TPM Vol. 32, No. S3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Facial Recognition with Advanced Emotional States: A CNN-RF and CNN-LSTM Hybrid Framework for Stress, Frustration, and Confidence

[1] P Sudhandradevi, [2*]V Bhuvaneswari
[1,2] Department of Computer Applications, Bharathiar University, Coimbatore, India
[1]psudhandradevi@gmail.com, [2]bhuvana_v@buc.edu.in

Abstract
Emotional health is important in human communication, affecting social relationships, decision-making, and social interactions. This study explores facial emotion recognition (FER) using Paul Ekman's basic emotion model, incorporating both the FER2013 dataset and an extended dataset, EmoFace (own dataset), which introduces "contempt" to analyze conflict resolution and social dynamics. This study leverages a Haar Cascade frontal facial classifier for efficient facial feature extraction, addressing the challenges posed by real-world images with varying orientations and backgrounds. The EmoNxtSeq hybrid fusion approach integrates CNN-RF to analyze confidence levels, whereas a CNN-LSTM fusion method is used to assess frustration and stress. The experimental results demonstrate that the model effectively identifies emotional states, highlighting the correlation between confidence, frustration, and stress during presentations. The findings offer insights for applications in mental health monitoring, public speaking training, and real-time feedback systems.

Keywords: ***Confidence Estimation, Convolution Neural Network, Early-Fusion, Facial Emotion Recognition, Frustration Recognition, Random Forest, Long Short-Term Memory, Stress Detection***

1. INTRODUCTION

Emotions play a role in shaping our behavior, decisions, and perception, which portrays the dynamic and complicated aspect of human expression through facial expressions. Human emotions and feelings play a crucial role in communication and lifestyle, where sentiment analysis establishes the polarity of a condition to categorize it as positive, negative, or neutral. However, it is restricted to these categories and cannot register the full variety of emotional experiences; for example, both anger and sadness fall under negative states. By contrast, Emotion Recognition (EI) registers the fine-grained distinctions between various emotional states (e.g., differentiating between anger and frustration) focuses on psychological states and takes account of cognitive and contextual considerations. [7] developed an emotion model categorized into a "wheel of emotions, universally recognizable across different cultures and are often conveyed through facial expressions and other physiological cues.

Facial expressions are classified into seven universal emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. The facial features associated with primary and basic emotions are given in Table 1. Primary emotions refer to fundamental emotional reactions triggered by specific situations. Addressing individuals based on their emotional state such as their confidence level, frustration and stress can potentially enhance their emotional well-being [3], such as their confidence level, frustration, and stress. When emotions are analyzed during a presentation confidence, frustration and stress are deeply interconnected. Confidence is an expression of self-confidence and faith in one's capability to deliver well. Frustration occurs when challenges or barriers interfere with effective communication. Stress is a reaction to pressure, which may either increase concentration (positive stress) or lead to anxiety (negative stress). These feelings are interrelated, and excessive stress can result in frustration, while confidence can change based on the management of stress. As stress levels rise, frustration also increases almost perfectly, suggesting that when individuals feel pressured, they are likely to experience frustration as well. Interestingly, confidence follows a similar pattern meaning that, a person may appear confident even when under stress or frustration, possibly as a coping mechanism. These insights are valuable in real-world scenarios such as public speaking training, workplace stress management, or mental health assessment, where understanding emotional dynamics can help improve well-being, performance, and real-time feedback systems [2].

Table 1 Emotion Categories and Corresponding Facial Features

S. No	Emotions	Description	Features
1	Anger	Anger is also manifested through facial expressions like frowning and glaring, Characterized by hostility, agitation, frustration, and antagonism towards others.	Eyebrows furrowed and drawn together; eyes glaring; lips narrowed.

2	Disgust	In social relationships, disgust can arise due to offensive acts or behaviors that go against social norms. Eg., foul tastes, odors, or sights.	Furrowed brows; nose wrinkled; upper lip raised; lips loosened.
3	Fear	In times of threat, people have the fear response.	Eyebrows furrowed; eyelids elevated; mouth open.
4	Happiness	A state of positive emotion experienced as joy, contentment, and welfare.	Eye muscles tightened; wrinkles near the eyes; cheeks raised; lip corners lifted.
5	Sadness	A temporary emotional state characterized by disappointment, grief, hopelessness, and low mood.	Eyebrow inner corners raised; droopy eyelids; lips pulled downward.
6	Surprise	Research shows that people are more likely to notice surprising events, making unusual news more memorable.	Eyebrows raised; eyelids widened; mouth open.

The fusion method unifies disparate data sources which attempt to identify and interpret human emotion via different types of data such as facial expressions, voice tone, gestures, and text content, thus capturing the emotional landscape. Since emotions are naturally multi-faceted and tend to be expressed through several channels at the same time, using a single source of information can lead to incomplete or incorrect interpretations. Such modalities are used as a lens to read emotions, and high-dimensional data from multiple modalities can make model training difficult and computational costs higher. The integration of several modalities (e.g., audio, visual, and physiological signals) poses synchronization, alignment, and representation challenges, which are crucial to the correct prediction of emotional responses. The CNN model is 60% accurate using real-time data, which reflects a moderate capacity for emotion classification. Early fusion involves merging raw features from different data modalities in an initial stage prior to learning or classification processes. Data-Level Fusion consolidates data from various sources to minimize inaccuracies and errors to enrich the data quality, and Feature-Level Fusion technique combines raw features before any decisions are made [19]. Late fusion involves combining the outcomes of (predictions or decisions) multiple models at a final stage. Model- level fusion focuses on extracting features from images using Neural Networks (NNs). Hybrid fusion combines both early and late fusion techniques. In Feature-level fusion, facial features extracted by a CNN can be concatenated into a framewise representation (e.g., audio tone or textual data) and passed to a classifier to make final predictions for each frame. This approach builds multiple decisions and combines them to increase the prediction accuracy.

2. RELATED METHODS

Research has shown that emotion recognition systems have evolved significantly, employing fusion techniques that integrate Facial Emotion Recognition (FER) [29] from image and video data, physiological signals, audio-visual data, and deep learning algorithms.

[4] reported that feature derivatives enhance facial emotion recognition, yielding better classification results on the AVEC2011 dataset. [18] introduced a CNN with Attention Mechanism (ACNN) to identify occluded facial regions and focus on key, unobstructed areas. [30] proposed a surface feature-based algorithm for estimating deformable models and fitting them to track 3D facial expressions. The 3D facial expression label map and motion vectors were input for classifiers, with local curve-based patches achieving FER accuracies above 60%, and over 70% in some cases. A genetic algorithm optimized weights, where a score-level fusion scheme with Support Vector Machine and Hidden Markov Model predicted expression labels for 3D and 4D scenarios. [5] proposed an automatic facial expression recognition system with higher accuracy for anger and surprise, but lower rates for disgust, fear, and happiness. The system includes facial detection, feature extraction with 3D Gabor filters, and emotion classification using Artificial Neural Network (ANN). [27] introduced the Regional Covariance Matrix (RCM) method, which uses Bayes Discriminant Analysis and Gaussian Mixture Models without requiring facial alignment. [11] developed a model that captures the spatial and temporal development of human appearances using Recurrent Neural Network (RNN) [6] and CNN. [13] reviewed deep FER systems addressing challenges such as overfitting and expression-unrelated variations to overcome this problem [14] proposed a hybrid method that combines a CNN and Long Short-Term Memory (LSTM).

