases until around epoch 30, after which it plateaus. The absence of significant divergence between training and validation loss

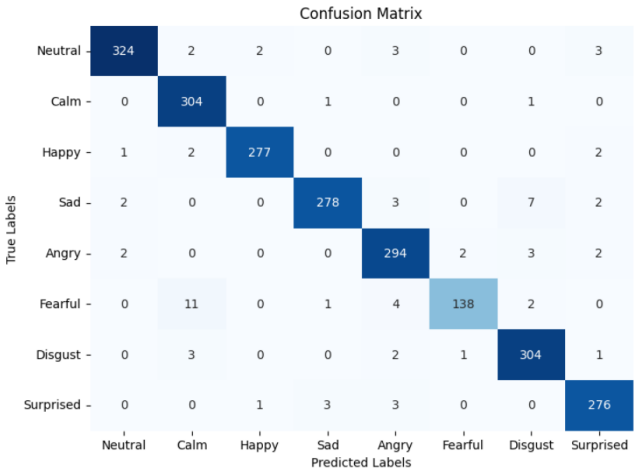

Fig. 7. Confusion Matrix-Audio

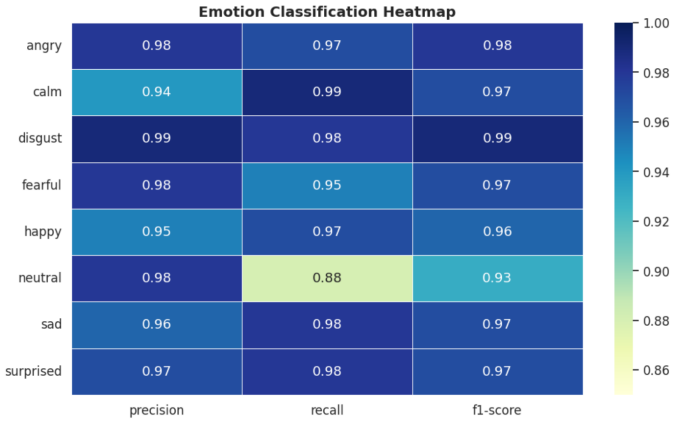

Fig. 8. Accuracy-Heatmap (Audio)

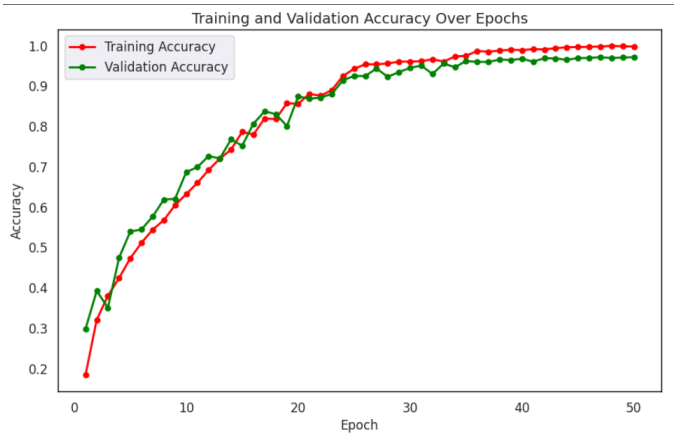

Fig. 9. Training-Validation Acuruacy Vs Epochs

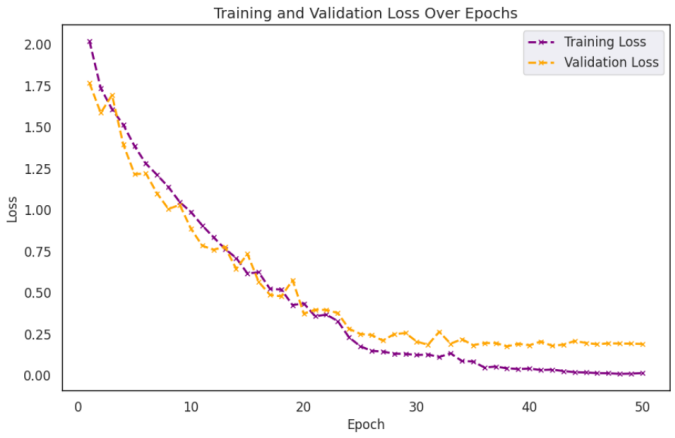

Fig. 10. Training-Validation Loss Vs Epochs

suggests that the model generalizes well and does not suffer from overfitting.

