Library Progress International
Vol.44 No.3, Jul-Dec 2024: P. 23073-23078

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

EmotiSense: Enhancing Information Accessibility and User Experience through Multimodal Emotion Recognition for Individuals with Disabilities

Dhruvil Shah[1], Isha Shah[1], Smriti Raman[1], Sahil Shah[1], Komal Patil[2], Aruna Gawade[2], Nilesh Rathod[2], Angelin Florence[2]

[1]Authors,Department of Artificial Intelligence and Machine Learning,
SVKM's Dwarkadas Jivanlal Sanghvi College of Engineering,
Mumbai,India,400056;
[2]Faculty,Department of Artificial Intelligence and Machine Learning,
SVKM's Dwarkadas Jivanlal Sanghvi College of Engineering,
Mumbai,India,400056;

How to cite this article: Dhruvil Shah, Isha Shah, Smriti Raman, Sahil Shah, Komal Patil, Aruna Gawade, Nilesh Rathod, Angelin Florence (2024) EmotiSense: Enhancing Information Accessibility and User Experience through Multimodal Emotion Recognition for Individuals with Disabilities. *Library Progress International,* 44(3), 26333-26352

ABSTRACT

Effective emotional recognition is critical for improving information access and user experience, particularly for individuals with disabilities who may face challenges in verbal communication. This study presents EmotiSense, a multimodal deep learning approach that integrates audio and visual data to enhance emotion recognition in individuals with special needs. By leveraging Long Short-Term Memory (LSTM) networks for speech patterns and Convolutional Neural Networks (CNN) for facial expressions, EmotiSense offers an innovative solution to monitor emotional well-being in non-verbal interactions. The system aims to support librarians, educators, and caregivers by automating the detection of emotional shifts, thus improving accessibility to digital library services and educational resources. The preliminary results (LSTM: 73.6%, CNN: 65.28%) demonstrate the potential of this approach to enhance user experience and provide tailored support, contributing to more inclusive information environments.

Keywords: Deep Learning, Artificial Intelligence, Emotion Recognition, Real Time Models, Multimodal Deep Learning, Assistive Technologies.

Introduction

Behavior analysis is a collection of approaches, procedures, and instruments for identifying and obtaining subjective data from language, including opinions and attitudes. Behavior analysis has historically focused on an individual's opinion polarity, or whether they have a favorable, neutral, or unfavorable opinion about something. Usually, a product or service whose online evaluation has been made public has been the subject of behavior study. This may help to explain why behavior analysis and opinion mining are sometimes employed interchangeably, even if feelings are better understood as highly charged opinions. One of the uses of behavior analysis was leveraged in EmotiSense in enhancing the lifestyle of the specially abled people catering to their needs. Behavior analysis is essential for monitoring the specially abled children as it aids in interpreting non-verbal cues, detecting emotional distress in real-time, and providing personalized support and interventions tailored to their specific needs, ultimately fostering emotional well-being and development.

1.1. The Evolution of Sentiment Analysis:

This section describes how Sentiment analysis has evolved over time and is one of the most significant technique in helping understand different perceptions. Traditionally, sentiment analysis has primarily focused on analyzing written text to determine the underlying sentiment conveyed by the words. Opinion mining, or sentiment analysis, first appeared as an area of NLP in the early 2000s. The terms "sentiment analysis" and "opinion mining" were first used in reputable contexts in 2003. Determining the polarity of textual data and categorizing it as positive, negative, or neutral was at the time the main objective of sentiment analysis. In the past, sentiment analysis systems were based on rule-based techniques that measured sentiment using pre-compiled lists of positive and negative terms. It became evident that human emotions are more complex than this three-way categorization, even while the fundamentals of sentiment analysis concentrated on identifying text as positive, negative, or neutral. The next development in sentiment analysis was the capacity to recognize emotion and emotional overtones in conversations. Beyond sentiment polarity, emotion identification uses specific textual emotions to be identified. These emotions can range from sarcasm, satire, humor, and joy to surprise, fear, and disgust; in short, they can incorporate any emotional undertones that textual data might contain(Mäntylä et al.,2018).

1.2. Role of Deep Learning Models

The advent of deep learning has revolutionized behavior analysis, enabling the development of more robust and accurate models. Deep learning models, particularly neural networks, have demonstrated superior performance in various NLP tasks, including sentiment analysis (Ranjan, R., & A.K., D., 2023). These models leverage multiple layers of interconnected neurons to automatically learn hierarchical representations of text data, capturing complex patterns and relationships within the deep learning.(Jemai, Fatma et al., 2021) Along with the success of deep learning in many other application domains, deep learning is also popularly used in sentiment analysis in recent years (Singh, N., & Jaiswal, U.C., 2023).

1. 1.2.1. Convolutional Neural Networks

Convolutional neural networks, or CNNs, are a unique kind of feedforward neural networks that were first used in computer vision. The human visual cortex, an animal brain's visual process, served as inspiration for its design.

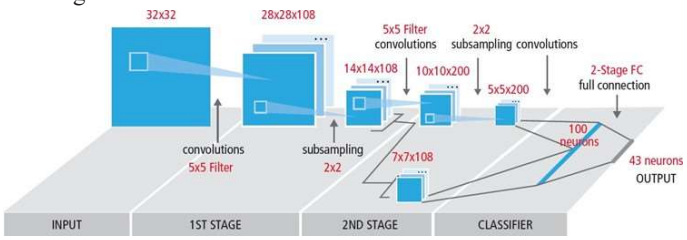


Figure 1. CNN Architecture

An image of 32 by 32 pixels for width and height and 1 pixel for input channel makes up the input. The image is scanned using the filter (size 5x5x1) in this initial step. Any region of the input image that the filter projects onto is known as a receptive field. Actually, a range of numbers (referred to as weights or parameters) make up the filter. Element-wise multiplications occur when the filter slides by multiplying its weight values by the original pixel values of the image. The single number obtained by adding all of the multiplications together is the representation of the receptive field. Every responsive field produces a number. Each field that is receptive generates a number. An array with dimensions of 28x28x1—also known as the activation map or feature map—is produced once the filter has finished sifting through the image.In Figure 1, 108 kinds of filters and thus have 108 stacked feature maps in the first stage, which consists of the first convolutional layer. Following the convolutional layer, a subsampling (or pooling) layer is usually used to progressively reduce the spatial size of the representation, thus to reduce the number of features and the computational complexity of the network(Zhang *et al*.,2018).

$$\eta_{i,j} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \Psi_{m,n} \cdot \chi_{i+m,j+n} + \omega$$

$$\zeta_{i,j} = \max(0, \eta_{i,j})$$

$$\rho_{i,j} = \max_{(m,n)\in R_{i,j}} \zeta_{m,n}$$

$$\lambda_i = \gamma \cdot \frac{\rho_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

$$\phi_k = \frac{e^{\lambda_k}}{\sum_{j=1}^{K} e^{\lambda_j}}$$

These equations form the backbone of a Convolutional Neural Network (CNN) processing pipeline. Starting with the convolution operation (η), which extracts spatial features using kernel Ψ on input χ, the signal then passes through a ReLU activation (ζ) to introduce non-linearity. The feature maps are then spatially downsampled through max pooling (ρ) to reduce dimensionality while retaining important features. Batch normalization (λ) stabilizes the learning process by normalizing intermediate outputs, and finally, the softmax function (φ) transforms the network's output into class probabilities. Together, these operations enable the CNN to hierarchically learn and extract meaningful features from input data (typically images) for tasks like classification, detection, or segmentation. These equations form the backbone of a Convolutional Neural Network (CNN) processing pipeline. Starting with the convolution operation (η), which extracts spatial features using kernel Ψ on input χ, the signal then passes through a ReLU activation (ζ) to introduce non-linearity. The feature maps are then spatially downsampled through max pooling (ρ) to reduce dimensionality while retaining important features. Batch normalization (λ) stabilizes the learning process by normalizing intermediate outputs, and finally, the softmax function (φ) transforms the network's output into class probabilities. Together, these operations enable the CNN to hierarchically learn and extract meaningful features from input data (typically images) for tasks like classification, detection, or segmentation.

1.2.2. Recurrent Neural Networks

A family of neural networks known as recurrent neural networks (RNNs) have connections between their neurons that create a directed loop. RNNs are popular for processing sequential information because, in contrast to feedforward neural networks, they can process a sequence of inputs by using an internal "memory." In other words, the RNN "memory" is the ability to execute the same operation for each element in a sequence, with each result depending on every prior computation—akin to "remembering" details of the past data that has been processed.

