

Optimized multi-layer self-attention network for feature-level data fusion in emotion recognition

Basamma Umesh Patil^{1,2}, Ashoka Davanageri Virupakshappa², Ajay Prakash Basappa Vijaya³

¹Department of Computer Science and Engineering, JSS Academy of Technical Education Bengaluru, Visvesvaraya Technological University, Belagavi, India

²Department of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru, Visvesvaraya Technological University, Belagavi, India

³Department of Artificial Intelligence and Machine Learning, Dr. Ambedkar Institute of Technology, Bengaluru, Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Dec 1, 2023

Revised May 2, 2024

Accepted Jun 1, 2024

Keywords:

Emotion recognition

Multimodalities

Feature level fusion

Multi-layer progressive dense

residual fusion network

Self-attention mountain gazelle

convolution neural network

ABSTRACT

Understanding human emotions across diverse data sources presents challenges in various applications including healthcare, human-machine interaction, security, marketing, and gaming. Prior research has explored fusion techniques to address multimodal data heterogeneity, yet often overlooks the importance of discriminative unimodal information and potential complementarity among fusion strategies. Recognizing emotions from video and audio data poses challenges such as non-verbal cues interpretation, varying expression, ambiguity in context, and the need for nuanced feature extraction to capture subtle emotional nuances accurately. To tackle these issues, it is imperative to employ efficient emotion representation and multimodal fusion techniques, as these tasks have significant importance within the realm of multifaceted recognizing study. This study introduced a novel approach, optimized multi-layer self-attention network for emotion recognition (OMSN-ER), focusing on feature-level data fusion. OMSN-ER precisely assesses emotional states by merging facial and voice data, utilizing a multi-layer progressive dense residual fusion network and a self-attention mountain gazelle convolution neural network. Implemented in Python with the RAVDESS dataset, the methodology achieves exceptional accuracy (0.9908), surpassing benchmarks and demonstrating efficacy in multimodal emotion recognition. This research represents promising advancements in the intricate field of emotion recognition.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Basamma Umesh Patil

Department of Information Science and Engineering, JSS Academy of Technical Education

Visvesvaraya Technological University

Dr. Vishnuvardhana Road, Bengaluru-560060, Karnataka, India

Email: bupatil25@gmail.com

1. INTRODUCTION

Emotions, intangible yet profound mental states expressed through written, visual, and verbal cues, lack clear physiological indicators [1], [2]. Understanding these emotions profoundly influences communication across domains like lie detection and surveillance, underscoring the need to explore emotional variability for insights into psychological well-being. In today's society, emotions play an integral role in our daily lives, prompting the crucial need for machines to comprehend and respond appropriately. For better outcomes, it is also crucial for machines to detect human emotions [3], [4]. Advancing emotion classification techniques is key for societal progress, integrating into daily life [5], [6]. Emotion recognition in conversations

(ERC), encompassing textual, visual, and audio cues, stands pivotal in analyzing multimedia information, essential for assessing user-content interactions [7]–[9]. Moreover, as healthcare concerns, such as air pollution, escalate, leveraging machine learning inspired by human emotion recognition becomes vital [10]–[12]. Additionally, precise recognition of human activities via multisensory modalities further motivates advancements in emotion recognition systems [13], [14]. Psychologists harness the strengths of textual, visual, and audio models in a combined approach to achieve a more comprehensive understanding of human emotions [15]–[17]. Notably, studies highlight the inadequacy of relying solely on one modality for emotion recognition, emphasizing the superiority of approaches integrating multiple modalities.

For a considerable period, state-of-the-art systems in speech emotion recognition (SER) faced challenges of low accuracy and high computational demands [18], [19]. However, recent advancements have introduced real-time systems exhibiting notable performances in emotion recognition. Anvarjon *et al.* [20] introduced a lightweight convolutional neural network (CNN) model with plain rectangular kernels and customized pooling layers, achieving remarkable accuracy on the interactive emotional dyadic motion capture (IEMOCAP) (77.01%) and emotional speech database (EMO-DB) datasets (92.02%). Challenges persist, particularly in the integration of additional multimodal input sources like facial expressions, as seen in studies such as Franzoni *et al.* [21], where partial facial images focusing solely on mouth movements maintained a 5% accuracy loss compared to full-face images across four emotions: neutral, happy, surprised, and angry.

