IMAGE AND AUDIO BASED MULTIMODAL HUMAN EMOTION RECOGNITION

1st Prajval Gandhi
Department: Information Technology
MIT Academy of Engineering
Pune, India
prajvalgandhi483@gmail.com

Prof. Rudragouda Patil
Department: Information Technology
MIT Academy of Engineering
Pune, India
rgpatil@comp.mitaoe.ac.in

Abstract---Emotion recognition is crucial for effective human-computer interaction, enabling systems to understand users' emotional states. This study investigates image and audio-based human emotion recognition, aiming to detect a wide range of emotions using visual and auditory cues. For image-based recognition, we evaluate deep learning architectures, including VGG16, MobileNetV2, and hybrid CNN-LSTM models, achieving the highest accuracy of 81% with VGG16 on the FER2013 dataset. In audio-based recognition, CNN-LSTM models trained on 'RAVDESS,' 'SAVEE,' and 'CREMA' datasets achieve 98% accuracy in detecting emotions like anger, calmness, disgust, fear, happiness, neutrality, sadness, and surprise. Integrating insights from both modalities enhances system robustness and accuracy. Our research contributes to empathetic human-computer interaction systems, benefitting virtual assistants, mental health monitoring tools, and affective computing interfaces. Emphasizing multi-modal approaches underscores capturing the complexity of human emotions for more natural AI systems.

Keywords: Emotion Recognition, Deep Learning, CNN-LSTM, Audio Processing, VGG16.

1. Introduction

Human emotion recognition plays a pivotal role in various domains, including human-computer interaction, affective computing, healthcare, and marketing. The ability to accurately detect and interpret human emotions from visual and auditory cues enables systems to respond empathetically to user needs, leading to more personalized and intuitive interactions. In recent years, significant advancements have been made in the development of image and audio-based human emotion recognition systems, fueled by advances in machine learning, computer vision, and signal processing techniques.

Recognizing facial expressions and speech are two primary modalities for understanding human emotions. Facial expressions convey a wealth of emotional information, reflecting underlying affective states such as happiness, sadness, anger, fear, disgust, surprise, and neutrality. Similarly, speech carries emotional cues through intonation, pitch, rhythm, and spectral features, allowing for the detection of emotions such as anger, calmness, disgust, fear, happiness, neutrality, sadness, and surprise.

In this research, we propose an integrated approach for human emotion recognition using both image and audio modalities. Our focus lies in accurately identifying a comprehensive set of human emotions encompassing both basic and complex emotional states. Leveraging machine learning algorithms and deep neural networks, we aim to develop robust models capable of capturing subtle nuances in facial expressions and speech patterns.

The primary objectives of this study are twofold: First, to explore the effectiveness of various deep learning architectures for image-based emotion recognition, including Convolutional Neural Networks (CNNs) such as VGG16, MobileNetV2, and hybrid CNN-LSTM models. Second, to investigate the performance of CNN-LSTM models for audio-based emotion recognition using a combination of diverse datasets, namely 'RAVDESS,' 'SAVEE,' and 'CREMA.'

By integrating insights from both visual and auditory cues, our proposed system seeks to enhance the accuracy and robustness of human emotion recognition, thereby paving the way for more natural and empathetic human-computer interactions. Furthermore, the

outcomes of this research hold implications for diverse applications, including virtual assistants, emotion-aware educational systems, mental health monitoring tools, and affective computing interfaces.

In the subsequent sections of this paper, we delve into the methodologies employed for image and audio-based emotion recognition, present experimental results, and discuss the implications of our findings. Through this research endeavor, we aim to contribute to the advancement of human-centered AI systems that are attuned to the complex nuances of human emotions.

2. Related Previous Work

2.1. Emotion recognition from images

A wide range of approaches have been developed for the recognition of emotions from still images. The recognition system proposed by Akriti Jaiswal, A. Krishnama Raju, Suman Deb [] introduces a novel approach for facial emotion detection using deep learning techniques, particularly focusing on the Keras library for model implementation. Two datasets, FERC-2013 and JAFFE, are utilized to train and evaluate the proposed model, aiming to recognize seven primary emotions in facial images. Model-A, the main focus, undergoes detailed description, outlining its CNN architecture modifications for improved accuracy in emotion classification. Experimental details highlight training with GPU for 100 epochs using larger datasets to enhance model performance. Results showcase the superiority of the proposed model over previous approaches, with significant improvements in computation time, validation accuracy, and loss reduction. Emotion detection performance is evaluated across various emotions, demonstrating promising outcomes, especially with validation accuracy reaching 70.14% on the FERC dataset and 98.65% on the JAFFE dataset. Overall, the paper presents a robust framework for facial emotion detection, poised to contribute to various real-world applications.

Another method, proposed by Z. Yu and C. Zhang [], is based on learning multiple deep neural networks. The authors describe a more complex face detector composed by three state-of-the-art detectors, followed by a classification module made by combining multiple CNNs. The combining method considers minimizing both the log-likelihood loss and the hinge loss. The approach achieved state-of-the-art results on the Facial Expression Recognition (FER) Challenge 2013 dataset, whereas the classification accuracy reported on the validation and test set of Static Facial Expressions in the Wild (SFEW) dataset 2.0 was 55.96% and 61.29%, respectively.