Reference [15] addresses occlusion and pose variations in FER with a RAN that captures the importance of facial regions. The RAN aggregates regional features and uses region-biased loss to prioritize significant areas. [2] explored Deep Convolutional Neural Networks for FER, analyzing ten emotions from the ADFES-BIV dataset, and tested them across multiple datasets. [20] introduced a facial expression recognition method using a multifeatured system with Gabor wavelets for local features and Haar wavelets for global features [1]. Dimensionality reduction is achieved through Nonlinear Principal Component Analysis (NLPCA) and a weighted fusion technique, enabling the

classification of six emotions with an SVM [28]. [22] developed a hybrid multimodal data fusion method on the DEAP video dataset, integrating audio and visual modalities through a latent space linear map and combining them with textual data using Dempster-Shafer (DS) theory. Marginal Fisher Analysis (MFA) for audio-visual fusion outperforms cross-modal factor analysis (CFA) and canonical correlation analysis (CCA). [10] used Principal Component Analysis (PCA) on video frame sequences to reduce the dimensionality and processing time. The CNN architecture with three convolutional layers and large kernels improves accuracy by 8% and 4%, respectively on the RML and eNTERFACE'05 databases addressing the FER problem.

[16, 17] discuss advancements in FER including the CNN-LSTM architecture which employs preprocessing techniques like data augmentation and normalization to enhance accuracy. The trend is shifting from unimodal to multimodal analysis in emotion recognition, emphasizing nuanced detection. Future research should focus on larger databases and robust multimodal architectures for a deeper understanding of human emotions. [21] examined emotion recognition in Human-Computer interactions and affective computing focusing on facial expressions within multimodal architectures that include speech and physiological signals. Their model achieves higher accuracies through hyperparameter tuning [3, 12] who proposed an efficient DCNN using transfer learning for emotion recognition from facial images in the KDEF and JAFFE datasets while considering profile views. Fine-tuning these views enhances classification accuracy for applications such as patient monitoring and security surveillance. [24] investigated features such as facial movements, speech, and hand gestures using fusion methods to combine datasets such as FER, CK+, and RAVDESS. [19] discussed a framework for assessing fusion methods in multimodal emotion recognition to identify emotions such as happiness, neutrality, sadness, and anger. While ten emotions are considered in the annotation process, the IEMOCAP dataset focuses on happiness/excitement, sadness, anger, and neutrality. [26] proposed a multimodal approach that classifies emotions from speech and images into categories. The speech utterances dataset IIT-R SIER contains images and labels like anger, happiness, hate, and sadness, emphasizing the value of using complementary information from multiple modalities for emotion recognition.

This literature review examines deep learning techniques, such as CNN-LSTM and DCNN, which achieve over 90% precision. Key advancements include robust CNN architectures with data augmentation for improved FER accuracy, feature derivatives for better classification, and PCA for video frame sequences. Multimodal analysis has replaced unimodal analysis in emotion recognition, highlighting the importance of subtle emotion detection. The integration of fusion techniques enhances the accuracy and robustness of emotion detection. Future research will likely focus on improving system adaptability to diverse emotional cues and contexts. This review emphasizes the need for larger databases and strong multimodal architectures to fully understand human emotions.

3. DATA AND METHODS

The objective of this work is to understand and predict human emotional states dynamically to help students overcome emotional challenges during presentations. Here emotional, and psychological health can be understood through the lens of primary emotions such as fear, anger, sadness, anger, disgust, contempt, and surprise. Through these emotional well-being states, stress (emotional response to external pressures), frustration (obstacle to achieving goals or desires), and confidence (feeling capable and competent) can be identified. The presenters face obstacles in their presentation, so it is important to evolve emotions over time and identify each emotion independently. The proposed hybrid fusion approach (EmoNxtSeq) is to identify the emotional well-being of the presenter, as shown in Fig. 1. The proposed methodology is organized into three phases: (i) Data collection, edge detection, and feature extraction using the Haar Cascade Classifier and Convolutional Neural Network; (ii) Prediction of emotional well-being using the EmoNxtSeq fusion approach, where confidence levels (low, medium, or high) are identified through feature-level fusion (CNNs + RF) and sequential frames of each emotional response are analyzed; (iii) Assessment of frustration and stress using temporal fusion (CNNs + LSTM), which effectively captures the evolution of emotions over time.

TPM Vol. 32, No. S3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

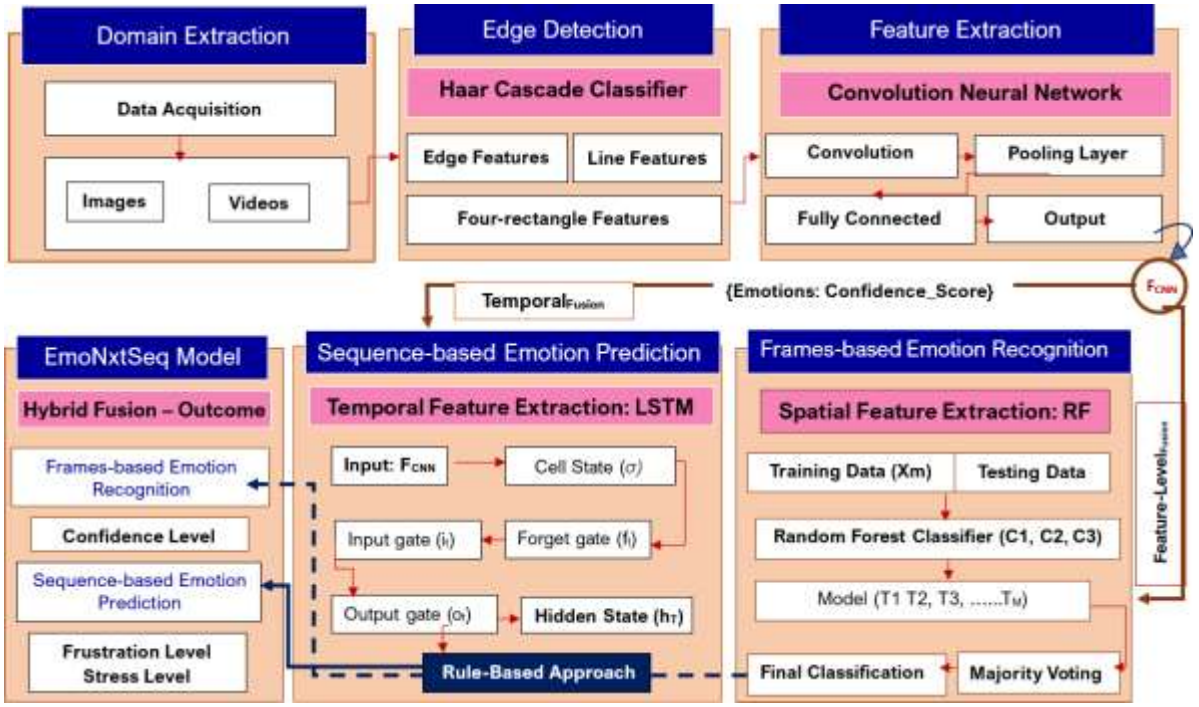


Fig. 1 Architecture of the proposed EmoNxtSeq hybrid fusion model

3.1 Data Acquisition

The basic emotion of Paul Ekman is recognized as expressing and experiencing emotional well-being. In this work, the FER2013dataset comprises 23,793 grayscale images of 48x48 resolution pixels for training and 5865 images for validation. To enhance this dataset, another emotion label 'contempt' from our own 'EmoFace' dataset (train: 284, validation: 39), is to observe conflict resolution, interpersonal relationships, and social dynamics. Table 2 shows the details of the FER2013 and along with our custom EmoFace dataset to ensure both benchmark coverage and domain specificity.

Table 2 FER2013 + EmoFace: Benchmark and domain-specific dataset fusion

Datasets	**Facial Expression Attributes**	**Size of FER2013 and EmoFace Dataset**	
		No. of images: Training	***No. of images: Validation***
FER2013	Anger	4005	958
	Disgust	436	111
	Fear	4107	1024
	Happiness	7214	1794
	Sadness	4850	1247
	Surprise	3181	831
EmoFace	Contempt	284	39
Total No. of Images (Train & Validation)		24,077	6004

3.2 Edge Detection

As the dataset do not exclusively contain frontal facial images; rather, they capture hierarchical features and spatial hierarchies that complicate the recognition of patterns within each class of facial data. The Haar Cascade frontal face classifier addresses this challenge by extracting facial features to classify potential facial regions that progress to the subsequent stage and filter-out only the nonface regions. This hierarchical approach facilitates faster computation by swiftly eliminating clearly nonface areas. The desired output is achieved when a sequence of images is fed into the neural network model.

