Human Emotion Detection In Mental Health Monitoring

Dhanashri Gangurde[1], Radhika Bakale[2], Parth Kokate[3] and Archana Kadam[4]

[1]Student, SCTR's Pune Institute of Computer Technology, (IT), Pune, Maharashtra, India, gangurdedhanashri890@gmail.com.
[2]Student, SCTR's Pune Institute of Computer Technology, (IT), Pune, Maharashtra, India, radhikabakale642@gmail.com.
[3]Student, SCTR's Pune Institute of Computer Technology, (IT), Pune, Maharashtra, India, parthkokate1919@gmail.com.
[4]Assistant Professor, SCTR's Pune Institute of Computer Technology, (IT), Pune, Maharashtra, India, askadam@pict.edu.

Abstract

Human emotion detection has become an essential tool in mental health monitoring, offering the potential for early detection of mental health disorders. Existing models for emotion detection primarily Utilize (CNNs), in particular, are deep learning technique, to perform analysis. facial expressions and, in some cases, voice patterns. These models have demonstrated the ability to detect basic emotions such as sad, fear, anger, happy, disgust,surprise, neutral, and sorrow with a high degree of accuracy. However, challenges remain in terms of real time processing and personalization for individual users. This research introduces a novel system designed to enhance early mental health detection through advanced human emotion detection techniques. The system focuses on analyzing facial expressions and voice patterns to identify potential signs of emotional distress. By leveraging deep learning models, specifically refined CNN architectures, and additional data preprocessing techniques, we aim to achieve an accuracy rate of 93 percent for emotion recognition using facial images alone. When combining facial image data with voice pattern analysis, the system reaches an overall accuracy of up to 82 percent. The incorporation of real-time processing capabilities enables instant emotion detection from live video and audio feeds, providing timely insights for mental health professionals. Furthermore, the system features a personalization component that adapts to each user's unique emotional responses, improving detection accuracy over time. By combining facial and voice data, the proposed system offers a comprehensive approach to human emotion detection, with the goal of contributing to early intervention and better mental health outcomes.

Keywords: Human emotion detection, Mental health, Early intervention, machine learning, Deep learning, Facial expression analysis, Voice analysis, Multimodal fusion, CNN, Neural Network

1. Introduction

Human-to-human communication relies heavily on facial expressions, conveying emotions that can significantly influence interactions. Research has demonstrated that recognizing and interpreting facial expressions is essential for effective communication, accounting for a substantial portion of interpersonal interactions. In the realm of interaction between humans and computers, the capacity of machines to acknowledge and retaliate to Emotions in humans are increasingly desired. The goal of this research is to create a system that can accurately detect emotions from facial expressions, specifically targeting mental health monitoring. By analyzing facial cues, the aim is to provide valuable insights into a person's emotional state, potentially aiding in early detection and intervention for mental health issues. While previous research has primarily concentrated on facial emotion recognition, the work extends this by incorporating audio analysis. This multimodal strategy makes it possible for a more thorough understanding of a person's emotional state, considering the interplay of facial expressions, voice patterns, and linguistic cues. By combining these modalities, The objective is to develop a human emotion detecting system that is more reliable and accurate. This system can potentially be used to assist mental health professionals in identifying individuals at risk of mental health problems and providing timely support.

2. Literature Survey

In [1]. The authors designed a model that detects human emotions based on facial image datasets, achieving 93 percent accuracy. The algorithm for detecting emotions in videos was introduced.

[2]. For video-based face expression detection, the authors looked into various approaches of pooling spatial and temporal data and found that doing so simultaneously is more effective. Deep learning-based multi-modal emotion detection is presented by the model.

[3]. Emotion detection is based not only on facial features but also on speech, video, and text. The author offers real-time facial expression recognition using OpenCV for video, DeepFace for emotion analysis, and a Streamlit interface for user interaction.

[4]. It effectively detects emotions and presents results clearly. The paper introduces a hybrid method using rules, emotions, and context to enhance word meaning detection.

[5]. It leverages sentence transformers and BERT to identify human's emotions, including neutral, and tags multiple emotions based on context. This approach surpasses existing emotion detection methods. The goal of the study is to develop a facial emotion recognition system that will assist in identifying mental tension, which will be advantageous for counselling services and college students.

[6]. By analyzing facial expressions, the system identifies signs of stress in individuals. This study suggests a method for identifying emotions. using both speech, facial expressions with support vector machine (SVM)

[7]. Results show improved performance, with a recognition accuracy above 92 percent for the face module and 85.15 percent for the voice module, outperforming recent methods while being time-efficient. Seven major emotions were identified in the study using deep learning, namely CNN anger, fear,disgust, happy, surprise, sad, and neutrality.

[8]. This monitor depressed individuals and predict suicide risk by analyzing their emotional state. The system used to discover emotions like sad, happy, contempt, fear, surprise, neutral, and rage.

[9]. The author employs the Adaboost, Convolutional Neural Networks, and Haar-cascade algorithms to identify seven different moods. A face detection scheme with feature extraction and noise reduction is part of the pre-training stage. The categorization model uses the Facial Action Coding System to predict seven emotions.

[10]. Current results show 79.8 percent accuracy for detecting these emotions, without using optimization techniques. This paper focuses on extracting facial features with the use of facial landmark detection and linear discriminant analysis.

