Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 22s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Deep learning Multimodal feature enhanced with cross-modal attention for emotion recognition

Shwetkranti Taware[*1], Anuradha Thakare[2], Manisha D. Kitukale[3]

[*1]Research Scholar, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune, and Maharashtra, India

[2]Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University Pune, India

Email ID: anuradha.thakare@pccoepune.org

[3]Professor, P. Wadhwani College of Pharmacy, Yavatmal, India

Email ID: kitukalemanisha5@gmail.com

***Corresponding Author:**

Shwetkranti Taware

Email ID: shweta.taware@gmail.com

Cite this paper as: Shwetkranti Taware, Anuradha Thakare, Manisha D. Kitukale (2025) Enrichment of honey with flavour of ginger. *Journal of Neonatal Surgery*, 14 (22s), 845-858.

ABSTRACT

Human emotion recognition is becoming very important in human-computer interaction applications and stress management applications. In these applications, multimodal features extracted from various modalities like visual, speech, text, and biomedical sensor readings etc. to classify various emotions. For effective emotion recognition there has been extensive research on various modalities and their features which are to be considered for classification of emotions. This work explores multiple modalities and proposes a novel deep learning based multimodal features which are enhanced by using cross modality attention for recognition of basic human emotions The proposed solution provides higher discriminative ability in three cases of emotions like basic emotion, activation of emotion (positive, negative) and arousal of emotion (high, low).

Keywords: *Multimodal Emotion Recognition, Deep Learning Multimodal Feature Extraction, Cross-Modal Attention, Multivariate LSTM*

1. INTRODUCTION

Human conversations are largely nonverbal, with studies indicating that 70-80% are nonverbal, distributing in visual and vocal with 55% and 38%, respectively. (A. Mehrabian and S. R. Ferris 1967). Emotional information is conveyed through nonverbal mechanisms like facial expressions, smiles, eye movements, variance in pitch, body postures, dysfluency in speech, loudness, etc. This emotion information can reveal true mental state of the observers in terms of anger, happiness, frustrated etc. It conveys more information than verbal communication. Understanding the true emotion of a person is important for better communication with him and making various conclusions about the person's mental health. Emotion recognition has become a very important application in areas like Human-Computer Interaction (HCI), psychiatric evaluation, candidate screening, human resource management etc. Affecting computing is the area, which covers the technique for automatic detection of emotion from various non-verbal cues and sensor measurements. The earliest work in affecting computing is the modeling of emotions using a nonlinear sigmoid function (R. W. Picard et al., 2001) by Rosalind Picard, who was a pioneer in the area of affecting computing. Over last two decades, many works have been proposed in the area of affective computing. Various modalities like visual cues, speech cues, text cues, physiological signals have used for emotion recognition. The emotion recognition approaches were in two categories like unimodal and multimodal. Visual cues mostly used Facial Action Coding System (FACS) which was developed on basis of Ekman's theory for emotion recognition. (P. Ekman and V. Wallace 2003; W. V. Friesen and P. Ekman). Features extracted from facial landmarks like eye contour, mouth regions, etc. were used to detect emotions. Various FACS approaches were proposed with different number of landmarks and correlating landmarks to different emotions. Among those approaches, most notable one is Ekman et al. work, which proposed 44 facial actions units and later revised it to 68 facial actions units. Speech modality approaches extracted spectral and acoustic features from speech and classified the features as emotions. Physiological modality approaches used the various measurements like heart rate, galvanic skin response, skin temperature, electrocardiogram (ECG), electromyogram (EMG) etc. Features extracted from these measurements are used to classify the emotions. Multi-modal

approaches used visual, speech, text, and physiological modalities in different combinations. The most challenging part of emotion recognition is that there is no standard set of features in different modalities with higher discriminative ability for emotions, emotion activation (negative/positive), and emotion arousal (high/low).

Addressing this gap, this research proposes a deep learning based multimodal feature with higher discriminative ability for emotions, emotion activation and emotion arousal. Feature extracted from modalities using existing methods enriched with a novel loss feedback controlled convolutional neural network. The discriminative ability of features in classifying emotions, emotion activation, and emotion arousal is tested using entropy and clustering analysis. The novel contributions of this work are listed below.

- Multimodal feature extraction with cross-modal attention to improve the cross-reference learning across modalities with higher discriminative ability for emotions, emotion activation, and emotion arousal.
- An enhanced Densenet to learn more intricate feature representation from images.
- Incremental cross-modality for learning from multiple modalities in a scalable way.

The rest of the paper is organized as follows. Section II presents the survey of various multimodal features for emotion recognition. Section III presents the proposed deep learning multi-modal feature. Comparison of results of the proposed solution with existing works and discussion on results presented in Section IV. Section V presents the conclusion and scope of future work.