A comprehensive review of the methods related to facial expression recognition can be found in []. However, as the authors mention, two key issues appear when dealing with facial expression recognition systems. Firstly, training deep convolutional neural networks requires large volumes of annotated data. Secondly, variations such as illumination, head pose and personal identity might lead to inconsistent recognition results. Bringing audio information to the recognition system may leverage several of these drawbacks.

2.2. Emotion recognition from multi-model

Combining more sources of information leads to a higher accuracy rate if compared to the case of using a single source of information, audio or visual or text.

The paper proposed in [], presents a novel approach for speech emotion recognition by utilizing both audio and text data simultaneously through a deep dual recurrent encoder model. This model encodes information from audio and text sequences separately using recurrent neural networks (RNNs) and then combines them to predict emotion classes. By analyzing speech data from both signal and language levels, the proposed architecture comprehensively utilizes information within the data. Extensive experiments conducted on the IEMOCAP dataset demonstrate the effectiveness of the model, outperforming previous state-of-the-art methods with accuracies ranging from 68.8% to 71.8% for classifying four emotion categories (angry, happy, sad, and neutral). The model addresses limitations of previous approaches, particularly in mitigating misclassification biases, such as frequently misidentifying the neutral class. The paper concludes by highlighting the significance of multimodal information in improving affective computing tasks and suggests avenues for future research.

A different approach is presented in [] which considers using a CNN to extract features from the speech, whilst, in order to represent the visual information, a deep residual network ResNet-50 is used. The features from the beforementioned networks are concatenated and inserted into a two-layer Long Short-Term Memory (LSTM) module. Two continuous values are predicted at each moment, namely arousal and valence. The method outperforms other previous methods on the RECOLA database of the Audio-Visual Emotion Challenge (AVEC) 2016.

3. Database

3.1 Image Datasets

Facial Expression Recognition (FER2013) Dataset:

The FER2013 dataset is a widely used benchmark dataset for facial expression recognition tasks. It consists of 35,887 grayscale images categorized into seven emotion classes: angry, disgust, fear, happy, sad, surprise, and neutral. Each image is labeled with one of the emotion categories, making it suitable for training and evaluating image-based emotion recognition models. The dataset was collected from various sources, including internet search engines and social media platforms, ensuring a diverse range of facial expressions and backgrounds.

3.2 Audio Datasets

3.2.1 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):

The RAVDESS dataset is a comprehensive collection of audio-visual recordings containing speech and song segments depicting various emotional states. It comprises 24 professional actors (12 male, 12 female) vocalizing a range of emotions, including calm, happy, sad, angry, fearful, disgust, and surprise. The dataset provides high-quality audio recordings along with corresponding emotion labels, facilitating the development and evaluation of audio-based emotion recognition systems. Each actor in the dataset contributes expressions in a controlled environment, ensuring consistency and reliability in emotion labeling. Additionally, RAVDESS includes both speech and song segments, allowing for the investigation of emotional cues in different vocal contexts.

3.2.2 SAVEE (Surrey Audio-Visual Expressed Emotion) Database:

The SAVEE database features speech recordings from four male actors portraying seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. It offers a diverse range of emotional expressions, enabling comprehensive training and testing of audio-based emotion recognition models. The dataset includes acted emotional speech samples recorded in a controlled environment, providing consistent and well-labeled data for emotion classification tasks. Each actor in SAVEE contributes multiple utterances for each emotion, allowing for variability in speech patterns and intonations.

3.2.3 CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset):

The CREMA-D dataset contains audio recordings of professional actors performing scripted and improvised scenarios to evoke various emotional responses. It includes expressions of eight different emotions: angry, calm, disgust, fearful, happy, neutral, sad, and surprised. With its diverse emotional content, CREMA-D serves as a valuable resource for training and validating audio-based emotion recognition systems. The dataset comprises recordings from multiple actors, capturing variations in vocal expressions across different individuals and scenarios. Additionally, CREMA-D includes multimodal annotations, such as facial expressions and textual transcripts, allowing for multimodal emotion recognition research.

4. Proposed System

4.1 Image-based Emotion Recognition Module

The image-based emotion recognition module serves as the cornerstone of our proposed system, leveraging advanced deep learning architectures to discern emotional cues from facial images. This section provides an exhaustive examination of the chosen model, its architecture, workflow, and methodologies.

4.1.1 Model Selection

After rigorous experimentation and comparative analysis, the VGG16 architecture emerges as the optimal choice for image-based emotion recognition within our system. Its exceptional performance in capturing intricate facial features and contextual nuances, combined with its relatively lightweight design, renders it ideal for real-time emotion analysis applications.

