Deep Learning-Based Emotion Recognition Using Face and Voice Analysis

Ms. Kavya L

PG Student, Department of Computer Science Engineering, Mangalam College of Engineering, Ettumanoor, kerala, India.e-mail: lkavya805@gmail.com

Ms. Merlin Mary James

Assistant Professor,Department of Computer Science Engineering, Mangalam College of Engineering, Ettumanoor, kerala, e-mail: merlin.james@mangalam.in

Abstract **-** Psychological challenges such as depression, anxiety, and other emotional disturbances are frequently observed among hospitalized individuals. However, most healthcare environments lack continuous monitoring systems to assess these emotional conditions. Early detection of depressive symptoms is essential for prompt and effective therapeutic response. This study introduces two sophisticated machine learning techniques for recognizing emotions through analysis of facial expressions and vocal characteristics. The facial emotion recognition approach utilizes both a custom Convolutional Neural Network (CNN) and a pre-trained ResNet model to interpret expressions. Meanwhile, voice-based emotion detection is carried out using hybrid models, specifically CNN-LSTM and CNN-GRU, to process audio inputs. Experimental findings indicate strong performance in accurately identifying emotional states, highlighting the potential of integrating these methods into clinical settings to enhance patient well-being and care quality.

Keywords: Facial expressions, Voice analysis/, CNN-LSTM, CNN-GRU, deep learning, image classification.

I.INTRODUCTION

Within hospital settings, individuals frequently encounter psychological conditions such as anxiety, sadness, and other emotional disturbances, which are often early indicators of depression and related mental health disorders. These emotional states are not always outwardly apparent to healthcare providers, making timely identification and intervention difficult. Conventional psychological evaluations largely depend on manual observations or patient-reported symptoms, both of which can introduce subjectivity and inaccuracies. Consequently, there is an urgent demand for advanced technologies capable of delivering real-time, continuous monitoring of a patient's emotional well-being. With significant strides in machine learning and image analysis, emotion recognition has emerged as a promising tool to support clinical decision-making by enabling accurate and timely assessments of emotional health. This research presents two primary methodologies for emotion detection: facial expression analysis and vocal emotion recognition. Both methods incorporate cutting-edge machine learning algorithms designed to accurately infer emotional states.

Facial Expression Analysis

Facial expressions serve as immediate and powerful indicators of emotional responses. The human face can express emotions such as happiness, sadness, anger, surprise, fear etc.. Accurately interpreting this expression has been a focal point of emotion recognition research. The introduction of deep learning—especially Convolutional Neural Networks (CNNs)—has brought significant improvements to the field.

In the proposed framework, facial emotion detection is performed using both custom-built CNNs and advanced pre-trained models such as ResNet. The custom CNN architecture is tailored to identify and classify vital facial features relevant to various emotions. In parallel, the ResNet model, fine-tuned for emotion classification, leverages its deep-layered architecture to capture complex and abstract patterns from facial imagery. Together, these models enable the system to effectively recognize and categorize emotions from facial inputs with high accuracy.

Vocal Emotion Recognition

The human voice is another rich channel for expressing emotions. Characteristics such as tone, pitch, speaking rate, and volume often mirror internal emotional states. Unlike facial cues, which are typically static, vocal features reflect dynamic and nuanced changes in emotion, making voice analysis an essential component of comprehensive emotion recognition systems. Recent progress in deep learning,

particularly the development of hybrid models combining CNNs with temporal sequence modeling architectures like Long Short Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), has lead to more effective analysis of speech signals. These hybrid models are especially suited for capturing the temporal dependencies in audio data, allowing for accurate identification of emotional content embedded in speech patterns.

II. Related Work

[1] The paper titled *"Emotion Recognition System via Facial Expressions and Speech"* by Aayushi Chaudhari (2023), published in *SN Computer Science* (Volume 4, Issue 363), introduces a multimodal emotion recognition framework that integrates facial expression detection and speech analysis. For facial recognition, the study uses Gabor filters in combination with Convolutional Neural Networks (CNNs), while for speech emotion analysis, mel-frequency cepstral coefficients (MFCCs) and CNNs are employed. The system is trained on various datasets including JAFFE, Kaggle, and EMOTIC for visual data, and RAVDESS, TESS, CREMA-D, and SAVEE for audio data. To manage high-dimensional features, Principal Component Analysis (PCA) is used, and classification is performed using Support Vector Machines (SVM). The system demonstrates high accuracy; however, limitations include the lack of real-time application, limited dataset diversity affecting generalizability, and minimal focus on model interpretability.