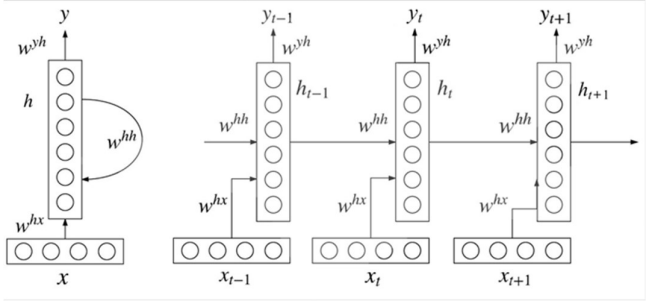


Figure 2. RNN Architecture

Figure 2: An illustration of an RNN. A folded sequence network with three time steps is shown on the right graph, while an unfolded network with cycles is shown on the left. The length of input determines the duration of time steps. For instance, the RNN would unfold into a neural network with six time steps or layers if the word sequence

to be processed consisted of a sentence of six words (Zhang *et al*.,2018).

$$\theta_t = \tanh(U_\theta x_t + W_\theta \theta_{t-1} + b_\theta)$$
$$\psi_t = \sigma(U_\psi \theta_t + b_\psi)$$
$$\mu_t = \tanh(U_\mu x_t + W_\mu \theta_{t-1})$$
$$\nabla_t = \prod_{k=1}^{t} W_\theta \mathrm{diag}(1 - \theta_k^2)$$
$$\Delta_t = \sum_{k=t}^{T} \nabla_k \cdot \frac{\partial \mathcal{L}}{\partial \theta_k}$$

These equations define a Recurrent Neural Network's (RNN) operation where temporal dependencies are captured through recursive processing. The network maintains a hidden state θ that gets updated at each time step using input x and previous state information. The output ψ is computed using the current hidden state through a sigmoid activation σ. The cell memory μ tracks information flow across time steps. The gradient equations (∇,Δ) handle the backward pass during training, addressing the challenges of learning long-term dependencies through backpropagation through time (BPTT). This architecture is particularly suited for sequential data processing, like text or time series, where temporal relationships are crucial for understanding the underlying patterns.

1.1

1.2 1.2.3. Long Short Term Memory Networks

A popular recurrent neural network (RNN) architecture in deep learning is called LSTM (Long Short-Term Memory). Identifying long-term dependencies is one of its strong points, which makes sequence prediction jobs ideal for it. Unlike standard neural networks, LSTM can handle entire data sequences rather than just individual data points because it has feedback connections. For this reason, it excels in recognizing and predicting patterns in sequential data, including time series, text, and voice. LSTM has grown into a powerful tool in deep learning and artificial intelligence that is enabling advancements across a wide range of domains because it can extract meaningful insights from sequential data.

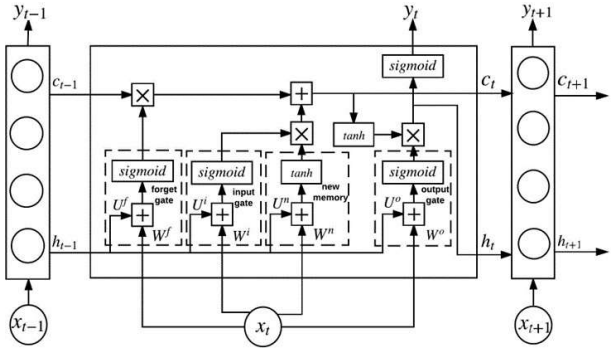


Figure 3. LSTM Architecture

Figure 3 shows an example of LSTM. At time step t, LSTM first decides what information to remove from the cell state. This decision is made by a sigmoid function/layer σ, called the forget gate. The function takes h_{t-1} (output from the previous hidden layer) and x_t (current input), and outputs a number in [0, 1], where 1 means keep and 0 means remove (Zhang *et al*.,2018).

$$\alpha_t = \Phi(U_\alpha z_t + R_\alpha S_{t-1} + K_\alpha)$$

$$\beta_t = \Phi(U_\beta z_t + R_\beta S_{t-1} + K_\beta)$$

$$\gamma_t = \Phi(U_\gamma z_t + R_\gamma S_{t-1} + K_\gamma)$$

$$M_t = \beta_t \odot M_{t-1} + \alpha_t \odot \tanh(U_m z_t + R_m S_{t-1} + K_m)$$

$$S_t = \gamma_t \odot \tanh(M_t)$$

These equations describe the core components of a Long Short-Term Memory (LSTM) network:

The equations represent three gates (α, β, γ) and two states (M, S), where:

- α_t (Input gate): Controls what new information is stored in memory
- β_t (Forget gate): Decides what information to discard from memory
- γ_t (Output gate): Determines what parts of memory are used in output
- M_t (Cell state): The network's long-term memory storage
- S_t (Hidden state): The network's current output state

Each gate uses a sigmoid function (Φ) to squash values between 0 and 1, acting as a filter. The equations combine current input (z_t), previous hidden state (S_{t-1}), and learned parameters (U, R, K) to make these decisions. The ⊙ symbol represents element-wise multiplication, allowing gates to selectively pass or block information flow. This architecture helps manage the vanishing gradient problem and enables learning of long-term dependencies in sequential data.

1.3. Leveraging Neural Networks in Behavior Analysis

Multimodal behavior analysis integrates information from different modalities such as text, audio, and visual cues to extract sentiment. Leveraging neural networks, specifically recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks for audio processing, and convolutional neural networks (CNNs) for facial feature analysis, has emerged as a promising approach in this domain.

Audio data, such as speech, contains valuable emotional cues that can contribute to behavior analysis (Russell Li and Zhidong Liu,2020). RNNs and LSTMs are particularly effective in processing sequential data like audio. In this context, spectrograms or other audio representations are fed into RNNs or LSTMs, which learn temporal dependencies and capture patterns in the audio data over time. These models can capture nuances in tone, intonation, and other acoustic features indicative of sentiment. Facial expressions are powerful indicators of emotion and sentiment. CNNs excel at processing spatial data and have been successfully applied to analyze facial features for sentiment analysis(Zhao Cheng Huang *et al*.,2019). By inputting facial images or video frames into CNN architectures, features such as facial expressions, eye movements, and gestures are automatically extracted and analyzed. CNNs can learn hierarchical representations of facial features, enabling them to discern subtle emotional cues and infer sentiment.

In summary, EmotiSense represents a paradigm shift in how behavior analysis is applied to support special needs children. By leveraging multimodal data processing, neural networks, and real-time monitoring (K. Vasanth *et al*.,2022), EmotiSense provides a holistic solution for monitoring and supporting the emotional well-being of special needs children, empowering caregivers and educators to provide personalized and timely assistance.

2. Related Work

Emotion detection, also known as emotion recognition, is a fascinating field that involves identifying and analyzing human emotions through multiple channels, such as facial expressions, speech patterns, and even body movements. Machine learning models for this purpose have become increasingly popular and advanced, utilizing various architectures to interpret human emotions effectively.

The paper "Facial Expression Recognition Through Machine Learning" explores the complexities of automatic facial expression recognition, emphasizing its significance in human-machine interfaces, behavioral science, and clinical applications. Classification is executed through the k-Nearest Neighbor algorithm, achieving a notable maximum accuracy of 90% (Qudoos, Abdul., 2016).

The paper "Stress and Anxiety Detection through Speech Recognition Using Deep Neural Network" delves into the impact of stress, a pervasive emotional tension affecting mental health. The paper outlines the utilization of a vocal/audio dataset from Kaggle to detect stress and anxiety using a developed deep neural network model, specifically a CNN (Divyashree, P. *et al.,* 2022).

The paper "Applying Image Processing Technology to Monitor the Disabilities" presents a novel approach to enhancing security for individuals with disabilities who live alone by introducing a home care service that detects potential accidents such as falls. It marks these regions of interest (ROI) to distinguish them from background or obstacles, and further assesses whether the human body is in a prone position and whether the area is safe.(Huang, Yu-Xian & Chung, Yi-Nung., 2014)

The aging population in Western societies has prompted the need for automated 24/7 surveillance models to ensure the safety of the elderly while respecting their privacy, presenting a significant challenge (Fleck, Sven & Straßer, Wolfgang, 2008).. A prototype model deployed in a home for assisted living has been operational round the clock for several months, demonstrating promising performance in addressing these challenges.

