I.J. Image, Graphics and Signal Processing, **2026, 1, 106-127**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2026.01.07

Weighted Late Fusion based Deep Attention Neural Network for Detecting Multi-Modal Emotion

Srinivas P. V. V. S.*
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Guntur District 522302, India
Email: cnu.pvvs@gmail.com
ORCID iD: https://orcid.org/0000-0002-4479-2594
*Corresponding Author

Shaik Nazeera Khamar
Department of Computer Science and Information Technology, K L University, Vaddeswaram, Guntur District 522302, India
Email: nazeerakhamar@gmail.com
ORCID iD: https://orcid.org/0009-0000-0242-5262

Nohith Borusu
Department of Computer Science and Information Technology, K L University, Vaddeswaram, Guntur District 522302, India
Email: nohithborusu56@gmail.com
ORCID iD: https://orcid.org/0009-0007-1387-6692

Mohan Guru Raghavendra Kota
Department of Computer Science and Information Technology, K L University, Vaddeswaram, Guntur District 522302, India
Email: mohanguru1816@gmail.com

Harika Vuyyuru
Department of Computer Science and Information Technology, K L University, Vaddeswaram, Guntur District 522302, India
Email: harikasiri2601@gmail.com
ORCID iD: https://orcid.org/0009-0004-1901-7504

Sampath Patchigolla
Department of Computer Science and Information Technology, K L University, Vaddeswaram, Guntur District 522302, India
Email: sampathpatchigolla@gmail.com
ORCID iD: https://orcid.org/0009-0002-0121-4564

Received: 09 July, 2025; Revised: 04 October, 2025; Accepted: 20 November, 2025; Published: 08 February, 2026

Abstract: In the field of affective computing research, multi-modal emotion detection has gained popularity as a way to boost recognition robustness and get around the constraints of processing a multiple type of data. Human emotions are utilized for defining a variety of methodologies, including physiological indicators, facial expressions, as well as neuroimaging tactics. Here, a novel deep attention mechanism is used for detecting multi-modal emotions. Initially, the data are collected from audio and video features. For dimensionality reduction, the audio features are extracted using Constant-Q chromagram and Mel-Frequency Cepstral Coefficients (MM-FC^2). After extraction, the audio generation is carried out by a Convolutional Dense Capsule Network (Conv_DCN) is used. Next is video data; the key frame extraction is carried out using Enhanced spatial-temporal and Second-Order Gaussian kernels. Here, Second-Order Gaussian kernels are a powerful tool for extracting features from video data and converting it into a format suitable for image-based analysis. Next, for video generation, DenseNet-169 is used. At last, all the extracted features are fused, and

emotions are detected using a Weighted Late Fusion Deep Attention Neural Network (WLF_DAttNN). Python tool is used for implementation, and the performance measure achieved an accuracy of 97% for RAVDESS and 96% for CREMA-D dataset.

Index Terms: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), 3D-Convolutional Neural Network (3D-CNN), Mel-Frequency Cepstral Coefficients (MFCCs)

1. Introduction

In recent years, the implementation of human-computer interaction systems has become more reliant on emotion recognition, which has attracted the interest of a wide range of industries. Emotions consist of multiple internal and external acts, which are complex in psycho-physiological processes [1]. In addition to physiological reactions such as skin temperature and heart rate, the appearance of emotions includes non-physiological behaviours such as body language and facial expressions. In general, there are two kinds of tasks related to emotion recognition: discrete (categorical) and continuous (dimensional) [2, 3]. The emotion space is typically divided into a number of fundamental emotion classes by discrete emotion recognition, including neutral, anger, sadness, and happiness. Arousal, valence, and dominance are three common characteristics used to characterize an emotional state. However, emotional state is seen by continuous emotion identification, which is distributed in a continuous space [4, 5].

Discrete emotion representation is easier to understand than continuous emotion representation. People communicate their emotions in a variety of ways, such as through spoken words and nonverbal signals like body language and facial expressions [6]. Understanding the emotional state is possible with such rich multi-modal information. Previous studies have demonstrated the complementary nature of many modalities for emotion perception [7, 8]. It contains various modalities, each conveying emotion-relevant information, and researchers are actively focusing on the best ways to integrate different modalities. Online social media platforms make it simple to gather vast amounts of emotional data [9]. Additionally, label ambiguity comes from the labour-intensive manual annotation method, which are used to annotate this data. The lack of high-quality supervised data has thus been a major barrier to the creation of robust and generalized emotion models [10].

Several algorithms have been considered in the past to recognize emotions in speech. Many people proposed machine learning techniques such as Support Vector Machines, Hidden Markov Models, Gaussian Mixture Models, and so on prior to the period of deep learning. Many speech-related domains, such as voice recognition, speaker recognition, speaker diarization, and so forth, have successfully used deep learning [11, 12, 13]. Deep learning has been employed in the majority of contemporary speech emotion detection efforts because it is effective at extracting very good high-level properties from audio, which helps to increase classification accuracy. In the field of computer vision, CNNs are widely recognized for their ability to extract high-quality features with great efficiency [14, 15, 16]. CNN has also been used to identify speech emotions, and studies have shown that they operate well on Speech Emotion Recognition (SER) tasks.

Recurrent neural networks (RNNs) are the most suitable for processing speech data since speech is a temporal sequence [17, 18]. According to a recent trend in the research, multi-modal emotion recognition models outperform unimodal emotion recognition models due to the additional information provided by the second modality [19]. Numerous research studies have demonstrated that training neural networks with raw audio characteristics yields better results than training them with hand-crafted features. However, there are also some limitations to handling emotion recognition. In this study, a novel deep-learning technique is used for detecting multi-modal emotions [20].

Global interest in multimodal real-time analysis of emotional states has lately grown significantly. Because dynamic multimodal analysis takes into account speech characteristics and employs variables like changes in facial expressions and eye movements over time, it performs better than static analysis of human emotions. It is still difficult to discover an effective multimodal model in this field of study that is not too complicated or computationally demanding. Therefore, an effective and reasonably light multimodal fusion model that takes into account both audio and visual data is put forth in this study to detect emotional states.

1.1 Motivation

In recent years, deep learning has been used more frequently in the field of multi-modal emotion identification when coupled with electroencephalograms. Some researchers have used deep learning to identify new features for emotion recognition because electroencephalogram recordings can be complicated. Convolutional neural network models were employed in previous studies towards automatic feature extraction as well as finish sentiment recognition, with some degree of success. So, these existing techniques motivate the creation of a novel DL method for detecting multi-modal emotion recognition. The major objectives of this proposed technique are:

- The data are collected from audio and video; in order to reduce the dimensionality, feature extraction is carried out.

- To extract the audio features, Constant-Q chromagram, and Modified Mel-Frequency Cepstral Coefficients are used.
- To perform audio modality generation, Conv_DCN is used.
- For extracting the video features, the keyframes are extracted utilizing Enhanced spatial-temporal and Second-Order Gaussian kernels. For video generation, DenseNet-169 is used.
- All the extracted features are fused and classified using a Weighted Late Fusion-based Deep Attention Neural Network.

1.2 Paper Organization

The remaining section of this work is explained below; Section 2 contains related works, and Section 3 contains the proposed methodology. Section 4 contains the result and discussion part, in which performance metrics and comparison analysis are mentioned. Section 5 mentioned the conclusion and future study.

2. Related Work

For Recognizing Multi-modal Emotion, Zhang et al. [21] suggested a Hierarchical Fusion Convolutional Neural Network. In this case, a hierarchical combination CNN technique is used to analyze potential information inside data by constructing multiple hierarchical system structures. For feature-level fusion, the global structures are created by merging weights through statistical features that are manually extracted from the final feature vector. DEAP and MAHNOB-HCI data sets were used for implementation. In this existing study, the quality of data is poor.

For MER, Krishna et al. [22] suggested a 1D Convolutional Neural Networks and Cross-Modal Attention. This existing method combines cross-modal attention networks between audio and text information with 1D convolutional models for raw audio processing to increase the performance of the emotion recognition system. This existing technique exhibits the cross-modal attention mechanism and raw audio features. This can compute the interactive information between audio and text sequences, which is particularly useful for emotion categorization. It causes trouble in managing lengthy sequences.

For Recognizing Multi-modal Emotion on the RAVDESS Dataset, Luna-Jiménez et al. [23] introduced Transfer Learning. A bi-LSTM with an attention mechanism after a pre-trained Spatial Transformer Network using saliency maps and facial images is included in this proposed system. Despite domain adaptation, the error analysis demonstrated that if frame-based systems were applied directly to a video-based task, they would still have some difficulties. This technique uses a limited range of datasets.

For identifying multi-modal emotion, Salama et al. [24] suggested an ensemble learning technique-based 3D-convolutional neural network framework. The suggested method starts by using the 3D-Convolutional Neural Network (3D-CNN) deep learning architecture to extract spatiotemporal features from human face video data and electroencephalogram (EEG) signals. The final fusion predictions are generated using a combination of data augmentation and ensemble learning algorithms. The multi-modalities of the suggested method are fused together using data and score fusion techniques. The main limitation of this existing work is that it is very sensitive to noise.

For identifying multi-modal emotion through spatiotemporal convolution, Sharafi et al. [25] introduce a novel neural architecture. In a novel multi-modal neural architecture, a hybrid network composed of two convolutional neural networks (CNNs) and a bidirectional long short-term memory (BiLSTM) network gets input from both audio and visual data. The energy features and Mel-Frequency Cepstral Coefficients (MFCCs) from audio signals and BiLSTM network outputs are merged with the spatial and temporal features taken from video frames. Finally, the extracted features are classified using the SoftMax layer. The main issue with this approach is the high computational cost. Table 1 shows the comparison analysis of the existing approach.

Table 1. Comparison analysis of existing approach

Authors	Techniques	Datasets	Parameters	Limitations
Zhang et al. [21]	FCNN	DEAP and MAHNOB-HCI	Accuracy – 84.71% for DEAP and MAHNOB-HCI accuracy – 89.00%	The quality of data is poor.
Krishna et al. [22]	1D CNN	IEMOCAP dataset	Accuracy- 65%	Trouble in managing long sequences.
Luna-Jiménez et al. [23]	BiLSTM	RAVDESS Dataset	Accuracy- 80%	Limited range of dataset
Salama et al. [24]	3D-CNN	DEAP dataset	Accuracy- 96.13%	Very sensitive to noise.
Sharafi et al. [25]	CNN+ BiLSTM	SAVEE and RML	Accuracy- 97% for SAVEE and RML accuracy – 94%	High computational cost.