[2] *"Emotion Recognition on Speech Processing Using Machine Learning"* by Vikrant Chole (2023) focuses on classifying emotional states from speech signals using machine learning methods. The model processes voice input by segmenting it into 60ms frames with a 10ms overlap and extracts features using Linear Predictive Coding (LPC) and MFCCs. Classification is carried out using the K-Nearest Neighbor (KNN) algorithm. The study use the Berlin and Spanish speech emotion databases. Although the research aims to enhance speech emotion recognition systems, it encounters issues such as overfitting, limited emotional depth in classification, insufficient data for training advanced models like RNNs, and the need for improved noise filtering and feature selection strategies.

[3] The study titled *"Automatic Recognition of Student Emotions from Facial Expressions During a Lecture"* by G. Tonguc (2020), published in *Computers & Education* (Volume 148, Issue 103797), investigates how student emotions vary throughout lecture sessions using facial expression analysis. The system captures webcam-based facial imagery and utilizes software developed in C, integrated with Microsoft's Emotion Recognition API. Statistical analysis is performed using the Manova test, examining emotional variations among 67 students from multiple academic disciplines. While the study offers valuable insights into student engagement and emotional fluctuations during lectures, it is limited by data loss due to face visibility issues, absence of long-term tracking, and a lack of qualitative feedback on students' learning experiences.

[4] In *"Speech Emotion Recognition Based on Emotion Perception"* by Gang Liu (2023), published in the *EURASIP Journal on Audio, Speech, and Music Processing*, a biologically inspired framework is proposed for speech emotion recognition. Drawing from concepts in neuroscience, the method introduces a human-like implicit emotional classification system supported by multi-task learning. The system is trained on the IEMOCAP dataset, which includes 12 hours of labeled emotional speech data. Results show notable improvement in recognition accuracy. Despite its contributions, the study identifies limitations such as the scarcity of high-quality, diverse datasets for speech emotion recognition, limited adaptability of existing SER networks, and underexplored variations in the stability of implicit emotional attributes across speakers.

[5] The paper *"Facial Emotion Recognition Using Deep Learning: Review and Insights"* by Wafa Mellouk (2020), published in *Procedia Computer Science*, offers a broad survey of deep learning based techniques in facial emotion recognition. The review outlines essential preprocessing steps like resizing, cropping, and normalization, and evaluates various architectures including CNNs, CNN-LSTMs, and 3D CNNs. Datasets such as FER2013, AffectNet, and Oulu-CASIA are reviewed alongside data augmentation methods for increasing model robustness. The paper presents comparative performance outcomes across models but falls short in highlighting specific research gaps, future research directions, and known limitations within current deep learning methodologies.

III. IMPLEMENTATION

A. System Architecture and Hardware Requirements

The overall system is structured around two core modules: Facial Emotion Recognition and Voice Emotion Recognition. These modules function by capturing real-time video input from a device's camera or webcam, which is then processed in sequential stages. Each module is organized into four fundamental phases: Image Preprocessing, Model Development, Model Training and Evaluation, and Model Deployment.

For development, the system leverages tools such as Python IDLE, Anaconda, and Visual Studio Code, all of

which contribute to an efficient and manageable coding environment. The platform runs on Windows 10, offering compatibility with various software libraries and frameworks. It supports both mobile and desktop application development—Flutter is utilized for frontend design, while backend development is handled in Python using web frameworks like Flask or Django. The system is operated on hardware equipped with an Intel i5 processor and 12GB of RAM, which provides reliable performance for programming, model training, and application testing.