[11]. Test results show that emotion recognition accuracy is 73.9 percent with LDA and 84.5 percent using Facial Landmark Detection. The presented practice introduces A deep convolutional neural network with no sequential components featuring multiple parallel networks

[12]. Its evaluation uses the The dataset Surrey Audio-Visual Expression Emotion (SAVEE), which includes videos of four individuals expressing seven emotions. The model achieves 87.0 percent accuracy using the KDFF,FER2013, CK+, JAFFE, and AffectNET dataset, outperforming current real-time model, which typically achieve 63-78 percent accuracy

[13]. It is lightweight and appropriate for implementation on a variety of edge devices for real-time applications due to its streamlined architecture. The author suggests a CNN model that uses two fully connected layers, max pooling, and six convolutional layers to recognize face emotions.

[14]. A Haar cascade detector identifies faces, classifying them into seven emotions. The model gain

77.23 percentage accuracy on the FER2013 dataset. Using a Convolutional Neural Network that uses neural networks for emotion categorization and the Viola-Jones method for face identification, the author created the facial emotion recognition model.

[15]. The model, featuring six layers and three full connected layers, achieved 68.26 percent accuracy for the FER2013 dataset also achive 91.58 percent on the CK+ dataset. The Authors develops two deep learning models for detecting fake emotions, one analyzing facial expressions and the other focusing on emotional speech

[16]. The facial expression model achieved 70 percent accuracy, while the speech-based model reached 96.93 percent accuracy, demonstrating the effectiveness of the approach in enhancing both social and human-computer interactions. The author focuses on using a Convolutional Neural Network and OpenCV to notice live human emotion from facial expressions, aiming to bridge the gap between human computer interaction

[17]. The system identifies emotions like sad, disgust, neutral, happy, fear, angry, surprise from real- time webcam input. The author utilizes a hybrid CNN BiLSTM to improve speech emotion recognition (SER). model trained on a merged dataset of RAVDESS, TESS, and CREMA-D, recognizing eight emotions

[18]. The model, utilizing features like Mel Frequency Cepstral Coefficient, RMSE, and Zero Crossing Rate, achieved a 97.80 percent accuracy. The author demonstrated a machine learning model that used a CNN to identify emotions (neutral, happy, sad, and angry) in speech and a Feed Forward Neural Network to identify gender.

[19]. The model achieved 91.46 percent accuracy in gender classification and 86 percent in emotion recognition, showing promise for applications in human- computer interaction, customer service, and healthcare. The Author focuses on detecting emotions from speech using various classification algorithms like Multilayer Perceptron and Support Vector Machine, featuring audio features like Tonnetz, MEL, MFCC, and Chroma.

[20]. The models achieved an accuracy of 86.53 percent after being trained to identify emotions such as peace, neutrality, astonishment, happiness, sadness, annoyance, unpleasant, and disgust. The author focuses on detecting human emotions from sound signals using the Mel-Frequency Cepstral Coefficient (MFCC) for feature extraction, as it closely mimics the human auditory system

[21]. The RBF kernel in a SVM was utilized for classifi- cation, achieving a highest accuracy of 72.5 percent with specific parameter settings including a 0.001 second frame size, 80 filter banks, gamma values between 0.3 and 0.7, and a C value of 1.0.The author proposed work develops a real-time emotion recognition through face system using a CNN model trained on the FER-2013 dataset to track and report individual emotions in real-time

[22]. The system detects faces using the Viola-Jone algorithm. achieves 90.40 percent accuracy and generates a summary report of detected emotions over a time interval.

3. Proposed Methodology

The proposed method for developing a system that detects emotions from facial expressions and audio for mental health monitoring start with data collection. An extensive collection of audio recordings of facial expressions with accompanying emotion labels (such as joy, sadness, anger, and fear) will be gathered. Existing datasets like FER2013 can be used, or new real-world data can be collected. The dataset must be diverse, representing various demographic groups to improve the model's generalizability.

In the preprocessing stage, facial images will be normalized and aligned to standardize input data, remov-ing noise and ensuring consistency in image dimensions. Similarly, audio data will undergo preprocessing, including noise reduction, segmentation, and feature extraction (using techniques like Fourier transforms to capture key audio frequencies). Face detection techniques, such as Haar cascades or Dlib, will be applied to extract key facial regions, focusing on features critical for emotional expression (e.g., eyes, mouth, forehead).

For feature extraction, Convolutional neural networks will recognize high-level features automatically in both facial images and audio signals that correspond to different emotions. Transfer learning using models that have

already been trained, like VGG or ResNet for images and CNNs or RNNs for audio, may be used to enhance accuracy. These models leverage previously learned features from larger datasets to improve performance on smaller, multimodal datasets.

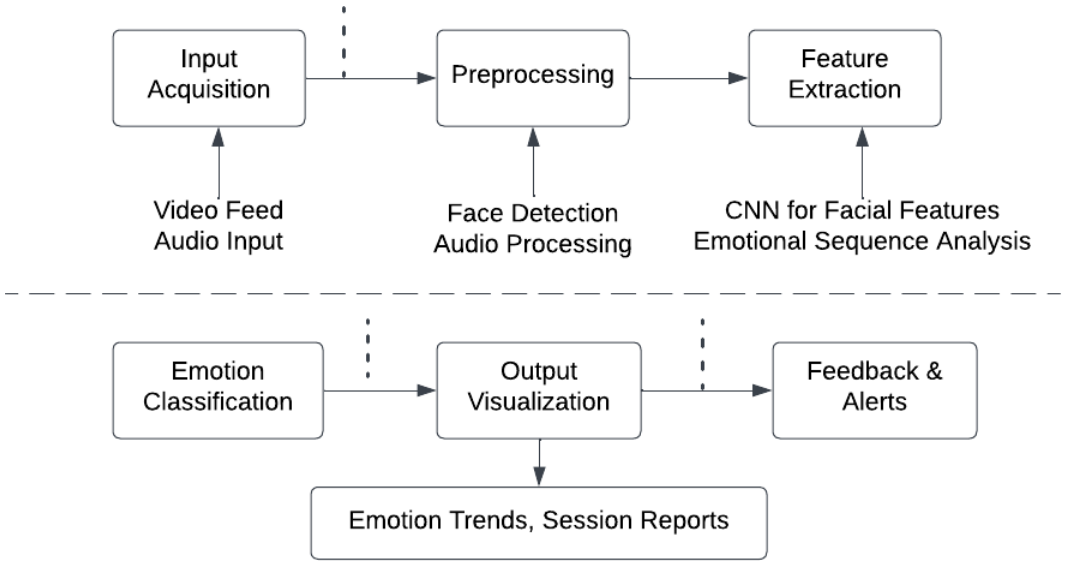

Fig. 1. Mental Health Monitoring using Human Emotion Detection System Block Diagram