2. RELATED WORK

(W. L. Zheng et al., 2019) extracted features from an Electroencephalogram (EEG) and an eye movement to recognize four emotions happy, sad, fearful, and neutral. Power spectral density and differential entropy extracted from the EEG signal as features. The features like pupil diameter, dispersion and statistics as fixation duration, blink duration, saccade extracted from eye movements. Though the method was able to achieve about 85% accuracy, the events were observed in a controlled environment while watching movies of different emotions. (D. Nguyen et al., 2017) extracted features from video and audio streams to classify six basic emotions of surprise, anger, disgust, happiness, sad, and fear. Spatio-temporal features are extracted from face regions of the video using a three-dimensional convolutional neural network. Spectrogram of speech signals extracted by using short time fast Fourier transform. In this work, the proposed spatio temporal features do not recognize spread of emotions and micro expressions. (P. Tzirakis et al., 2017) extracted spatiotemporal features from video and audio streams using deep learning networks. Resnet 50 network extracts spatiotemporal features from the face regions of the video. A recurrent network with LSTM (Long Short Term Memory) cells extracts spatiotemporal features from audio segments. This approach has not detected temporal emotional variance. (Zhang et. al. 2020) extracted features from EEG, EMG, Galvanic Skin Response (GSR), and Respiration Signals (RES) physiological signals to classify emotions. Power spectral density extracted from EEG. Power and statistical moments extracted from EMG. Number/Amplitude of peaks, rise time, and statistical moments extracted from GSR. Main frequency, power spectral density, and statistical moments extracted from RES. With physiological signals, only 57 % of emotion recognition accuracy has been obtained. (B. Chen et al. 2021) extracted acoustic and textual features to classify emotional state, Speech text embedding is optimized and fine-tuned using the learning process. Speech and text feature were learned and refined by cross modal semantic interaction and temporal alignment. (Y. Cimtay et al. 2020) combined facial expression analysis, GSR, and EEG analysis for emotion recognition. Face image passed to Convolution Neural Network (CNN) to predict seven emotion classes (disgust, angry, afraid, neutral, happy, surprised and sad). GSR and EEG signals are windowed and passed as a whole to CNN to extract CNN features. The decision tree classifies the CNN features which are extracted from facial expression, EEG and GSR. The results were snapshot based without temporal analysis of emotions. (Zhou et al. 2021) Extracted audio and video features and used it for emotion classification. Audio features extracted from a fully convolutional network. Video features extracted from the visual geometry group (VGGNet). The features are fused using bilinear pooling. The fused features are used for emotion classification. The method was able to achieve 63% accuracy. (Wu et al., 2021) extracted EEG and eye movement features and used for emotion classification. EEG topological features of strength, clustering coefficient, and Eigenvector centrality were extracted using the brain connectivity toolbox. Eye movement features like pupil diameter, fixation duration, blink duration etc. were extracted. The extracted features were classified using Canonical correlation analysis model. (Dai et al. 2021) proposed a multi-modal emotion recognition system with audio, video, and text modality. Glove embedding features extracted from text. Seventy-four different features like fundamental frequency, spectral parameters, wavelet responses, etc. extracted from speech. Thirty-five facial action units were extracted from faces in videos. The method was able to achieve only about 60% accuracy. (Lee et. al. 2021) extracted hand-crafted features from video, audio, and text and fused them to a high-level representation for emotion recognition. Facial landmarks-based features and VGG16 features extracted from videos. Loudness, pitch, jitter, and Mel Frequency Cepstral Coefficient (MFCC) extracted from the audios. Word tokens are extracted from text using bidirectional encoder representation from transformers (BERT). (Noroozi et al., 2017) proposed a multimodal emotion recognition system based on audio and visual cues. Facial landmarks distance and angles features extracted from a video. MFCC, filter bank energies, and prosodic features extracted from a video. The CNN choose frame

from video for feature extraction. (K. Zhang et al., 2021) extracted deep learning features from each of the modalities of audio, video, and text using gated recurrent unit layers. Deep canonical correlation analysis based on the encoder-decoder network was used to extract cross-modal correlations. However, the correlation is done in a smaller time window in this method. (Li et al., 2018) extracted features from multichannel EEG signals and used for emotion recognition. Discrete wavelet transformation was applied to extract four frequency bands. From these bands, entropy, and energy are calculated as features. The features are classified into emotions using the k- nearest neighbor (k-NN) classifier. (Cai et al. 2020) combined speech and facial features to classify emotion. Facial expression features extracted from multiple small-scale kernel convolutional blocks. CNN with LSTM is used to extract speech features. Deep neural networks are used to fuse both facial and speech features. (Nath et al., 2020) extracted band power features from the EEG signal and used it for classifying valence and arousal using the LSTM classifier. (Guo et al., 2017) extracted time domain and discrete wavelet transform features from the EEG signals and classified them into two classes of valence and arousal using a combined support vector machine (SVM) and Hidden Markov Model (HMM) classifier. (Pandey et al., 2019) applied variation mode decomposition to extract features from EEG. The features are classified into two labels of valence and arousal using a deep neural network (DNN) classifier. (W. Zheng et al., 2019) extracted differential entropy features from EEG signals and classified them using an extreme learning machine. The method is classified to three labels of positive, negative, and neutral emotions.

From the survey, it inferred that compared to unimodal approaches, multimodal approaches have higher accuracy. However, in most work, features were extracted individually without exploiting the cross dependencies for avoiding false positives in emotion classification. Most approaches were snapshot based without consideration for the temporal variance of emotion over a period. Most of the approaches used a fixed time duration of window, which missed out micro-expression, which are strong indicators of true emotions. This work addresses these problems and proposes effective multimodal features for fine-grained emotion classification.

3. DEEP LEARNING MULTIMODAL FEATURE

The architecture of the proposed Multimodal feature extraction framework is given in Figure 1. The multimodal inputs considered in this work are video, audio, text, and EEG. However, the work can be extended for other features.