The paper "Analysis of Trustworthiness Recognition models from an aural and emotional perspective introduces a novel deep learning-driven multimodal fusion for automated deception detection, it incorporates audio cues alongside visual and textual cues. The proposed deep convolutional neural network (CNN) approach outperforms state-of-the-art methods, achieving a remarkable 96% accuracy compared to the recent literature's 82%(Luna-Jiménez *et al*.,2022).

The paper "A Study of Sentiment Analysis: Concepts, Techniques, and Challenges" describes how Sentiment Analysis (SA) involves extensively exploring web-stored data to categorize expressed views in text, aiming to evaluate the author's attitude toward a specific topic, movie, or product. Results are classified as positive, negative, or neutral(Aqlan *et al*.,2019).

The paper by Cristina Luna-Jiménez, David Griol, Zoraida Callejas, and others proposed a model for emotion recognition using both speech and facial cues. They employ transfer learning techniques for speech-based emotion recognition, finding that fine-tuning the CNN-14 model from the PANNs framework yields the best performance. For facial emotion recognition, they introduce a framework involving a pre-trained Spatial Transformer Network and a bi-LSTM with an attention mechanism. Combining these modalities with a late fusion strategy, they achieve an 80.08% accuracy on the RAVDESS dataset, classifying eight emotions through subject-wise 5-fold cross-validation. However, they note limitations, such as challenges faced by frame-based models in video-based tasks, indicating the need for further research to address these issues and optimize the use of pre-trained models(Luna-Jiménez C *et al*.,2021).

The paper "Personalized Emotion Detection using IoT and Machine Learning" focuses on personalized emotion detection using IoT and machine learning techniques, achieving up to 92% accuracy in detecting emotions under normal resting conditions(Jothiraj *et al*.,2022).

The paper "Emotion Regulation and Intellectual Disability. Journal of Developmental Disabilities" discusses the absence of literature at the intersection of intellectual disability and emotion regulation despite a substantial body of research on each individually. It highlights the practitioners' observations of the importance of emotion regulation in therapy sessions with individuals having mild mental retardation or borderline intellectual functioning(Mcclure *et al*.,2009).

The paper " Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions" proposes a model that integrates emotion recognition technologies, commonly used in human-computer interaction (HCI).It utilizes cutting-edge technology, including tangible user interfaces

(TUIs) and facial expression recognition, enabling natural interaction for children by grasping objects and using facial expressions. The model was evaluated in collaboration with an association for children with ASD, with results indicating positive outcomes and validating the effectiveness of the approach(Garcia-Garcia *et al*.,2022).
The paper “Modalities of monitoring: Evidence from cameras and recorders in policing” addresses principal-agency problems in policy implementation environments where authority is delegated to field agents, focusing on the challenges of adverse selection and moral hazard. This literature review provides insights into the complex dynamics of monitoring strategies and their implications for addressing principal-agency problems in various policy contexts(Andrew B. Whitford *et al*.,2023).
The paper "Real-Time Abnormal Object Detection for Video Surveillance in Smart Cities" presents a lightweight subclass detection method based on convolutional neural networks (CNNs) to detect guns and knives in multiview camera setups. By employing a multiclass subclass detection CNN, the method accurately classifies object frames into different subclasses, distinguishing abnormal from normal behaviors. Extensive experiments across various datasets demonstrate high mean average precision scores, with a notable precision score of 85.5% in multiview camera setups(Ingle *et al*.,2022).
Thus from the literature survey it was concluded that using CNN architecture for visual component yields better results as compared to other architectures and using LSTM architecture for audio component.

Methodology

EmotiSense represents an advancement in the field of behavior analysis tailored specifically for special needs children (Lee C. M. et al.,2002). Recognizing the unique emotional and behavioral challenges faced by these children, EmotiSense combines technologies such as neural networks, multimodal data processing, and real-time monitoring to provide comprehensive support and assistance.

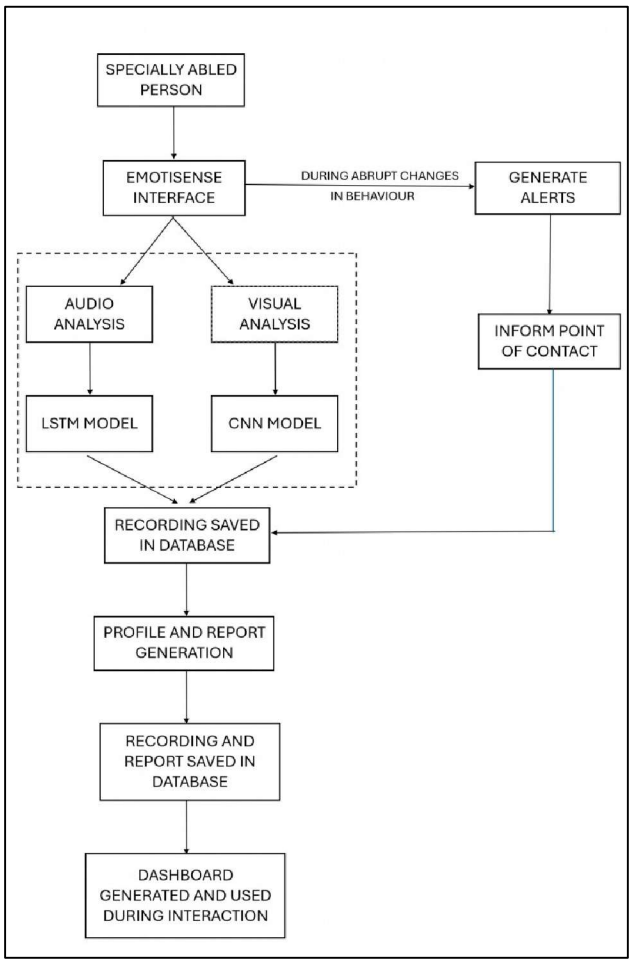


Figure 4. Overall Workflow

The model features a frontend interface for capturing audio and visual data from specially abled people. Utilizing LSTM for audio and CNN for visual analysis, it extracts pitch, intensity, and facial expressions in real-time. A comprehensive report will be generated and saved to each child's profile, documenting instances of distress over time (Rane et al.,2020). This report will serve as a valuable tool for evaluating the child's emotional well-being, tracking progress, and informing future interventions and support strategies tailored to their individual needs. Integrated backend processing generates comprehensive reports, saving recordings and analyses in a database for future reference and analysis, ensuring efficient interactions.

The proposed model was developed using the following steps: -

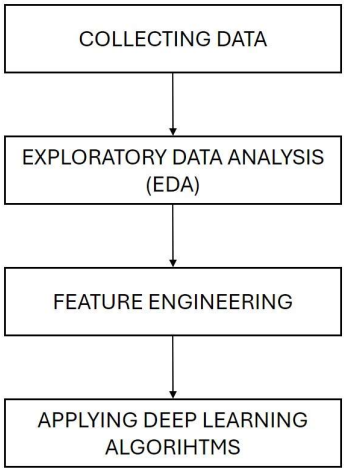


Figure 5. Chronology of audio and visual processing

3.1. Data Collection

Three audio datasets were gathered: CREMA-D (a data set comprising 7,442 original clips from 91 actors) and RAVDESS-emotional speech audio (containing 1440 files: 60 trials per actor x 24 actors = 1440 featuring 24 professional actors, 12 females, 12 male). The 48 male and 43 female actors in these films, who ranged in age from 20 to 74 and represented a range of racial and ethnic backgrounds, conveyed emotion with SURREY Audio-Visual (480 files, 120 utterances by 4 speakers) from KAGGLE. On request, the University of Oulu supplied their OULU-Casia dataset for the visual model (G. Zhao *et al*., 2011). Six facial expressions—surprise, happiness, sorrow, anger, fear, and disgust—from 80 individuals between the ages of 23 and 58 are included in the Oulu-CASIA NIR&VIS facial expression database. 73.8% of the subjects are males(G. Zhao *et al*., 2011). The subjects were asked to sit on a chair in the observation room in a way that he/ she is in front of camera

3.2. Exploratory Data Analysis

3.2.1. Audio analysis

For audio analysis, RNNs and LSTMs are employed to process speech data, extracting emotional cues from tone of voice, pitch variations, and speech patterns (S. Saraswat *et al*.,2023). This enables EmotiSense to detect subtle changes in the child's mood or emotional state, providing early intervention in case of agitation or distress (Anderson R. Avila *et al*.,2019) .