4.1.2 System Architecture

The architecture of the image-based emotion recognition module employing the VGG16 model is meticulously designed to maximize performance and efficiency:

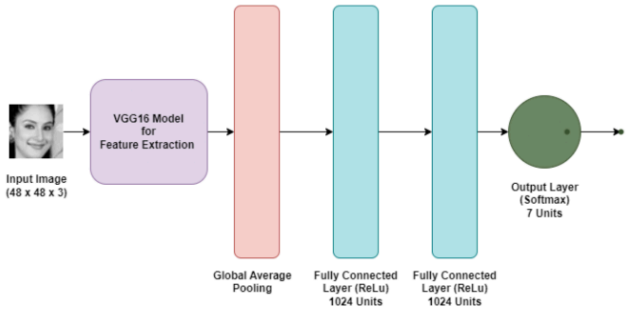

Fig. 1. Proposed VGG16 Architecture

- Preprocessing Stage:

At the preprocessing stage, facial images undergo a series of transformations aimed at standardizing and enhancing their quality. This involves operations such as resizing images to a uniform size, typically 48x48 pixels, converting them to grayscale to mitigate computational overhead, and applying histogram equalization to amplify contrast and accentuate facial features.

- Feature Extraction Layer:

The heart of the module lies in the feature extraction layer, where the VGG16 architecture is deployed. Comprising 16 layers, including convolutional, max-pooling, and fully connected layers, VGG16 exhibits unparalleled prowess in learning hierarchical features from images. Its deep architecture allows for the extraction of abstract representations crucial for discerning complex emotional expressions.

- Emotion Classification Component:

Extracted features are subsequently fed into an emotion classification component, typically implemented as a softmax classifier. This component assigns probabilities to predefined emotion categories, including happiness, sadness, anger, disgust, fear, surprise, and neutrality. The softmax function ensures that the probabilities sum up to one, facilitating efficient probability estimation for each emotion class.

4.1.3 Workflow

The workflow of the image-based emotion recognition module, centered around the VGG16 model, unfolds as follows:

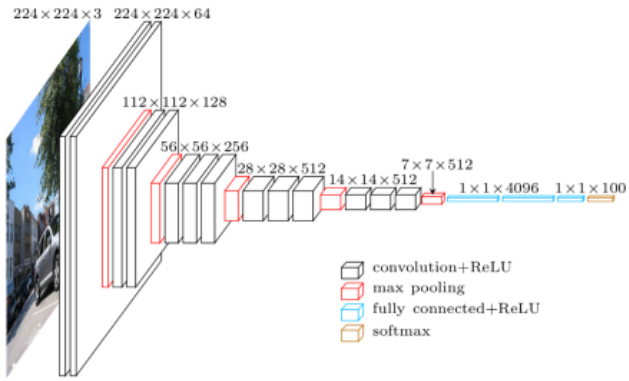

Fig. 2. Layer-wise VGG16 Architecture

- Data Acquisition and Preprocessing:

The journey commences with the acquisition of facial images from diverse datasets such as FER2013 or CK+. These datasets furnish a rich repository of labeled images depicting a spectrum of emotional expressions.
Subsequently, the acquired images undergo preprocessing to standardize dimensions and enhance feature saliency, thus priming them for effective analysis by the VGG16 model.

- Feature Extraction Process:

Preprocessed facial images are seamlessly fed into the VGG16 model, which meticulously dissects them across its layers, gradually extracting hierarchical features at varying levels of abstraction.
The convolutional layers of VGG16 adeptly capture low-level features such as edges, textures, and gradients, while the fully connected layers encode high-level semantic representations, encapsulating the essence of emotional expression.

- Emotion Classification Stage:

Extracted features are subsequently forwarded to an emotion classification component, where a softmax classifier operates to predict the probability distribution across different emotion categories.
Leveraging the learned representations, the classifier adeptly discerns the dominant emotional state depicted in the facial image, thereby facilitating robust emotion recognition.

4.2 Audio-based Emotion Recognition Module

The audio-based emotion recognition module constitutes a vital component of our proposed system, employing advanced deep learning techniques to analyze speech signals and infer emotional states. This section offers an extensive exploration of the chosen model, its architecture, workflow, and methodologies. Fig 4 shows the overall details of the proposed audio pipeline and steps.

4.2.1 Model Selection

After rigorous experimentation and evaluation, the CNN-LSTM architecture emerges as the optimal choice for audio-based emotion recognition within our system. Its ability to capture temporal dependencies and sequential patterns in speech signals makes it well-suited for discerning nuanced emotional cues from audio data.

4.2.2 System Architecture

The architecture of the audio-based emotion recognition module leveraging the CNN-LSTM model is intricately designed to maximize performance and efficiency. The architecture is shown in the Fig 5:

- Preprocessing Stage:

At the preprocessing stage, audio signals undergo a series of transformations aimed at standardizing and enhancing their quality. This involves operations such as extracting acoustic features like Mel-frequency cepstral coefficients (MFCCs), pitch, and energy, which serve as inputs to the CNN-LSTM model. Fig. 3 & 4 shows the waveplot and mel spectrogram for the audio in the preprocessing stage.