B. Facial Emotion Recognition

[i] Image Preprocessing

Preprocessing plays a critical role in Facial emotion recognition, Images are properly formatted for learning and generalization by deep learning models. Initially, datasets such as FER2013 or CK+, which contain labeled facial expression images, are loaded into the system. Images are then resized to uniform resolution, commonly 224x224 pixels, to ensure compatibility across training batches and neural network architectures.

Face detection is subsequently performed using MTCNN (Multi-Task Cascaded Convolutional Networks), a widely used deep learning-based face detector. MTCNN identifies facial landmarks and provides bounding boxes around detected faces, allowing the system to crop and isolate the relevant facial regions. After cropping, normalization is applied—scaling pixel values to standard range (typically [0, 1] or [-1, 1])—which ensures consistency with the input expectations of CNN model.

The dataset is split into three subsets: training, validation, and testing. The training set is used for learning model parameters, the validation set for tuning hyperparameters, and test set for assessing the model's performance on previously un seen data. This preprocessing pipeline ensures standardized, high quality input for optimal model performance.

[ii] Model Development

In building facial emotion recognition models, two primary strategies are commonly employed. One approach involves the design and implementation of custom Convolutional Neural Network (CNN) specifically for facial expression classification. This network includes multiple convolutional layers that extract hierarchical spatial features from facial inputs, along with pooling layers that reduce spatial dimensions and computational load, helping to mitigate overfitting. Following feature extraction, fully connected layers consolidate learned features to perform final emotion classification. The architecture is optimized by adjusting parameters such as the number of filters, kernel sizes, activation functions, and dropout rates. This flexibility allows the custom CNN to effectively capture emotion-specific patterns in facial expressions and deliver accurate predictions.

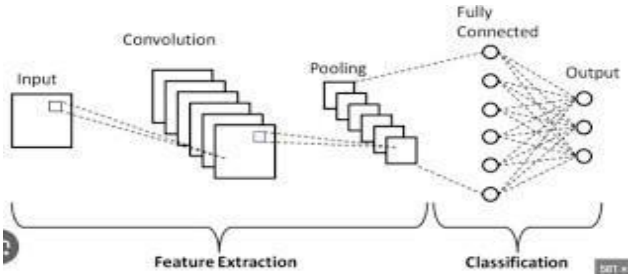


Fig 1. Convolutional Neural Network (CNN)

The second modeling approach utilizes transfer learning, specifically through the fine-tuning of a pre-trained ResNet architecture (e.g., ResNet-50). Originally trained on extensive datasets such as ImageNet, ResNet models have already developed robust and generalizable feature extraction capabilities across a wide range of visual patterns. In the context of facial emotion recognition, transfer learning involves modifying and retraining only the final layers of the ResNet network, while keeping the earlier convolutional layers frozen. These fixed layers retain the model's foundational knowledge, while the new layers adapt specifically to the facial emotion dataset. This technique significantly accelerates the training process and enhances model accuracy, as the base model already possesses a deep understanding of image features. Additionally, it reduces the dependency on large, labeled datasets, making it a practical and efficient choice for emotion recognition tasks where annotated data may be limited.

[iii] Model Training and Evaluation

The training and evaluation phase is essential in ensuring that the facial emotion recognition model performs effectively and reliably. This stage begins with the model compilation process, where key parameters such as the optimizer, loss function, and evaluation metrics are defined. One of the most widely used optimizers is Adam, as it dynamically adjusts learning rates during training, enabling the model to converge efficiently and identify optimal weight values. For classification tasks involving multiple emotion categories, the categorical cross entropy loss function is typically selected. This function quantifies the divergence between the predicted output probabilities and the actual class labels, guiding the model to reduce its prediction errors.

While accuracy provides a basic indicator of how often the model predicts correctly, it doesn't always offer a complete picture—especially in cases where the dataset may have

imbalanced class distributions. Therefore, additional evaluation metrics such as precision, recall, and F1 score are employed for a more detailed analysis:

- Precision: The proportion of correct positive predictions among all predicted positives.
- Recall : The model's ability to identify all relevant instances of a particular emotion class.
- F1 score: Calculated as the harmonic mean of precision and recall. Offers a balanced performance measure, which is especially valuable while dealing with skewed datasets.