MFCC represents the spectral characteristics of sound. The vertical axis of the graph in the image represents the MFCC index, while the horizontal axis represents time in seconds. Here, a darker color indicates a higher value of coefficient, while lighter colors represent lower values. This would often be accompanied by a color bar on the side of the graph indicating the range of values represented by the color spectrum

MFCC (Mel Frequency Cepstral Coefficients) is a condensed representation of the waveforms represented by accumulating the huge number of sinusoids of an audio signal. MFCC coefficients contain details about the variations in the spectrum bands.

In Fig. 3, MFCC Spectrogram representation, the x-axis represents time, spanning from 0 to 3.5 seconds. The

color intensity across the timeline depicts the power spectrum of the sound at each point in time, with darker colors indicating lower coefficient and brighter colors signifying higher coefficient value. A color bar beside the spectrogram provides a scale of intensity ranging from -600 to 100. A notable finding from the spectrogram is the significant change in intensity around the 0 to 0.5-second mark, after which it remains relatively consistent. This indicates a significant event or feature in the audio signal at that time.

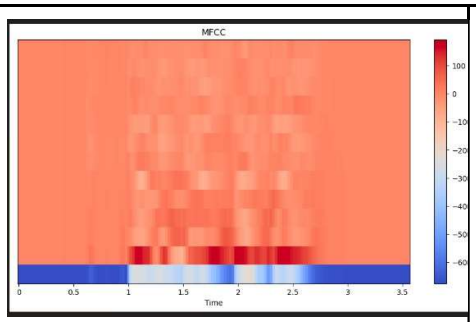
Mel Spectrograms and Hz Spectrograms are visualized with time in seconds on the x-axis and frequency in Hertz on the y-axis, with color intensity indicating amplitude in decibels. Chroma plots also use time in seconds on the x-axis but represent pitch classes on the y-axis, with amplitude in decibels as the color intensity. Contrast features are plotted similarly to Mel Spectrograms, with time in seconds on the x-axis, frequency in Hertz on the y-axis, and contrast in decibels as the color intensity. MFCC and MFCC Spectrogram plots use time in seconds on the x-axis, MFCC index on the y-axis, and amplitude in decibels as the color intensity. Finally, the Amplitude Waveform plot uses time in seconds on the x-axis and amplitude in decibels on the y-axis.

In Amplitude Waveform image, the horizontal axis, labeled “Time” in the image, shows the time progression of the sound wave. The vertical axis, labeled “Amplitude” in the image, represents the intensity or loudness of the sound wave. The higher the vertical displacement from the center line (horizontal axis), the greater the amplitude and the louder the sound. In this specific image, the waveform shows a complex periodic sound wave. This means the sound has a regular repeating pattern (periodic), and consists of multiple frequencies at the same time (complex).

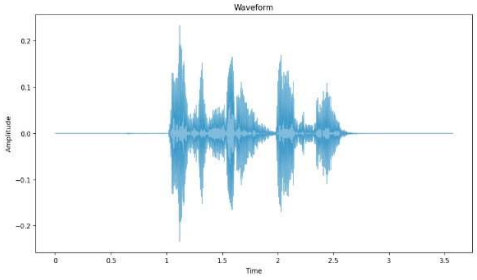
In the graph for chroma and contrast, the horizontal axis, like before, acts as a timeline, showing the speech unfolding over time (usually in seconds or milliseconds). The vertical axis still represents pitch, which conveys intensity levels. Darker areas indicate stronger or more prominent tonal regions within a specific time window of the speech. Lighter areas represent weaker or less prominent regions.

These features represent the speaker's vocal tract characteristics, which are helpful for distinguishing speech sounds (Alim *et al*.,2018). It can benefit in learning about the unique tonal characteristics reflected by the intensity patterns where variations in pitch can be visualized and the speech's emotional state (e.g., anger or sadness) can be reflected.

Table 1: Feature extraction

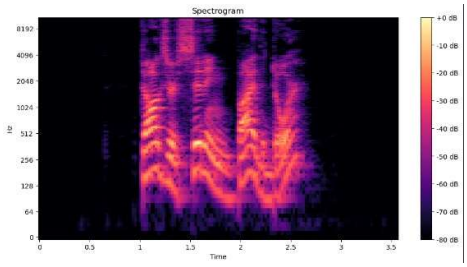


2. **MFCC Spectrogram**

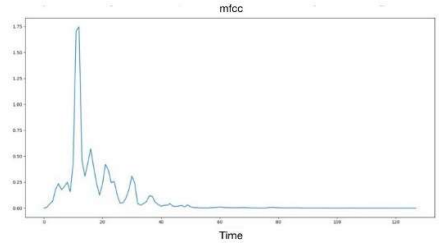


3. **Amplitude Waveform**

4.

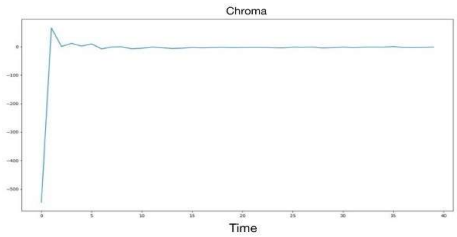


5. **Hz Spectrogram**

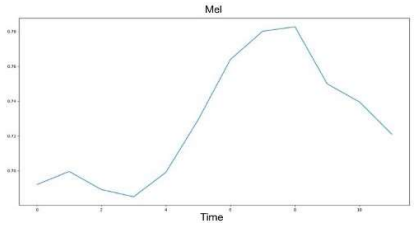


Graph for MFCC

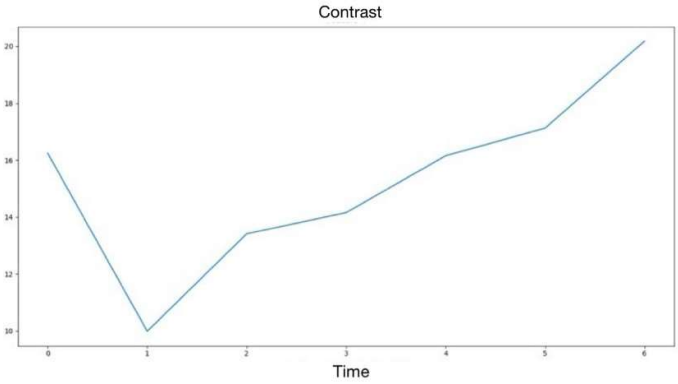
6.



Graph for Chroma



Graph for Mel



Graph for Contrast

In all the above graphs, on vertical axis amplitude of the features plotted and time interval on the horizontal axis. All the features were extracted and plotted as a part of the EDA to analyze how audio signals at different times cause a fluctuation in the features which is used to accurately determine the emotion of the speaker.

3.2.2. Visual Analysis

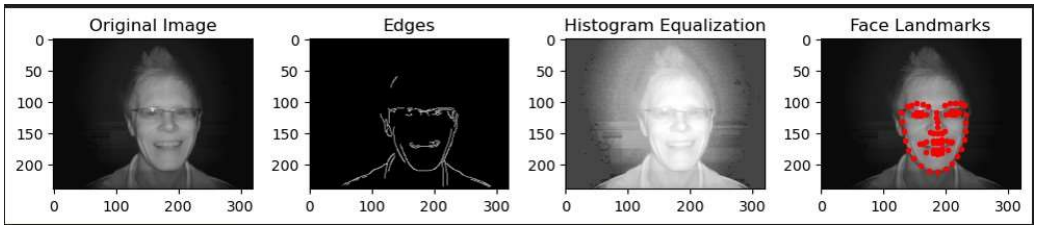


Figure 6. Visual Feature Extraction

In Fig.6 each section of the image shows a grayscale image of a human face with a box drawn around the face. The "Edges" section appears to show an edge detection filter applied to the image, highlighting the outlines of facial features. The "Histogram Equalization" section appears to show the image after histogram equalization, a technique used to improve the contrast in an image. The "Face Landmarks" section appears to show facial recognition software that has identified and marked key points on the face such as the eyes, nose, and mouth.

3.3. Feature Engineering

This research tried different set of features and finally got the best results with features MFCC, chroma, contrast, centroid, zero crossing rate, pitch and intensity.