2.1 Problem Statement

In modern civilization, human-robot interaction (HRI) technologies are essential. However, most HRI systems still have disagreement issues, which render human-robot communication ineffective. Some of the limitations caused by previous techniques are very sensitive to noise, high computational cost, the quality of data, etc. To overcome this limitation, novel deep learning techniques are used for the detection of multi-modal emotion recognition.

3. Proposed Methodology

Since emotion recognition is a dynamic process that focuses on the emotional state of the individual, now two people's activities will always produce the same sentiments. To develop a unified human-computer edge, novel researchers in this field focus on emotional state through account-based reasoning. Human feelings are utilized to define a selection of methodologies, like neuroimaging tactics, facial expressions, etc. However, some existing DL methods contain some issues like computational complexity, processing time delay, etc. In order to overcome this limitation, novel deep learning techniques are used to detect multi-modal emotion recognition. Fig. 1 shows the architecture proposed methodology.

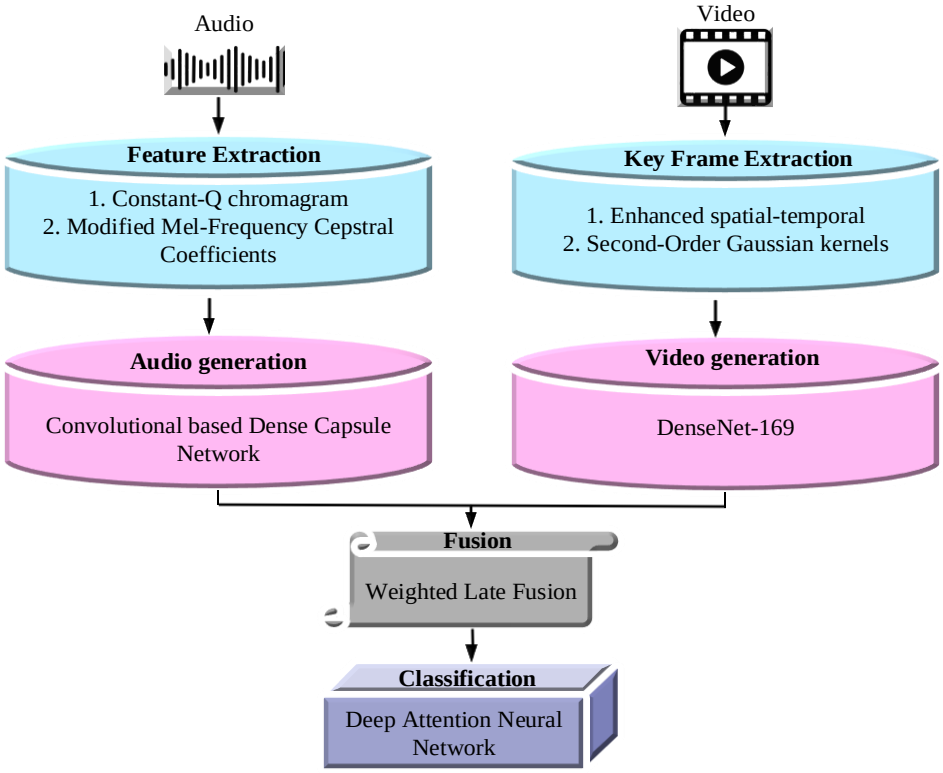


Fig. 1. Architecture of proposed method

The above fig. clearly mentions the flow of the proposed technique. First, the data are collected from the audio signal, and the features are extracted using a new deep-learning technique, and then audio modality generation takes place. After audio extraction, video data is recovered using a novel deep learning technique and DenseNet for video modality generation. Finally, all of the retrieved features are fused together using a unique fusion method, and classification is performed utilizing a deep attention mechanism.

3.1 Audio feature extraction using Constant-Q chromagram

In audio processing, audio feature extraction is the process of converting unprocessed sound waves into useful information. It concerns the alteration or processing of audio signals. It reduces unwanted noise and maintains the time-frequency ranges in balance by converting digital and analog signals.

3.1.1 Constant-Q chromagram

A method for musical signal analysis called a constant-Q chromagram is used to quickly and easily evaluate pitch and chord information. C-Q chromatograms use the Constant-Q Transform (CQT) to analyze audio signals, such as music. CQT splits the sound spectrum into frequency bits that are spaced logarithmically apart in order to imitate how individuals perceive pitch. Equations (1) and (2) define the center frequency, which is used to create constant-Q chromatograms by modifying each speech.

$$g_l = g_1 \bullet 2^{l/a} \tag{1}$$

$$b_l(m) = \frac{1}{D}\left(\frac{m}{M_l}\right)\exp\left[j\left(2\pi m\frac{g_l}{g_p}+\phi_l\right)\right] \tag{2}$$

Where, ϕ_l indicates the phase shift and $\omega(r)$ is a window function related to the Hanning window.

$$D = \sum_{k=-\left\lfloor M_l/2 \right\rfloor}^{\left\lfloor M_l/2 \right\rfloor} \omega\left(\frac{k+M_l/2}{M_l}\right) \tag{3}$$

This section uses the audio processing tools Librosa to finish the extraction of audio files that represent Constant-Q chromatograms in order to carry out the previously mentioned procedure. 120 × 120×3 is the dimension of the chromatogram, and 512 is the number of samples among consecutive chromatic frames. All Constant-Q chromatograms are normalized at this stage to give Constant-Q chromatogram data for later feature extraction and model training.

3.1.2 Modified Mel-Frequency Cepstral Coefficients (MM-FC²)

Mel-Frequency Cepstral Coefficients (MFCCs) are a class of features that are frequently utilized in music information retrieval (MIR), automated speech recognition (ASR), and other audio processing applications; these features are extracted from audio signals. By framing, windowing, and running a series of triangle filters evenly spread out beside the Mel-frequency scale, MFCC characteristics are generated. After taking the logarithm of each filter's energy output and performing a Discrete Cosine Transform (DCT), a bit of constants known as vectors $N(j)$ on the frame j was produced. Change the Mel-frequency using the following formula to increase the linearity of the RCS signal's frequency range.

$$g_{tw} = \frac{g_a}{\log_{10} 2}\cdot\log_{10}\left(1+\frac{g}{g_a}\right) \tag{4}$$

$$g = g_a\cdot\left(10^{\frac{\log_{10} 2.g_{tw}}{g_a}}-1\right) \tag{5}$$

Where, g_a represent frequency defined through the spectral collective contribution r on the amount of the RCS indications, where, g_{tw} and g represent modified MF as well as linear frequency rulers.

$$\frac{\sum_{l=1}^{L'} abs\left[FFT\left(\sum_{m=1}^{M} p(m)\right)\right]}{\sum_{l=1}^{\lfloor l/2 \rfloor} abs\left[FFT\left(\sum_{m=1}^{M} p(m)\right)\right]} \geq r \tag{6}$$

Where, $1 \leq m \leq M$ represents the total amount of signal classes intended from micro-drones and $p(m)$ represents the arrangement of RCS during motion level m. The M-MFCC method value differs depending on the micro-drone's mobility condition. As a method of numerically calculating these variations, the Pearson correlation coefficient was proposed.

$$\rho(n,m) = \frac{Conv(N_n, N_m)}{\sigma_{M_n}\cdot\sigma_{N_m}} \tag{7}$$

Where, $\sigma_{M_n} and \sigma_{N_m}$ are the standard deviation values of N_n *and* N_m, and $Conv(\cdot)$ represent the covariance operation. The outcomes show that the M-MFCC characteristics can accurately capture the varied movements of micro-drones.

3.2 Audio modality generation using Convolutional based Dense Capsule Network

After feature extraction, the extracted features are securely stored in modality generation. Audio modality generation takes place by using a convolutional-based dense capsule network [26]. In the CNN technique, there may be

9 levels in this network, such as 3 convolutional, 3 fully connected, as well as 3 max-pooling layers. The below equation is used to convolve each convolution layer (layers 1, 3, and 5) with its corresponding kernel size (3, 4, and 4).

$$y_m = \sum_{l=0}^{M-1} x_l g_{m-l} \tag{8}$$

Where, x, g, and M stand for the number of elements in x, the filter, and the signal, respectively. Where, x represent the output vector's representation. Fig. 2 shows the architecture of CNN.

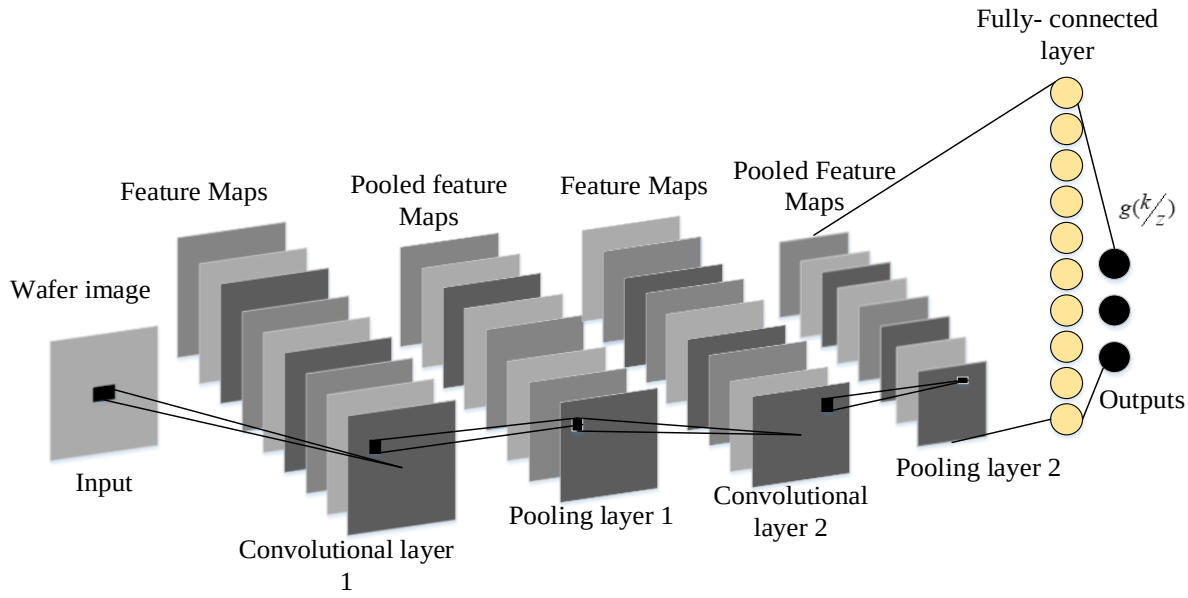


Fig. 2. Architecture of CNN

After each convolution layer, the feature maps go through a max-pooling process. Max-pooling was performed to reduce the size of the feature map. The FC layers have 30, 20, and 5 output neurons to reach the final layer (layer 9). Softmax function uses N, S, V, F, and Q to give separate outputs for each class. The weight and bias are updated in the calculations below.

$$\Delta V_k(r+1) = -\frac{y_\lambda}{t} V_k - \frac{y}{m}\frac{\partial D}{\partial V_k} + n\Delta V_k(r) \tag{9}$$

$$\Delta A_k(r+1) = -\frac{y}{m}\frac{\partial D}{\partial A_k} + n\Delta A_k(r) \tag{10}$$

Where the symbols V, A, k, λ, y, m, n, and r, stand for the corresponding weight, bias, layer number, regularisation parameter, learning rate, momentum, update step, and cost function.