TPM Vol. 32, No. S3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

3.3 Proposed Methodology: EmoNxtSeq Hybrid Fusion Framework

The objective of this study lies in its potential to identify the emotional states of students, thereby enabling mentors and counselors to provide appropriate guidance and support to mitigate issues related to diversion and anxiety. The proposed methodology, a hybrid EmoNxtSeq approach, integrates CNN for feature extraction, and spatial feature selection to identify the confidence-level of the presenter using Random Forest and temporal patterns to identify stress and frustration using LSTM.

$$Feature\_Level_{fusion} = F_{CNN} \oplus F_{RF}$$

$$Temporal_{fusion} = F_{CNN} \oplus F_{LSTM}$$

where $\oplus$ denotes the concatenation of feature vectors.

3.3.1 Feature Extraction: Convolutional Neural Network (CNN) Architecture

Convolutional Neural Networks, extract features from the input image (X) and the numerical form of the image are fed into the convolutional layer where each numerical value corresponds to the intensity of the respective pixel. The significant variations in the pixel values occur with those of neighboring pixels near the edges, as shown in Fig. 2.
[[[[

Fig. 2 Representation of Pixel Intensity Matrix

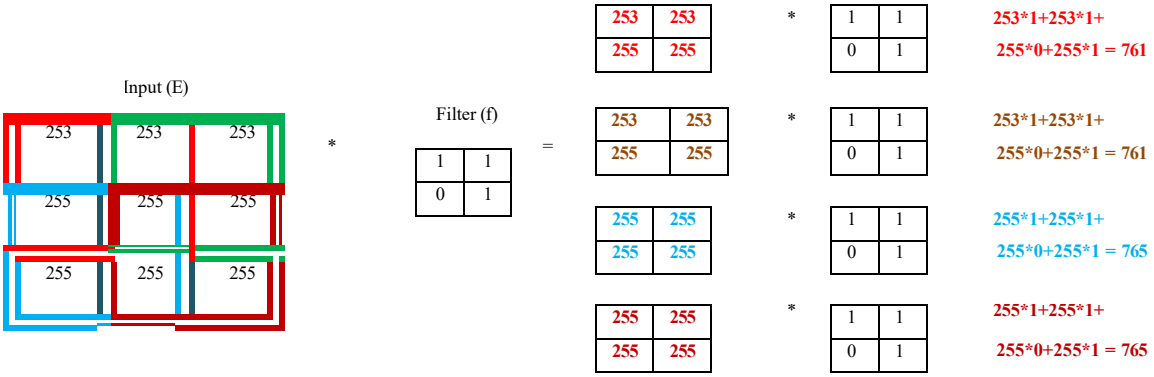
Let the input image be represented as a three-dimensional matrix (i, j, k) in Eq. (1).

$$X \in R^{h \times w \times c} => \{X_{ijk} \mid i \in [1, h], j \in [1, w], k \in [1, c]\} \quad \textbf{(1)}$$

Here h, w, c represents height, width, and number of channels respectively.
In convolution operation, each convolution layer applies a series of filters (l), to the input image and the filter at layer l be denoted by,

$$w_f^{(l)} \in R^{k_h \times k_w \times c} \quad \textbf{(2)}$$

In Eq. 2, k_h and k_w are the height and width of the filter respectively, whereas (f) represents a specific filter within the layer. The dimensions of the feature maps start at 64, 128, and 256 and increase to 1024 in later layers, which is common for deeper feature extraction. The filter traverses the image and processes patches around each pixel to perform elementwise multiplication, after which the resulting values are summed as follows:



TPM Vol. 32, No. S3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

The same filter is then applied to an image. The convolution operation at a specific location (i, j) for filter f at layer l is defined in Eq. 3:

$$(X * W_f^{(l)})_{ij} = \sum_{u=1}^{k_h} \sum_{v=1}^{k_w} \sum_{d=1}^{c} w_f^{(l)}(u, v, d).\, x\,(i + u, j + v, d) \quad \textbf{(3)}$$

In Eq. (4) above * represents the convolution operation (in Fig. 3), which generates an activation map for each filter applied. To maintain the clarity of the image, the input must be provided to the convolutional layer in the dimensions of (48, 48, 1). An increase in image size may lead to a potential loss of information.

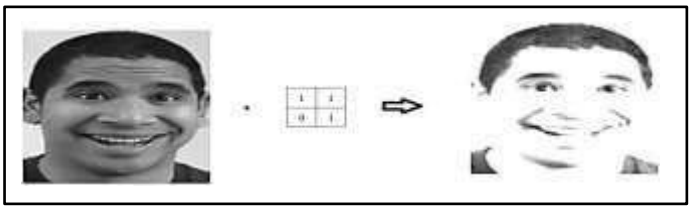


Fig. 3 Feature extraction through Convolution Operation

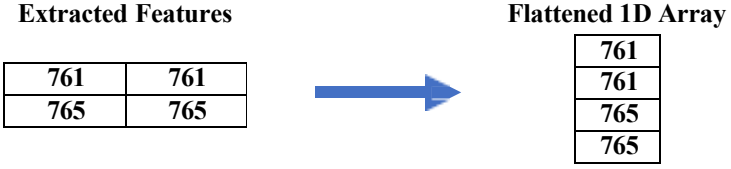
Initially, a linear transformation is applied to all the neurons of the activation function and here, Rectified Linear Unit (ReLU) activation function σ (·) captures patterns by selectively activating neurons. Neurons are deactivated only when the output of the linear transformation is less than zero and consequently inactivate neurons with outputs below the threshold to capture the complex patterns. Following the convolution process, the resulting output is subsequently processed through a nonlinear ReLU activation function as bias,

$$A_f^{(l)}(i, j) = \sigma\left((X * W_f^{(l)})_{ij} + b_f^{(l)}\right) \quad \textbf{(4)}$$

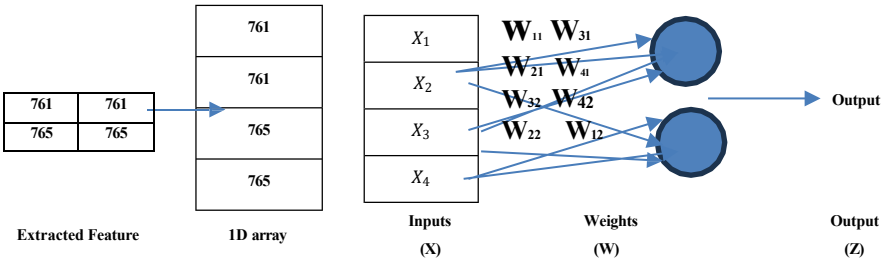
Eq. (4), ReLU(z)=max (0, z), where $A_f^{(l)}$ is the output feature map after applying filter *f* at layer *l*. To decrease the spatial dimensions of the feature map, max-pooling with a window of size $p_h \times p_w$ selects the maximum value from the corresponding region within the feature map. This pooling window is applied to:

$$P_f^{(l)}(i, j) \quad \max_{(u,v) \in [1,p_h] \times [1,p_w]} \quad A_f^{(l)}(i + u, j + v) \quad \textbf{(5)}$$

The output produced by the convolutional layer is a two-dimensional matrix (Eq. 5); however, the fully connected layer operates exclusively with one-dimensional data. Max pooling layers reduce the spatial dimensions of the feature maps, helping reduce computation and controlling overfitting. After a certain number of layers, Dropout layer to prevent overfitting by randomly setting a fraction of the input units to 0 during training. Once the data are converted into a one-dimensional flattened array, the convolutional layer extracts valuable features from the data and transmits them to the fully connected layer, which subsequently generates the results.