MFCC: Using a method similar to the Discrete Fourier Transform (DFT), the raw audio signal is first transformed into the frequency domain to create MFCCs. Next, the mel-scale is applied to simulate how the human ear perceives sound frequency. Ultimately, the mel-scaled spectrum is used to compute cepstral coefficients.

Chroma: A descriptor called the chroma feature condenses the tonal content of a musical audio signal.

Contrast: Small variations in speech sounds can have a significant impact on how listeners perceive a sound, which can change the way words are mentally lexically entered.
Centroid: The "center of mass" of the spectrum can be found using a measurement called the spectral centroid. Because of its strong perceptual association with the perceived "brightness" of a sound, musical timbre is described by it.
Zero crossing rate: The speed at which a signal changes from positive to negative to zero or from negative to positive to zero.
Pitch: The amount of vibrations per second that the vocal cords produce determines how high or low a tone appears to the human ear.
Intensity: It is perceived as the loudness of the sound.

3.4. Applying Deep Learning Algorithms

CNN: CNN architectures are great at extracting local features from speech spectrograms that are basically visual representations of audio. These features capture the short-term variations in pitch, intensity, and energy which are crucial for emotion recognition.
The CNN sequential model utilizes a 14-layer architecture for speech emotion recognition. It consists of convolutional layers (conv1d, conv1d_1, conv1d_2) which extracts local features from the speech data. Activation layers (activation, activation_1, activation_2, activation_3, activation_4) introduce non-linearities, allowing the model to learn more complex patterns. Pooling layers (max_pooling1d, max_pooling1d_1, max_pooling1d_2) efficiently reduce the data size while retaining key information. Dropout layers (dropout, dropout_1, dropout_2, dropout_3) help prevent overfitting by randomly dropping neurons during training. A Flatten layer transforms the data into a single vector, which is then fed into dense layers (dense with 512 units and dense_1 with 7 units for emotion prediction). This architecture boasts 1.54 million trainable parameters.

CNN & RNN: CNN & RNN architecture leverages the strengths of both CNNs and RNNs. CNNs extract local features, while RNNs handle the sequential nature of speech. This combination can lead to superior emotion recognition accuracy by capturing both short-term variations and long-term emotional flow within speech.
The CNN & RNN sequential model utilizes a 11-layer for architecture speech emotion recognition. It consists of convolutional layers (conv1d, conv1d_1) which extracts local features from the speech data. The training process is stabilized by the batch normalization layers (batch_normalization, batch_normalization_1). While keeping important information, the max pooling layers (max_pooling1d, max_pooling1d_1) effectively minimize the amount of the data. Overfitting is prevented via dropout layers (dropout, dropout_1). A Long Short-Term Memory (LSTM) layer is incorporated into this design to record long-term dependencies in speech sequences, which are essential for the gradual emergence of emotions. The model finalizes with dense layers (dense with 64 units and dense_1 with 7 units for emotion prediction). This sequential architecture ensures ordered processing, with a total of 114,247 trainable parameters for the model to learn the complexities of speech emotions.
LSTM: LSTM's are adept at capturing long-term dependencies in speech sequences. As emotions unfold over time, LSTMs learn these temporal patterns, considering not just the current sound but also the emotional context of preceding speech.
The LSTM sequential model utilizes a 6-layer for architecture speech emotion recognition. The core layer is a Long Short-Term Memory (LSTM) layer with 128 units, adept at capturing long-term emotional trends within speech sequences. Following the LSTM layer, a sequence of densely connected layers (dense_2, dense_3, dense_4) progressively transforms the features into a lower-dimensional space (64, 32, and finally 7 units) suitable for emotion classification. Dropout layers that are inserted between dense layers help prevent overfitting during training. This architecture achieves a balance between performance and 77,127 trainable parameters.
The two components of the model are:

3.4.1. Audio Component

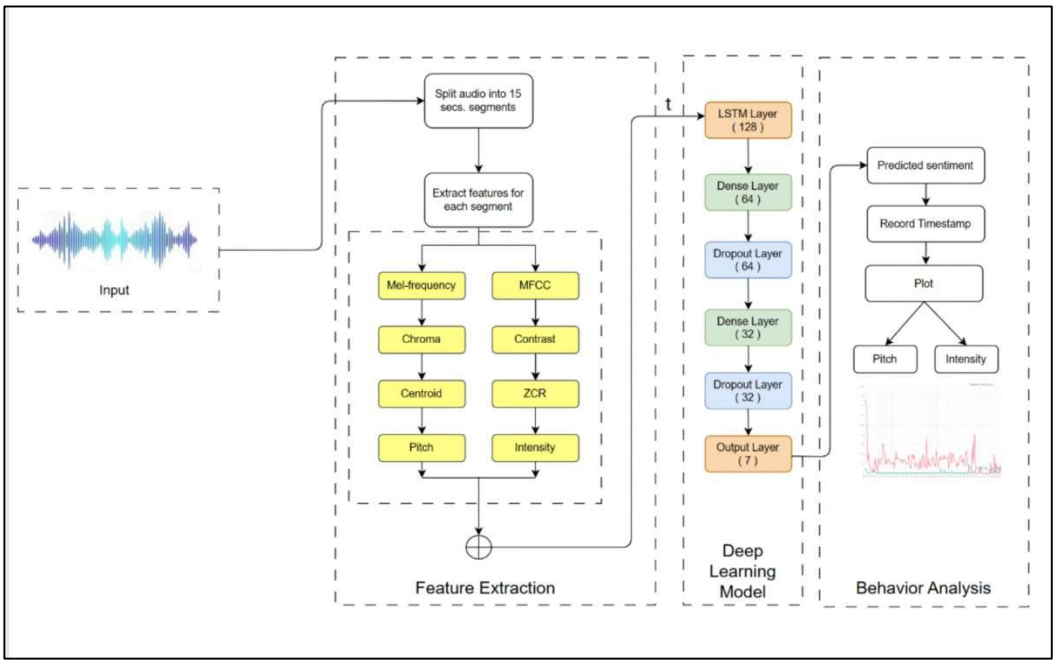


Figure 7. Audio analysis workflow

In the initial phase, audios were collected and meticulously preprocessed them, employing techniques such as noise reduction and normalization. Feature extraction followed, with a focus on essential audio features like MFCCs and RMS, crucial for capturing meaningful patterns in the dataset (Garima Sharma *et al*.,2019). Next step was transformation of these features into a format suitable for deep learning models, like ANN, CNN, and RNN. The subsequent step involved implementing these algorithms and rigorously testing their accuracy in identifying audio patterns (Maghilnan S and Rajesh Kumar M,2017). Evaluation metrics such as accuracy, precision, and recall were employed to assess model performance. Algorithm 1 describes the audio analysis process

Algorithm 1: Audio analysis

Input:

1. s : Audio file
2. a : Audio emotion recognition model
3. b : Array consisting audio emotion labels

Audio Emotion Recognition (s , a , b) **:**

1. Split s into segments each of duration 15 seconds
2. Extract features from segments , predict emotion e' and append it to predicted emotion array E
    a. Predicted emotion array E : $E \rightarrow [\,]$
    b. Time-stamps array T : $T \rightarrow [\,]$
    c. **For** each segment **Do** :
        I. $r \rightarrow [\,]$
        II. Extract following features : *chroma , mfcc , mel-frequency , contrast ,centroid , zcr , pitch , intensity*
        III. Append value of each feature to r
        IV. Reshape r such that it becomes a column vector with single depth layer
        V. $e' = a\,(r)$
        VI. $e' = b\,(\text{argmax}\,(e'))$
        VII. Append e' to E
        VIII. Record time-stamp t and append it to T
    d. **End For**

3. **Return** E , T

Argmax is a mathematical operation that identifies the argument that corresponds to the maximum value obtained from the array of predicted emotions.

LSTM layer analyzes the sequence of sounds to grasp the speaker's emotion. Final layers process this information and predicts the sentiment (happy, angry, disgust, fear, sad, surprised and neutral) of the audio (Kevin Tomba *et al*.,2018).

3.4.2. Visual Component

A multimodal model processes audio and visual inputs simultaneously, leveraging neural networks like CNNs for visual and RNNs for audio information. These networks independently extract features from each modality. Post-processing, the features are fused, combining both streams through techniques like concatenation or attention mechanisms. This fusion enhances the model's ability to capture intricate correlations between audio and visual cues. The joint representation undergoes further refinement through neural network layers, resulting in a comprehensive multimodal model. This approach proves valuable in diverse applications, offering a more nuanced understanding of input data and overcoming the limitations of individual modalities.