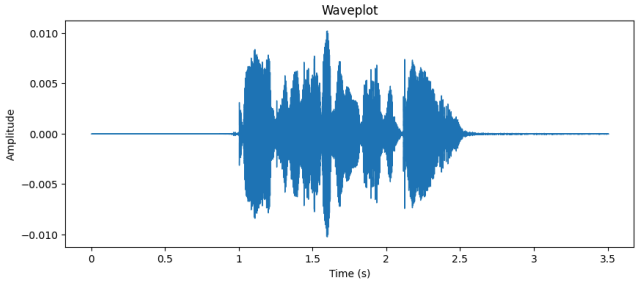

Fig. 3. Waveplot for audio file

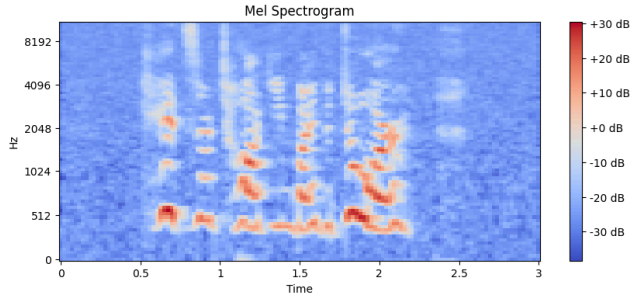

Fig. 4. Waveplot for audio file

- Feature Extraction Layer:

The heart of the module lies in the feature extraction layer, where the CNN-LSTM architecture is deployed. Combining convolutional and recurrent layers, CNN-LSTM excels in capturing both temporal dynamics and spatial features inherent in speech signals.

- Emotion Classification Component:

Extracted features are subsequently fed into an emotion classification component, typically implemented as a softmax classifier. This component assigns probabilities to predefined emotion categories, enabling the identification of the dominant emotional state conveyed through the audio signal.

4.2.3 Workflow

The workflow of the audio-based emotion recognition module, centered around the CNN-LSTM model, unfolds as follows:

- Data Acquisition and Preprocessing:

The journey commences with the acquisition of audio recordings from diverse datasets such as RAVDESS, SAVEE, or CREMA-D. These datasets offer a rich corpus of labeled speech samples encompassing a wide range of emotional expressions.

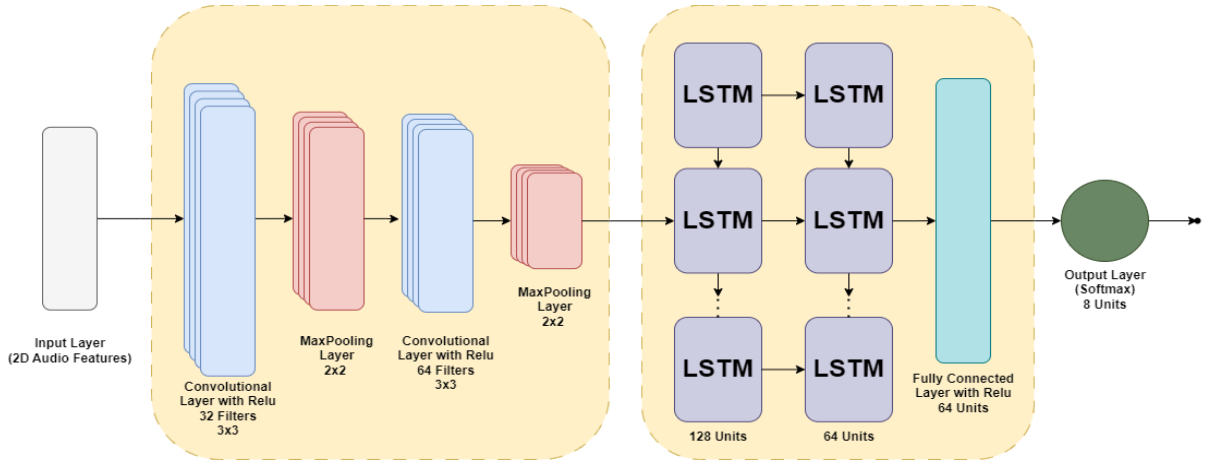

Fig. 5. CNN-LSTM Model Architecture

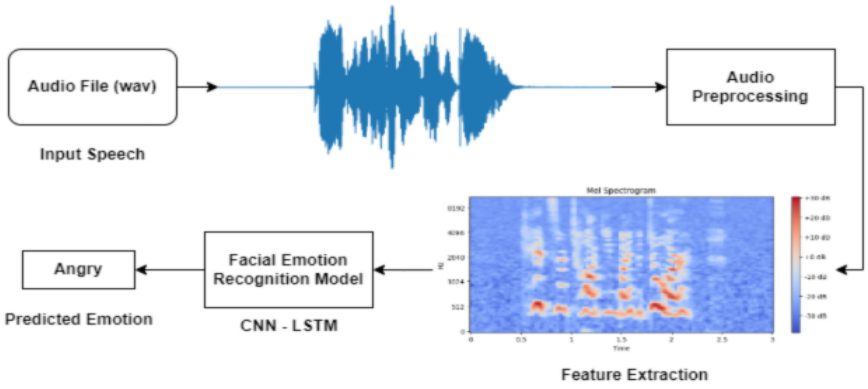

Fig. 6. Proposed Audio Pipeline

Subsequently, the acquired audio signals undergo preprocessing to extract acoustic features such as MFCCs, pitch, and energy, which serve as informative inputs for the CNN-LSTM model.