In this all values are considered distinct features. The fully connected architecture executes both linear and non-linear transformations.



After multiple layers (*l*) of convolution and pooling, the fully connected neural network (FCNN) is transformed into a one-dimensional vector for processing the resulting feature maps.

$$F_{CNN} = Flatten\,(P_f^{(L)}) \quad \textbf{(6)}$$

The flattened vector (Eq. 6) from the fully connected neural network (FCNN) with 1250 units for high-level feature extraction, followed by dropout of the final dense layer has 7 units, likely corresponding to the 7 output classes for classification to achieve final classification. The output is calculated as follows:

$$\hat{y} = softmax\,(W_{fc} \cdot F_{CNN} + b_{fc}) \quad \textbf{(7)}$$

Eq. 7, W_{fc} represents the weight matrix of the fully connected layer, b_{fc} is the bias vector, and $\hat{y}$ is the predicted class of the corresponding emotional state. Fig. 4 shows the architecture of the CNN model.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 46, 46, 1024)	10,240
batch_normalization (BatchNormalization)	(None, 46, 46, 1024)	4,096
max_pooling2d (MaxPooling2D)	(None, 23, 23, 1024)	0
dropout (Dropout)	(None, 23, 23, 1024)	0
conv2d_1 (Conv2D)	(None, 21, 21, 128)	1,179,776
batch_normalization_1 (BatchNormalization)	(None, 21, 21, 128)	512
max_pooling2d_1 (MaxPooling2D)	(None, 10, 10, 128)	0
dropout_1 (Dropout)	(None, 10, 10, 128)	0
conv2d_2 (Conv2D)	(None, 8, 8, 256)	295,168
batch_normalization_2 (BatchNormalization)	(None, 8, 8, 256)	1,024
max_pooling2d_2 (MaxPooling2D)	(None, 4, 4, 256)	0
dropout_2 (Dropout)	(None, 4, 4, 256)	0
flatten (Flatten)	(None, 4096)	0
dense (Dense)	(None, 1024)	4,195,328
batch_normalization_3 (BatchNormalization)	(None, 1024)	4,096
dropout_3 (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 7)	7,175

Total params: 5,697,415 (21.73 MB)
Trainable params: 5,692,551 (21.72 MB)
Non-trainable params: 4,864 (19.00 KB)

Fig. 4 Structural design of Proposed EmoNxtSeq Model – CNN Architecture

The parameters of the CNN architecture designed for an image classification task with 7 classes. The large number of parameters, especially in the dense layers, suggests that the model is designed to handle high-dimensional data. The overall parameter count of the CNN model is 5,697,415 (indicating a relatively large model), with 5,692,551 trainable parameters (all parameters are trainable with two optimizers) and 4,864 nontrainable parameters (nontrainable weights).

3.3.1 Spatial Feature Extraction with Random Forest: Frames-based Emotion Recognition

The work focuses on specific and increasingly important setting of the presentation and offers unique challenge where nonverbal cues are more limited than face-to-face interactions. Recognizing emotion during presentation is important for understanding the dynamic emotional state of individuals throughout their confidence level. Emotions directly influence confidence, and being able to monitor and manage these emotions ensures that individuals can perform at their best. The features extracted from the CNN model generate classification decisions, where emotions are often conveyed through limited or subtle cues due to the digital and sometimes constrained nature of the video. The output feature vector (F_{CNN}), is input into Random Forest (RF) model for feature selection and dimensionality reduction. To overcome this issue, spatial features are extracted by focusing on visual cues from facial expressions and gestures. Let $F_{CNN} \in R^n$ represent the output from the CNN, where each element of the FCNN corresponds to significant emotional cues from the image.

$$F_{CNN} = [f_1, f_2, \ldots, f_n] \quad \textbf{(8)}$$

In Eq.8, *n* extracted features by CNN such as edges, textures, and facial landmarks. The decision trees within the Random Forest are labeled as $\{T_1, T_2, \ldots, T_m\}$, indicating the total number of trees (m) in the ensemble. The final output of the Random Forest model is derived from the aggregated predictions of these individual trees.

$$\hat{y}RF = Mode\ (T_1(F_{CNN}), T_2(F_{CNN}), \ldots, T_M(F_{CNN})) \quad \textbf{(9)}$$

In Eq.9, $\hat{y} \in \{1, 2, \ldots, C\}$, where C represents the total number of emotional categories used to reduce the feature dimensions and select the most informative features for emotion prediction. To integrate historical data for the prediction of future emotional states, let F^t_{CNN} denote the feature vector at time step t. The historical data for the preceding k time steps are represented as follows:

$$H_t = [F^{t-k}_{CNN}, F^{t-k+1}_{CNN}, \ldots, F^{t}_{CNN}] \quad \textbf{(10)}$$

In Eq.10, H_t encapsulates the temporal dynamics associated with emotional fluctuations over time to concatenate the historical feature vector H_t to forecast the emotional state at the subsequent time step t+1.

$$\hat{y}_{t+1} = mode(T_1(H_t), T_2(H_t) \ldots, T_M(H_t)) \quad \textbf{(11)}$$

The anticipated emotional state $\hat{y}_{t+1}$ at time t+1, is derived from historical data collected from preceding time steps and the feature vectors extracted through CNN.

$$\hat{y}_{t+1} = RandomForestPrediction\ (H_t) \quad \textbf{(12)}$$

In this hybrid fusion technique, CNN extract hierarchical and discriminative features from images of facial expressions, whereas Random Forest classifiers predict future emotional states based on the evolving sequence of these features (Eq.10 & Eq.11). The confidence level is predicted based on the outcome of each class of Random Forest. Based on Eq.13, the predicted classes, and maps confidence levels such as low (0.0 and 0.33), medium (0.34 and 0.66), and high (0.67 and 1.0) confidence intervals are given in Eq.14. [23, 25].

$$P_{max} = max\ (P_{Anger}, P_{Contempt}, P_{Disgust}, P_{Fear}, P_{Happy}, P_{Sad}, P_{Surprise}) \quad \textbf{(13)}$$

Based on the maximum probability $P_{max,}$ the confidence level (CL) is categorized into three ranges:

$$CL = \begin{cases} Low & if\ P_{max} < 0.33 \\ Medium & if\ P_{max} \leq P_{max} < 0.66 \\ High & if\ P_{max} \geq 0.66 \end{cases} \quad \textbf{(14)}$$

This helps improve their emotional regulation and presentation performance. By understanding and regulating emotions such as fear, sadness, and happiness presenters can increase their confidence levels and ultimately their effectiveness during presentations.

3.3.2 Temporal Feature Extraction with LSTM: Sequence-based Emotion Prediction

The features selected from the CNN are subsequently input into the LSTM network, to capture the temporal dependencies to predict future emotional states. Let F^t_{CNN} represent the feature vector at time step t, which computes the hidden state h_t as follows,

The input gate (i_t) determines the new emotion to be stored in the cell state C_t, in Eq.15

$$i_t = \sigma\ (W_i \cdot [h_t - 1, F^t_{CNN}] + b_i) \quad \textbf{(15)}$$

In Eq.16, The forget gate (f_t) is responsible for determining the information from the previous cell state C_{t-1}.

$$f_t = \sigma\ (W_f [h_t - 1, F^t_{CNN}] + b_f) \quad \textbf{(16)}$$

The cell state is updated by applying the forget gate to the preceding cell state and incorporating the new candidate values (Eq.17), which are scaled by the input gate.

$$C_t = f_i \odot C_{t-1} + i_t \odot C_t \quad \textbf{(17)}$$

Here, C_t is the updated cell state, and $\odot$ denotes the elementwise product. In Eq. 18, the output gate (o_t) is responsible for determining the subsequent hidden state (h_t), which will be utilized in the following time step and may also serve as the output.

$$ot = \sigma\ (Wo \cdot [ht - 1,\ F^t_{CNN}\ ] + bo)$$

$$h_t = o_t \odot tanh\ (C_t) \quad \textbf{(18)}$$

In Eq. 19, updated hidden state (h_T) is determined by the output gate and the current cell state. Upon completion of all the time steps, the final hidden state h_T (at the last time step - T) encapsulates the temporal features where it determined by the output gate and the current emotional state.

$$F_{LSTM} = h_T \quad \textbf{(19)}$$

The temporal features extracted using FLSTM are subsequently employed to predict the subsequent emotional state by understanding the relationship between emotions and stress/frustration, that influence and interact with basic emotions such as anger, sadness, surprise, happiness, fear, disgust, and contempt. The impact factors on frustration and stress are given in Table 3.