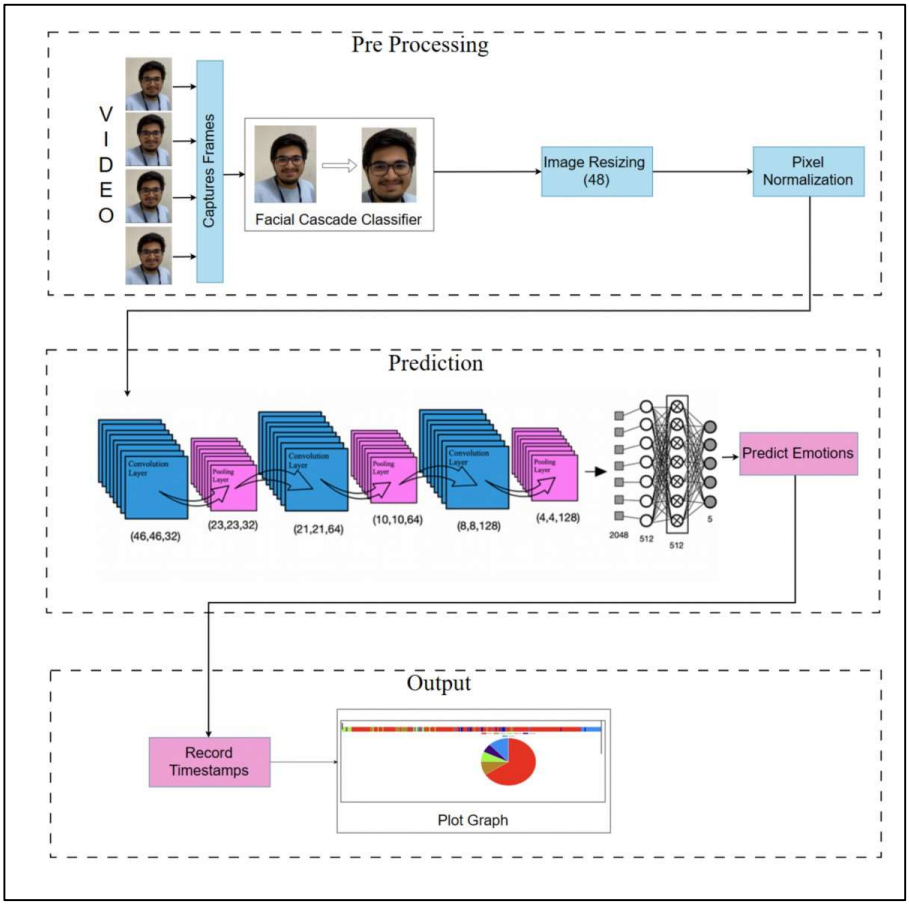


Figure 8. Visual Analysis workflow

A system called "facial emotion detection" makes use of computer vision algorithms to read facial expressions and recognize different emotional states. Through the utilization of facial landmarks and traits, such as mouth shape and eye movements, the model is able to precisely identify a wide range of emotions, including happy, sadness, rage, surprise, and more (Ghadage *et al*.,2023). Applications for this technology can be found in a number of industries, such as market research, mental health evaluations, and human-computer interface. Convolutional neural networks (CNNs), one type of deep learning model, is frequently used to obtain high accuracy in emotion

classification based on face clues. From improving user experiences to supporting the diagnosis and treatment of emotional disorders, facial emotion recognition has a broad range of applications.

First, frames are taken out of the visual input by EmotiSense. After that, a Convolutional Neural Network (CNN) architecture intended for visual analysis processes these frames. A number of convolutional layers make up the model, which gathers features from the face photos. Max-pooling layers are placed after these convolutional layers in order to down sample the feature maps and collect pertinent data. Additional convolutional and max-pooling layers are added during this process to further hone the recovered features. Feature maps from the last convolutional layer are flattened into a vector form. The model is then able to discover intricate patterns in the data by passing this vector across fully connected dense layers. In order to reduce overfitting, dropout layers are used, which randomly remove units during training. In the end, the model generates a sentiment prediction based on the learned features extracted from the facial images (K. Patel *et al*.,2020).

Fig.8 showcases a Convolutional Neural Network (CNN) architecture designed for facial sentiment analysis.The image enters first, and convolutional layers find key facial features like eyes and mouth. Max pooling shrinks the data, keeping important details. Flattened data is then interpreted by dense layers, which associate features with emotions (happy, sad, etc.). Finally, the output layer predicts the most likely sentiment based on these interpretations.

Algorithm 2 describes the visual analysis process.

Algorithm 2: Visual analysis

Input:

1. f : Live frame
2. c : Facial detection cascade classifier
3. v : Facial emotion recognition model
4. l : Array consisting facial emotion labels
5. i , j : Arbitrary offsets

Facial Emotion Recognition (*f, c, v, l, i, j*)**:**

1. Convert f to gray-scale
2. Obtain f' from f such that f' contains facial area in f :
    a. Predicted emotion array E : $E \rightarrow [\,]$
    b. Time-stamps array T : $T \rightarrow [\,]$
        I. $d = c(f)$
    c. **For** *x,y,w,h* in d **Do** :
        I. $f' = f[y+i : y+h , x+j : x+w-j]$
        II. Re-size f' to size 48
        III. Normalize f' : $f' = \frac{f'}{255}$
        IV. Predict facial emotion e :
            i. $e = v(f')$
            ii. $e = l(e)$
            iii. Append e to E
        V. Record time-stamp t and append it to T
    d. **End For**
3. **Return** E , T

3. Results

4.1. Audio Results

Three distinct models—a CNN & RNN model, a CNN model, and an LSTM network—were used for prediction tasks during the encounter. In cross-validation studies, the CNN model performed better in terms of test accuracy, but the LSTM model performed better when making predictions in real time. Hence LSTM model was finally concluded.

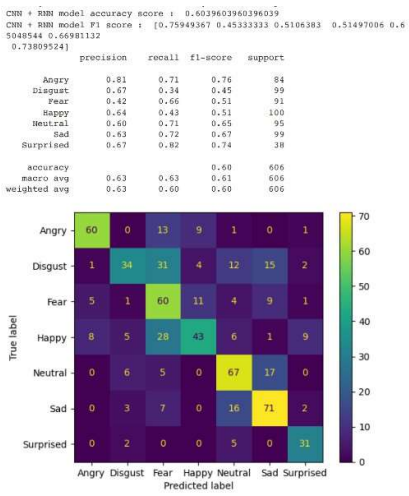
Table 2: Audio evaluation result

	CNN and RNN	CNN	LSTM
TEST ACCURACY	60.40%	87.62%	73.6%

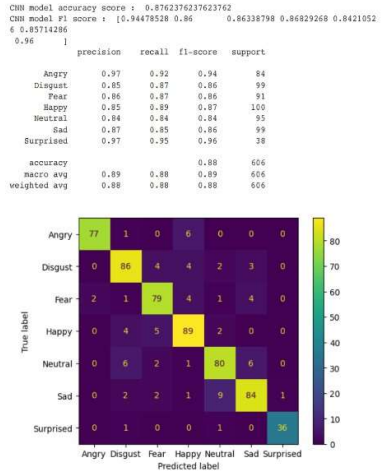
The labels audio model of EmotiSense is trained on are Angry, Disgust, Fear, Happy, Neutral, Sad and Surprised.

The following is the confusion matrix of the CNN-RNN, CNN and RNN model. The confusion matrix table compares the model's predicted emotions to the actual emotions present. It has rows representing the actual emotions and columns representing the predicted emotions. Each cell shows how often the model confused one emotion for another. Rows show true emotions and columns show emotions the model predicted. Numbers in each box tell you how often the model was right (happy predicted for happy face) or wrong (predicted happy for a sad face). This helps identify emotions the model struggles and is unable to predict.

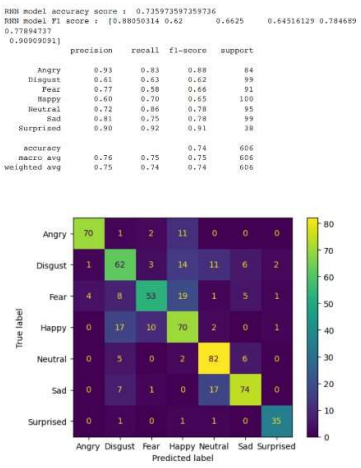
Table 3: Confusion Matrix of audio component



CNN and RNN



CNN



RNN

From the above analysis it was concluded that CNN model had the best accuracy but RNN(LSTM) model performed better on real time scenarios.