By combining these metrics, developers can ensure that the emotion recognition model not only achieves high accuracy but also performs equitably across all emotion categories.

Emotion	Precision	Recall	F1-Score	Support
Happiness	0.91	0.92	0.91	250
Angry	0.87	0.85	0.86	180
Sad	0.89	0.90	0.89	200
Relaxed	0.88	0.87	0.87	220
Overall Accuracy	0.89			850
Macro Average	0.89	0.88	0.88	
Weighted Average	0.89	0.89	0.89	

Fig 2: Analysis of the Classification Report

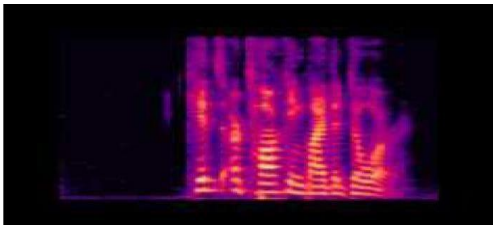
[iv] Model Deployment

Following the performance evaluation, the model that delivers the best results—whether it's the custom CNN or the fine-tuned ResNet—is selected for deployment based on key evaluation metrics like accuracy or F1-score. The selected model is then saved to preserve its trained parameters, ensuring it can be seamlessly integrated into real-world systems without requiring additional training. To facilitate this integration, the model is typically saved in widely supported formats such as HDF5 (.h5), TensorFlow SavedModel, or PyTorch's .pth format. These formats enable easy loading and deployment across platforms, making the emotion recognition model readily usable in various practical applications, including healthcare, education, or customer service environments.

C. Voice Emotion Recognition

[i] Audio Preprocessing

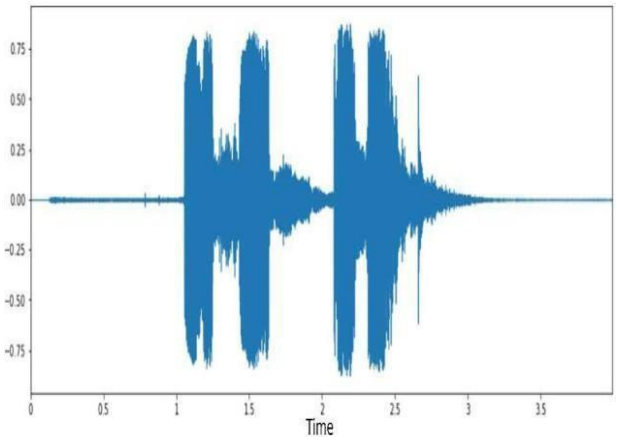
The first step in voice emotion recognition involves organizing and processing audio datasets like RAVDESS, which include not only the audio recordings but also metadata such as speaker details and labeled emotions. To improve the models ability to generalize and perform well under different conditions, data augmentation techniques are employed. These includes pitch shifting, time stretching, and adding background noises. All of which simulate variations found in real world audio inputs. After augmentation, the next critical phase is featuring extraction, where raw audio signals are transformed into representations suitable for deep learning. Commonly used features include Mel-spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs). These formats effectively capture the essential properties of speech, such as frequency content and energy distribution, which are vital for accurately identifying emotional states from vocal cues.



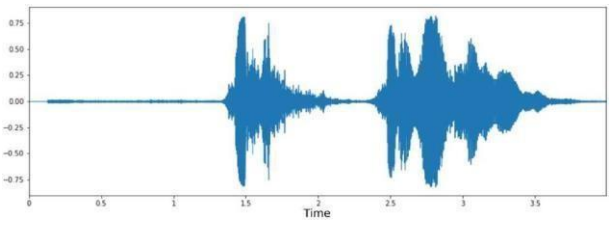
Mel Spectrogram

MFCC Spectrogram

Fig 3: Representation of speech frequencies in two different forms



B) Speech signal data of happy voice



Speech signal data of angry voice

Fig 4: Representation of speech signals for different classes: a happy and b angry