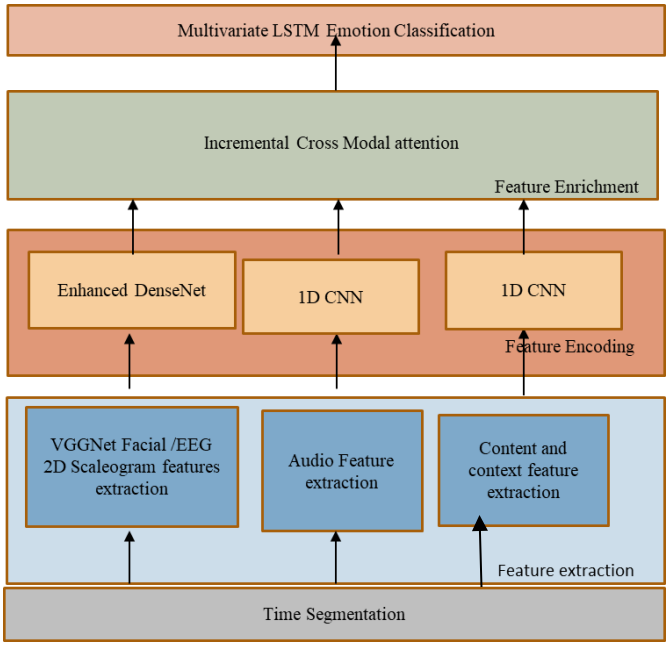

Figure 1 Multimodal Feature Extraction Framework

Cross-dependencies between multimodal inputs are exploited using cross-modal attention. Instead of fixed time duration, the time duration for segmenting the inputs is found by marking salient regions and time segmenting the inputs. The proposed multimodal feature extraction technique has the following stages, (i) time segmentation of inputs (ii) feature extraction (iii) cross-modality feature enrichment (iv) emotion classification. Each of the stages is detailed in the below subsections.

3.1 Time Segmentation

The difference in face region, silence region in audio segments, etc., unusual amplitude or frequency shift in EEG, etc. are the probable segmentation markers in the inputs. From the video face, regions are extracted frame-by-frame using the Voila Jones algorithm. From the facial regions, the following features were extracted.

1. Table 1
2. Facial features

Feature	Description of facial feature
$f1$	Distance between anchor landmarks and mouth contour landmarks
$f2$	Distance between anchor landmarks and mouth corner landmarks
$f3$	Eye contour region area
$f4$	Distance between anchor and eyebrow landmark
$f5$	Area of a polygon drawn around external landmarks
$f6$	The average value of the Euclidean norm of a set of landmarks compared to the last frame.

The Euclidean distance between the features between frames is calculated and when the distance is greater than a threshold, then the time of the frame is marked for segmentation. The facial features provide different texture changes while depicting the emotions. These features provide the impact of the emotion on the different facial region that helps to characterize the specific emotion.

3.2 Feature Extraction

In most existing works, spatiotemporal features from face regions are extracted using VGGNet. Compared to VGGNet, Densenet can learn more intricate features. Densenet is a deep learning model recently proposed to solve the problem of vanishing gradients in Resnet models. In this model, the convolutional features extracted at each layer are passed as input to subsequent layers. This improves the learning ability of Densenet to learn more intricate features. CNN performance generally increased by increasing the number of layers. However, increasing layers create a vanishing gradient problem. As depth increases, the features vanish in the longer path travel, and this reduces the discriminating ability of features. Densenet has many dense blocks which have various numbers of filters per block.In every block dimensions are unique. A transition layer is placed between blocks for batch normalization. Downsampling is done to match the dimensions of the subsequent layer. Densenet becomes overfit and computational complexity is higher. This work enhances the Densenet and uses it for deep feature representation from images. A few modifications to the original Densenet model are introduced in terms of a fully connected layer and strides for better performance. The architecture of the modified Densenet model is given in Figure 2.

The fully connected layer is replaced with a fully convolutional layer. The pooling layer was replaced with a stride layer. The convolutional layer has a kernel of size 1×1 and a channel depth of 21. An additional accumulation layer is added to accumulate convolutional features.

With the selected period, the significant frames were obtained using structural similarity. The significant frames within the time segment are the frames with significant differences compared to other frames. It was found by calculating the structural similarity of each frame to its previously found significant frame and when the structural similarity is higher than threshold, the frame selected as significant frame. From each of the significant frames, faces were extracted. The faces passed for spatiotemporal feature extraction.

The audio samples extracted from the time segments marked are provided in Table II. The prosodic features provide the details about the intonation, speech variations, and prosodic changes over the speech signal, The spectral features depict the frequency domain representation of the different emotions using MFCC and LPCC. The pitch features show the increase or decrease in the speech tone during the emotional depiction.

Table 2 Audio features

Audio Feature	**Features Description**
Prosody	Pitch, Formant, Intensity, Energy

features	
Spectral features	MFCC, LPCC
Pitch statistics	Mean/median, maximum/upper quartile, minimum/lower quartile, and Range/interquartile range