Table 3 Emotion Weights and Impact Factors on Frustration & Stress

	Emotions (E)	***Weights (W)***	**Frustration Impact_Factor**	**Stress Impact_Factor**
E_1	Anger	w_1	9	8
E_2	Disgust	w_2	8	7
E_3	Contempt	w_3	9	6
E_4	Fear	w_4	6	9
E_5	Happy	w_5	3	2
E_6	Sad	w_6	7	9
E_7	Surprise	w_7	4	5

While facing the unbearable or ethically challenging situations, Frustration will induce disgust and another attribute contempt arises when people feel difficulties in their responsibility and observe others as incompetent [8, 9]. Based on these factors, maximum weights for frustration are assigned as anger (9), contempt (9), and disgust (8). In Eq. (20) frustration scores is calculated as per the emotion score and weight(w) of the respective emotions(E) and based on the score frustration level is calculated in Eq. (21).

$$Frustration\_score = [\ P(E_1) * w_1] + [\ P(E_2) * w_2] + \ldots + [P(E_3) * w_3] \quad \textbf{(20)}$$

$$Frustration\_level = (Frustration\ Score/\ (w_1 + w_2 + w_3)) * 100 \quad \textbf{(21)}$$

Stress usually associated with situations where an individual feels overwhelmed or under pressure, they often exhibit anger due to obstacles preventing goal achievement or a perceived lack of control. Typically stress correlates with emotions such as anger(E_1), fear (E_4), and sadness (E_7). Higher values for fear and sadness in Table 3, shows where the emotions are closely linked to stress in Eq. (22).

$$Stress\_Level = Frustration\_score + [P(E_1) * w_1] + [P(E_4) * w_4] + [P(E_6) * w_6] \quad \textbf{(22)}$$

The temporal features extracted using Fusion Long Short-Term Memory (FLSTM) are leveraged to predict an individual's subsequent emotional state, understanding the relationships among emotions, stress, and frustration for modeling emotional well-being.

Algorithm 1: Hybrid Fusion Determine Emotional Sequence State: EmoNxtSeq

Step 1: CNN Initialization

Initialize CNN() # Input: Image data

conv_layer_1 = ConvolutionLayer(filter_size=(2, 2), num_filters=32)
pooling_layer_1 = MaxPoolingLayer(pool_size=(2, 2))
conv_layer_2 = ConvolutionLayer(filter_size=(2, 2), num_filters=64)
pooling_layer_2 = MaxPoolingLayer(pool_size=(2, 2))
flatten_layer = FlattenLayer()
RETURN conv_layer_1, pooling_layer_1, conv_layer_2, pooling_layer_2, flatten_layer

Step 2: Forward Pass-through CNN

Input: Image data -- Perform a forward pass through the CNN:
Apply conv_layer_1a; ReLU (conv_layer_1**)** activation; pool_1(.ReLU (conv_layer_1))
Apply conv_layer_2; ReLU (conv_layer_2) activation; pool_2(.ReLU (conv_layer_2))
flattened_output = flatten_layer(pool_2); RETURN flattened_output

Step 3: Random Forest Initialization

Initialize RandomForestClassifier() # Input: Feature vectors from CNN
rf_model = RandomForest(n_estimators=7) #No. of estimators (trees): 7 emotional labels
RETURN rf_model

Step 4: Predict Emotion and Calculate Confidence Level Using RF

Initialize rf_model, and fused_features_list
Calculate confidence level for each prediction: (highest confidence)
if P_max< 0.33: "Low" elif 0.33 <= P_max< 0.66: "Medium" else "High"
RETURN confidence_levels

Step 5: Train Random Forest

Input: Combined feature vector and corresponding label
train_RF(fused_features_list, labels): #Train the RF on fused features along with the labels.
rf_model.FIT(fused_features_list, labels)

Step 6: LSTM Initialization

Initialize LSTM() # Input: Sequential data (e.g., video or image)
lstm_layer = LSTM(input_size, hidden_size, num_layers) #No. of layers: depth (LSTM network)
RETURN lstm_layer

Step 7: Forward Pass-through LSTM

Input: Sequential data (e.g., video, image)
Perform a **forward pass** through the LSTM:
lstm_output = lstm_layer(sequence_input)
RETURN lstm_output

Step 8: Calculate Stress_frustration

Input: Anger, Contempt, Disgust, Fear, Sad, Happy, Surprise): # Weights, Confidence_Level

Perform calculate_stress_frustration (weights, conf_lev):
frust_score= (w_frust['anger'] * anger) + (w_frust ['contempt'] * contempt) + (w_frust ['sad'] * sad)
frust_level = (frust_score) / (w_anger+w_disgust+w_contempt)
stress_level=frust_score (w_surprise * surprise) + (w_fear * fear) + (w_sad* sad)
RETURN stress_level, frust_level

Step 9: Feature Fusion

Input: CNN feature vector from Step 2 and LSTM feature vector from Step 6.
fused_features = CONCATENATE (CNN_features, RF_features) #single combined feature vector.4
fused_features = CONCATENATE (CNN_features, LSTM_features)
RETURN fused_features

Step 10: Prediction using the Hybrid Fusion Model

Hybrid_fusion_predict(input1, input2): # Input: i) Test image and Video
CNN_features = forward_CNN(test_image)
LSTM_features = forward_LSTM(test_sequence)
fused_features = feature_fusion(CNN_features, LSTM_features)
prediction = rf_model.PREDICT(fused_features)
Confidence_levels = predict_emotion_and_confidence(list(prediction))
stress_frustration = calculate_stress_frustration(stress_level, frust)
RETURN prediction

In Algorithm 1, an input set of emotions with corresponding confidence scores (emot, values) indicates the likelihood of the presence of these emotions. The CNN extracts the feature, where spatial features are extracted using Random Forest to find the dominant emotions based on the frames, and the LSTM extracts the temporal features where past emotional states influence future stress and frustration levels.

3.4 Performance Evaluation Metrics

The model proposed is evaluated based on performance metrics such as precision, recall, accuracy, and f1-score. These measures are used to find True Positives, i.e., emotions predicted correctly as being of the actual emotional state, True Negatives, refers to emotions predicted correctly as not being of a specific emotional state, False Positives is emotions predicted wrongly as existing when they do not, and False Negatives is emotions predicted wrongly as non-existent when they exist. Accuracy (Eq. 23), is the proportion of instances predicted correctly out of all instances.

$$Accuracy = (True\ Positive\ (TP) + True\ Negative\ (TN)) / (Total\ Instances) \quad \textbf{(23)}$$

Precision (Eq. 24) measures the precision of positive predictions for a given class.

$$Precision = (True\ Positives\ (TP)) / (True\ Positives\ (TP) + False\ Positives\ (FP)) \quad \textbf{(24)}$$

Recall, which evaluates how well the model can pick up all relevant cases correctly in each class (Eq. 25)

$$Recall = (True\ Positives\ (TP)) / (True\ Positives\ (TP) + False\ Negatives\ (FN)) \quad \textbf{(25)}$$

The F1 score (Eq. 26) is the harmonic mean of recall and precision, effectively averaging these two measures together to improve performance, especially in situations marked by skewed class distributions. This makes it particularly applicable when working with imbalanced data sets, where some classes, like rare occurrences or minority classes, are more significant for identification.