3.2.1 Dense Capsule Network

Tensor capsule y (1–5) is an example of a tensor capsule layer, where the routing process is accomplished by using tensors as units and convolutions as transformation operations. The term "convolution" refers to the conventional two-dimensional convolutional layer. Dense capsule blocks are sub-network modules made up of tensor capsule layers coupled by dense connections. The dense connectivity, which is the structure's foundation, is formed by direct connections from one layer to every other layer. Dense connectivity can be calculated by,

$$y_k = G_k([y_0,...,y_{k-1}]) \tag{11}$$

Every feature map in the layer k is represented by the values $y_0,...,y_{k-1}$. $[y_0,...,y_{k-1}]$ Denotes the feature map concatenation of these maps. Convolutions, rectified linear units, and other operations in the k^{th} layer combine to form $G_k(\bullet)$. Fig. 3 shows the Block diagram of the Dense Capsule Network.

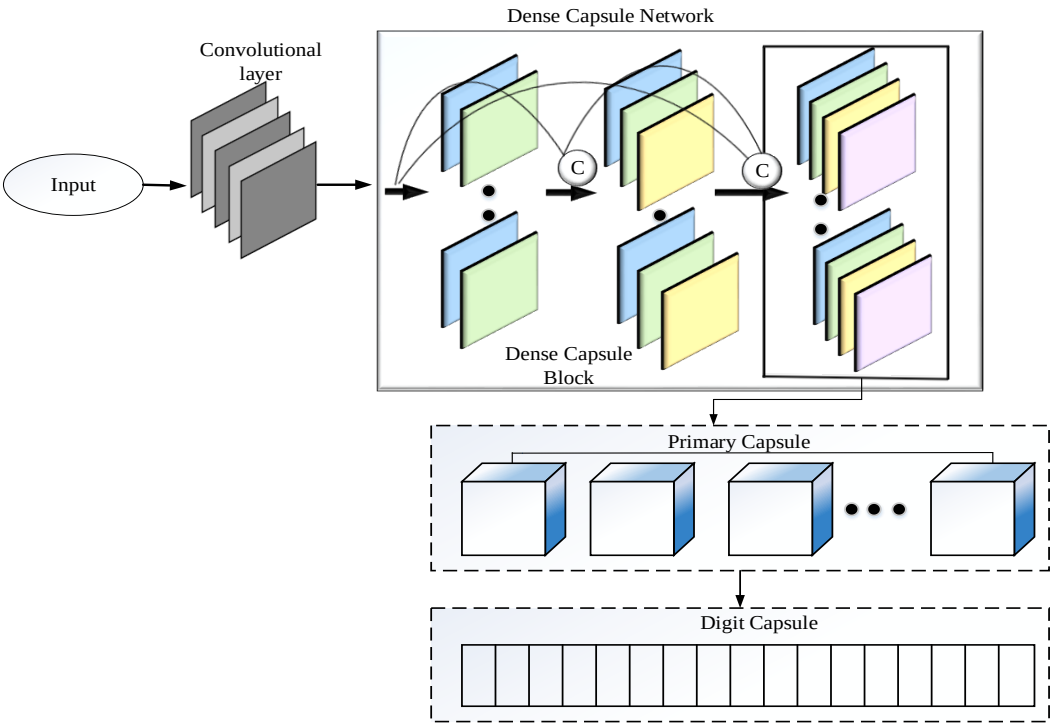


Fig. 3. Block diagram of Dense Capsule Network

In this study, the concept of dense connection is extended to the field of capsule networks. The dense connectedness can then be estimated by,

$$v_j^k = G^k\left(\left[v_j^0,\ldots,v_j^{k-1}\right]\right) \tag{12}$$

Where, $\left[v_j^0,\ldots,v_j^{k-1}\right]$ represent the concatenation of every tensor capsule following layer k. Tensor capsule-specific dynamic routing functions are represented by $G_k(\bullet)$. The audio modality is generated from this dense capsule network.

3.3 Video feature extraction

A variety of granularities, including pixel, frame, and segment levels, can be used to extract features from videos. Frame-level features, such as edges or histograms, capture the global attributes of each frame, whereas pixel-level features record the raw intensity values of each pixel.

3.3.1 Enhanced spatial-temporal

Spatial and temporal dimensions are the two fundamental components of comprehending video data in the context of video feature extraction. The suggested method employs spatial feature extraction to determine the best sites for temporal feature computations, which are extremely effective at distinguishing the contents of video segments. To accomplish the associated goals, spatial features should perform quite well. Therefore, instead of looking for local interest locations, a global spatial feature within a frame will be computed. The below equation shows the mean removed block in spatial features.

$$Y = \sum_{n=1}^{N} \lambda_n^{\frac{1}{2}} v_n u_n^R \tag{13}$$

Where, the eigenvectors of YY^R and Y^RY are $v_n u_n$. The eigenvalues of a diagonal matrix are Λ, then $\lambda_1 \geq \lambda_2 \geq \ldots. \geq \lambda_N$, on the diagonal line. Fig. 4 shows the flowchart of spatial features.

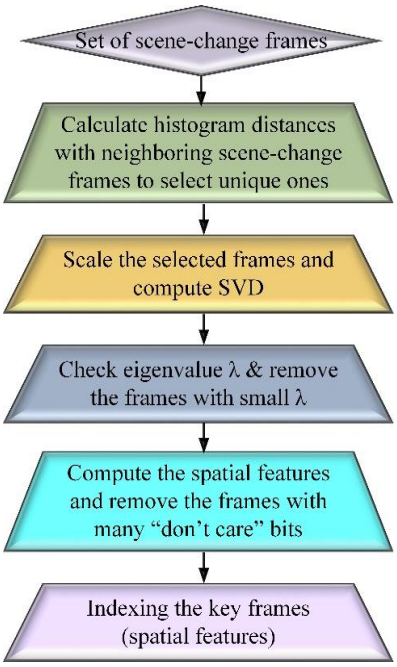


Fig. 4. Flowchart of spatial features.

Similar to taking a picture and examining its contents, spatial features are able to collect the visual information contained in a single frame of a video. First, the frames are altered; next, the histogram distance between the altered frames and nearby scene changes is calculated to identify the unique ones. After scaling the chosen frames, calculate the Singular Value Decomposition (SVD). Examine the eigenvalue λ and eliminate the frames with low λ. Determine spatial features and discard frames with a lot of "don't care" bits. The final step is Keyframe indexing (spatial characteristics).

Temporal features capture the dynamics or movement of information between frames in a video. A variety of tasks, including action recognition, object tracking, and anomaly detection, are made possible by these features, which are essential for comprehending the "movement" and "action" inside the video data. Since shot lengths may efficiently maintain a video's distinctive qualities, they are used as temporal features.

$$K_j = r(T_j) - r(T_{j-1}) \tag{14}$$

Where the matching "time unit" of the shot-change frame T is denoted by $r(T)$, the chosen keyframes, and the duration of each shot for an input video will be recorded.

3.3.2 Second-Order Gaussian kernels

The scale-space architecture has been widely employed in computer vision for feature extraction, including ridge detection and edge identification. An image's scale-space representation is produced by convolving the data through Gaussian kernels of varying scales. The process of convolving an image with a Gaussian kernel yields the scale-space representation $J_p(y;\sigma)$ at a scale σ for a given two-dimensional (2D) image signal $J(y)$, where $y = [y, x]^R$ indicates the planar coordinates.

$$J_p(y;\sigma) = J(y) * h(y;\sigma) \tag{15}$$

Where the scale in scale-space is denoted by σ, and the 2D Gaussian kernel $h(y;\sigma)$ is represented as:

$$h(y;\sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{y^R y}{2\sigma^2}\right) \tag{16}$$

Lindeberg developed a multiplication factor, and the kernel is normalized as follows because the scale parameter is inversely related to the magnitude of a non-normalized second-order Gaussian kernel:

$$h_K^{"}(y;\sigma) = (-1)^{\eta} \cdot \sigma^2 \cdot \frac{\partial^2 h(y;\sigma)}{\partial y^2} \tag{17}$$

Where a parameter η indicates that the kernel can be applied to both local dark $(\eta = 0)$ and local brilliant $(\eta = 1)$ blob formations.

$$h_K^{"}(y;\sigma) = -\frac{y^2 - \sigma^2}{2\pi\sigma^4} \exp\left(-\frac{y^R y}{2\sigma^2}\right) \tag{18}$$

From the above equation, the local contrast between the two lateral regions and the central region is fundamentally measured by the second-order Gaussian kernel.

3.4 Video Generation using DenseNet-169

One convolutional neural network (CNN) architecture called DenseNet-169 was created by Huang et al. (2017) in an effort to improve CNN efficiency and performance in object identification applications. DenseNet-169 [27] is a deep convolutional neural network design with dense layer connections that promotes information flow and feature reuse. This makes it an effective tool for extracting features from images and videos, resulting in excellent accuracy across a range of jobs. These attributes include visual features like color, texture, and shape that could indicate the image types. Fig. 5 shows the architecture of DenseNet-169.