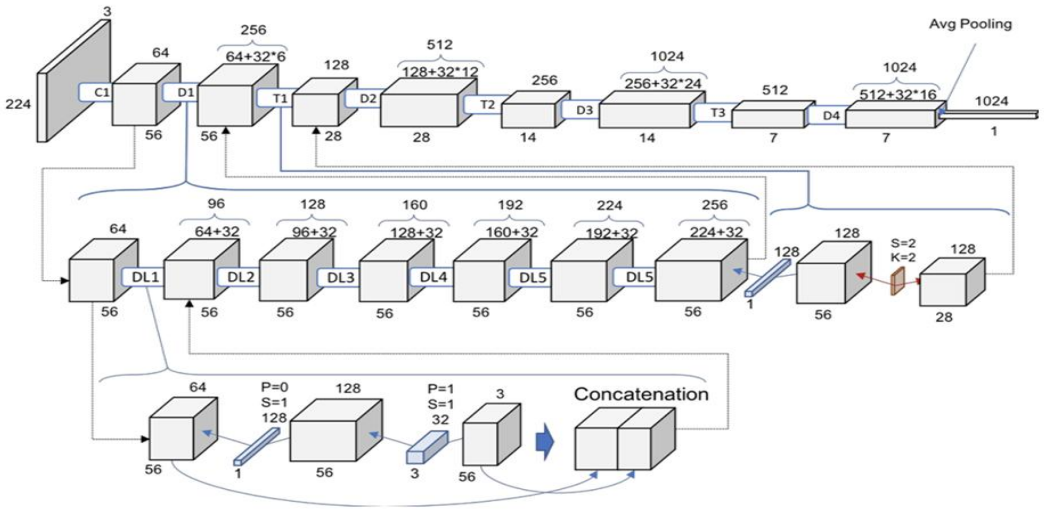

Figure. 2 Enhanced Densenet

The extracted audio features passed to 1D-CNN for a higher-level representation of the audio features.

Most of the text feature extraction models are based only on content and do not consider the context. Due to these even sarcastic comments are processed as genuine and incorrect emotions inferred. This paper suggests a feature that integrates both content and context features from text.

Texts within the time window are processed to extract the following content and context features.

Content features: These features are extracted from the distribution of words and special symbols like emoticons. Word distribution and semantic relation between words learned using Glove embedding (S. Huang et al., 2014). Glove is a powerful word-embedding algorithm. This method of unsupervised learning learned a word vector representation for a text corpus. It is done by reducing the dimension of the co-occurrence matrix. The vector for the word is constructed in such a way that similar words cluster together and different words repel each other. Compared to other word embedding models like word2vec, glove embedding better captures local and global statistics.

Emoticons are special symbols, which carry information about positive, negative, and sarcastic expressions. The frequency of positive, negative, and sarcastic symbols in the document is counted and the emoticon feature vector is constructed.

Context features: Sentiment information is the measure of sentiment expressed in texts. It has two components intrinsic and extrinsic. Intrinsic refers to the person's internal sentimental state. Extrinsic refers to an event topic, which acts as a risk indicator. Intrinsic sentiment information was extracted for text using the domain-specific sentiment lexicon approach proposed by (S. Huang et al., 2014). The output of intrinsic sentiment information is a word vector with polarity score (+1/-1) for each of domain specific word in the text. Extrinsic sentiment information is extracted using Latent Dirichlet Allocation (LDA) topic modeling approach proposed by (N. Förster et al., 2021). The output of topic modeling is a vector of scores with each element representing the topic score.

Sentence incongruity is the concept of polarity contrast between the positive candidate term and a negative phrase or negative candidate term with the positive phrase. The order of occurrence is not important. Camp (S. Cohan, 2005) detailed the incongruity patterns in English language sentences and the summary presented in Table 3.

Table 3 Sentence incongruity rules

Candidate term (Verb positive/negative)	Positive/Negative patterns
Verb	Verb followed by Verb
Verb present participle	Verb followed by Adverb
Verb Gerund	Adverb followed by Verb
Verb past participle	Verb followed by a proposition
Verb past form	Verb followed by an adjective
Verb present participle third person singular	Verb followed by a noun

The sentences are POS tagged and the count of several patterns as defined in the Table above is found and given as a sentence incongruity (si) feature.

The content and context features in the time window are joined to create the text feature vector. The text features passed to One Dimensional CNN (1D-CNN) to get a high-level representation of the feature. The configuration of the 1D-CNN used in this work is given in Table 4.

Table 4 1D-CNN for feature representation

Layer	Configuration
Convolutional 1D	10*128, ReLU, Stride=2
MaxPool 1D	Size=2, Stride=2
Convolutional 1D	10*128, ReLU, Stride=2
MaxPool 1D	Size=2, Stride=2
Convolutional 1D	8*128, ReLU, Stride=2
Convolutional 1D	8*128, ReLU, Stride=2
Flatten	-
Dense	1*512, ReLU

The raw EEG signal is preprocessed to remove noises. A notch filter was used to remove the electricity-induced noises and preprocessing given in Table 5 was done over the EEG signal.

Table 5 EEG Pre-processing

Process	Details
Filter data	Low pass at 30Hz and high pass at 1Hz.

Baseline correction	[—0.5 s, —0.1 s] latency around fixation events
Artifact removal	Independent Component Analysis on EEG Epochs.

The EEG signals are time segmented based on segments marked by procedure in Section 1.1. Each of the segments is applied continuous wavelet transform as

$$X_w(a,b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{\infty} x(t)\varphi(\frac{t-b}{a})dt$$

$\varphi(t)$ is the mother wavelet with a scale factor of a translation factor of b. The application of continuous wavelet transforms on an EEG signal results in a 2D scalogram, which provides detailed information about the state space of the system. A sample scalogram for the EEG time segment is given in Figure 4.

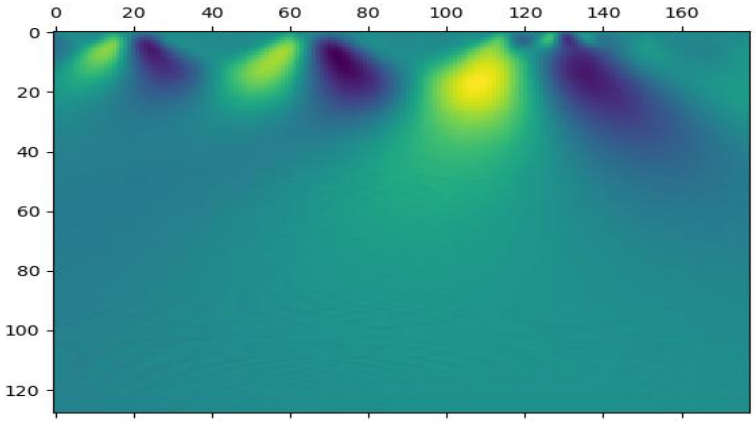

Figure 2 2D Scalogram image of EEG

The scalogram image was passed to Enhanced Densenet to get the high-level representation of EEG features.