[ii] Model Development

The CNN-LSTM architecture integrates two advanced deep learning components to effectively recognize emotions from audio input.Convolutional Neural Network (CNN) layers, which are responsible for extracting spatial features from input representations such as spectrograms. These spectrograms visualize how the audio signal's frequency content changes over time. The CNN layers are particularly effective at identifying localized acoustic features—such as tone, pitch, and vocal intensity—that are often indicative of emotional states. Once spatial features are extracted, they are passed to Long Short Term Memory (LSTM) layers. LSTMs are a type of Recurrent Neural Network (RNN) specifically designed to learn and retain patterns over extended sequences. This makes them well-suited for analyzing the temporal dynamics of speech, allowing the model to understand how emotional cues unfold throughout the audio clip. By combining CNNs for spatial analysis with LSTMs for sequential modeling, this hybrid architecture captures both instantaneous audio features and their temporal progression, providing a comprehensive framework for accurate emotion recognition from speech.

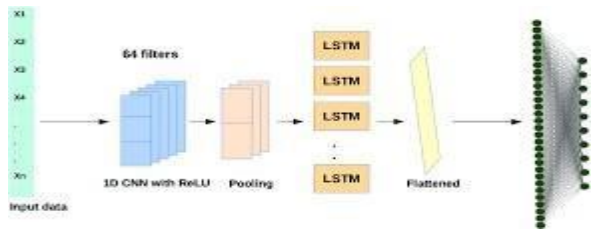


Fig 5: CNN-LSTM Architecture

[iii] Model Training and Evaluation

The training and evaluation phase begins by compiling the model, which includes selecting an appropriate loss function and optimizer. For tasks involving multiple emotion categories, categorical cross entropy is the standard loss function, as it measures the predicted class probabilities align with the true emotion labels. Optimizers such as Adam or Stochastic Gradient Descent (SGD) are employed to adjust model weights, facilitating efficient learning and convergence during training. Once compiled, the model is trained using the training dataset, where it learns to map audio features to corresponding emotional labels through backpropagation. A validation dataset is used simultaneously to monitor the model learning progress and to fine tune hyperparameters, reducing the risk of overfitting by ensuring the model performs well on unseen data.

After training, the model effectiveness is tested using a separate test set. Performance is evaluated using metrics like accuracy, precision, recall, and F1 score. While accuracy gives an overall success rate, precision indicates how many predicted emotions were correct, and recall shows how well the model detects actual emotional instances. The F1 score provides a balanced measure of both precision and recall, which is especially important in imbalanced datasets, ensuring that less frequent emotions are also recognized effectively.

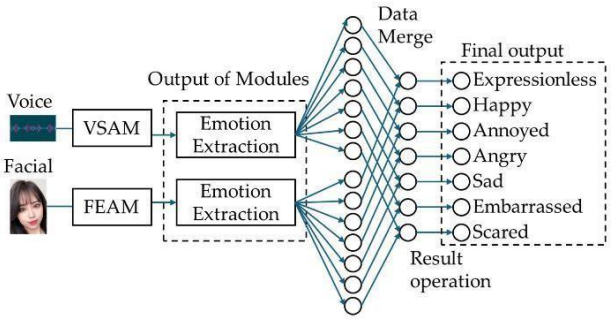


Fig 6: Architectural diagram for facial and voice analysis

D. Key Observations are as follows:

The model demonstrates high precision for class 0 (0.85), indicating that when it predicts this class, it is correct most of the time. In contrast, precision for class 3 is lower (0.57), implying that predictions for this class include a greater number of incorrect classifications or false positives. On the other hand, the recall for class 3 is relatively strong (0.60), meaning the model is effective at identifying actual instances of this class and retrieving relevant samples. This suggests the model is more sensitive to detecting class 3 but may also misclassify other classes as class 3. The F1 score for class 3 is notably high (0.90), indicating a well balanced performance between precision and recall for that class. Despite this, the noticeable differences in precision and recall across other classes reveal a disparity in the

model's performance, which might be the result of class imbalance within the dataset. Such imbalance can lead the model to favor certain classes, making it less reliable for underrepresented emotions.