In Figure 9, the training and validation accuracy curves demonstrate a consistent improvement in classification performance. Training accuracy starts around 18% and climbs steadily to 99% by epoch 50. Validation accuracy closely follows this trend, reaching above 97%, indicating that the model performs reliably on unseen data. The narrow gap between the two curves further confirms good generalization.

To gain insights into the model's behavior on a class-by-class basis, we examined the confusion matrix shown in Figure 7. The matrix reveals high accuracy across all emotion classes, particularly for Neutral, Calm, Happy, Disgust, and Surprised, each having over 275 correctly predicted samples. Some confusion is observed for the Fearful class, which is often misclassified as Calm, Angry, or Disgust. This overlap is likely due to subtle similarities in facial expressions or vocal tones between these emotions.

Further quantitative analysis is presented in the heatmap of precision, recall, and F1-score for each emotion class (Figure 8). The model achieved an average precision, recall, and F1-score above 0.95 across most emotions. Disgust shows perfect classification with a precision and F1-score of 0.99. Calm and Happy also demonstrate strong performance, although Calm has a slightly lower precision (0.94), possibly due to confusion with Neutral or Fearful expressions. Notably, Neutral has a recall of 0.88, indicating some missed classifications, but still maintains a high F1-score of 0.93.

Overall, these evaluation metrics confirm that the audio classifier model achieves robust emotion recognition across a wide range of emotional states.

B. Model Performance Evaluation-Video

The performance of the proposed visual emotion recognition model was evaluated using training and validation loss, accuracy curves, a confusion matrix, and a precision-recall-F1 score heatmap. It achieved an accuracy of 96.88% in testing data.

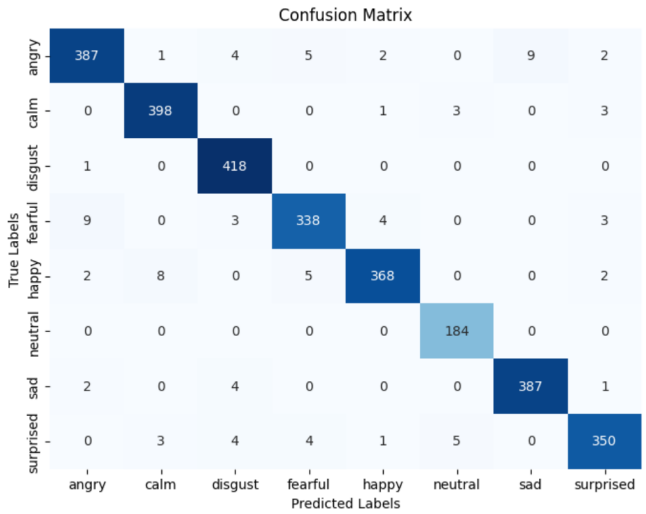

Fig. 11. Confusion Matrix-Video

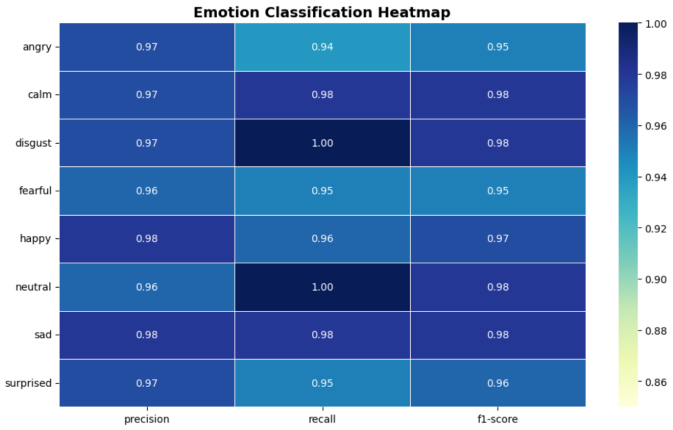

Fig. 12. Accuracy-Heatmap (Video)