4.2. Visual Result

During interaction, two different models were employed, a CNN & improved CNN model. The observed 100% accuracy achieved by the Convolutional Neural Network (CNN) model suggests a strong likelihood of overfitting. Hence the second one which has good live prediction accuracy was selected for the visual analysis.

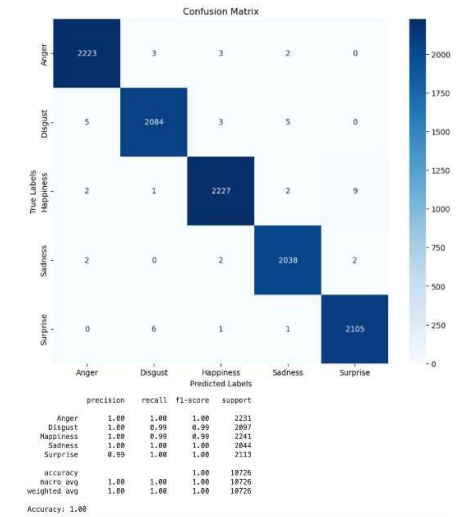
7.

Table 4: Visual evaluation Result

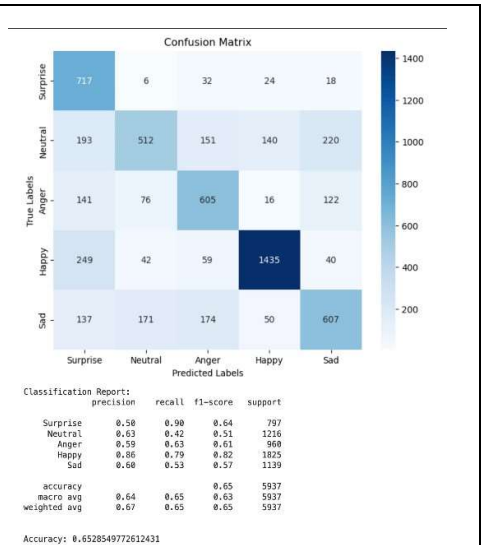
	CNN	Improved CNN
TEST ACCURACY	100%	65.28%

The labels visual model of EmotiSense is trained on are Angry, Sad, Happy, Neutral and Surprised. The following is the confusion matrix of the CNN, improvised CNN model.

Table 5: Confusion Matrix- finetuned CNN



CNN



Improved CNN

The first section of UI has a start and stop capture button after which the video and speech is recorded

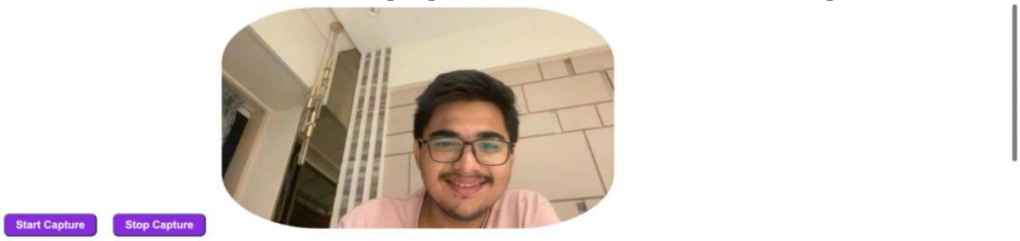


Figure 9. GUI interface

It then shows a timeline and pie chart of the emotion prediction of audio analysis as well as visual analysis.

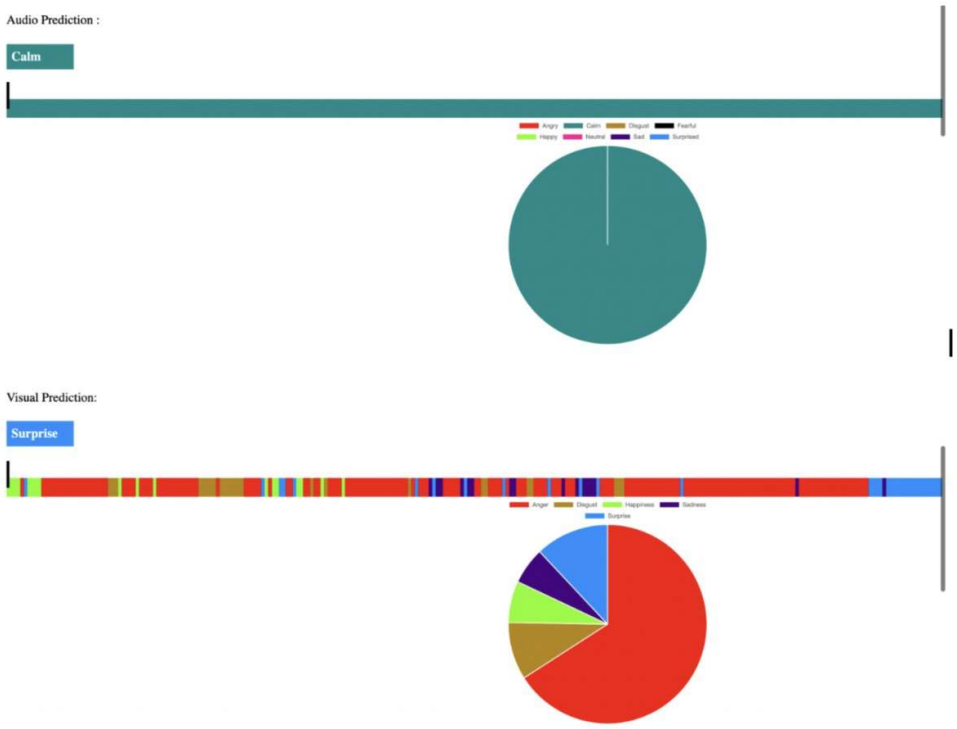


Figure 10. Audio and Visual analysis interface

The time frame is depicted as a slider which represents different emotion generated at intervals of time. All these generated emotion are represented in a form of pie chart for both audio and visual model.

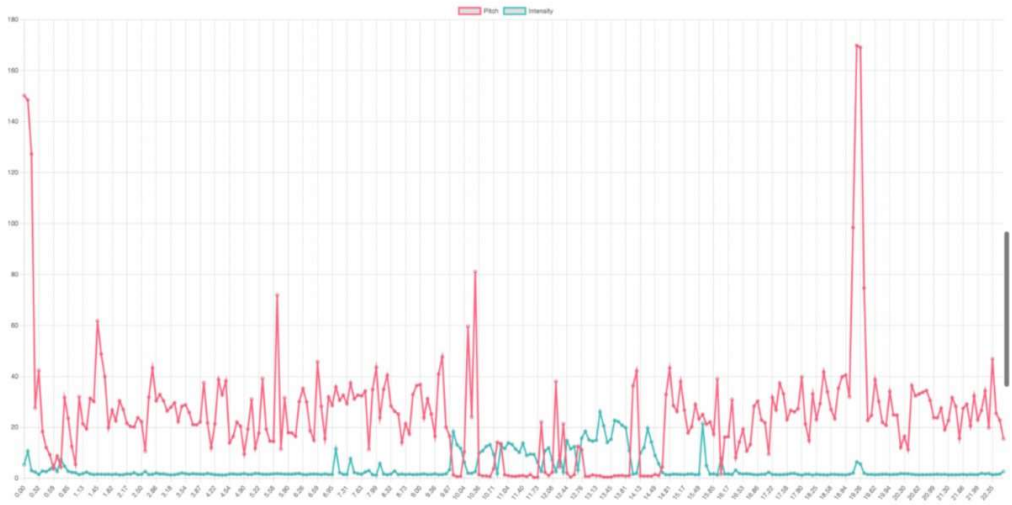


Figure 11. Pitch and Intensity against time stamp

Pitch and intensity which are the extracted features of speech on which the final prediction is done is also depicted in the form of a graph where colour red is pitch and green is intensity.