3.3 Cross-modality feature enhancement

The feature represented by multiple modalities is enhanced using cross-modal attention. Multi-head attention proposed by (N. Vaswani et al., 2017) extended to multiple modalities of attention in this work. Say there are two modalities$\{m_1, m_2\}$.

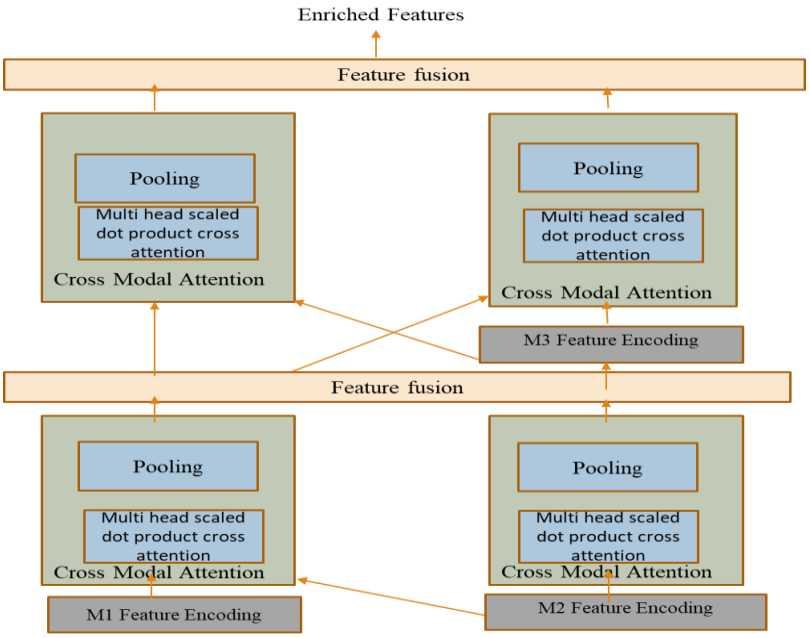

Figure 5 Incremental cross-modal attention(ICMFE)

The cross-modal attention for modal m_1 takes the output of its feature-encoding layer as a query vector and the output of m_2 feature encoding layers as key and value vectors. It then applies multi-head scaled dot product attention. It helps each modality to learn cross-reference information from other modalities. Finally, features from the cross-modality attention of m_1 and m_2 pool and passed to prediction using a softmax classifier. This cross-modal attention proposed by (N. Vaswani et al., 2017) was enhanced using a novel incremental cross-reference approach for multiple modalities as shown in Figure 5. In this the cross modalities feature learned from two modalities $\{m_1, m_2\}$ is passed to the next cross-modal attention taking the output of $\{m_{2+i}\}$ feature encoding layer for learning cross reference information between $\{m_1, m_2\}$ and $\{m_{2+i}\}$. This process is repeated until all N modalities are covered. The final feature was then passed for temporal emotion classification.

2.1. Emotion Classification

The obtained multimodal features passed to multivariate LSTM for temporal classification of emotion. There are different multivariate LSTM trains. One for seven basic emotions (neutral, happiness, sadness, anger, disgust, surprise, and fear), second for emotion activation (negative/positive), and third for emotion arousal (high/low).

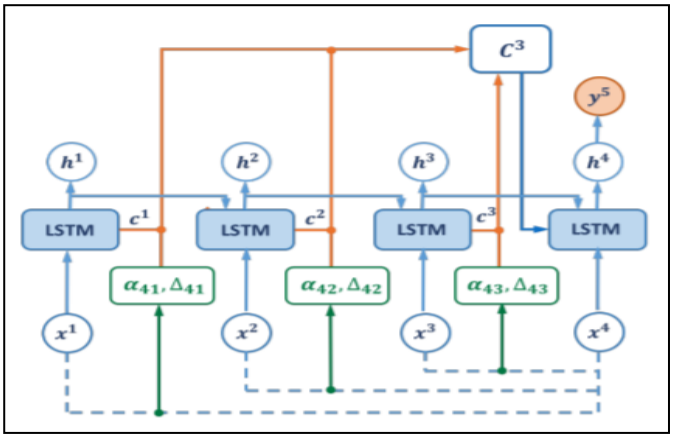

Figure 6. LSTM architecture

The architecture of the multivariate LSTM is used to predict the emotion classes from the series of feature vectors shown in Figure 6. As shown in Figure 6, each LSTM node takes the current input vector x and the previous hidden state as input. With this input, it calculates the cell activation as a weighted sum of inputs ($W_c x_t$) along with the bias (b_c). The cell activation got as a result, is then processed with a hyperbolic tangent activation function ($\emptyset_t$) as below

$$c_t = \emptyset_t(W_c x_t + U_c h_{t-1} + b_c)$$

In the above equation, h_{t-1} is the cell activation result of the previous LSTM node in the sequence. The values W_c and U_c are the weights for input and the hidden state vector. The level of activation to be retained or forgotten is done by controlling the gates.