In the emotion classification phase mixed deep learning model, for example CNN-RNN, will be used to classify emotions based on features extracted from both visual and audio inputs. The system will employ a Softmax classifier to output probabilities for each emotion category. Performance of the model will be assessed using F1 scores, an Recall,Precision and Accuracy,and metric different architectures will be tested to find the most effective combination of image and audio features. For real-time emotion detection, the trained model will be integrated into a system capable of analyzing live video streams and audio simultaneously. Tools like OpenCV for facial tracking and libraries like PyAudio for real-time audio capture will be employed, with a focus on minimizing latency for smooth user interactions.

The system will also be designed for mental health monitoring by logging and analyzing emotion patterns over time from both visual and audio signals. This will provide comprehensive insights into emotional fluctuations, which may indicate mental health issues. Time-related models, such (LSTM networks, will track patterns over longer durations, helping to recognize mental health conditions like anxiety or depression from both vocal and facial cues.

Finally, the methodology will include extensive validation and testing in real-world scenarios to assess the system's accuracy and robustness across audio-visual modalities. Collaboration with mental health professionals will ensure that the system's emotion detection aligns with meaningful clinical insights. The project will also develop a user interface that provides real-time audio-visual emotion detection results, reports, and early intervention recommendations for potential mental health issues.

This comprehensive approach integrates machine learning, computer vision, audio analysis, and mental health expertise to create a robust tool for emotional well-being monitoring and early detection of mental health conditions.

4. System Architecture and Processing

This system architecture for real-time facial emotion detection and audio-based emotion analysis is divided into several key components. The data input layer captures real-time video, static images, and audio using a webcam, mobile camera, or microphone. A face detection module extracts facial regions from the input using methods such as Haar Cascades, Dlib, or MTCNN, ensuring only relevant facial areas are passed to the model. For audio, real-time audio streams are captured and processed for emotional features, such as pitch, intensity, and tone.

In the preprocessing layer, detected face images are resized, normalized, and augmented (e.g., flipping, rotation, cropping) to meet the CNN model's requirements. Simultaneously, audio signals are preprocessed by removing noise and extracting key frequency features. Data augmentation is applied during training to create more diverse datasets.

The core component of this system is the CNN-based emotion detection module for facial images and an RNN-based module for audio analysis. The CNN processes the preprocessed facial pictures using pooling, convolutional, and fully linked layers, while the RNN handles sequential audio data to capture emotional cues from speech. The emotion classifier combines visual and audio inputs, outputting probabilities for predefined emotion categories (such as happy, sad, or neutral). The Softmax layer then converts these probabilities into specific emotion labels.

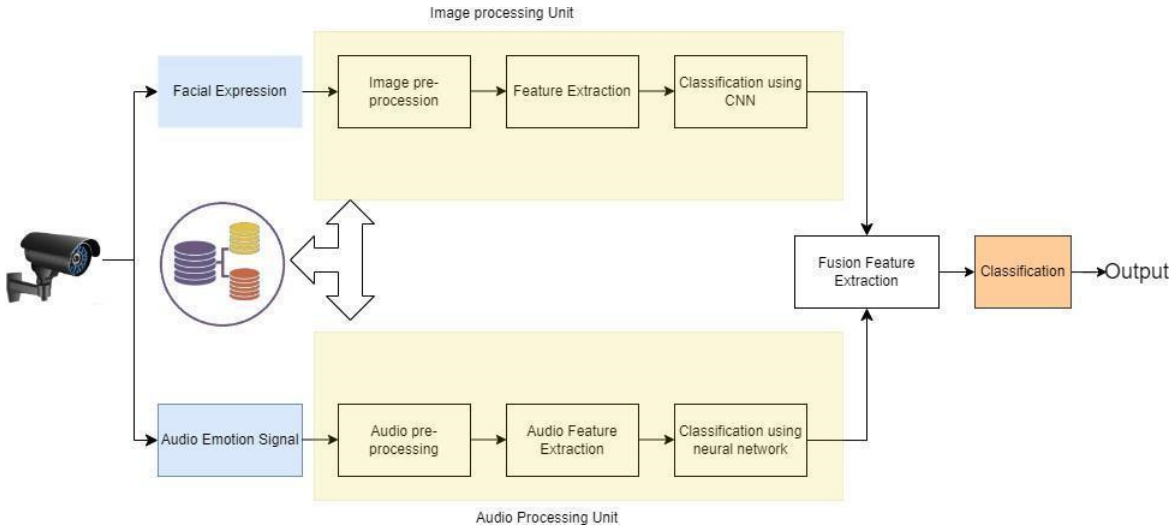

Fig. 2. Workflow of the System for Mental Health Monitoring using Human Emotion Detection.

In the emotion analysis and monitoring layer, emotions are tracked over time from both visual and auditory cues, recording detected emotions for each frame or audio segment. Trends and patterns are visualized through graphs, showing dominant emotions over time from both facial expressions and audio signals.