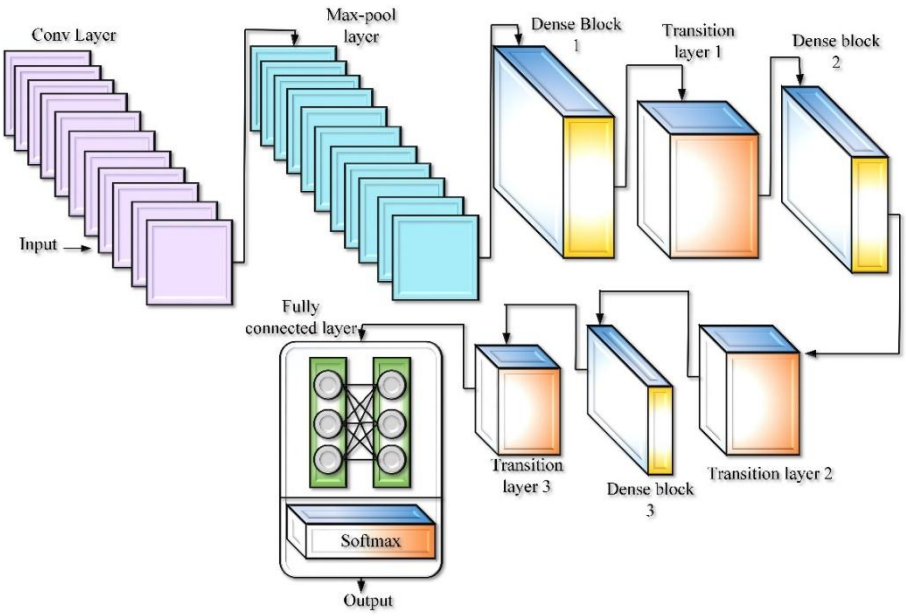


Fig. 5. DenseNet-169

Convolutional layers, max-pool layers, dense layers (completely connected layers), and transition layers are part of the architecture. The last layer of the model employs SoftMax activation, whereas the rest of the design makes use of the ReLU activation function. The max-pool layers reduce the dimensionality of their inputs, whilst the CN layers extract the features of the image. The flattening layer, which serves as an Artificial Neural Network (ANN) with a single array input from the flattening layer, is followed by the fully linked layers.

3.5 Weighted Late Fusion

Weighted late fusion is a technique for merging the findings of many feature extraction methods for both audio and video analysis. The capacity to accommodate missing data, flexibility in adopting additional modalities, and increased accuracy through the combination of several models are just a few advantages of Weighted Late Fusion. In contrast to early fusion, it also permits lower dimensionality and autonomous optimization of individual models. Compared to conventional deep neural networks, deep attention neural networks provide a number of important advantages, chief among them being the ability to lower computing costs and allow models to concentrate on pertinent portions of the input. This makes it possible to digest information more effectively and efficiently, especially when working with sequential data or when contextual awareness is essential. To improve feature fusion, the suggested model incorporates these two benefits. This benefit set the suggested fusion model apart from other models that were already in use.

Weights are allocated to each algorithm based on its performance or contribution, and the weighted outputs are then merged to produce a final, possibly more accurate forecast. To establish the significance of each biomarker, the model was trained independently for each of the three physiological signals using video signals. The detection ratio is used by the framework to assign weights instead of F-scores. The below equation displays the detection ratio (DR).

$$DR = {tp}/{(tp + tn + fp + fn)} \tag{19}$$

Where, fp represent False Positive, fn represent False Negative, tp indicates True Positive, as well as tn represent True Negative. Each output class's weight V , is determined by:

$$V = 1 - DR \tag{20}$$

To determine the class's predictive score, the weight of every class is then multiplied with the probability vector, Q , which is associated with each model.

$$Q = V * Q \tag{21}$$

The model's score Q , is determined by aggregating the results of each class. The final fusion output is given in below equation,

$$output\,class = \max(P_{audio}, P_{video}) \tag{22}$$

The audio and video features are fused using this weight late fusion approach based on the above equation. After fusion, the classification stage is utilized to detect multi-modal emotions.

3.6 Deep Attention Neural Network for classification

The self-attention mechanism uses a non-local operation to weight each pixel based on its correlation in order to model long-distance interdependence. Weights can be used to determine a region's significance. An area has greater significance than its weight. The operation that is not localized is stated as follows:

$$x_j = \frac{1}{D(y)} \sum_{\forall i} g(yj, yi) h(yi) \tag{23}$$

When input and output are represented by the same size variables, y and x , respectively. Where, j denotes the output location index, while i stands for the index of every potential location. Fig. 6 shows the architecture of the self-attention network.

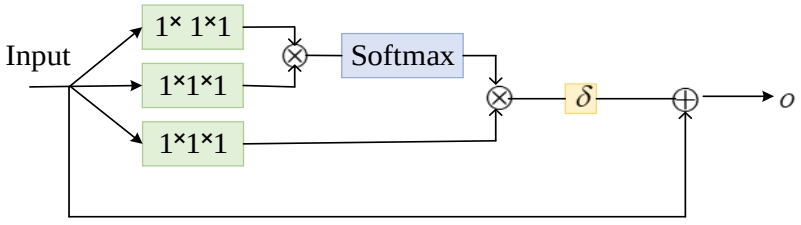


Fig. 6. Self-attention neural network

The relationship between j and all potentially associated places i , which may be stated as weight, is determined using the pairwise function g . Where, g produces a scalar value, the eigenvalue of the input signal at position i is determined by the unary function h , a map function, and its output is a vector. The concatenation function g is given by,

$$g(yj, yi) = \mathrm{Re}\,LU\left(v_g^R\left[\theta(yj), \phi(yi)\right]\right) \tag{24}$$

The concatenation opera is indicated by $[.,.]$. In order to acquire the final self-attention output, this method additionally connects the result of the non-local operation with the input features.

$$c_j = V_c xj + yj \tag{25}$$

Where a residual connection is represented $b + yj$, the weight matrix V_c has an increased number of channels equal to the input y. The self-attention mechanism can be introduced to pre-trained models in a flexible way without affecting the original model's performance due to the residual connection. By using a self-attention neural network, the multi-modal emotion is recognized.

4. Result and Discussion

A Python tool is used to assess the performance using the Kaggle dataset. The evaluation employs accuracy, recall, precision, and F1 score as a performance measure. The proposed technique is compared with existing techniques like LSTM+attention, CNN+BiLSTM+attention, and ResNet-50.

4.1 Dataset Description

RAVDESS: One of the datasets used in the SER tasks is RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) [28]. A total of twelve male and twelve female performers, representing eight different emotional states, are captured on audio and video recordings reciting English sentences. Only the spoken audio samples were used in this investigation. There are 1440 audio files in total, each with 60 trials and a 48 kHz sample rate. The spoken audio samples, such as sad, happy, angry, calm, fearful, surprised, disgusted, and neutral, is categories in this study. This dataset contains 80% training samples and 20% testing samples.

CREMA-D: The CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) dataset is the most difficult to utilize because it incorporates 7442 recordings from ninety-one actors and actresses (43 female and 48 male) representing a variety of racial and cultural backgrounds [28]. Actors recited twelve words that represented six different emotional states: happy, neutral, disgusted, fearful, and sad. In this case, 80% of samples are taken for training and 20% for testing. The hyperparameter setting for the proposed model is discussed in Table 2.

Table 2. Hyperparameter setting

Methods	hyperparameter	Settings
WLF	Adaptation parameter	0.25
	Activation function	ReLU
	dropout	0.5
	Momentum	0.9
	Mini batch size	128
	L2 regularization	10^{-4}
	Output size	512
	loss	Cross entropy function
DAttNN	Batch size	100
	Learning rate	0.001
	Number of hidden layers	3
	Activation function	ReLU
	optimizer	Adam
Conv_DCN	L2 regularization term	0.001
	Activation function	ReLU
	Batch size	100
	Learning rate	0.001
	optimizer	adam
	filter	32
	Kernel size	3

4.2 Evaluation metrics

The calculation metrics of the proposed method, like precision, accuracy, etc., were compared and evaluated. Table 3 shows the comparison table of performance metrics.

Table 3. Performance metrics

Metrics	Formula
Accuracy- It is a term for a quantitative metric used to evaluate how well a model or system predicts or estimates the intended result.	$Acc = \left(Tru_{po} + Tru_{ne}\right) \Big/ \left(Tru_{po} + Fal_{po} + Tru_{ne} + Fal_{ne}\right)$
Precision-Precision indicates the degree of accuracy of your model when it makes positive predictions.	$pre = Tru_{po} \Big/ Tru_{po} + Fal_{po}$
Recall- Recall is defined as the ratio of true positives (TP) to the total of false negatives (FN).	$recall = Tru_{po} \Big/ Tru_{po} + Fal_{ne}$
F1 score- It balances the significance of the classifier's precision and recall by combining them into a single metric.	$F1 = \left(\Pr e * recall\right) \Big/ \left(pre + recall\right)$

Where, Tru_{po} indicates true positive rate, Tru_{ne} represent true negative rate, Fal_{po} represent false positive, and Fal_{ne} indicates false negative rate.

4.3 Performance Evaluation and Comparison Analysis

The comparison analysis and performance measures of proposed and existing techniques of the RAVDESS dataset are mentioned in this section. Fig. 7 shows the training testing accuracy of the proposed method.

RAVDESS dataset: From the above fig., when the epoch value is 200, the training accuracy range starts rising, and for all other epochs, the accuracy remains constant. For testing accuracy, when the epoch value is 200, the accuracy starts rising and remains constant for all other epochs. Testing accuracy makes a slight difference from training. Fig. 8 shows the training testing loss of the proposed technique.