The hidden state information is calculated at the final state. The gates control how much activation must be retained and how much must be forgotten. Input gate control how activation must be retained and forget gate decided how much cell activation must be forgotten. The final gate is incorporated to calculate the hidden state. The final gate takes two pieces of information, forgot vector (f_t) and input vector ((i_t) as input to provide the output vector (o_t).

$$f_t = \emptyset_s(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \emptyset_s(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \emptyset_s(W_o x_t + U_o h_{t-1} + b_o)$$

f_t The forgotten gate vector. i_t is The input gate vector. o_t The output gate vector.

It takes the $Z = (Z_1, Z_2, \ldots Z_T)$, where T emotion features observation is used to predict the emotion at time T+1 and each Z_i is the input embedding of the transformed original sequence $X = (X_1, X_2, \ldots X_T)$. The final LSTM layer output is passed to a Softmax classifier in the regression setting [22]. In the regression setting, softmax classifier the LSTM output to one of the possible values of emotions. The output of the softmax classifier is the emotion class prediction for the given feature values. The loss function for training the softmax regression classifier is given as

$$L = -[\sum_{i=1}^{m}\sum_{k=0}^{1} 1\{y^{(i)} = k\} \log P(y^{(i)} = k|z^{(i)}; \theta)]$$

Where

$$P(y^{(i)} = k|z^{(i)}; \theta) = \frac{\exp(\theta^{(k)} z^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(k)} z^{(i)})}$$

Where $\theta^{(1)}, \theta^{(2)}, \ldots \theta^{(k)}$ are the parameters of the model, and $\exp(\theta^{(k)} z^{(i)})$ is the normalization of the parameter with the input feature values.

The overall flow for cross-modality-based emotion classification is described as pseudo-code below.

Algorithm: Detect Emotion

Input: Face frames, audio, text streams

Output: Emotion

1. Sigframes← {first frame}
2. Base← first frame
3. for each frame f
   If SSIM (f, Base)> 0.6
      Sigframe← {Sigframe, f}
      Base←f
   End
4. Emotion←null
5. For each frame x in Sigframe
   . As← Segment audio till the time of x.
   Tx← Collect text till the time of x.
   Eg← creates an EEG scaleogram till the time of x.
   F1<-Extract Feature f1-f6 (from Table 1)
   F2<-Extract Table 2 features from as
   F3<=Extract Context and content features from Tx
   F4←Extract CNN features from EEG.
   F← cross-modal attention (F1, F2, F3, F4).
   Emotion←Invoke_LSTM (F)
   End
6. Return Emotion

3. RESULTS

The performance of the proposed deep learning multimodal features was tested against the K-EmoCon dataset (C. Y. Park et al., 2020). It is a multimodal dataset with audio-visual recording, metadata, EEG, and peripheral physiological signals. Different from other datasets, it has comprehensive annotations of emotions, emotion activation, and emotion arousal. The performance of the proposed solution compared against Deep Canonical Correlation Analysis proposed by (K. Zhang et al., 2021), multitask learning-based emotion recognition proposed by (Dai et al., 2021), and BERT-based emotion recognition proposed by (Lee et al., 2021). The performance was measured in terms of accuracy, precision, recall, F1-score, mean absolute error (MAE), and correlation coefficient.

Table 6 presents the comparison results for overall seven basic emotions.

Table 6. Comparison of results for seven emotions

Measures	Proposed	Zhang et al	Dai et al	Lee et al
Precision	0.87	0.79	0.77	0.71
Recall	0.9	0.92	0.8	0.66
F1-score	0.85	0.81	0.74	0.7
Accuracy	0.87	0.82	0.78	0.77
MAE	0.7233	0.862	0.889	0.901
Correlation Coefficient	0.83	0.79	0.76	0.75

The proposed solution can provide at least 5% higher accuracy compared to existing works. The accuracy has improved in the proposed solution due to feature enrichment with cross-modality attention. Cross-modality, reference in the feature has been added with attention and this ensured the consistency of results with multi-modality feedback. Though deep learning has been used for feature learning in existing works, the features are learned separately from each modality, and fusion is done only at the last stage by concatenating. However, hierarchical cross-modality learning allowed learning cross-reference information in each combination of modalities in the proposed solution.

The activation plays an imperative role in the depiction of the speech's emotion. It describes the intensity of the voice intonation, timbre, and prosody. The results for emotion activation are given in Table 7. The accuracy of the proposed solution is at least 4% higher compared to existing works. Cross-modality attention enriched features, and this improved the accuracy of the proposed solution.

Table 7. Comparison of results for emotion activation

Measures	Proposed	Zhang et al	Dai et al	Lee et al
Precision	0.87	0.8	0.78	0.74
Recall	0.9	0.92	0.81	0.67
F1-score	0.85	0.82	0.74	0.72
Accuracy	0.87	0.83	0.79	0.79
MAE	0.69	0.74	0.77	0.78
Correlation Coefficient	0.85	0.81	0.79	0.78

The outcomes of the system is evaluated for the different arousal value of the emotion such as high and low arousal. The overall results for the emotion recognition system for the different arousal is given in Table 8.

Table 8. Comparison of results for emotion arousal

Measures	Proposed	Zhang et al	Dai et al	Lee et al
Precision	0.89	0.83	0.81	0.81
Recall	0.92	0.93	0.78	0.79
F1-score	0.86	0.84	0.76	0.77

Accuracy	0.89	0.84	0.81	0.8
MAE	0.52	0.6	0.61	0.58
Correlation Coefficient	0.9	0.87	0.81	0.8

The accuracy in the proposed solution is at least 5% higher compared to existing works. Accuracy is higher in the proposed solution for emotion arousal classification compared to basic emotions. Proposed cross-modality features worked best for emotion arousal. The ICMFE-based(Proposed) scheme shows an increase of 4.65% for neutral, 1.04% for happy, 1.60% for sadness, 1.13% for anger, 4.44% for disgust, 4.94% for surprise, 6.71% for fear, and 3.5% for overall emotions in emotion recognition over simple feature fusion. The accuracy rate measured with and without cross-modality for different emotion classes and the result given in Fig. 7.