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad \textbf{(26)}$$

4. RESULTS AND DISCUSSION

The proposed EmoNxtSeq model identifies the stress, frustration, and confidence levels of a presenter during the presentation. CNN processes images or video frames to extract features by employing Haar wavelets or Haar features to detect facial landmarks, muscle movements, and facial expression patterns, such as the eyes, eyebrows, nose, and lips. The convolutional layer extracts various filters, whereas the pooling layer retains the most significant features from the extracted data. The output of these extracted features is then fed as input to the dense layers, which require a 1D array format. These layers perform linear or nonlinear transformations to capture patterns within the data. The output layer produces results in the form of probability values, with each value representing the likelihood of corresponding class labels. Each model processes its data independently, and their outputs are combined to make the final decision, a process known as decision-level fusion. Related techniques include majority voting, where the final prediction is the one that receives the most votes, and probability averaging, where the probabilities from all models are summed and then averaged.

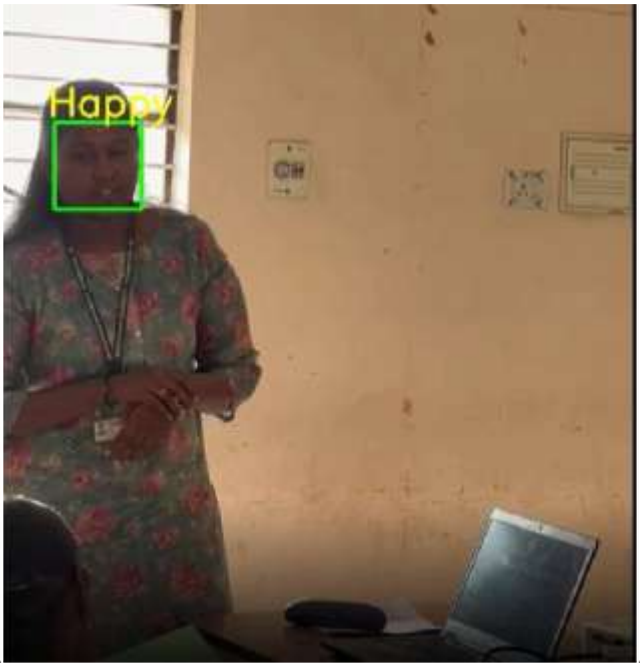


Fig. 5 Emotion detection output from the proposed model, identifying the facial expression as “Happy.”

In Fig.5, the detected emotion is "Happy", as indicated by the bounding box and the label over the presenter’s face. This approach enables real-time tracking of the presenter’s emotional state, helping to assess confidence, stress, or frustration levels dynamically. The insights derived from this analysis could be useful for improving presentation skills, audience engagement, and mental well-being monitoring.

Table 4 Performance Metrics of CNN Model for Emotion Classification

Class Labels	***Precision***	***Recall***	***F1-Score***	***Average***
0	0.52	0.52	0.52	
1	0.81	0.45	0.57	
2	0.51	0.33	0.40	
3	0.80	0.82	0.81	**0.60**
4	0.42	0.64	0.51	
5	0.86	0.69	0.76	
6	0.54	0.48	0.51	

Happy (0.81 F1-score) and Surprise (0.76 F1-score) are the highest-performing categories, which suggests that the model can predict these emotions reliably. Contempt has low recall (0.45) but high precision (0.81), which means that when the model identifies this emotion, it is most likely accurate, but it fails to capture many true instances. Fear, Sad, and Contempt Classes have low recall rates, meaning it has trouble accurately identifying these emotions.

4.1 Results of Frame-based Emotion Recognition

The frame-based emotion recognition analyzes the presenter's emotions across multiple video frames. Each frame was classified into one of the predefined emotion categories based on the dominant emotion ‘Surprise’ detected across consecutive frames, and the confidence level was categorized into low, medium, and high. A high confidence level was assigned when the presenter consistently exhibited positive emotions with strong prediction probabilities, a medium confidence level when there was a mix of neutral and slightly negative emotions, such as mild sadness or slightly anxious expressions; and a low confidence level, when the negative emotions were dominant, indicating a lack of confidence in the presentation. The maximum value of the emotion is used to predict the emotion, as shown in Fig. 6 and the level of confidence is identified as 17.4% as the dominance of positive emotion. The results provide valuable insights into how emotions impact the confidence of a speaker, enabling real-time feedback and potential enhancements in public speaking performance.

TPM Vol. 32, No. S3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

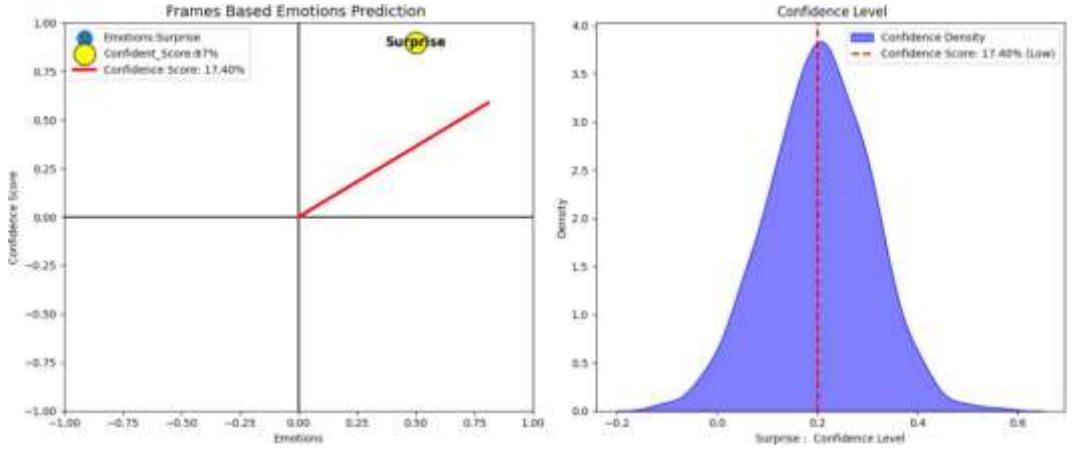


Fig. 6 Confidence analysis of the presenter: CNN-based emotion recognition detects “Surprise” with low confidence

4.2. Results of Sequence-based Emotion Prediction

Emotional fluctuations during a presentation are inherently dynamic and evolve over time. The output labels from the convolutional neural network are fed as inputs from the LSTM, and the inputs are split into X (the sequence of emotion values), and y (the corresponding sequence of the next value). These emotional transitions (e.g., shifts from confidence to nervousness or excitement to anxiety) are sequential and often rely on past emotions to predict future states, which aligns with the LSTM network. In the case of a presentation, emotional data might be captured at regular time intervals (e.g., every second or every few sentences spoken). The model predicts the emotion based on the previous frames according to the window size(n=50), which means that it predicts the previous 50 emotions. The input is fed into the LSTM layer in the shape of (50,1) with the activation function ‘ReLU’ after the LSTM process and the model output is fed as an input for the output layer to predict the emotions using the Softmax function. It returns the sequence values of each emotion.