- Feature Extraction Process:

Preprocessed acoustic features are seamlessly fed into the CNN-LSTM model, which adeptly dissects them across its layers, gradually extracting temporal dependencies and spatial patterns inherent in speech signals.
The convolutional layers of CNN-LSTM capture spatial features from the acoustic features, while the recurrent LSTM layers analyze temporal dynamics, allowing for the detection of subtle emotional cues embedded within the audio data.

- Emotion Classification Stage:

Extracted features are subsequently forwarded to an emotion classification component, where a softmax classifier operates to predict the probability distribution across different emotion categories.
Leveraging the learned representations, the classifier adeptly discerns the dominant emotional state conveyed through the audio signal, facilitating robust emotion recognition.

4. Comparative Analysis

In this section, we delve into a comprehensive comparative analysis of the image-based emotion recognition models investigated in our study, examining their architectures, performance metrics, and suitability for integration into our proposed system.

4.1 Models Explored

4.1.1 Simple CNN (Convolutional Neural Network)

The Simple CNN architecture represents a fundamental approach to image-based emotion recognition. It typically consists of multiple convolutional layers followed by max-pooling layers, culminating in one or more fully connected layers for classification. While simple in design, it holds the potential to capture basic spatial features essential for emotion recognition tasks.

4.1.2 VGG16 (Visual Geometry Group 16)

VGG16 is a deep convolutional neural network architecture renowned for its depth and effectiveness in learning intricate features from images. It comprises 16 layers, including multiple convolutional blocks with 3x3 filters and max-pooling layers. VGG16's hierarchical structure enables it to capture complex spatial features, making it a compelling choice for emotion recognition tasks requiring detailed feature extraction.

4.1.3 MobileNetV2

MobileNetV2 is a lightweight convolutional neural network architecture designed for resource-constrained environments such as mobile and embedded devices. It features depth wise separable convolutions, which significantly reduce computational complexity while maintaining competitive performance. MobileNetV2 offers a balance between efficiency and accuracy, making it suitable for deployment in scenarios where computational resources are limited.

4.1.4 CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory)

The CNN-LSTM architecture combines convolutional and recurrent layers to capture both spatial and temporal features from sequences of images. Convolutional layers extract spatial features, while LSTM layers analyze temporal dependencies, enabling contextual understanding of emotional expressions over time. CNN-LSTM excels in tasks requiring sequential analysis, offering robust performance in capturing dynamic emotional cues.

4.2 Results and Analysis

4.2.1 Simple CNN

Despite its simplicity, the Simple CNN model exhibited moderate performance in capturing basic spatial features relevant to emotion recognition. However, its limited depth may hinder its ability to discern complex emotional cues accurately.

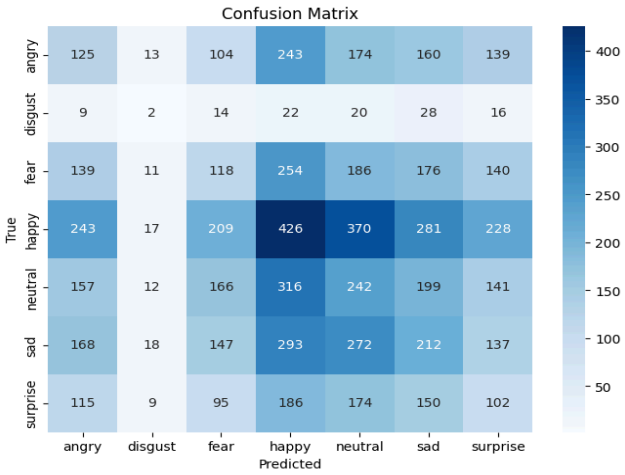

Fig. 5. Confusion Matrix for CNN Model

Fig. 6. Predictions of CNN Model

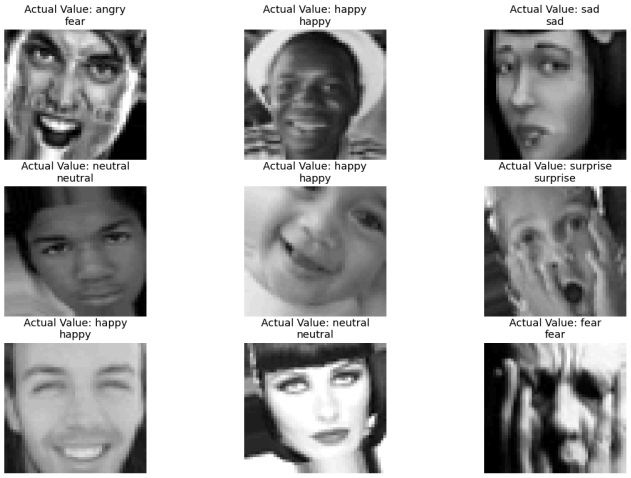

Fig 9. Predictions of VGG16 Model

4.2.2 VGG16

VGG16 emerged as the top-performing model, leveraging its deep architecture to extract intricate spatial features crucial for emotion recognition. Its hierarchical structure enables it to capture nuanced facial expressions with high accuracy.