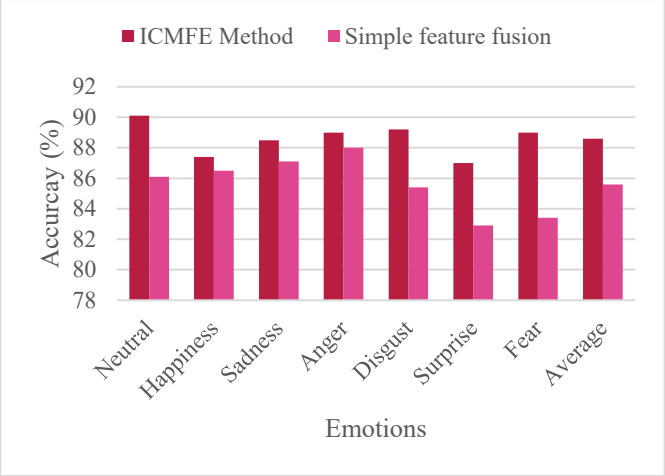

Figure 7 Accuracy for different emotions

Cross-modality, attention has increased the accuracy by at least 3% compared to simple feature fusion. Cross-modality attention was the salient part of the proposed solution, which has increased the accuracy compared to existing works.

The accuracy rate measured with and without cross-modality attention for different emotion activation and the result is given in Figure 8.

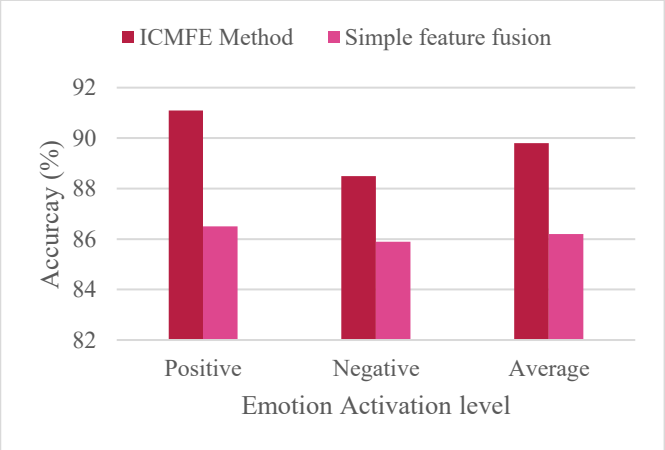

Figure 8. Emotion activation accuracy comparison

The ICMFE-based(Proposed) emotion recognition depicts an improvement of 5.31% and 3.02% for positive and negative

emotion activation over the simple feature fusion. For the case of emotion activation, cross-modality attention gives an overall improvement of 3.6% compared to simple feature fusion. The accuracy rate was measured with and without cross-modality attention for different emotion arousal and the result is given in Figure 9.

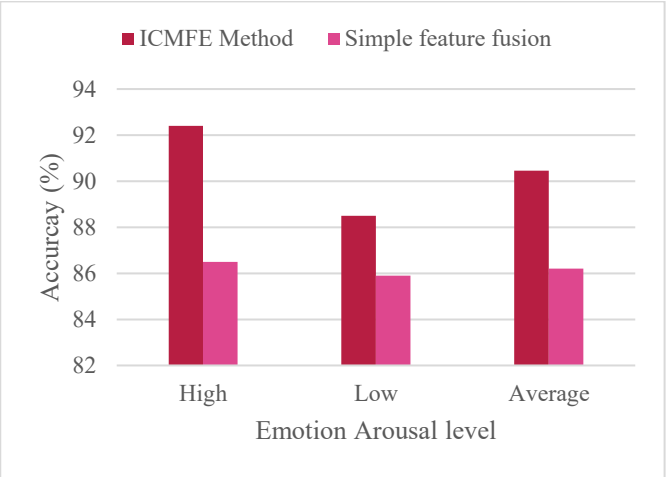

Figure 9 Emotion Arousal Accuracy Comparison

The ICMFE-based(Proposed) emotion recognition shows an improvement of 6.82% and 3.02% for high and low arousal respectively over the simple feature fusion. For comparison of emotion arousal, cross-modality attention has shown an overall increase of 4.25% compared to simple feature fusion.

The data points are clustered based on the features (cross-modality attention and simple feature fusion). The effectiveness of clusters was measured in terms of average cohesion, average separation, and silhouette coefficient. The results for clustering are given in Table 9.

The cross-modality attention features achieve an average cohesion of 636, an average separation of 469, and a silhouette coefficient of 0.82. However, the simple coefficient attains an average cohesion of 536, average separation of 365, and silhouette coefficient of 0.78.

Table 9 Clustering analysis results

Clustering metrics	**Cross-modality attention feature**	**Simple feature fusion**
Average cohesion	636	536
Average separation	469	365
silhouette coefficient	0.82	0.78

The average cohesion and average separation are higher in cross-modality attention compared to simple features. The higher values demonstrate the data points are better clustered leading to better discriminating ability.