IV. Results and Discussion

The developed emotion recognition system, which combines image and audio processing techniques, has shown encouraging results throughout various stages from data preprocessing and model development to evaluation and deployment. The applications effectiveness was validated using standard performance metrics such as accuracy, precision, recall, and F1 score, all of which affirm the systems ability to accurately interpret emotional states through both facial expressions and vocal cues. Despite these positive outcomes, certain limitations emerged during evaluation. A significant challenge was the system's sensitivity to background noise in audio inputs. Although techniques like pitch shifting and noise injection were employed for data augmentation, the model still encountered difficulties when handling poor- quality or distorted audio samples. Future improvements could involve integrating advanced noise suppression algorithms and expanding the training set with more diverse and real-world audio data to enhance generalizability and resilience.

Similarly, while facial emotion recognition yielded reliable results, its performance tended to fluctuate in the presence of inconsistent lighting, partial occlusions, or low- resolution images. Addressing this could involve adopting more advanced face detection algorithms, enhancing image normalization processes, and incorporating adaptive preprocessing techniques to boost the system's accuracy in real-world conditions.

VI. Conclusion and Future Work

The developed emotion recognition system effectively integrates both image and audio processing models to analyze and classify human emotions from facial expressions and speech. Leveraging advanced neural network architectures such as CNN-LSTM and CNN- GRU, the application captures spatial and temporal features with high precision, resulting in reliable emotion predictions. Comprehensive testing and evaluation using accuracy, precision, recall, and F1 score. The system capability to perform effectively in real-world conditions. The architecture benefits from a React-based front-end seamlessly connected to a Python Flask/Django back-end, with deployment on cloud platforms like AWS ensuring scalability, security, and responsiveness. RESTful APIs facilitate efficient communication between the interface and backend, enabling real-time emotion detection with a smooth user experience. While the system shows strong overall performance, areas for improvement remain. In particular, enhanced preprocessing techniques could mitigate the impact of noisy audio inputs, and more robust facial recognition approaches may be needed to handle variable lighting and occlusions.

In summary, the application delivers a scalable and dependable solution for emotion recognition, with promising use cases in mental health monitoring, customer engagement, and human-computer interaction. Continued efforts to improve data diversity, model resilience, and feature capabilities will help expand the system's effectiveness and adaptability across broader domains.

References

[1] Aayushi Chaudhari,Chintan Bhatt, Thanh Thi Nguyen, Nisarg Patel,Kirtan Chavda, Kalind Sarda. "Emotion Recognition System via Facial Expressions and Speech Using Machine Learning and Deep Learning Techniques". Vol 2023

[2] Sonawane B, Sharma P. "Deep learning based approach of emotion detection and grading system. Pattern Recognit Image Anal". Vol 2020.

[3] Prof. Vikrant Chole, Ms. Namrata Yerande, Ms. Aarti Shinde," EMOTION RECOGNITION ON SPEECH PROCESSING USING MACHINE LEARNING",Vol 2023

[4] Güray Tonguç, Betul Ozaydın Ozkara," Automatic recognition of student emotions from facial expressions during a lecture ", Volume 2020

[5] Gang Liul, Shifang Cai1 and Ce Wang," Speech emotion recognition based on emotion perception", volume 2023

[6] V. Sreenivas, V. Namdeo, and E. Vijay Kumar, "Modified deep belief network-based human emotion recognition with multiscale

features from video sequences," Software: Practice and Experience, vol. 51, no. 6, pp. 1259-1279, Jun. 2021.

[7] H. Liao et al., "Deep learning enhanced attributes conditional random forest for robust facial expression recognition," Multimedia Tools and Applications, vol. 80, no. 19, pp. 28627- 28645, 2021.

[8] F. Wang, H. Chen, and L. Kong, "Real-time facial expression recognition on robots for healthcare," in 2018 IEEE International Conference on Intelligence and Safety for Robotics, Shenyang, China, pp. 402-406, 2018.

[9] L. M. Martinez and R. Valenti, "Deep learning for non-human emotion recognition: A review," Artificial Intelligence Review, vol. 53, no. 1, pp. 1-20, 2020.

[10] S. Yin, Y. Wang, and J. Luo, "Recognizing animal emotions via deep learning: A survey," Pattern Recognition Letters, vol. 131, pp. 128-135, 2020.