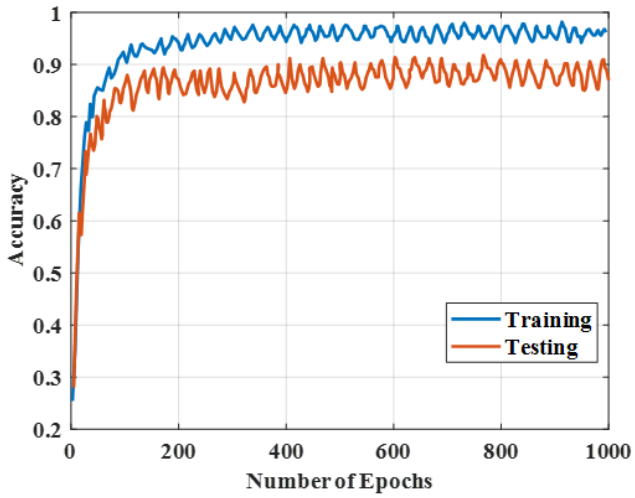


Fig. 7. Training testing accuracy

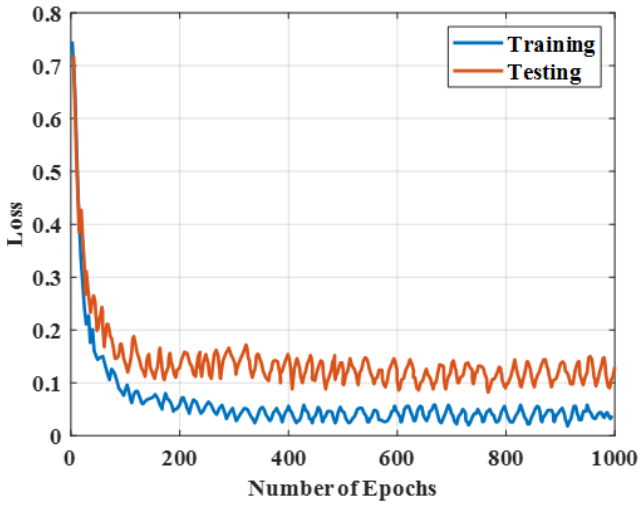


Fig. 8. Training testing loss

From the above graph, when the value of epoch is 0 to 200, the training loss starts to fall down. When the epoch value is above 200, the loss becomes very low and remains constant for all other epoch values. For testing loss, the epoch value is 0 to 200, and the loss value starts decreasing and remains constant for all other epochs. Fig. 9 indicates the confusion matrix of the RAVDESS dataset.

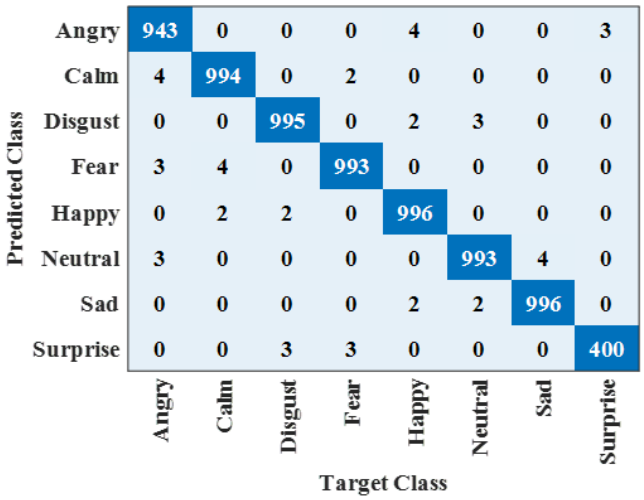
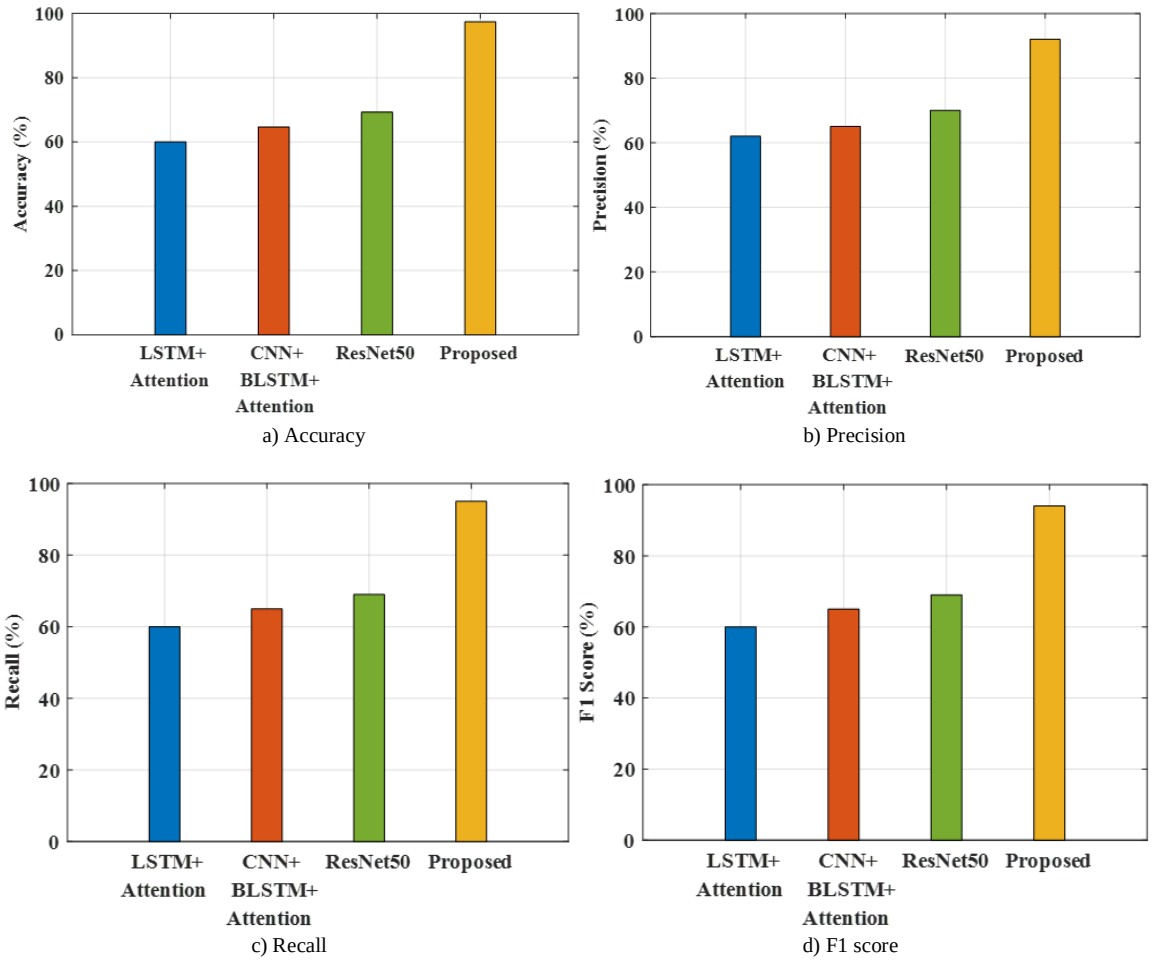


Fig. 9. Confusion matrix of the RAVDESS dataset

From the above graph, the RAVDESS dataset contains 8 classes, from which 943 classes are totally predicted for anger. Here, 4 classes are correctly predicted for calm, 3 classes for fear, 3 classes for neutral, and no classes are predicted for disgust, happy, sad, and surprise. For calm, a total of 994 classes are correctly predicted; from that 4 classes are predicted for fear, 2 classes are predicted for happiness, and all other classes are wrongly predicted. Fig. 10 shows the comparison of both proposed and existing techniques.



a) Accuracy b) Precision

c) Recall d) F1 score

Fig. 10. (a)-(d): Evaluation analysis of suggested and existing approach

Fig. 10(a) shows the accuracy range of other existing approaches. The accuracy value of existing techniques is 60%, 64.6%, and 69.33%, and for the proposed method, it is 97.38%. Fig. 10 (b) indicates the precision range of existing approaches is 62%, 65%, and 70%, and the proposed method achieves 92%. Fig. 10 (c) represents the recall value of existing techniques at 60%, 65%, 69%, and proposed is 95%. Fig. 10 (d) indicates the f1 score of the proposed is 94%, and for existing, it is 60%, 65%, and 69%. Compared with other existing methods, the proposed method achieves better results. Table 4 shows the comparison analysis of the RAVDESS dataset. Fig. 11 indicates the batch size.

Table 4. Comparison analysis

Metrics (%)	Proposed	LSTM+attention	CNN+BiLSTM_attention	ResNet-50
Accuracy	97.3	60	64.6	69.3
Precision	92	62	65	70
Recall	95	60	65	69
F1 score	94	60	65	69

The above graph indicates that the accuracy value can be found by altering the batch size. When the batch size is 16, the accuracy starts to fall down, and when the batch size is 32 to 64, it slightly rises. When the batch size is 64 to 128, the accuracy value increases. Compared to the existing approach, the proposed approach yields lower outcomes. Fig. 12 shows the learning rate of the existing and proposed approach.

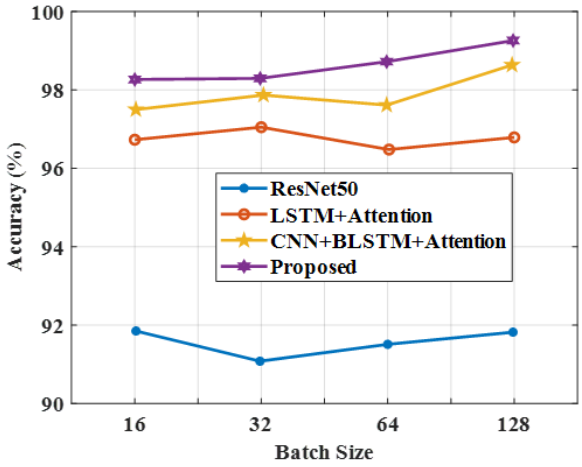


Fig. 11. Batch size vs Accuracy

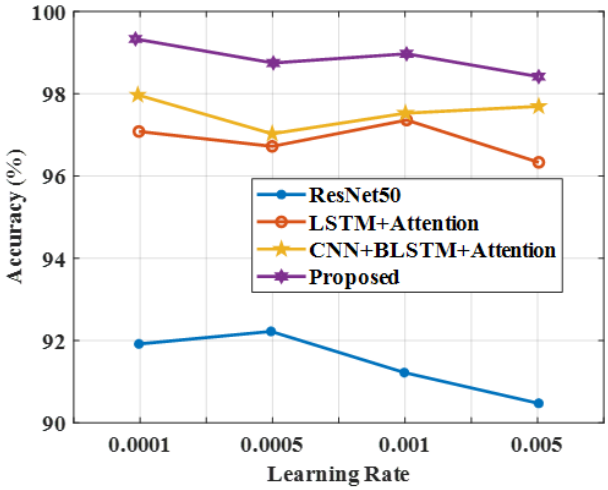


Fig. 12. Learning rate

The above graph indicates by varying the learning rate, the accuracy range is identified. As compared with the existing approach, the proposed received a low learning rate. Fig. 13 shows the training testing accuracy of the CREMA-D dataset.