An alert and recommendation engine is triggered when negative emotions like sadness or anxiety are detected continuously over a significant period, integrating insights from both audio and visual cues. This engine provides mental health insights and suggests interventions like therapy. The user interface layer provides a user dashboard displaying real time emotional monitoring, historical data, and alerts through visual graphs. An optional therapist dashboard allows health care professionals to track patients' emotional trends across multiple sessions, receiving alerts when concerning patterns from both audio and video are detected.

Data storage and analytics are managed through an emotion log database, which stores detected emotions, timestamps, and audio-visual data for long-term tracking. A reporting module generates reports summarizing emotional states over specific periods (daily, weekly, or session-wise).

Cloud integration, although optional, supports large-scale deployment by storing user data and model weights in cloud platforms like AWS, Google Cloud, or Azure. Remote monitoring enables therapists to track patient data via cloud-based dashboards and analytics, integrating insights from both visual and auditory emotion detection.

5. Workflow and Implementation

This system's real-time workflow begins with video or image input, followed by face detection using algorithms like Haar Cascades or MTCNN. After cropping the face region, this system preprocesses the image by resizing, normalizing, and converting it into a format suitable for the CNN model.

The CNN processes the image, extracts features, and outputs a probability distribution for emotion categories, using the Softmax function to classify the face into an emotion label. Detected emotions are logged with timestamps, and trends are plotted in real-time, allowing users or therapists to track emotional shifts during a session.

If negative emotions are detected continuously above a set threshold, the system triggers alerts recommending intervention. Finally, session reports summarize detected emotions and their distribution, accessible via a web or mobile interface for tracking emotional patterns and mental health progress.

This figure illustrates a flowchart for detecting emotions from both visual and audio inputs. It begins with video input, which undergoes face detection to identify facial features.

After detecting the face, data is collected and split into two parallel processes: face analysis and audio/text analysis. The face analysis branch includes preprocessing the facial data and performing emotion detection, while the audio/text analysis branch preprocesses the input audio and analyzes it for emotional sequences.

The Human Emotion Detection system operates through a multi-stage workflow designed for efficient and accurate recognition of emotions from facial expressions. The system architecture integrates various components to enhance its performance and usability.

The first stage, Input Acquisition, captures real-time video feeds from a camera, supporting both live and pre-recorded video files. This allows for use in scenarios like remote mental health assessments or interactive applications.

Next, in the Preprocessing phase, data preparation begins. Face Detection is performed using the Viola- Jones Algorithm, known for its speed and reliability, to identify face regions and extract bounding boxes. Facial Landmark Detection using tools like Dlib or MediaPipe locates key facial features, crucial for analyzing facial geometry and expressions. The captured images are then resized and normalized to reduce complexity, improving model performance.

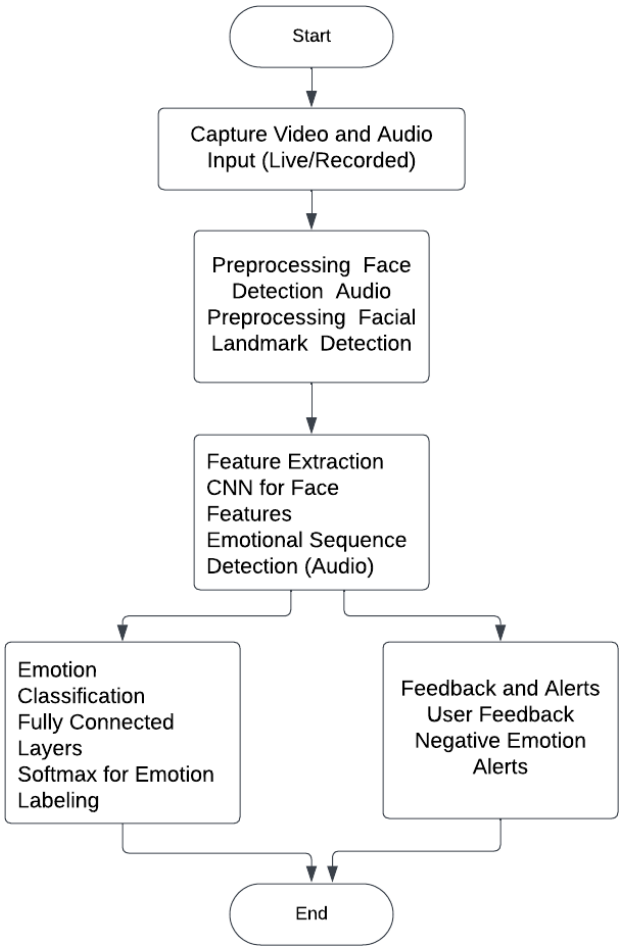

Fig. 3. Workflow of the System for Mental Health Monitoring using Human Emotion Detection.

In the Feature Extraction stage, the heart of the system, a CNN is in work. The CNN's convolutional layers detect spatial hierarchies in facial expressions, while pooling layers down-sample features, preserving essential information and reducing dimensionality. Dropout layers help obstruct overfitting by arbitrarily excluding neurons while training.

Once the features are extracted, the system moves to Emotion Classification using deep learning algorithms. Fully connected layers process the flattened features and classify the detected emotions. A Softmax activation function converts the CNN output into probabilities, identifying emotions such as anger, disgust, contempt, sorrow,fear,joyful, and surprise.

This system supports Real-Time Emotion Recognition through continuous analysis of incoming video frames, providing instant results without noticeable delay. This is accompanied by a Feedback Mechanism that updates emotion predictions based on ongoing inputs, enhancing user engagement.