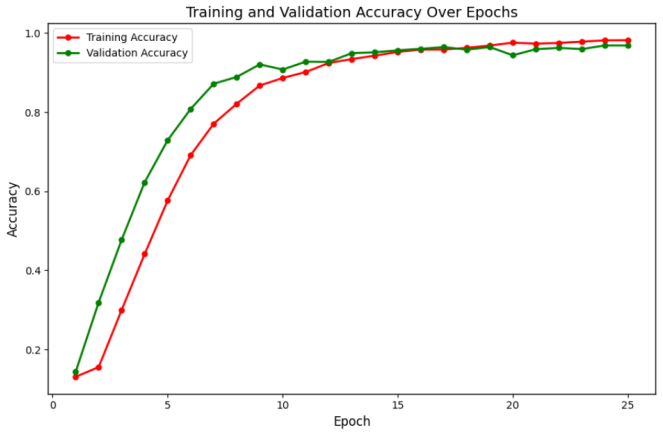

Fig. 13. Training Validation Accuracy Vs Epochs

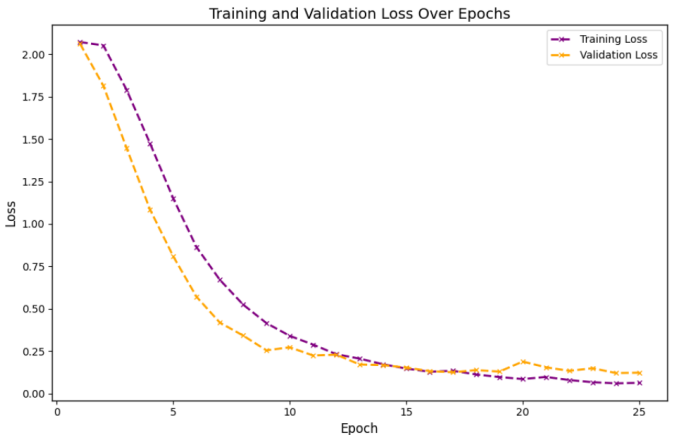

Fig. 14. Training Validation Loss Vs Epochs

Figure 14 shows the training and validation loss over 50 epochs. Both losses decrease rapidly during initial epochs, indicating effective learning. After epoch 30, the validation loss plateaus while training loss continues to decline slightly. The lack of divergence between the curves suggests minimal overfitting and good generalization.

Figure 13 presents the training and validation accuracy. Accuracy improves consistently, with training accuracy reaching 99% and validation accuracy exceeding 97% by epoch 50. The close alignment between the two curves further supports the model's ability to generalize well to unseen data.

The confusion matrix in Figure 11 provides insight into class-wise performance. The model shows strong accuracy for Neutral, Calm, Happy, Disgust, and Surprised—each with over 275 correct predictions. Some misclassifications are noted for Fearful, often confused with Calm, Angry, or Disgust. These errors may stem from overlapping features in facial expressions or speech tone.

Figure 12 presents a heatmap of precision, recall, and F1-score for each emotion. Most classes achieved scores above 0.95. Disgust yielded the highest metrics, with precision and F1-score of 0.99. Calm and Happy also performed well, though Calm showed slightly lower precision (0.94), possibly due to overlap with Neutral or Fearful. Neutral had a recall of 0.88, indicating occasional missed classifications but maintained a high F1-score (0.93).

The user interface Figure 15 of our emotion recognition system demonstrates the capabilities of our developed multi-modal model through three distinct modules: audio, visual, and combined input analysis. In the Audio Emotion Recognition section, users can initiate a voice recording by clicking the "Start Recording" button, which processes the tone, pitch, and speech patterns to predict the emotional state — for instance, identifying the emotion as "sad." The Visual Emotion Recognition component enables users to analyze emotions based on facial expressions captured via webcam, with a deep learning model detecting features and classifying the emotion, such as "neutral" in the displayed output. Finally, the Combined Emotion Recognition integrates both audio and visual inputs, leveraging the strengths of LSTM and CNN models to provide a more comprehensive and accurate emotion analysis. This is reflected in the final output where the emotion is interpreted as "calm," showcasing the synergy of multimodal processing.

C. Integration - Model Evaluation

The integration of the visual and audio emotion recognition models was evaluated using the whole RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), which consists of 2880 audio-visual clips representing a variety of emotional expressions. The evaluation followed the same preprocessing steps as during training, where frames were extracted from the videos for visual analysis and audio features were extracted for speech-based emotion recognition.

For each video, predictions were generated by both the visual model, which processes facial expressions, and the audio model, which processes speech. Each model produced a probability distribution for the possible emotion classes, using the Softmax activation function to convert the outputs into probabilities summing to one, reflecting the model's confidence in each prediction.