Another emerging source of information for real-time systems is textual data. With the advent of transformers [22] and bidirectional encoder representations from transformers (BERT) models [23], numerous publications have highlighted the advantages of using natural language due to smaller file sizes and the abundance of available datasets. This has led to its application in various tasks, including sentiment and emotion recognition [24], [25], showcasing its versatility across domains. Jiménez *et al.* [26] proposed an emotion recognition system utilizing transfer learning techniques. The framework involved a pre-trained spatial transformer network on saliency maps and facial images, followed by a bidirectional long short-term memory (BiLSTM) with an attention mechanism. Employing a late fusion strategy yielded an accuracy of 80.08% on the ryerson audio-visual database of emotional speech and song (RAVDESS) dataset. However, direct application to video-based tasks encountered domain adaptation issues.

Jiménez *et al.* [13] proposed an emotion recognition system employing aural transformers and action units for the RAVDESS dataset. This comprised a speech emotion recognizer (SER) and a facial emotion recognizer (FER). The SER utilized a pre-trained xlsr-Wav2Vec2.0 transformer with transfer-learning techniques such as embedding extraction and fine-tuning. The FER involved the extraction of action units from videos. A late fusion technique combining these modalities achieved an accuracy of 86.70%. Limitations included the incapacity of action units to capture crucial information from video frames. Radoi *et al.* [27] proposed an end-to-end emotion recognition framework based on the temporal aggregation of multimodal information. Temporal audio-visual information was extracted using the temporally aggregated audio-visual network (TA-AVN), achieving an overall accuracy of 84% for the CREMA-D dataset and 78.7% for the RAVDESS dataset. However, using limited annotated data, obtained through a random selection of temporal windows within individual video fragments, proved insufficient for emotion recognition.

Existing methodologies in emotion recognition often grapple with challenges related to the integration of diverse modalities, leading to inefficiencies in information processing and inaccuracies in emotion detection. Traditional approaches may lack precise control over information flow, resulting in the loss of valuable data and unexpected noise from individual modalities. Additionally, these methods tend to consume substantial processing time without achieving optimal accuracy. To address these limitations, this research proposes the adoption of feature-level data fusion techniques. By integrating features extracted from multiple modalities at a feature-level representation, this approach offers several advantages. Feature-level fusion facilitates a more effective combination of information from different sources, enabling enhanced classification accuracy and reduced noise. Moreover, it allows for a more comprehensive utilization of available data while minimizing processing time. The utilization of feature-level data fusion in this study aims to overcome the shortcomings of existing methods, striving for enhanced multimodal fusion and classification efficiency. By leveraging feature-level fusion techniques, the proposed approach endeavors to achieve superior accuracy, comprehensive information utilization, reduced noise, minimized processing time, and robust fusion. Ultimately, this innovative methodology seeks to revolutionize emotion recognition systems, offering nuanced insights and accurate assessments of emotional states across diverse domains, thereby reshaping human-computer interactions and emotional understanding.

Effective emotion recognition across multiple modalities remains a formidable challenge due to limitations in traditional feature extraction methods and the need for enhanced generalization and cross-modal information integration. Current approaches often encounter complexity limitations and inefficiencies in capturing temporal dependencies. This leads to suboptimal performance in recognizing nuanced emotional states.

In this study, a novel approach has been proposed to address key challenges in emotion recognition across various modalities. The proposed contributions encompass several innovative methodologies: i) handcrafted feature extraction is time-consuming. BiLSTM allows end-to-end learning for better performance, avoiding complexity limitations; ii) to boost generalization, the model employs BiLSTM for audio-visual tasks, capturing temporal dependencies effectively. It enhances efficiency by prioritizing keyframes; iii) the multi-layer progressive dense residual fusion network (MPDRF-net) enhances emotion classification by integrating multi-modal information using feature-level fusion, addressing unimodal shortcomings for improved cross-modal information exchange and recognition performance; and iv) in existing methods, a total 8 classes in the RAVDESS dataset are not classified, but the self-attention Mountain Gazelle CNN (SMGCNN) accurately categorizes happy, angry, disgusted, fearful, calm, sad, surprised, and neutral states.

2. METHOD

In the realm of emotion identification through multimodal data like video and audio, feature representation and fusion stand out as crucial yet challenging tasks. This research introduces an enhanced optimized multi-layer self-attention network for emotion recognition (OMSN-ER) aimed at accurately assessing the participant's emotional states across different modalities. This technique amalgamates information from audio and video sources, enabling a comprehensive approach to audio-video emotion identification. The workflow of this research is depicted in Figure 1. Initial data acquisition involves gathering visual and audio data from the RAVDESS dataset. Pre-processing techniques are applied to minimize undesirable effects, enhancing the quality of video data and refining speed signals. The pre-processed output feeds into the BiLSTM-based feature extraction process. Audio feature extraction involves capturing time and frequency domain features along with Mel-based features. For video, contextual information within word-level visual features is extracted. The extracted features undergo fusion using the MPDRF-Net. These fused features are then inputted into the SMGCNN to accurately classify the emotion states such as neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. The efficacy of this proposed method is assessed and compared against existing methods using various evaluation metrics. The audio and video data are preprocessed to remove noise and prepare them for further analysis.