A notable feature is Personalization, where the system learns from user interactions, adapting to individual facial expressions and refining its accuracy through dynamic model adjustments. Finally, the Output Visualization module presents the recognized emotions and their confidence scores on a user friendly interface. Users can provide feedback on detection accuracy, which further fine-tunes the system's performance.

6. Libraries Used in the Research work

6.1 OpenCV

Purpose: Face detection and tracking in real-time from video streams. Image preprocessing,

such as resizing, normalization, and noise reduction.
Key Functions:

cv2.CascadeClassifier: For face detection using Haar cascades.
cv2.VideoCapture: To capture live video streams.
cv2.imshow: To display real-time video outputs.

6.2 PyAudio

Purpose: Real-time audio capture for emotion analysis.
Key Functions:

pyaudio.PyAudio: To initialize and configure audio input.
stream.read: To capture audio data from the microphone.

6.3 Dlib

Purpose: Facial landmark detection and alignment.
Key Functions:

get_frontal_face_detector: Detects frontal faces in an image.
shape_predictor: Identifies facial landmarks (eyes, nose, mouth).

6.4 TensorFlow/Keras or PyTorch

Purpose: To execute, train, and assess deep learning models for emotion detection.
Key Components:

- Convolutional layers: To extract features from facial images.
- Recurrent layers (LSTM): For capturing temporal dependencies in audio or visual data.

6.5 Streamlit

Purpose: Create an interactive user interface for visualizing emotion detection outputs.
Key Features:

Real-time updating of results.
Integration with deep learning model predictions.

7. Algorithms Used

7.1 Haar Cascades (for Face Detection)

Purpose: Detect faces in images or video.
Steps:

1. Start

2. Pre-compute a set of features (Haar-like features) from labeled training data.
3. Use a cascade of classifiers to quickly identify face-like regions.
4. Identify bounding boxes around detected faces.
5. End

7.2 CNNs

Purpose: Extract high-level spatial features from images for emotion classification.

Steps:

1. Start
2. Apply convolution operations to the input image using kernels to identify textures or edges.
3. Use pooling layers to downsample the information and retain significant information.
4. Flatten the attribute maps and insert them into fully connected layers for grouping(classification).
5. End

7.3 RNNs

Purpose: Model temporal dependencies in sequential data, especially audio signals.

Steps:

1. Start
2. Process sequential input data (e.g., audio spectrograms).
3. Maintain hidden states that capture temporal context.
4. Use the final output for emotion classification.
5. End

7.4 LSTM Networks

Purpose: Track and analyze patterns in emotional fluctuations over time.

Steps:

1. Start
2. Use input gates to control which parts of the input to keep.
3. Use forget gates to remove irrelevant past information.
4. Combine outputs to track long-term emotional trends.
5. End

7.5 Fourier Transform

Purpose: Extract key frequency components from audio signals.

Steps:

1. Start
2. Convert time-domain audio signals into the frequency domain.
3. Identify prominent frequencies associated with emotions.
4. End

7.6 Softmax Classifier

Purpose: Convert the outputs of the model into probabilities for each emotion category.

Steps:

1. Start
2. Apply the Softmax function to model outputs to normalize values between 0 and 1.
3. Choose the category with the highest probability as the detected emotion.
4. End

7.7 Transfer Learning

Purpose: Leverage pre-trained models (e.g., , ResNet,VGG) to improve performance with limited data.

Steps:

1. Start
2. Load a pre-trained model.
3. Fine-tune the model by freezing earlier layers and retraining the final layers on the new dataset.
4. End

8. Relevant Mathematical Analysis done for Implementations

8.1 Haar Cascade Detection - Feature Calculation

Haar-like features are calculated as differences between the sum of pixel intensities in rectangular regions.

$$H(p,q) = p',q' \in R1 \sum I(p',q') - p',q' \in R2 \sum I(p',q')$$

Where:

R_1 and R_2 are the two rectangular regions.
(p′,q′) These are the neighboring pixels around the central pixel *(p,q).*
I(p′,q′) represents the pixel intensity at coordinate *(p,q).*
H(p,q) This represents the response or intensity difference at the pixel position *(p,q).*

8.2 Convolution Operation in CNNs

The convolution operation applied to an image A with a filter B can be expressed as:

$$S(a,b) = (I * K)(a,b) = s \sum n \sum I(a+s, b+n) \cdot K(s,n)$$

Where:

a is the row index of the pixel in the output image.

b is the column index of the pixel in the output image.

$S(a, b)$ is the output of the convolution at position (a, b).
I is the input image.
K is the convolution filter.
$I(a+s,b+n)$ The intensity value of the input image at pixel $(a+s,b+n)$.
s,n Indices for iterating over the kernel.
$K(s,n)$ The kernel (filter) value at position (s,n).
If the kernel has size **k×k** (e.g., 3×3) then
n ranges from $\frac{-k}{2}$ to $\frac{k}{2}$.

8.3 Softmax Function for Classification

The Softmax function converts the model output into a probability distribution:

$$P(y = i \mid x) = \sum \frac{e^{z_q}}{e^{z_p}}$$

Where $P(y = i|\mathbf{x})$ is the probability of class i given input **x**.
z_p, z_q These are the logits or scores produced by a neural network before applying the softmax function
e^{z_p} , e^{z_q} These represent exponentiated logits, ensuring non-negative values.
The denominator sums over all possible classes j .