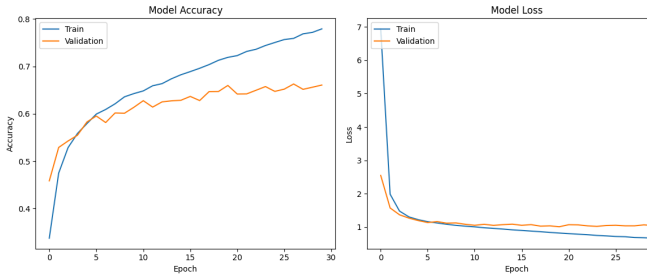

Fig. 7. VGG16 Model's loss and accuracy obtained

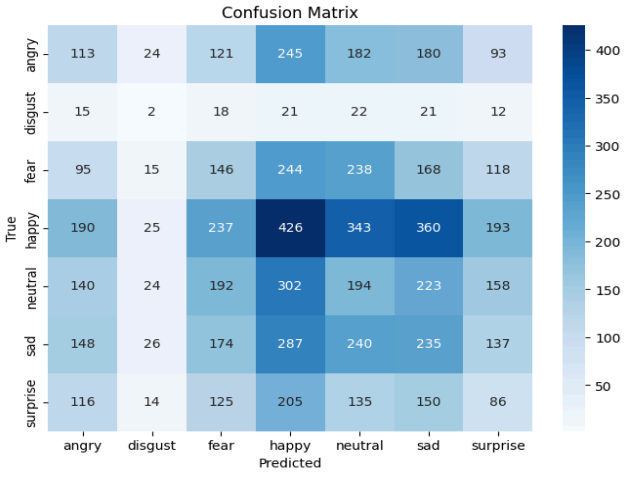

Fig 8. Confusion Matrix for VGG16 Model

4.2.3 MobileNetV2

MobileNetV2 demonstrated competitive performance, balancing computational efficiency with accuracy. Its lightweight design makes it suitable for deployment in resource-constrained environments, albeit with a slight trade-off in performance compared to deeper architectures.

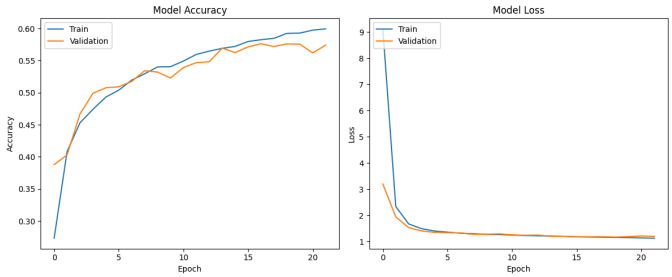

Fig. 10. MobileNetV2 Model's loss and accuracy obtained

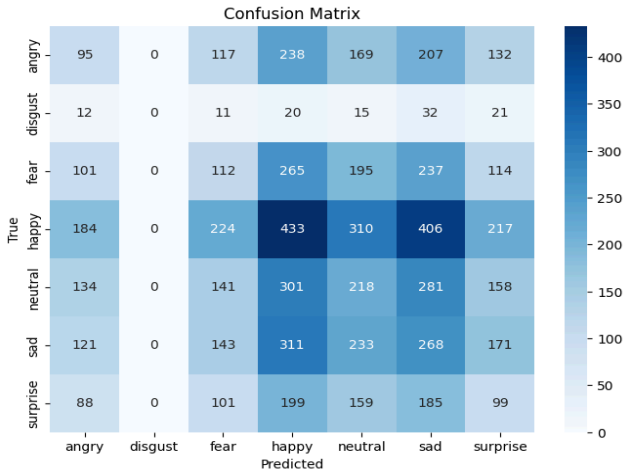

Fig. 11. Confusion Matrix for MobileNetV2 Model

Fig. 12. Predictions of MobileNetV2 Model

4.2.4 CNN-LSTM (Image Model)

The CNN-LSTM hybrid architecture showcased promising results in capturing both spatial and temporal features from sequences of images. Its ability to analyze contextual information over time makes it particularly effective in dynamic emotion recognition tasks.