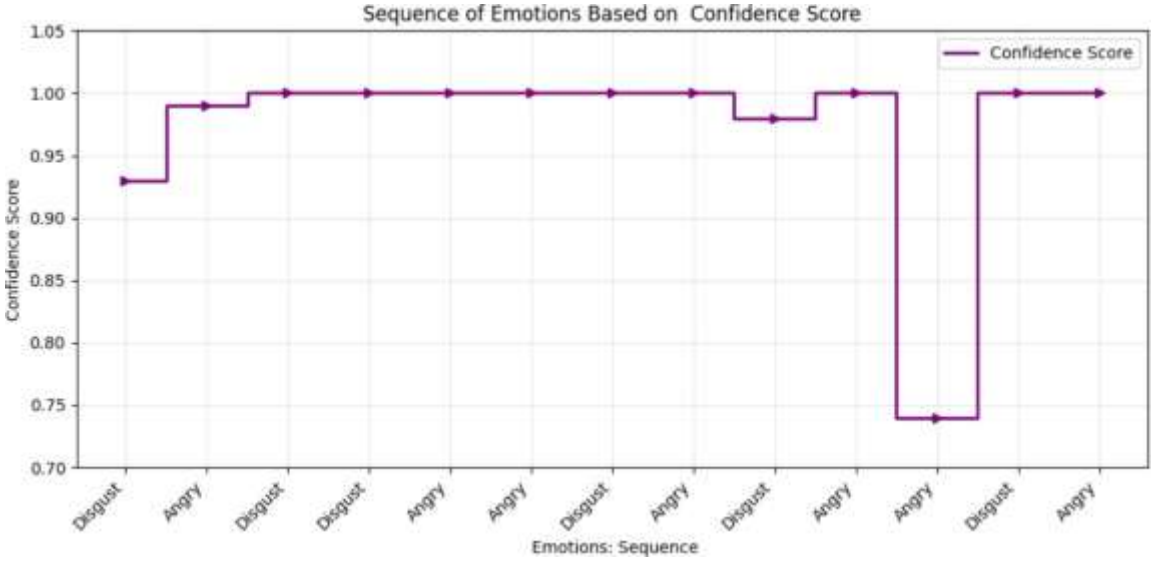


Fig. 7 Sequence of the presenter’s emotions based on confidence scores.

Fig. 7 displays the results of a sequence-based emotion prediction model that uses temporal features to analyze emotions over time. The model appears to process video frames sequentially, extracting facial features and analyzing temporal variations to increase the accuracy of emotion recognition. This approach helps in capturing subtle emotional transitions rather than relying solely on static image classification. By incorporating temporal dynamics, the model can provide a more comprehensive understanding of emotional expressions.

TPM Vol. 32, No. S3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

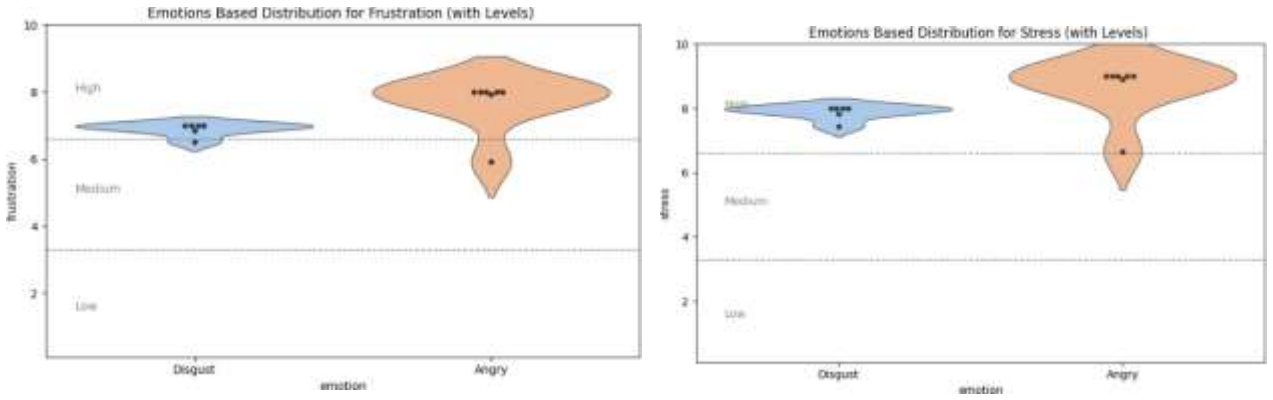


Fig. 8 Emotion-based distribution of frustration and stress levels. The violin plots illustrate the dominance of Disgust and Anger, with both emotions contributing to high levels of frustration

The emotions that predicted in Fig. 8 are 'Disgust' and 'Angry' as inferred negative emotional states and are categorized into low, medium, and high levels of frustration. By aggregating these emotion weights over a sequence of frames, the model can compute a frustration score and accuracy of frustration. Similarly, the stress level can be inferred based on the dominance of negative emotions such as anger, contempt, and sadness over time and categorized into low (0--30), medium (31--60), high (61--80), and critical levels (81--100).

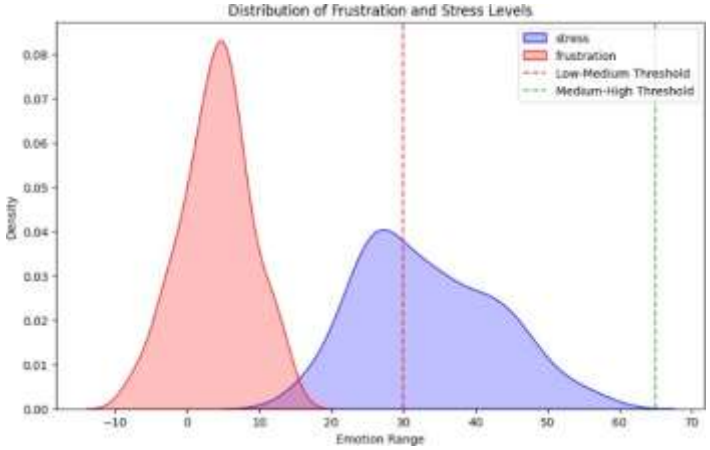


Fig. 9 Distribution of frustration and stress levels across the emotion range. The density plots show frustration (red) and stress (blue), with thresholds indicating low–medium and medium–high intensity boundaries.

If the stress increases, frustration also occurs, as shown in Fig. 9. A high positive correlation, means that as the stress increases, confidence tends to increase rather than decrease. This suggests a possible resilience mechanism in which individuals maintain confidence under pressure. Similarly, high correlation, implies that frustration does not necessarily mean a loss of confidence but rather an emotional reaction tied to high levels of engagement or effort.

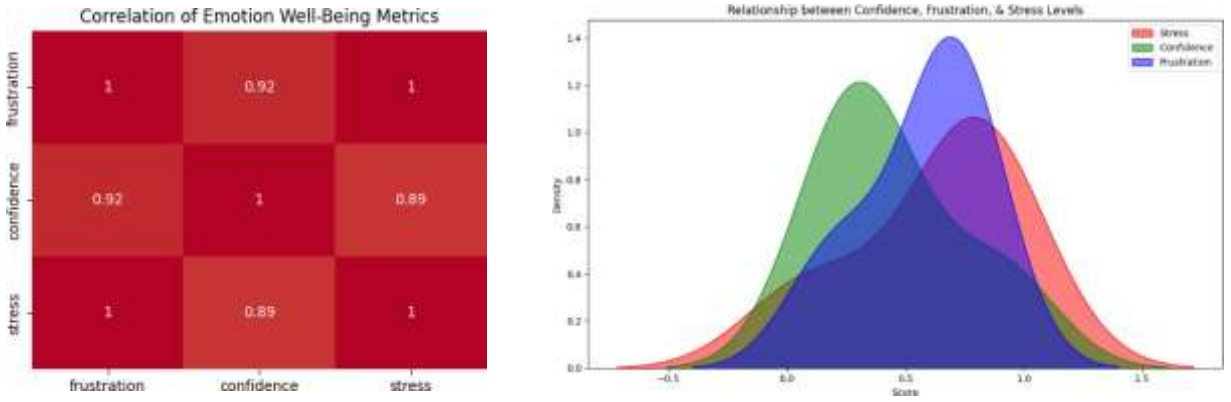


Fig. 10 Relationships between confidence, frustration and stress levels vary in intensity and interact with one another

In Fig. 10, visualization of the presenter reveals strong correlations among confidence, frustration, and stress levels (correlations close to 1). This suggests that variations in one metric significantly impact the others. The distinct emotion-based distributions, particularly for "Disgust" and "Angry" emotions. High-stress values with minimal variation mean that there is a lot of stress present, which may influence their level of confidence and frustration. Positive relationship between frustration and stress suggests that when stress level is higher, frustration level too rises. Third, confidence, while positively correlated, appears to be in an unstable relationship with stress and frustration, perhaps showing moments of excess confidence or confidence deficiency during presentation. In any case, emotional state of presenter brings out this complicated interaction of these three as needing emotional regulating techniques to heighten performance as well as level of engagement.