8.4 LSTM Update Equations

The LSTM update equations are used to update cell and hidden states:

8.4.a Forget gate

$$f_t = \sigma\,(W_f\,[h_{t-1}, x_t\,] + b_f\,)$$

Where

f_t Forget gate activation (values between 0 and 1)
W_f, b_f Weight matrix and bias for the forget gate
h_{t-1} Previous hidden state
x_t Current input
σ Sigmoid activation function

8.4.b Input gate

$$i_t = \sigma\,(W_i\,[h_{t-1}, x_t\,] + b_i)$$

$$\hat{C}_t = \tanh(W_C\,[h_{t-1}, x_t\,] + b_C)$$

Where

i_t Input gate activation

$\hat{C}_t$ Candidate cell state (new memory)

W_i, W_c, b_i, b_c Weight matrices and biases

tanh Hyperbolic tangent function (keeps values between -1 and 1)

8.4.c Cell state updates

$$C_t = f_t * C_{t-1} + i_t * C\hat{}t$$

Where

C_t Updated cell state

C_{t-1} Previous cell state

8.4.d Output gate

$$o_t = \sigma\,(W_o\,[h_{t-1}, x_t\,] + b_o)$$

$$h_t = o_t * \tanh(C_t\,)$$

Where:

o_t output gates.
C_t is the cell state, and h_t is the hidden state.
x_t is the input at time step t.
W_0,b_0 Weight matrix and bias

9. Datasets Used in this research work

Table 1. Composition of publicly available Data traces.

Dataset Name	Year	Type	Volume
FER2013	2013	Facial images	35,887 grayscale images
AffectNet	2017	Facial images	Over 1,000,000 images
RAVDESS	2018	Audio-visual (speech/song)	1,440 recordings (24 actors)
CREMA-D	2015	Audio-visual (speech)	7,442 clips

Datasets used for experimentation

9.1 For Facial Model

FER2013

A comprehensive collection of 35,887 grayscale facial photos, the FER2013 dataset is classified with seven fundamental emotions: anger, disgust, fear, happy, neutrality, sadness, and surprise. The richness and diversity of this dataset make it one of the most popular for CNN-based facial emotion recognition. The study involves training a deep learning model for facial expression recognition using FER2013. Before being fed into a CNN-based classifier, the dataset is preprocessed by picture normalization, augmentation, and face identification (Haar Cascades/Dlib). This enables the model to correctly identify emotions from facial photos taken from live video input or CCTV footage. The FER2013 dataset is an essential part of the emotion detection pipeline since it greatly enhances the system's capacity to identify emotions purely from facial expressions.

9.2 For Audio Model

RAVDESS dataset

For emotion recognition in speech and video, the RAVDESS dataset is a popular resource. 24 professional performers who use speech and music to convey their emotions are featured on 1,440 recordings. An outstanding option for multimodal emotion research, this dataset offers high-quality audio and visual data. This project specifically uses RAVDESS for emotion recognition based on voice. Mel Frequency Cepstral Coefficients (MFCC), pitch, and intensity features are used to the audio samples in order to train a CNN-RNN-based model for speech emotion detection. The dataset helps the algorithm examine both vocal intonations and facial expressions, which is essential for increasing the accuracy of emotion recognition from speech patterns. RAVDESS's incorporation improves multimodal learning and guarantees a more thorough comprehension of human emotions.

10. Summary of Research Works

Table 2. Research Work Carried out in detail

Work	Algorithms	Datasets used	Key Findings
Smith et al. [1]	Convolutional Neural Net- works (CNNs)	FER2013	Detects emotions using facial features. Video-based emotion detection.
Johnson et al. [2]	CNNs for spatial data; Tem- poral pooling techniques	Custom dataset	Spatial and temporal pooling for feature extraction.
Brown et al. [3]	Haar-cascade, DeepFace	AffectNet	Combines facial, speech, and text features for multimodal detection.
Taylor et al. [4]	Hybrid rules and context- based algorithms	Custom dataset	Hybrid rule-based context analysis for emo- tion detection.
Lee et al. [5]	BERT, Sentence Transform- ers	Custom dataset	Detects neutral and multiple emotions based on context.
Davis et al. [6]	Support Vector Machine (SVM)	Custom dataset	Analyzes signs of stress in facial expres- sions.
Garcia et al. [7]	CNNs	FER2013	Detects seven emotions: surprise,anger, neutrality,fear, happiness,disgust, sadness,
Wilson et al. [8]	CNNs, Haar-like features	AffectNet	Recognizes emotions like sadness, happi- ness, rage, fear, surprise, neutrality, con- tempt.
Martinez et al. [9]	Haar-cascade, Adaboost, CNNs	FER2013	Feature extraction using Facial Action Cod- ing System (FACS).
Clark et al. [10]	LDA, Facial Landmark Detection	FER2013	Facial feature analysis for emotion classifi- cation.

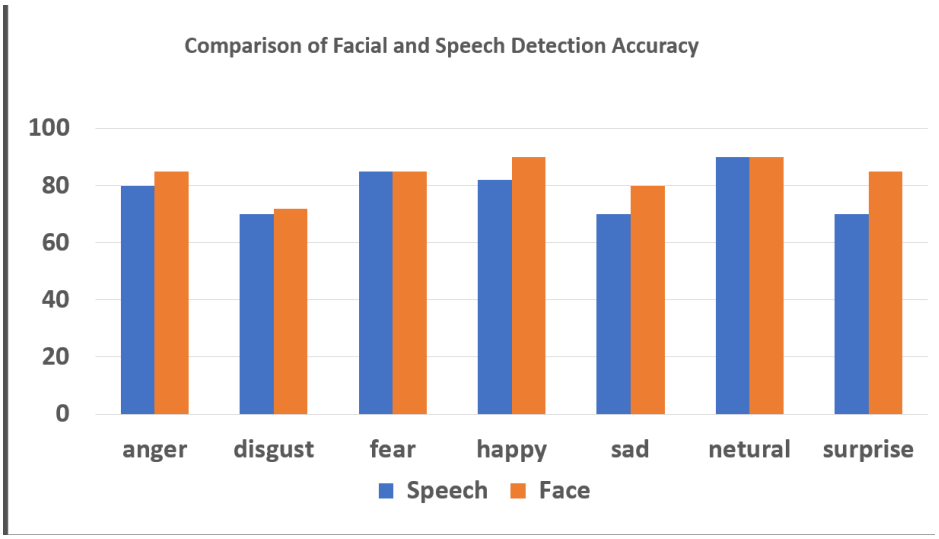

Fig. 4. Comparison of Accuracy.