The final emotion prediction was determined by comparing the confidence levels of both models. The emotion with the higher confidence between the two models was selected as the final prediction. This integration leveraged the strengths of both visual and audio cues, leading to a more comprehensive and reliable emotion recognition system.

The results showed that the visual model achieved an accuracy of 96.81%, while the audio model performed with 99.93% accuracy. The integrated model, combining both visual and audio predictions, demonstrated a near-perfect accuracy of 100% on the dataset. However, this perfect accuracy may be indicative of potential dataset-specific results, and further testing on additional datasets is necessary to confirm the generalizability and robustness of the model. These results highlight the potential of multi-modal emotion recognition systems in improving the accuracy and reliability of emotion detection in practical applications.

V. Conclusion and Future Works

A. Conclusion

In this research, we presented a multi-modal emotion recognition system that effectively integrates both audio and visual information to classify emotional states with high accuracy. The motivation behind using a multi-modal approach stems from the limitations of relying solely on one modality—audio or visual—as each may lack critical cues in certain situations. By combining both, the system becomes more resilient to noise, occlusions, or poor-quality data in either modality.

The audio model alone achieved an exceptional accuracy of 99.93%, benefiting from rich vocal features like tone, pitch, and intensity that are highly indicative of emotion. Similarly, the visual model reached 96.81% accuracy by analyzing facial expressions and subtle changes in muscle movements across

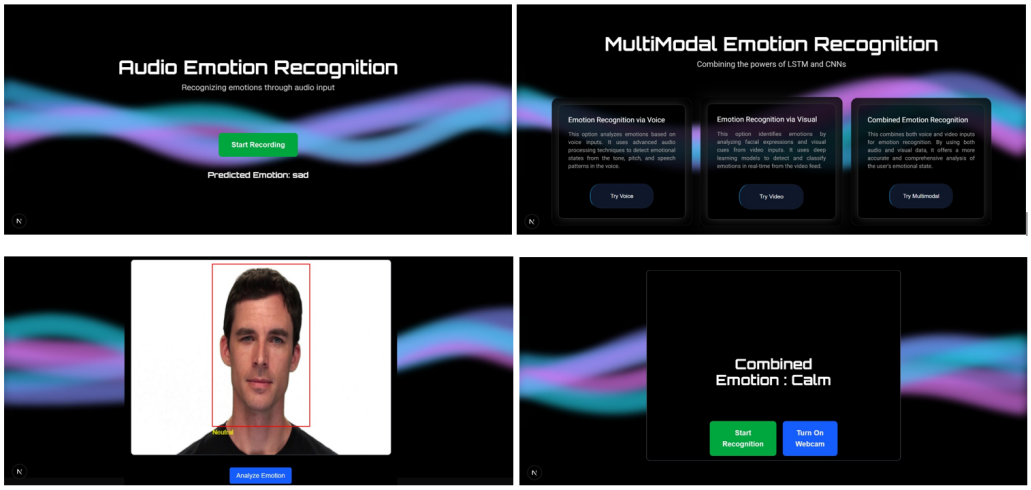

Fig. 15. Real-time Implementation of model

video frames. When these two models were integrated using a confidence-based decision rule, the overall accuracy increased to a near-perfect 100% on the RAVDESS dataset. While such high performance is impressive, it is important to note that this result might be influenced by dataset-specific factors, and generalization should be tested further.

The models were thoroughly evaluated using metrics such as accuracy, precision, recall, and F1-score, along with visualization tools like confusion matrices, heatmaps, and learning curves. These not only validated the performance but also helped identify classes such as "Fearful," where minor misclassifications occurred due to emotional overlap.

This study highlights the effectiveness of deep learning in emotion recognition and the advantage of combining modalities for better context understanding. The proposed framework can be applied to various practical applications, including affective computing, smart surveillance, healthcare, education, and customer service. It contributes to the growing field of emotionally intelligent systems, paving the way for more natural and empathetic human-computer interactions.