5. CONCLUSION

The EmoNxtSeq model offers a holistic solution to real-time emotion detection through the analysis of a presenter's level of stress, confidence, and frustration based on a CNN for feature extraction and an LSTM for sequential emotion forecasting. The model accurately detects states and changes in emotions over time, which can be useful insights for understanding the role emotions play in confidence-building during presentations. The strong correlation among stress, frustration, and confidence indicates that emotional dynamics are very important in the performance of a speaker, so emotional regulation techniques are significant. Through frame-based and sequence-based emotion recognition, the model supports real-time tracking and evaluation, which can benefit public speaking ability improvement, audience interaction, and mental well-being monitoring. Its 60% accuracy level, however, leaves room for improvement and most notably in the detection of emotions such as fear, sadness, and contempt, with lower recall values. To achieve improved performance, future enhancements should concentrate on multimodal data integration, improving model accuracy using sophisticated deep learning architectures and incorporating XAI techniques to improve transparency.

Credit Authorship Contribution Statement:

Sudhandradevi P: Data Curation, Methodology, Visualization, Writing & Editing - Original Draft.

V. Bhuvaneswari: Review & Editing - Conceptualization, Methodology, Resources, Supervision, Investigation

Funding Information:

I (SUDHANDRADEVI P), thank the Department of Computer Applications, Bharathiar University, Coimbatore for providing financial support under the promotion of the University Research Fellowship, Bharathiar University.

Conflict of Interest: The authors declare that they have no competing financial interests.

Data Availability Statement: As per request.

Research Involving Humans and /or Animals: Human – Facial Images

Informed Consent: Yes; Certified from the **Human Ethics Committee (BUHEC), Bharathiar University**

REFERENCES

[1] Barbhuiya, A. H. M. J. I., & Hemachandran, K. (2013). Wavelet transformations & its major applications in digital image processing. *International Journal of Engineering Research & Technology (IJERT), 2*(3).

[2] Abdulsalam, W. H., et al. (2019). Facial emotion recognition from videos using deep convolutional neural networks. *International Journal of Machine Learning and Computing, 9*(1), 14–19. https://doi.org/10.18178/ijmlc.2019.9.1.772

[3] Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. *Electronics, 10*(9), 1036. https://doi.org/10.3390/electronics10091036

[4] Cruz, A., Bhanu, B., & Thakoor, N. (2012). Facial emotion recognition in continuous video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. https://doi.org/10.1109/CVPRW.2012.6210342

[5] Anwar, S. M., Majid, M., & Khan, B. (2017). Facial expression recognition using stationary wavelet transform features. *Mathematical Problems in Engineering, 2017*, 1–9. https://doi.org/10.1155/2017/6742530

[6] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2016). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 274–282). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1011

[7] Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

[8] Ekman, P., & Friesen, W. V. (1986). A new pan-cultural facial expression of emotion. *Motivation and Emotion, 10*(2), 159–168. https://doi.org/10.1007/BF00992253

[9] Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology, 39*(3), 281–291. https://doi.org/10.1017/S0048577201393198

[10] Hajarolasvadi, N., & Demirel, H. (2020). Deep facial emotion recognition in video using eigenframes. *IET Image Processing, 14*(14), 3536–3546. https://doi.org/10.1049/iet-ipr.2019.1566

[11] Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., & Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters, 115,* 101–106. https://doi.org/10.1016/j.patrec.2018.04.010

[12] Khaireddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. *arXiv*. https://arxiv.org/abs/2105.03588

[13] Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors, 18*(2), 401. https://doi.org/10.3390/s18020401

[14] Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

[15] Li, Y., Zeng, J., Shan, S., & Chen, X. (2019). Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing, 28*(5), 2439–2450. https://doi.org/10.1109/TIP.2018.2886767

[16] Tan, L., Zhang, K., Wang, K., Zeng, X., Peng, X., & Qiao, Y. (2017). Group emotion recognition with individual facial emotion CNNs and global image based CNNs. *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 549–554). ACM. https://doi.org/10.1145/3136755.3143008

[17] Mellouk, W., & Handouzi, W. (2020). Facial emotion recognition using deep learning: Review and insights. *Procedia Computer Science, 175,* 689–694. https://doi.org/10.1016/j.procs.2020.07.101

[18] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[19] Peña, D., Aguilera, A., Dongo, I., Heredia, J., & Cardinale, Y. (2023). A framework to evaluate fusion methods for multimodal emotion recognition. *IEEE Access, 11,* 3240420. https://doi.org/10.1109/ACCESS.2023.3240420

[20] Reddy, C. V. R., Reddy, U. S., & Kishore, K. V. K. (2019). Facial emotion recognition using NLPCA and SVM. *Traitement du Signal, 36*(1), 13–22. https://doi.org/10.18280/ts.360102

[21] Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems, 2*(1), 53–79. https://doi.org/10.33969/AIS.2020.21005

[22] Nemati, S., Rohani, R., Basiri, M. E., Abdar, M., Yen, N. Y., & Makarenkov, V. (2019). A hybrid latent space data fusion method for multimodal emotion recognition. *IEEE Access, 7,* 172948–172964. https://doi.org/10.1109/ACCESS.2019.2955637

[23] Sharma, N., & Bansal, R. (2018). Fusion-based emotion recognition from facial expressions: A hybrid deep learning approach. In *Proceedings of the 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2203–2208). IEEE. https://doi.org/10.1109/ICACCI.2018.8554415

[24] Subramanian, G., Cholendiran, N., Prathyusha, K., Balasubramanain, N., & Jagatheesan, A. (2021). Multimodal emotion recognition using different fusion techniques. In *Proceedings of the 2021 International Conference on Bio-Signals, Images, and Instrumentation (ICBSII)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICBSII51839.2021.9445146

[25] Tao, J., & Tan, T. (2003). Affective computing: A review. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (pp. 981–995). IEEE. https://doi.org/10.1109/ACII.2003.1238035

[26] Wang, Y., Gu, Y., Yin, Y., Han, Y., Zhang, H., Wang, S., Li, C., & Quan, D. (2023). Multimodal transformer augmented fusion for speech emotion recognition. *Frontiers in Neurorobotics, 17,* 1181598. https://doi.org/10.3389/fnbot.2023.1181598

[27] Wu, J., Lin, Z., Zheng, W., & Zha, H. (2017). Locality-constrained linear coding based bi-layer model for multi-view facial expression recognition. *Neurocomputing, 239,* 143–152. https://doi.org/10.1016/j.neucom.2017.01.004

[28] Zhang, X., Mahoor, M. H., & Mavadati, S. M. (2015). Facial expression recognition using p-norm MKL multiclass-SVM. *Machine Vision and Applications, 26*(4), 467–483. https://doi.org/10.1007/s00138-015-0677-6

[29] Zhao, G., & Zhong, Z. (2017). Facial expression recognition: A comprehensive review. *ACM Computing Surveys, 50*(2), 1–36. https://doi.org/10.1145/3038924

[30] Zhen, Q., Huang, D., Wang, Y., & Chen, L. (2016). Muscular movement model-based automatic 3D/4D facial expression recognition. *IEEE Transactions on Multimedia, 18*(7), 1438–1450. https://doi.org/10.1109/TMM.2016.2559543