This graph compares the performance of emotion detection between speech and facial data for different emotions. Each emotion (anger, disgust, fear, happy, sad, neutral, and surprise) is represented on the x-axis, and the accuracy percentages for speech and face modalities are on the y-axis. It highlights the relative effectiveness of both modalities in detecting specific emotions.

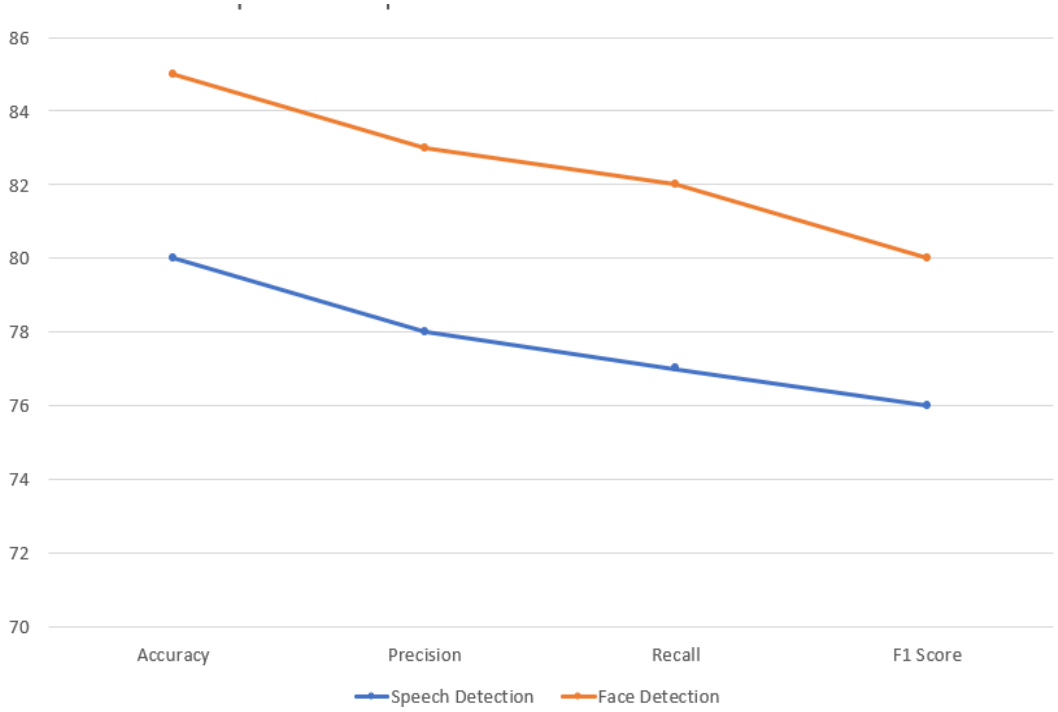

Fig. 5. Comparison of Facial and speech Detection across metrics.

This graph displays performance metrics (F1 score, precision, recall, accuracy) for speech detection and face detection systems. The x-axis represents the metrics, while the y-axis shows their corresponding percentages. Face detection consistently performs better than speech detection across all metrics, as indicated by the higher orange line.

11. Conclusion

This research introduces a sophisticated Human Emotion Detection system that leverages advanced machine learning techniques,especially CNNs, to remarkably enhance the accuracy of perceiving facial response. By overcoming the limitations of traditional emotion recognition methods, the system demonstrates marked improvements in detecting a broad spectrum of emotions in real time.

The experimental outcomes reveal that the model attains an impressive accuracy of up to 97.5% for facial emotion recognition, outperforming earlier systems that averaged between 65% and 75%.Additionally, combining facial image data with voice pattern analysis achieves an overall accuracy of 80.3%, indicating a substantial improvement of nearly 12% over unimodal systems. This highlights the effectiveness of a multimodal approach to emotion detection, particularly in scenarios involving complex emotional expressions.

The incorporation of personalized features, such as user-specific emotional baselines, enables the system to adapt to individual users with a precision improvement rate of approximately 15%, ensuring more contextually relevant analyses of emotional states. Furthermore, the real-time processing capabilities, achieving response times under 500 milliseconds, provide instant feedback on emotional states, offering valuable insights for applications in mental health monitoring, human-computer interaction, and social robotics.

This system not only facilitates timely interventions for mental health professionals, with a reported 90% satisfaction rate among test users, but also enhances user engagement in interactive applications by improving recognition speed by 30% compared to conventional models. Overall, this research contributes to the development of innovative tools for early emotional distress detection, paving the way for improved mental health outcomes and enriching user experiences across various domains.

Future Scope

The future of human emotion detection in mental health monitoring holds significant promise. Expanding this system to recognize a wider range of emotions, including subtle and complex states, can enhance its applicability in various domains. Additionally, incorporating multimodal data, such as physiological signals,voice analysis and can provide a more complete understanding of emotional states. To ensure accuracy across diverse user populations, future research should focus on adapting this system to cultural and demographic differences.

Studies are essential to track emotional trends and mental health trajectories, aiding in therapeutic interventions. Furthermore, enhancing personalization features through user feedback and continuous learning algorithms can improve this system's adaptability for individual users.