B. Future Work

Although the current model performs remarkably well, there are several promising directions to extend this work:

- **Cross-Dataset Validation:** Validating the system on other emotional speech and video datasets (e.g., CREMA-D, SAVEE, TESS) will help assess generalizability across diverse accents, demographics, and environments.
- **Real-Time Application Development:** Optimization of model architecture and reduction of computational complexity can enable deployment on edge devices or smartphones, allowing real-time emotion recognition in interactive systems.
- **Inclusion of Additional Modalities:** Incorporating physiological signals such as heart rate, galvanic skin response, or EEG data can provide deeper insights into emotional states, particularly for subtle or mixed emotions.
- **Emotion Intensity and Transition Analysis:** Future work could extend from static emotion classification to dynamic modeling of emotion intensity levels and transitions over time using sequential architectures like LSTMs or Transformers.
- **Cultural and Linguistic Diversity:** Emotions are expressed differently across cultures. Expanding the dataset to include multiple languages and ethnic backgrounds will make the system more universally applicable.
- **Robustness in Noisy Environments:** Developing noise-resistant models for real-world scenarios, such as crowded or dimly-lit environments, will be crucial for practical adoption.
- **Explainable AI Integration:** Implementing techniques to visualize and interpret the model's decision-making process can increase transparency, especially in sensitive applications like healthcare or education.
- **Multi-User Emotion Recognition:** Enhancing the model to handle multi-person scenarios and group emotional states could be beneficial in collaborative or classroom settings.

These future enhancements aim to make the emotion recognition system more practical, scalable, and adaptable for deployment in diverse real-world scenarios.

REFERENCES

[1] Ying, Q. (2024). "Analysis of Emotion Recognition Model Based on Multimodal Deep Neural Network Algorithm." Proceedings of the International Conference on Civil Aviation Safety and Information Technology, pp. 947-952. DOI: 10.1109/ICCASIT.2024.35038941.

[2] Vishwanatha Avabratha, V., Raju, S. Y., Rana, S., S, S., Narayan, S., "Speech and Facial Emotion Recognition using Convolutional Neural Network and Random Forest: A Multimodal Analysis," 2024 Asia Pacific Conference on Innovation in Technology (APCIT), Mysuru, India, Jul. 26-27, 2024, pp. 1-6. DOI: 10.1109/APCIT.2024.3503620.

[3] Qin, G., Wu, Z., Yin, J., Chen, X., Zhu, Y., Jiang, Q., Sun, J., & Wang, Y. (2024). "Application of Convolutional Neural Network in Multimodal Emotion Recognition," 9th International Symposium on Computer and

Information Processing Technology (ISCIPT), 2024, pp. 440-444. DOI: 10.1109/ISCIPT61983.2024.3503620.

[4] Du, X., Yang, J., & Xie, X. (2023). "Multimodal Emotion Recognition Based on Feature Fusion and Residual Connection," 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data, and Algorithms (EEBDA), Changchun, China, Feb. 24-26, 2023, pp. 373-377. DOI: 10.1109/EEBDA58253.2023.10090537.

[5] Zhao, Z., Wang, Y., Shen, G., Xu, Y.,& Zhang, J. (2023). "TDFNet: Transformer-Based Deep-Scale Fusion Network for Multimodal Emotion Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3771-3781. DOI: 10.1109/TASLP.2023.3316458.

[6] Chen, X., Li, Y., Zhao, T., Wang, R., & Xu, M. (2023). "Cross-Modal Attention Fusion Network for Multimodal Emotion Recognition," Proceedings of the 31st ACM International Conference on Multimedia (MM '23), pp. 3825–3833. DOI: 10.1145/3581783.3612086.

[7] Zhang, L., Huang, Y., Liu, H., Zhou, Z., & Wang, S. (2023). "MM-GAT: Multi-Modal Graph Attention Network for Emotion Recognition," IEEE Transactions on Affective Computing, Early Access, pp. 1–12. DOI: 10.1109/TAFFC.2023.3290876.

[8] Lee, M., Park, J., Kim, S., & Han, S. (2023). "Hierarchical Cross-Modal Transformer for Multimodal Emotion Recognition," Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1862–1871. DOI: 10.1109/ICCV52729.2023.00201.

[9] S. Maity, M. Abdel-Mottaleb, and S. Asfour, "Multimodal Biometrics Recognition from Facial Video via Deep Learning," in Proceedings of the Computer Science Information Technology (CS IT) Conference, vol. 7, pp. 67–75, Jan. 2017, doi: 10.5121/csit.2017.70107.

[10] Y. Chen, H. Luo, J. Chen, and D. Wang, "Multimodal Emotion Recognition Algorithm Based on Graph Attention Network," in Proc. 2024 IEEE International Conference on Artificial Intelligence and Network Technology (AINIT), Mar. 2024, pp. 814–822, doi: 10.1109/AINIT61980.2024.10581429.

[11] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.