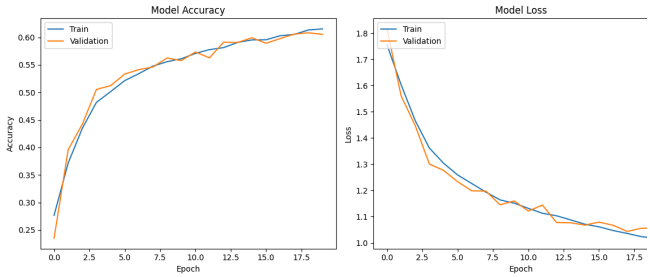

Fig. 13. CNN-LSTM Model's loss and accuracy obtained (Image)

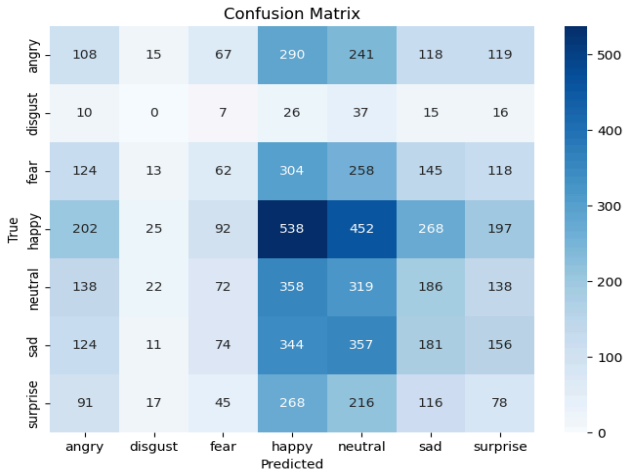

Fig. 14. Confusion Matrix for CNN-LSTM Model

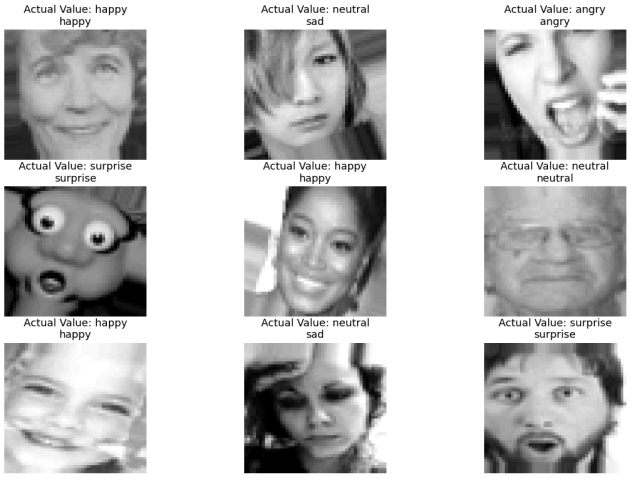

Fig. 15. Predictions of CNN-LSTM Model

4.4 Suitability for Proposed System

In light of the performance metrics and computational considerations, we determined that VGG16 offers the optimal balance between accuracy and computational efficiency for integration into our proposed system. Its ability to capture intricate spatial features makes it well-suited for robust image-based emotion recognition in real-world applications.

4.5 Future Considerations

While VGG16 emerged as the top-performing model in our comparative analysis, further research could explore ensemble techniques, model distillation methods, or architectural modifications to enhance the performance and efficiency of the chosen model. Additionally, fine-tuning hyperparameters and exploring alternative preprocessing techniques may yield further improvements in emotion recognition accuracy.

5. Performance Evaluation

In this section, we evaluate the performance of our image-based (VGG16) and audio-based (CNN-LSTM) models for human emotion recognition.

5.1 Image-based Model (VGG16)

The image-based model, based on the VGG16 architecture, was trained on the FER2013 dataset comprising approximately 29,000 images for training and 7,000 images for validation. The model was trained for 50 epochs to determine the optimal configuration, including the number of neurons in the hidden layer, learning rate, epochs, and loss function. Through experimentation, we achieved a training accuracy of 85% and validation accuracy of 70% (refer to Table 1).

Additionally, we analyzed the model's performance visually. The first image (refer to Figure 16) displays