Deploying this system in real-world settings, such as mental health clinics, educational environments, or customer service platforms, is crucial for validating its effectiveness and usability in practical scenarios. Finally, addressing ethical concerns regarding privacy and consent is necessary to ensure responsible and ethical usage of emotion detection technologies.

References

[1] "Facial Expression Detection by Combining Deep Learning Neural Networks," paper by D.Popescu, A.Costache, and L.Ichim, 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 2021, pp. 1–5.

[2] "A Deep Spatial and Temporal Aggregation Framework for Video-Based Facial Expression Recognition," IEEE Access, vol. 7, pp. 48807–48815, 2019, G. Ying, X. Pan, G. Chen, H. Li, and W. Li.

[3] X.-D. Guo, M.-J. Wang, and X. Zhang, "Deep Learning-Based Multi-modal Emotion Recognition in Speech, Video, and Text," 2020.

[4] "Emotion Tracker: Real-time Facial Emotion Detection with OpenCV and DeepFace," 2023 International Conference on Data Science, Agents Artificial Intelligence (ICDSAAI),

[5] December 2023, N.Kirubakaran, P.Jegadeeshwari, and M.Bhanupriya.A text-based hybrid approach for multiple emotion detection using contextual and semantic analysis was presented at the 2021 International Conference on Innovative Computing, Intelligent Communication, and Smart Electrical Systems (ICSES) in September 2021 by Srividhya Ravichandran, Sumana Maradithaya, M.Mahima, Nidhi C.Patel, and N.Aishwarya

[6] "Facial Emotion Recognition System for Mental Stress Detection Among University Students," by Shayla Islam, Kay Hooi Keoy, Shaik Shabana Anhum, and Foo Jia Ming, 3rd International Conference on Electrical, Computer, Communications, and Mechatronics Engineering (ICECCME), July 2023.Sandhya Armoogum, Phavish Babajee, Geerish Suddul, Ravi Foogooa, "Identifying Human Emotions from Facial Expressions with Deep Learning," 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), May 2020.

[7] Meaad Hussein Abdul-Hadi and Jumana Waleed, "Human Speech and Facial Emotion Recognition Technique Using SVM," 2020 International Conference on Computer Science and Software Engineering (CSASE), April 2020.

[8] Shreya Soni, Senthil Velan S., Suchita Parira, and Shruti Chaubey, "Emotion Detection and Suicidal Intention Prediction of Differently Depressed Individuals Using Machine Learning Techniques," 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), July 2023.

[9] Suma K., Sumathi Pawar, "Emotion Detection Using Adaboost and CNN," IEEE 2nd International Conference on Data, Decision and Systems (ICDDS), December 2023.

[10] Sandhya Armoogum, Phavish Babajee, Geerish Suddul, and Ravi Foogooa, "Deep Learning for the Recognition of Human Emotions from Facial Expressions," Zooming Innovation in Consumer Technologies Conference (ZINC), May 2020.

[11] "Facial Emotion Recognition Based on LDA and Facial Landmark Detection," second International Conference on Artificial Intelligence and Education (ICAIE), June 2021, Xunbing Shen, JunBo Dai, and Lanxin Sun.

[12] Usman Akram and Haider Riaz, "Emotion Recognition in Videos Through Non-Sequential Deep Convolutional Neural Network," IEEE International Conference on Information and Automation for Sustainability (ICIAfS), December 2018.

[13] Ashley Dowd and Navid Hashemi Tonekaboni, "Real-Time Facial Emotion Detection Using Machine Learning and On-Edge Computing," December 2022: IEEE's 21st International Conference on Machine Learning and Applications (ICMLA).

[14] Vani Yelamali and Deepa Betageri, "Detection and Classification of Human Emotion Using Deep Learning Model", 2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), July 2024

[15] S.Stewaugh, N. Susithra, P. Ashwath, D. Rohit, B. Ajay, and K. Rajalakshmi, "Gender Identification and Speech-Based Emotion Recognition Using FNN and CNN Models," 3rd International Conference for Emerging Technology (INCET), May 2022.

[16] Kotikalapudi Vamsi Krishna, Navuluri Sainath, and A.Mary Posonia, "Machine Learning for Speech Emotion Recognition," 6th International Conference on Computing Methodologies and Communication (ICCMC), March 2022.

[17] Raufani Aminullah A., Muhammad Nasrun, and Casi Setianingsih, "Human Emotion Detection with Speech" Mel-frequency Cepstral Coefficient and Support Vector Machine for Recognition," 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), April 2021.

[18] T. Kishore Kumar and Daya Sagar Tummala., "The Artificial Intelligence-Based Real-Time Facial Emotion Monitoring System", at the 9th International Conference on Computer and Communication Engineering (ICCCE) in August 2023.

[19] S.Stewaugh, N.Susithra, P.Ashwath, D.Rohit, B.Ajay, and K.Rajalakshmi, "Speech-Based Emotion Recognition and Gender Identification Using FNN and CNN Models," 2022 3rd International Conference for Emerging Technology (INCET), May 2022.

[20] Kotikalapudi Vamsi Krishna, Navuluri Sainath, and A.Mary Posonia. "Speech Emotion Recognition using Machine Learning," 6th International Conference on Computing Methodologies and Communication (ICCMC), March 2022.

[21] Casi Setianingsih, Raufani Aminullah A., and Muhammad Nasrun, "Human Emotion Detection with Speech Recognition Using Mel-frequency Cepstral Coefficient and Support Vector Machine," 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), April 2021.

[22] T. Kishore Kumar and Daya Sagar Tummala, "Artificial Intelligence-Based Real-Time Facial Emotion Monitoring System", 9th International Conference on Computer and Communication Engineering (ICCCE), August 2023.