CREMA-D dataset: When the epoch value is 0 to 200, the training accuracy starts rising. When the epoch value is above 200, the accuracy rises very high and remains constant for all other epoch values. Testing is similar to training, but there are some differences. Fig. 14 shows the loss of training and testing.

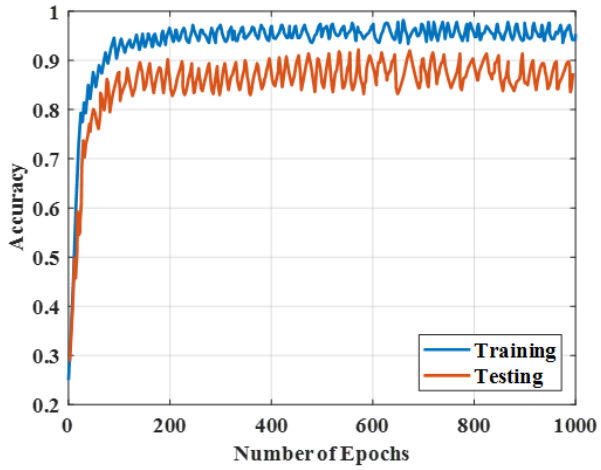


Fig. 13. Training testing accuracy of CREMA-D dataset

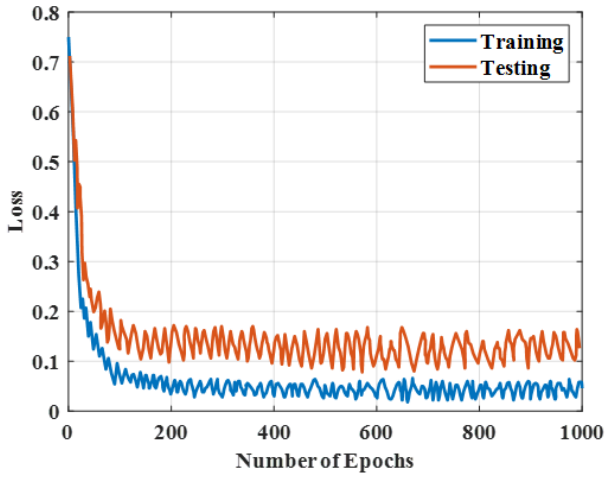


Fig. 14. Training testing loss of CREMA-D dataset

When the epoch value is 0, the loss of training starts falling off. When the number of epochs is 200, the training loss starts decreasing and remains constant for every other number of epochs. Testing is similar to training, except testing includes some variation in loss values. Fig. 15 indicates the confusion matrix of the CREMA-D dataset.

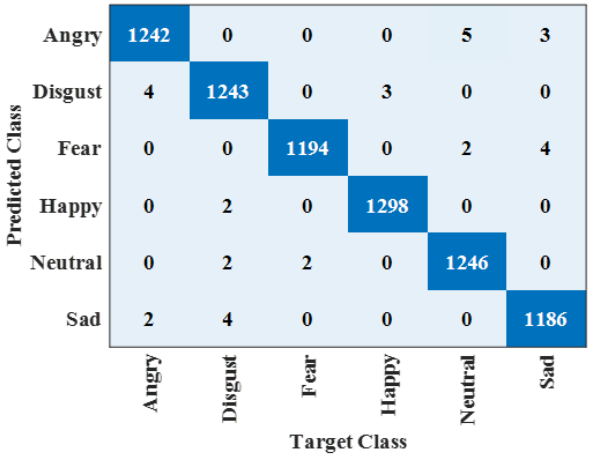


Fig. 15. Confusion matrix of CREMA-D dataset

The confusion matrix contains 6 classes such as angry, disgusted, fearful, happy, neutral, and sad. For angry, the total predicted class is 1242, 4 predicted class for angry, 2 predicted class for sad, and all are not correctly predicted. The total predicted class for disgust is 1243; from that, 2 classes are predicted for happy, 2 for neutral, 4 predicted for sad, and all are wrongly predicted. Fig. 16 indicates the comparison analysis of the CREMA-D dataset.

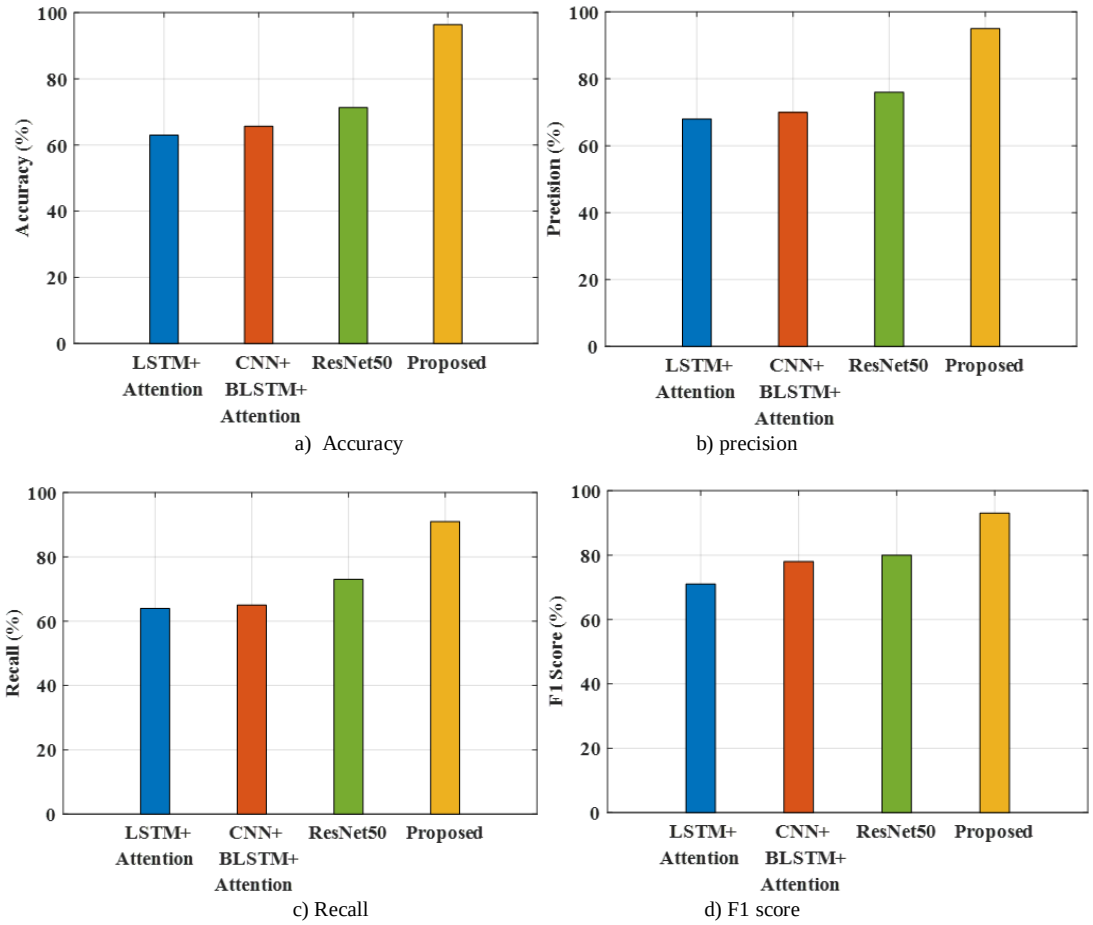


Fig. 16. (a)-(d): Comparison analysis of existing as well as proposed approach

Fig. 16(a) shows that the accuracy of the proposed technique is 96.3%, while for other existing techniques, it is 63%, 65%, and 71%. Fig. 16 (b) indicates the precision of existing techniques is 68%, 70%, 76%, and for the proposed, it is 95%. Fig. 16 (c) indicates the recall range of the proposed is 91%, and for existing are 64%, 65%, and 73%. Fig. 16 (d) represents the f1 score of the proposed, which is 93%, and for existing, it is 71%, 78%, and 80%. Table 5 shows the comparison analysis of the CREMA-D database. Fig. 17 shows the batch size of the proposed technique.

Table 5. Comparison analysis of existing and proposed approach

Metrics (%)	Proposed	LSTM+attention	CNN+BiLSTM_attention	ResNet-50
Accuracy	96	63	65	71
Precision	95	68	70	76
Recall	91	64	65	73
F1 score	93	71	78	80

The above fig. indicates that, by varying batch size, the accuracy range is calculated. When the batch size is 16, the accuracy value starts increasing. When the batch size is 32, accuracy starts to rise, and at 128, it increases at the highest value. Compared with the existing, the proposed achieves a low outcome. Fig. 18 shows the learning rate.

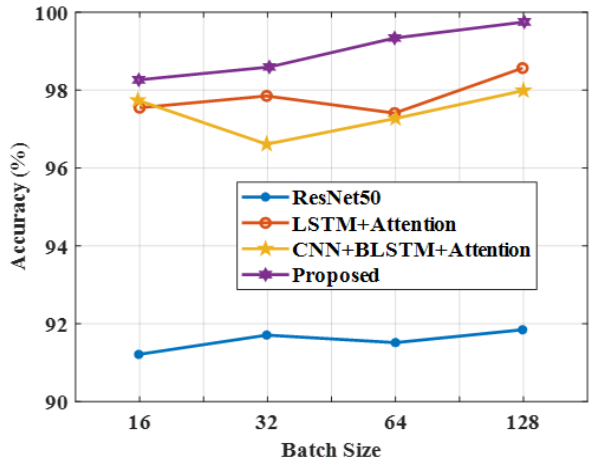


Fig. 17. Batch size vs accuracy

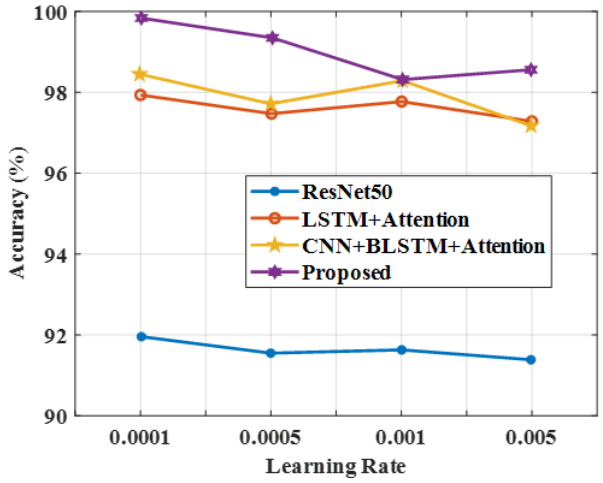


Fig. 18. Learning rate

From the above graph, accuracy is determined by varying learning rates. The learning rate of the proposed method receives low outcomes when compared with the existing approach. Fig. 19 indicates the processing time of the proposed technique.