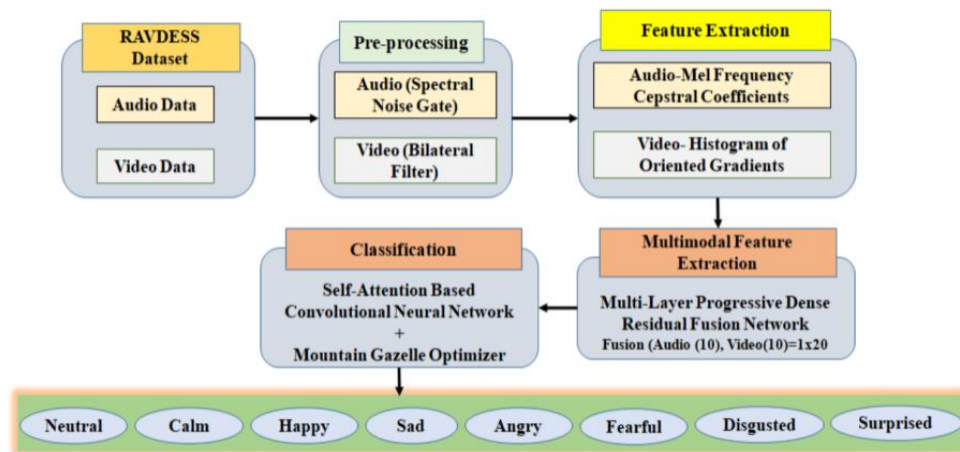


Figure 1. The workflow of OMSN-ER for emotion detection using the RAVDESS dataset

2.1. Dataset description

The RAVDESS database [28] is a validated multimodal dataset containing emotional speech and song recordings. It is designed specifically for recognizing emotions and is used extensively in the field of computing and audio analysis. It consists of 24 professional actors (12 female and 12 male) performing scripted speech and song in 8 different emotional states such as happy, fearful, calm, angry, sad, disgusted, neutral and surprised. There are 7,356 files in the RAVDESS. Three modality forms, including audio-only (16 bit, 48 kHz .wav), audio-video (720 p H.264, AAC 48 kHz, .mp4), and video-only (no sound), are present in the dataset. The audio files are typically stored in the WAV format, which is a standard uncompressed audio format. Actors expressed emotions at normal and strong intensity levels.

2.2. Pre-processing techniques

The initial step in preparing the audio and video data from the RAVDESS dataset involves comprehensive pre-processing [29]. This stage is crucial to mitigate issues like incompleteness, inconsistency, and noise that commonly afflict raw data. Through techniques such as bilateral filtering for video and spectral noise gate (SNG) for audio, unwanted artifacts are removed to enhance the quality of the visual frames and speech signals.

Noise handling: managing noise in audio and video data is pivotal to ensuring the robustness of subsequent models against overfitting. Addressing sensor noise, compression artifacts, and environmental conditions, a denoising threshold of 0.1 is applied to strike a balance between preserving data integrity and noise removal. Storage efficiency and tracking: videos inherently demand extensive storage due to their high-dimensional nature. Employing pre-processing techniques aids in reducing video file sizes without compromising quality. Furthermore, object detection and tracking mechanisms enable the identification and monitoring of specific objects or regions of interest within the frames. Audio separation and language detection: the process involves transcribing speech content, identifying spoken languages, and delving into the emotional or sentimental aspects of the audio. Simultaneously, segregating speech from background noise is essential to capture the intended content effectively. Features are extracted from the preprocessed audio and video data.

2.3. Multimodal feature extraction with bidirectional long short-term memory

Feature extraction is a crucial step in data recovery, aiming to extract pertinent audio and video features for improved classification outcomes. Patterns are typically extracted from raw audio data through hand-crafted features, essential as raw audio data is inherently in the time domain, with vital information unveiled in the frequency domain. This paper introduces a feature vector encompassing self-engineered features from both domains. The BiLSTM-based encoder is employed for extraction.

Audio feature extraction [30]: this phase involves extracting various features like Mel-frequency cepstral coefficients, energy, pitch, and fast fourier transform (FFT) parameters. Transforming audio signals from the time to frequency domain reveals substantial insights. Features such as amplitude envelope, zero, and mean crossing rate are computed to enrich the time domain characteristics.