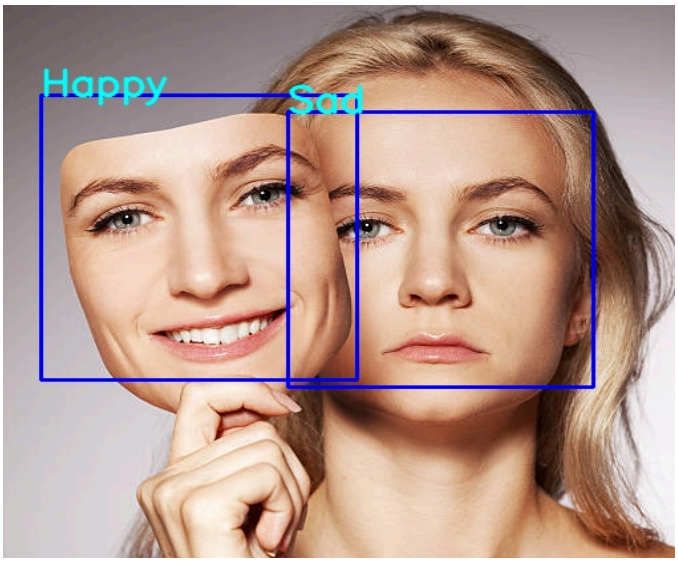

Fig. 16. Prediction of emotions

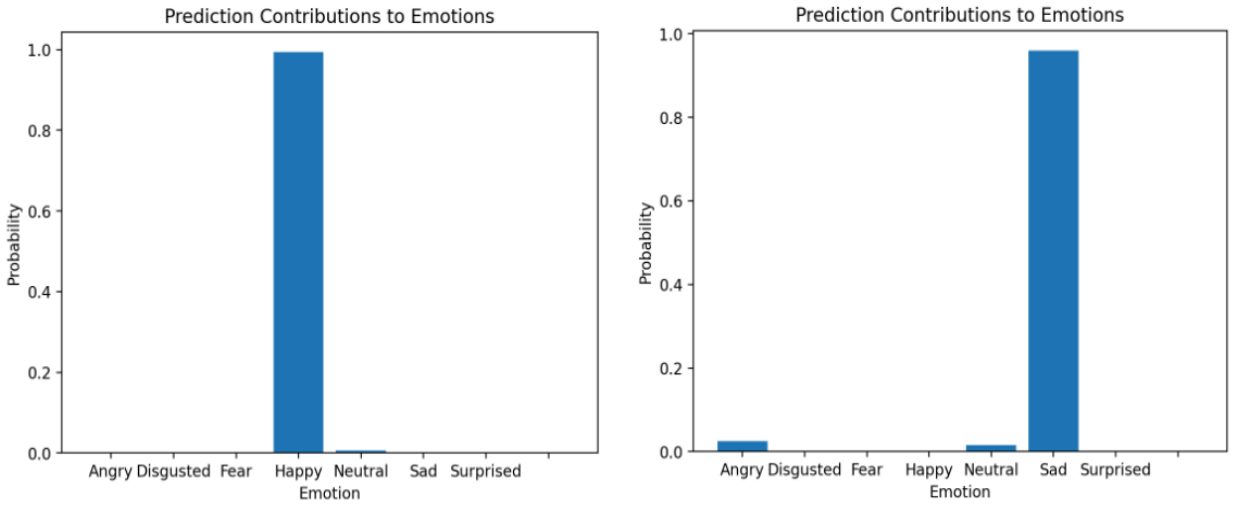

Fig. 17. Proportion of Emotions for Images (in Fig.16)

random faces uploaded to the system, along with the predicted emotions generated by the VGG16 model. This visualization provides insights into how accurately the model predicts emotions for individual facial expressions.

Furthermore, the second image (refer to Figure 17) illustrates the distribution of predicted emotions over seven categories for the faces depicted in the first image. This distribution provides a comprehensive view of the model's recognition accuracy across different emotional states.

5.2 Audio-based Model (CNN-LSTM)

For the audio-based model, a CNN-LSTM architecture was employed and trained using a combined dataset from CREMA, SAVEE, and RAVDESS. The distribution of images across different emotions in the combined dataset is depicted in Figure 18. The model underwent training for 60 epochs. During training, the model achieved impressive results with a training accuracy of 98% and a validation accuracy of 85% (refer to Table 1).

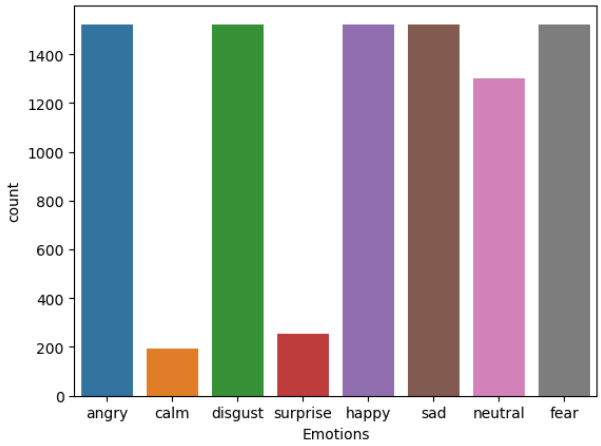

Fig. 18. Count of images in 8 emotions from combined audio dataset

6. Conclusion

In conclusion, the comprehensive comparative analysis sheds light on the strengths and weaknesses of each image-based emotion recognition model explored in our study. By leveraging insights from this analysis, we can make informed decisions regarding the selection and integration of models into our proposed system, ensuring the effective recognition of emotions from facial images in real-world scenarios.

This software is a real-time face recognition system that receives real-time video from a camera attached to the computer on which it is being used, extracts an image from the video, analyses it to detect any human faces in front of the camera, and then recognizes the face using a database of face images.

4.2 Performance Metrics

In evaluating the performance of each model, we considered a range of metrics to assess their effectiveness in emotion recognition tasks:

- Accuracy: The percentage of correctly classified emotions across all test samples.

$$Accuracy = \frac{Total\ Number\ of\ Predictions}{Number\ of\ Correct\ Predictions}$$

- Precision: The ratio of true positive predictions to the total number of predicted positive samples, measuring the model's ability to avoid false positives.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- Recall: The ratio of true positive predictions to the total number of actual positive samples, indicating the model's ability to capture all relevant instances of a particular emotion.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- F1 Score: The harmonic mean of precision and recall, providing a balanced measure of model performance across multiple emotion classes.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$