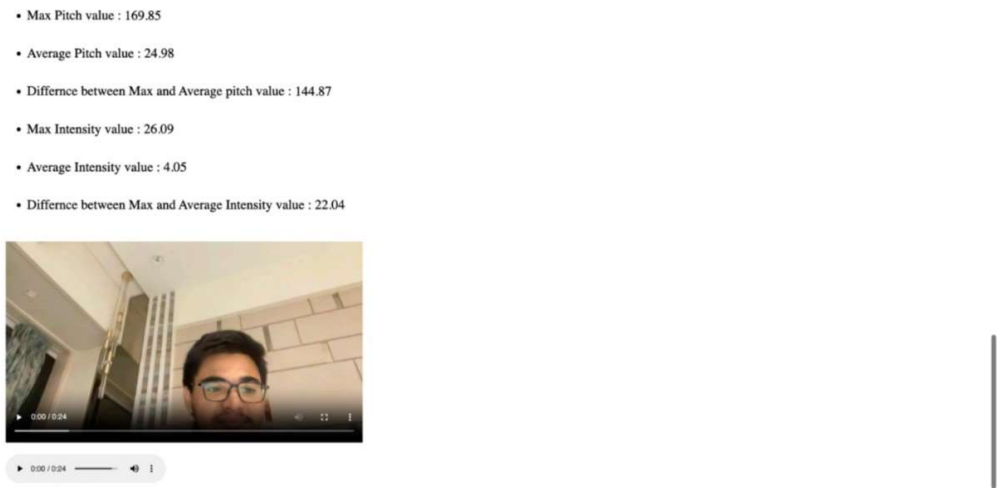


Figure 12. Highlights of Pitch and Intensity distribution

The minimum, average and maximum values of pitch and intensity was measured and shown on the UI to do further analysis of the speaker

5. Conclusions

The project concludes with the successful development of an emotion monitoring model for specially abled children, employing voice and facial recognition techniques powered by convolutional neural networks (CNNs).One of the key features of EmotiSense is its real-time monitoring and alerting model. In the event of sudden changes in the child's emotional state or agitation, EmotiSense triggers alerts to notify caregivers, parents, or teachers, enabling timely intervention and support. These alerts can be customized based on the severity of the situation, ensuring that appropriate action is taken to address the child's needs. By amalgamating voice and facial data analysis, the model adeptly detects various emotional states in real-time, facilitating timely interventions to mitigate sudden agitation. The envisaged future scope entails the integration of additional sensors and advanced algorithms to incorporate a trigger model capable of automatically activating alarms or alerts upon detecting distress. Further research endeavors will focus on refining real-time data analysis and decision-making algorithms, alongside exploring machine learning techniques for personalized responses based on individual behavioral patterns and preferences. Overall, the project's outcomes signify a promising step towards enhancing emotional support and care for specially abled children and their caregivers.

References

- Alim, S.A. and Rashid, N.K.A., 2018. *Some commonly used speech feature extraction algorithms* (pp. 2-19). London, UK:: IntechOpen.
- Avila, A.R., Kshirsagar, S.R., Tiwari, A., Lafond, D., O'Shaughnessy, D. and Falk, T.H., 2019, September. Speech-based stress classification based on modulation spectral features and convolutional neural networks. In *2019 27th European Signal Processing Conference (EUSIPCO)* (pp. 1-5). IEEE.
- Whitford, A.B. and Whitford, A.M., 2023. Modalities of monitoring: Evidence from cameras and recorders in policing. *Government Information Quarterly*, *40*(4), p.101882.
- Aqlan, A.A.Q., Manjula, B. and Lakshman Naik, R., 2019. A study of sentiment analysis: concepts, techniques, and challenges. In *Proceedings of International Conference on Computational Intelligence and Data Engineering: Proceedings of ICCIDE 2018* (pp. 147-162). Springer Singapore.
- Divyashree, P., Yadav, A.G., Jayadev, N. and Chidaravalli, S., 2022. Stress and anxiety detection through speech recognition using deep neural network. *International Journal of Innovative Research in Technology*, *8*(11), p.5.
- Fleck, S. and Straßer, W., 2008. Smart camera based monitoring system and its application to assisted living. *Proceedings of the IEEE*, *96*(10), pp.1698-1714.
- Zhao, G., Huang, X., Taini, M., Li, S.Z. and PietikäInen, M., 2011. Facial expression recognition from near-infrared videos. *Image and vision computing*, *29*(9), pp.607-619.
- Garcia-Garcia, J.M., Penichet, V.M., Lozano, M.D. and Fernando, A., 2022. Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions. *Universal Access in the Information Society*, *21*(4), pp.809-825.
- Sharma, G., Umapathy, K. and Krishnan, S., 2020. Trends in audio signal feature extraction methods. *Applied Acoustics*, *158*, p.107020.
- Ghadage, D.I. and Gaikwad, N.P., 2023. Sentimental Analysis of People Using Facial Expression.
- Huang, Y.X. and Chung, Y.N., 2014. Applying image processing technology to monitor the disabilities' security. In *Proceedings of the 2nd International Conference on Intelligent Technologies and Engineering Systems (ICITES2013)* (pp. 503-509). Springer International Publishing.
- Ingle, P.Y. and Kim, Y.G., 2022. Real-time abnormal object detection for video surveillance in smart cities. *Sensors*, *22*(10), p.3862.
- Jemai, F., Hayouni, M. and Baccar, S., 2021, June. Sentiment analysis using machine learning algorithms. In *2021 International Wireless Communications and Mobile Computing (IWCMC)* (pp. 775-779). IEEE.
- Jothiraj, F.V.S. and Mashhadi, A., 2022. Personalized Emotion Detection using IoT and Machine Learning. *arXiv preprint arXiv:2209.06464*.
- Patel, K., Mehta, D., Mistry, C., Gupta, R., Tanwar, S., Kumar, N. and Alazab, M., 2020. Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges. *IEEE access*, *8*, pp.90495-90519.
- Tomba, K., Dumoulin, J., Mugellini, E., Abou Khaled, O. and Hawila, S., 2018, July. Stress detection through speech analysis. In *ICETE (1)* (pp. 560-564).
- Lee, C.M., Narayanan, S.S. and Pieraccini, R., 2002, August. Classifying emotions in human-machine spoken dialogs. In *Proceedings. IEEE International Conference on Multimedia and Expo* (Vol. 1, pp. 737-740). IEEE.

- Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M. and Fernández-Martínez, F., 2021. Multimodal emotion recognition on RAVDESS dataset using transfer learning. *Sensors*, *21*(22), p.7665.
- Luna-Jiménez, C., Kleinlein, R., Lutfi, S.L., Montero, J.M. and Fernández-Martınez, F., 2022. Analysis of Trustworthiness Recognition models from an aural and emotional perspective. *Proc. IberSPEECH*, *2022*, pp.81-85.
- Maghilnan, S. and Kumar, M.R., 2017, June. Sentiment analysis on speaker specific speech data. In *2017 international conference on intelligent computing and control (I2C2)* (pp. 1-5). IEEE.
- Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32.
- Mcclure, K.S., Halpern, J., Wolper, P.A. and Donahue, J., 2009. Emotion regulation and intellectual disability.
- Perveen, N., Ahmad, N., Khan, M.A.Q.B., Khalid, R. and Qadri, S., 2016. Facial expression recognition through machine learning. *International Journal of Scientific & Technology Research*, *5*(03).
- Rane, A.L. and Kshatriya, A.R., 2020. Audio Opinion Mining and Sentiment Analysis of Customer Product or Services Reviews. In *ICDSMLA 2019: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications* (pp. 282-293). Springer Singapore.
- Ranjan, R. and Daniel, A.K., 2022. An Optimized Deep ConvNet Sentiment Classification Model with Word Embedding and BiLSTM Technique. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *11*(3), pp.309-329.
- Li, R. and Liu, Z., 2020. Stress detection using deep neural networks. *BMC Medical Informatics and Decision Making*, *20*, pp.1-10.
- Saraswat, S., Bhardwaj, S., Vashistha, S. and Kumar, R., 2023, March. Sentiment Analysis of Audio Files Using Machine Learning and Textual Classification of Audio Data. In *2023 6th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1-5). IEEE.
- Singh, N. and Jaiswal, U.C., 2023. Sentiment Analysis Using Machine Learning: A Comparative Study. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *12*, pp.e26785-e26785.
- Zhang, L., Wang, S. and Liu, B., 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), p.e1253.
- Huang, Z., Epps, J., Joachim, D. and Sethu, V., 2019. Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection. *IEEE Journal of selected topics in Signal Processing*, *14*(2), pp.435-448.

6. **Conflicts of Interest Declaration**

The author(s) declare that there are no conflicts of interest regarding the publication of this paper.