4. CONCLUSION

A multimodal deep learning feature enhanced with cross-modal attention is proposed in this work. The modality features are enriched cross-reference from other modality features and due to this discrimination, the ability in emotion recognition is increased. The proposed multi-modal feature-learning framework is extensible and can be integrated with other modalities like eye movements, ECG, etc. The proposed multi-modal features were able to achieve more than 87% accuracy for seven basic emotions, emotion activation, and emotion arousal. Compared to existing works, the proposed solution's accuracy is at least 5% higher.

ACKNOWLEDGEMENT

Funding

Not Applicable

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Authors' contributions

She supervised every step of the work and provided critical review and valuable input. All authors read and approved the final manuscript.

REFERENCES

[1] Mehrabian and S. R. Ferris, "Inference of attitudes from nonverbal communication in two channels." *J. Consult. Psychol*., vol. 31, no. 3, pp. 248, 1967.

[2] B.Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu and D. Zhang, "Multimodal Emotion Recognition With Temporal and Semantic Consistency," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 29, pp. 3592-3603, 2021.

[3] Y. Park, N. Cha, S. Kang, et al., "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations." *Scientific Data*, vol. 7, 2020.

[4] Dai, Wenliang, Samuel Cahyawijaya, Yejin Bang, and Pascale Fung. "Weakly supervised Multi-task Learning for Multimodal Affect Recognition." *arXiv preprint arXiv*: *2104.11560*, 2021.

[5] Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatiotemporal features for multimodal emotion recognition." *in Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, IEEE*, pp. 1215–1223, 2017.

[6] Guo, Kairui, Henry Candra, Hairong Yu, Huiqi Li, Hung T. Nguyen, and Steven W. Su. "EEG-based emotion classification using innovative features and combined SVM and HMM classifier." *In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE,* pp. 489-492, 2017.

[7] K. Zhang, Y. Li, J. Wang, Z. Wang and X. Li, "Feature Fusion for Multimodal Emotion Recognition Based on Deep Canonical Correlation Analysis." *in IEEE Signal Processing Letters*, vol. 28, pp. 1898-1902, 2021.

[8] L. Cai, J. Dong, and M. Wei, "Multi-Modal Emotion Recognition From Speech and Facial Expression Based on Deep Learning," *2020 Chinese Automation Congress (CAC),* pp. 5726-5729, 2020.

[9] Lee, Sanghyun, David K. Han, and HanseokKo. "Multimodal Emotion Recognition Fusion Analysis Adapting BERT with Heterogeneous Feature Unification." *IEEE Access*, vol. 9, 2021.

[10] Li, Mi, HongpeiXu, Xingwang Liu, and Shengfu Lu. "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification." *Technology and health care*, vol. 26, no. S1, pp. 509-519, 2018.

[11] Nath, Debarshi, Mrigank Singh, DivyashikhaSethia, Diksha Kalra, and S. Indu. "An efficient approach to EEG-based emotion recognition using LSTM network." *In 2020 16th IEEE international colloquium on signal processing & its Applications (CSPA), IEEE,* pp. 88-92, 2020.

[12] N. Förster, and A. Mehler, "Twitter Author Topic Modeling-Comparative and Classifactory Topic Analysis Using Latent Dirichlet Allocation." 2021.

[13] Noroozi, Fatemeh, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and GholamrezaAnbarjafari. "Audio-visual emotion recognition in video clips." *IEEE Transactions on Affective Computing,* vol. 10, no. 1, pp. 60-75, 2017.

[14] N. Vaswani, N. Shazeer, J. Parmar, et al., "Attention is all you need." *in Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[15] Pandey, Pallavi, and K. R. Seeja. "Emotional state recognition with EEG signals using the subject independent approach." *In Data Science and Big Data Analytics, Springer, Singapore*, pp. 117-1, 2019.

[16] P. Ekman, V. Wallace, "Unmasking the Face." *Malor Book, Cambridge*, 2003.

[17] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks." *in IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301-1309, Dec. 2017.

[18] R. W. Picard, E. Vyzas, J. Healey, "Toward machine emotional intelligence: analysis of the affective

physiological state." *IEEE Trans. Pattern Anal. Mach. Intell*. vol. 23, no. 10, pp. 1175–1191, 2001.

[19] S. Cohan, "Incongruous Entertainment: Camp, Cultural Value, and the MGM Musical." 2005.

[20] S. Huang, Z. Niu, Z., and C. Shi, "Automatic construction of domain-specific sentiment lexicon based on constrained label propagation." *Knowledge-Based Systems*, Vol. 56, pp. 191–200, 2014.doi:10.1016/j.knosys.2013.11.009

[21] W. L. Zheng, W. Liu, Y. Lu, B. -L. Lu and A. Cichocki, "EmotionMeter: A Multimodal Framework for Recognizing Human Emotions." *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110-1122, March. 2019.

[22] Wu, Xun, Wei-Long Zheng, and Bao-Liang Lu. "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition." *arXiv preprint arXiv: 2004.01973*, 2020.

[23] W. V. Friesen, P. Ekman, "Emfacs-7: Emotional facial action coding system." *Unpublished manuscript, University of California at San Francisco*.

[24] W. Zheng, J. Zhu, and B. Lu, "Identifying Stable Patterns over Time for Emotion Recognition from EEG," *in IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417-429, 1 July-Sept. 2019, doi 10.1109/TAFFC.2017.2712143.

[25] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion," *in IEEE Access,* vol. 8, pp. 168865-168878, 2020.

[26] Zhang, Xiaowei, Jinyong Liu, JianShen, Shaojie Li, et al. "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine." *IEEE transactions on cybernetics,* 2020.

[27] Zhou, Hengshun, Jun Du, Yuanyuan Zhang, Qing Wang, Qing-Feng Liu, and Chin-Hui Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.