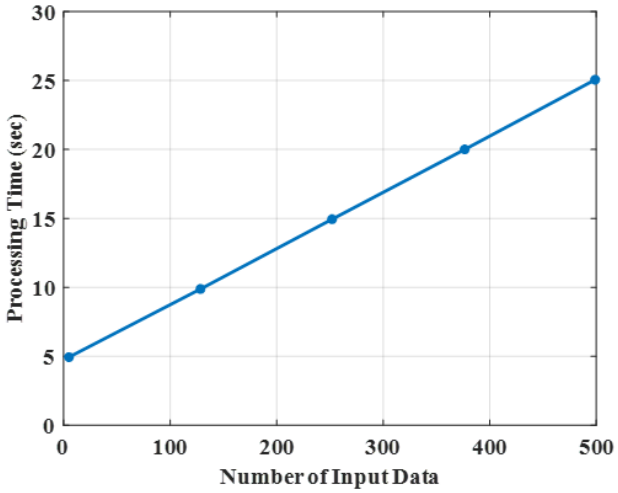


Fig. 19. Processing time

From the above fig., the processing time of the proposed achieves better results. When the number of input data is 0, the processing time is 5 seconds. When the number of data is 100, the processing time increases up to 10s. When the number of data increases, the processing time also increases. Fig. 20 shows the ROC curve of the proposed approach.

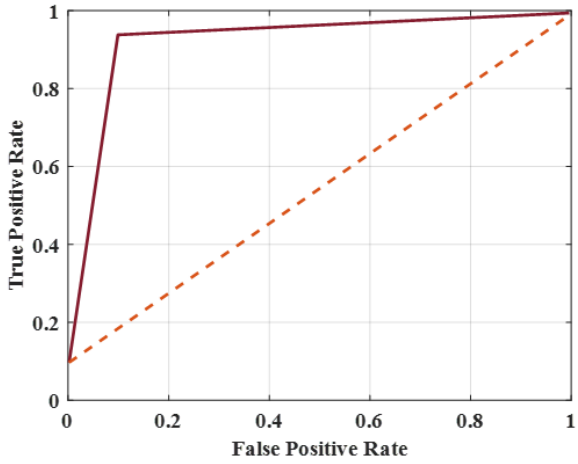
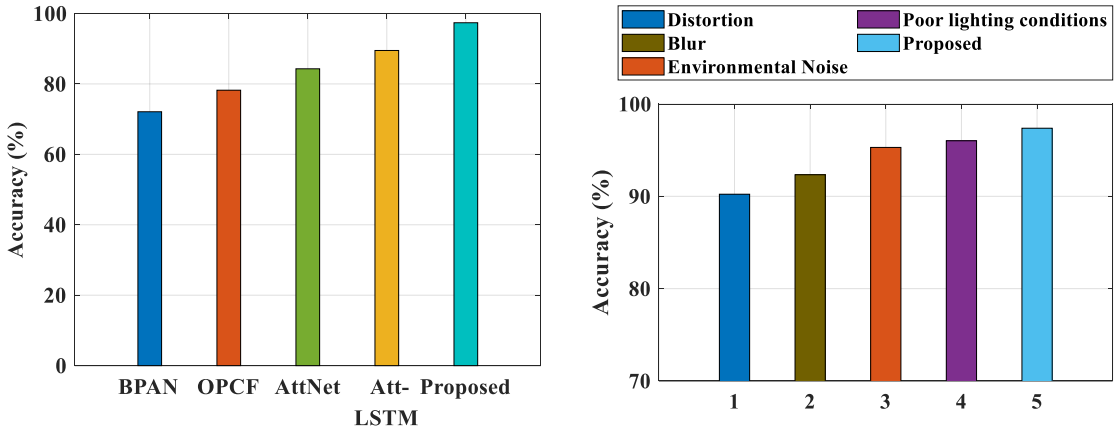


Fig. 20. ROC curve

The Receiver Operating Characteristic Curve (ROC) calculates the true and false positive range for various edge settings to identify the precise effectiveness of a recommended strategy. The area under the Curve (AUC) is selected and correlated with ROC. The suggested approach performed better with a higher ROC value than the existing ones.

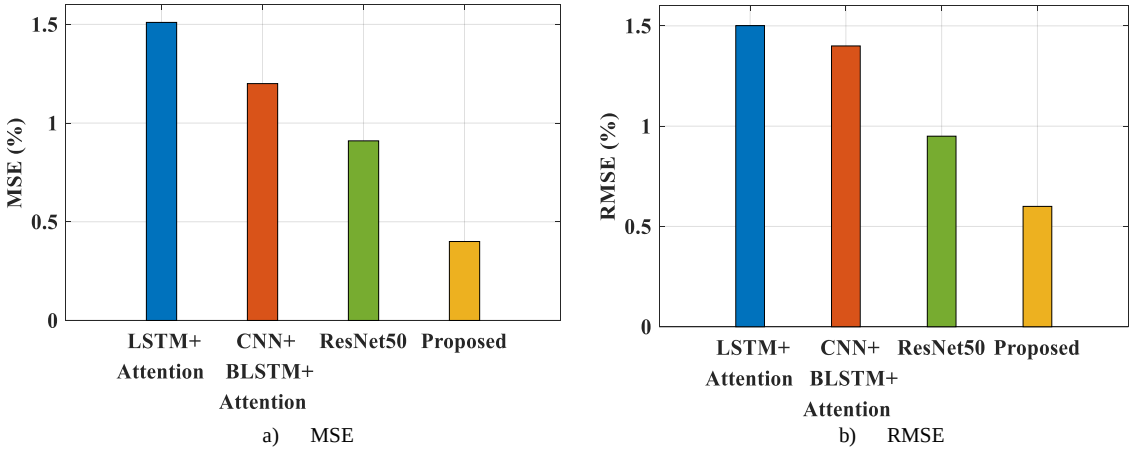


a) Feature fusion model based comparison

b) Noise sensitivity analysis

Fig. 21. Feature fusion based comparison, noise sensitivity analysis

The feature fusion based comparison analysis is shown in Fig. 21 (a). The noise sensitivity analysis for the proposed and existing models is shown in Fig. 21 (b). the existing feature fusion models that are taken for comparison are BPAN (Bilinear Pooling and Attention Network), OPCF (over parameterized convolutional fusion model), AttNet (attentional fusion network), and AttNet-LSTM (attention based LSTM fusion network).



a) MSE

b) RMSE

Fig. 22. MSE and RMSE analysis

The MSE and RMSE analysis for proposed and existing models is shown in Fig. 22 (a and b). The existing models that are taken for comparison are LSTM+Attention, CNN+BiLSTM+Attention, and ResNet50. The MSE is computed by dividing the number of observations by the sum of squared discrepancies between the expected and observed values. RMSE is a popular statistic for assessing the precision of prediction models in data analysis and machine learning. The average size of the variations between the expected and actual values is basically what it measures.

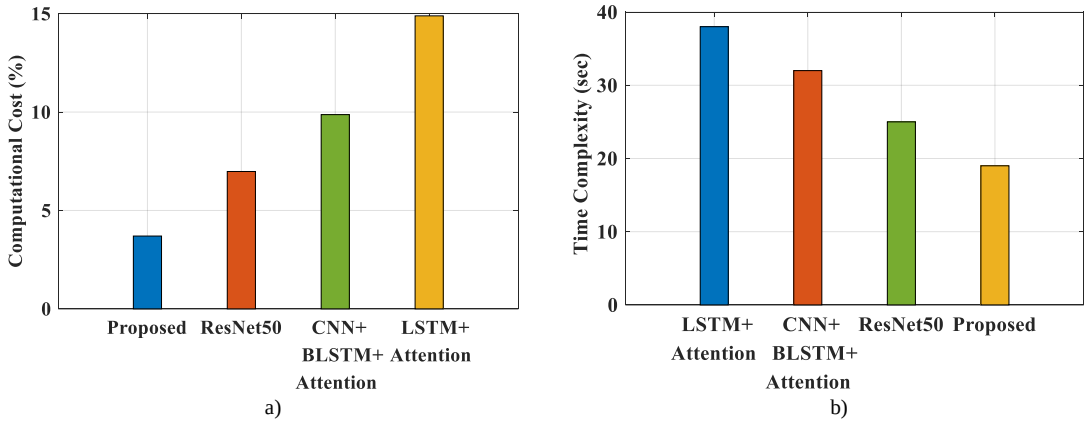


Fig. 23. Computational cost and computational time analysis

The computational cost and time analysis are shown in Fig. 23 (a & b). The existing models that are taken for comparison are LSTM+Attention, CNN+BiLSTM+Attention, and ResNet50. The cost obtained by the proposed model is found to be less than the other existing models. This is because of the feature fusion model integrated within the proposed framework. The fusion strategy has the ability to fuse the most relevant features. Therefore, the learning process can be accomplished in less time duration.

Table 6. Ablation study analysis

	Module1	Module2	Module3	Module4	Module5
Accuracy	97.38	93.4	90.25	88.26	82.10
Precision	92	90.87	84.78	79.45	75.23
Recall	95	92.27	89.99	87.23	81.01
F1-score	94	91.23	89.15	82.78	79.41

The ablation study analysis for the proposed model is shown in Table 6. Module 1 (proposed), module 2 (without DenseNet-169), module 3(without Dense Capsule Network), module 4 (without Modified MFCC), and Module 5 (without Constant-Q chromagram).

The integrity of feature extraction from multimodal data is preserved while feature extraction efficiency is increased using the technique. Additionally, the information from several senses is combined through a multimodal attention process. An efficient allocation of information processing resources is made possible by a multimodal attention mechanism, which accurately gathers crucial information. Based on the weighted distribution of attention to specific local information within the input data, this method increases the attention scores to locally important information while selectively ignoring lower-weighted local details. This approach enhances the model's robustness and scalability, fostering adaptability to shifting input conditions. Despite depending on frame-based and keyframe extraction, the architecture might not be able to handle longer temporal sequences since it does not explicitly describe long-range temporal dependencies. Because of this, the suggested model might not be able to accurately record or learn the links between frames that are separated in time, which could result in errors in tasks that call for comprehension of the complete sequence. This is considered the major limitation of the proposed model, which can be overcome in the future by integrating the temporal sequence analysis model.

5. Conclusion and Future Study

In this study, the audio features are extracted using Constant-Q chromagram as well as Mel-Frequency Cepstral Coefficients (MM-FC^2) model, and for performing audio generation, Convolutional based Dense Capsule Network (Conv_DCN) is used. Next is video data; here, the keyframes were extracted using the Enhanced spatial-temporal and Second-Order Gaussian kernels method. To perform video generation, DenseNet-169 is used. Finally, all the extracted features are concatenated using weighted late fusion. To detect the emotions, a Deep Attention Neural Network (WLF_DAttNN) is used. The performance metrics of the RAVDESS dataset accuracy is 97%, and the accuracy range of the CREMA-D dataset is 96%. Further, the complexities in the dataset further reduce the efficiency of the proposed model. This will also be overcome in the future by introducing an efficient pre-processing model. This model restricted evaluation over demonstrating robustness across diverse real-world conditions, including spontaneous emotions, varied

cultural expressions, and different languages beyond English. Therefore, in future work, cross-domain research can be carried out by integrating the MED4 database with other databases utilizing adaptive techniques in order to construct algorithms that are resilient to varied environmental settings.

References

[1] W. Liu, J. L. Qiu, W. L. Zheng, and B. L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 715-729, 2021.

[2] Z. Ma, F. Ma, B. Sun, and S. Li, "Hybrid mutimodal fusion for dimensional emotion recognition," In Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, pp. 29-36, 2021.

[3] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, and D. Jiang, "Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models," Knowledge-Based Systems, vol. 235, p. 107598, 2022.

[4] P. Bhattacharya, R. K. Gupta, and Y. Yang, "Exploring the contextual factors affecting multimodal emotion recognition in videos," IEEE Transactions on Affective Computing, 2021.

[5] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," In Proceedings of the 28th ACM international conference on multimedia, pp. 2852-2861, 2020.

[6] C. M. A. Ilyas, R. Nunes, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Deep Emotion Recognition through Upper Body Movements and Facial Expression," In VISIGRAPP (5: VISAPP), pp. 669-679, 2021.

[7] S. Liu, P. Gao, Y. Li, W. Fu, and W. Ding, "Multi-modal fusion network with complementarity and importance for emotion recognition," Information Sciences, vol. 619, pp. 679-694, 2023.

[8] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," Biomedical Signal Processing and Control, vol. 70, p. 103029, 2021.

[9] J. Chen, C. Wang, K. Wang, C. Yin, C. Zhao, T. Xu, and T. Yang, "HEU Emotion: a large-scale database for multimodal emotion recognition in the wild," Neural Computing and Applications, vol. 33, pp. 8669-8685, 2021.

[10] Y. Yao, M. Papakostas, M. Burzo, M. Abouelenien, and R. Mihalcea, "Muser: Multimodal stress detection using emotion recognition as an auxiliary task," arXiv preprint arXiv:2105.08146. 2021.

[11] M. Martinc, and S. Pollak, "Tackling the ADReSS Challenge: A Multimodal Approach to the Automated Recognition of Alzheimer's Dementia," In Interspeech, pp. 2157-2161, 2020.

[12] A. Nirmal, D.Jayaswal and P. H. Kachare, "A Hybrid Bald Eagle-Crow Search Algorithm for Gaussian mixture model optimization in the speaker verification framework," Decision Analytics Journal, p. 100385, 2023.

[13] P. Baki, A Multimodal Approach for Automatic Mania Assessment in Bipolar Disorder. arXiv preprint arXiv:2112.09467. 2021.

[14] C. Li, Z. Bao, L. Li and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," Information Processing & Management, vol. 57, no. 3, p. 102185, 2020.

[15] Q. Wang, M. Wang, Y. Yang and X. Zhang, "Multi-modal emotion recognition using EEG and speech signals," Computers in Biology and Medicine, vol. 149, p. 105907, 2022.

[16] B. Xie, M. Sidulova and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," Sensors, vol. 21, no. 14, p. 4913, 2021.

[17] N. H. Ho, H. J. Yang, S. H. Kim and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," IEEE Access, vol. 8, pp. 61672-61686, 2020.

[18] L. Sun, Z. Lian, J. Tao, B. Liu and M. Niu, "Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism," In Proceedings of the 1st international on multimodal sentiment analysis in real-life media challenge and workshop, pp. 27-34, 2020.

[19] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu and D. Zhang, "Multimodal emotion recognition with temporal and semantic consistency," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3592-3603, 2021.

[20] M. F. Carbonell, M. Boman and P. Laukka, "Comparing supervised and unsupervised approaches to multimodal emotion recognition," PeerJ Computer Science, vol. 7, p. e804, 2021.

[21] Y. Zhang, C. Cheng and Y. Zhang, "Multimodal emotion recognition using a hierarchical fusion convolutional neural network," IEEE access, vol. 9, pp. 7943-7951, 2021.

[22] D. N. Krishna and A. Patil, "Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks," In Interspeech, pp. 4243-4247, 2020.

[23] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero and F. Fernández-Martínez, "Multimodal emotion recognition on RAVDESS dataset using transfer learning," Sensors, vol. 21, no. 22, p. 7665, 2021.

[24] E. S. Salama, R. A. El-Khoribi, M. E. Shoman and M. A. W. Shalaby, "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," Egyptian Informatics Journal, vol. 22, no. 2, pp. 167-176, 2021.

[25] M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," Biomedical Signal Processing and Control, vol. 78, p. 103970, 2022.

[26] S. Alagarsamy, V. Govindaraj, M. Irfan, R. Swami and N. M. Kumar, "Smart recognition of real time face using convolution neural network (CNN) Technique," vol. 83, pp. 23406-23411, 2021.

[27] A. Vulli, P. N. Srinivasu, M. S. K. Sashank, J. Shafi, J. Choi and M. F. Ijaz, "Fine-tuned DenseNet-169 for breast cancer metastasis prediction using FastAI and 1-cycle policy," Sensors, vol. 22, no. 8, p. 2988, 2022.

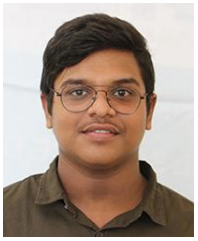
[28] M. R. Ahmed, S. Islam, A. M. Islam and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," Expert Systems with Applications, vol. 218, p. 119633, 2023.

Authors' Profiles

P. V. V. S. Srinivas is an Associate Professor at K L University with significant expertise in computer science and engineering, particularly in deep learning and data science. His work focuses on applying advanced computational approaches to medical diagnostics, renewable energy, and machine learning optimization problems. Dr. Srinivas has published in several prominent journals, contributing to research on topics like facial emotion recognition, photovoltaic systems, and distributed computing. His research has garnered over 400 citations, reflecting the impact of his collaborative studies. Beyond publishing, he actively mentors students, participates in peer review, and fosters interdisciplinary academic networking, positioning himself as a prominent figure in advancing computational research in academia.

Shaik Nazeera Khamar is an information technology professional currently working as a Fullstack web Developer, with prior experience at Codetree. Nazeera is an alumna of KL University, which equipped her with both theoretical knowledge and practical skills relevant to the fast-evolving tech landscape. Through her career, she has demonstrated expertise with various platforms and digital transformation initiatives. Nazeera is recognized among her peers for collaborative teamwork, adaptiveness, and a rapid learning curve that enables her to solve complex business and technical challenges efficiently.

Nohith Borusu completed his B.Tech in Computer Science and Information Technology (CSIT) from KL University, specializing in Machine Learning, Data Science, and Artificial Intelligence. During my studies, he worked on a research paper on multimodal emotion recognition, focusing on how multiple data sources like text, speech, and visuals can improve emotion detection. Currently, he is working at Microland as a ServiceNow Developer in the ServiceNow Center of Excellence, where he designs catalog items, build workflows, and develop automation solutions. His goal is to bridge AI research and enterprise IT innovation.



Mohan Guru Raghavendra Kota is an emerging professional with strong dedication to continuous learning and growth. With a foundation built on technical expertise and problem-solving skills, he demonstrates the ability to adapt quickly and contribute effectively in dynamic environments. His focus on collaboration, innovation, and results makes him a valuable team player. Mohan is passionate about applying his knowledge to practical challenges, ensuring measurable impact in every role he undertakes. Driven, detail-oriented, and future-focused, he is committed to building a career that blends skill development with meaningful contributions.

Harika Vuyyuru completed her B.Tech in Computer Science and Information Technology (CSIT) from Koneru Lakshmaiah (KL) University. During her academic journey, she specialized in Machine Learning, Data Science, and Artificial Intelligence, areas that fascinated him because of their immense potential to transform the future of technology. As part of my academic research, she worked on a paper focused on multimodal emotion recognition, which explores how machines can better understand human emotions by combining multiple inputs such as speech, facial expressions, and text. This project allowed her to experiment with advanced deep learning models and multimodal data processing techniques. It not only strengthened her technical expertise but also gave her valuable research experience with real-world applications in fields such as healthcare, interactive systems, and digital well-being.

Sampath Patchigolla is a student with academic experience at KL University and GITAM Deemed to be University, where he pursued studies in engineering and technology. His academic engagement has included project work and internships, notably at Elitelogix Exim Agency, gaining practical exposure in finance and management along with his technical education. Sampath has been involved in research activities related to machine learning, facial emotion recognition, and computational intelligence as part of collaborative academic projects, contributing to innovations in deep learning applications. He embodies a commitment to interdisciplinary learning and research, aspiring to advance in technical fields by leveraging his educational foundation and practical experiences to solve real-world problems effectively.

How to cite this paper: Srinivas P. V. V. S., Shaik Nazeera Khamar, Nohith Borusu, Mohan Guru Raghavendra Kota, Harika Vuyyuru, Sampath Patchigolla, "Weighted Late Fusion based Deep Attention Neural Network for Detecting Multi-Modal Emotion", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.1, pp. 106-127, 2026. DOI:10.5815/ijigsp.2026.01.07