Video feature extraction: the utilization of histogram of oriented gradient models facilitates the extraction of local shape and texture information from video content. BiLSTM aids in acquiring contextual information from the frames, contributing to a deeper understanding of the visual data. The process involves power spectra computation, mel-filter bank application, logarithmic transformation, and FFT. The extracted audio and video features are then fused to create a multimodal representation of the emotional state.

2.4. Feature-level multimodal feature fusion with multi-layer progressive dense residual fusion network

Following the extraction of features from both audio and video sources, the integration of these diverse features assumes paramount importance. Feature-level data fusion, facilitated by the MPDRF-Net, becomes pivotal in harmonizing the extracted features to minimize information loss [31]–[33]. Through the integration of features using down-sampling aggregated residual (DAR) modules and a residual transformer, the MPDRF-Net effectively combines feature-level representations to extract long-distance information. The resulting fused feature reconstruction enhances semantic information, contributing to improved emotion classification.

2.4.1. Progressive feature fusion

MPDRF-Net's architecture is structured to adeptly combine the rich audio and video features. Through progressive feature fusion, it strategically amalgamates the nuanced details captured by both modalities. The integration process involves a series of transformations that refine the fused features. The architecture of the MPDRF-Net model, depicted in Figure 2, encompasses bilateral channel pixel weighted (BCPW) modules, specifically designed to leverage the rich array of fine-grained features present in both audio and visual data. By incorporating DAR modules alongside residual transformers, the model adeptly captures long-distance information, facilitating the generation of fused features that encapsulate the essence of both modalities. These fused features are subsequently inputted into a classification model, such as a self-attention mountain gazelle CNN, to predict the emotional state of the individual in the dataset. Through the MPDRF-Net, the extracted features from audio and video sources are fused, utilizing progressive feature fusion techniques to mitigate information loss effectively. The audio and video features, each represented by a (10,10) matrix, are fed into input layers 1 and 2, respectively, where the PMC-Net in the feature fusion layer progressively integrates multimodal features while learning the intricacies of each layer. The fusion process integrates collected features using a combination of DAR modules and residual transformers, facilitating the extraction of crucial long-distance information. During decoding, the fused feature information is reconstructed using the extracted semantic information, with aggregated residual operations upsampling features across

stages 1 to 3. Employing dense skip connections for input features, nodes in stages 1 to 4 enhance the movement of feature information at each layer, ultimately amalgamating features from all modalities to generate fused features.

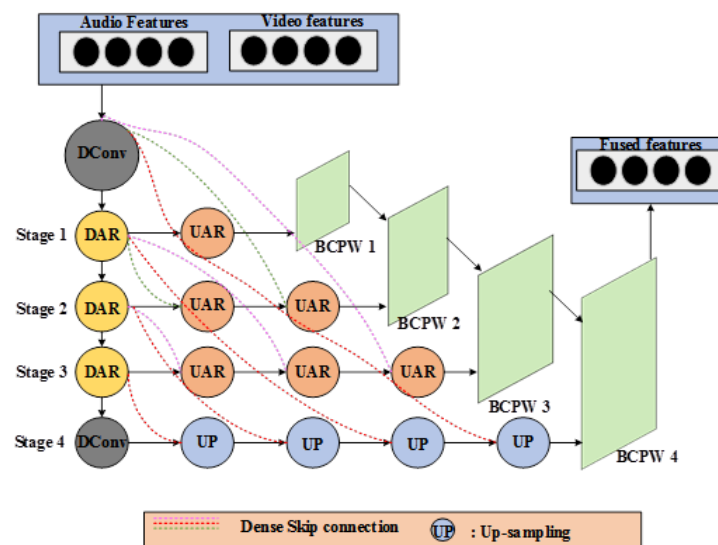


Figure 2. Architecture of the proposed fusion model

2.5. Self-attention mountain gazelle convolutional neural network-based classifier for 8 emotional states

The subsequent phase involves deploying the SMGCNN, an attention mechanism-integrated deep CNN, specifically tailored for emotion recognition from facial expressions and audio cues. Deep learning performance scales up by initially downsampling and extracting expression features, applying them to a self-attention residual module. The final expression is classified through a fully connected layer.

2.5.1. Residual network and self-attention module

The adoption of a residual network addresses the challenge of the gradient vanishing problem encountered when deepening the network layers, making it difficult to effectively train a model. By employing a residual network, the convolutional layer learns the residual between input and output, facilitating the identification of the mapping between the input and output by establishing connections between them. The output of the residual network is defined mathematically, representing the output of the residual connection as a function of the input vector and the mapping between input and output. Additionally, the self-attention module is introduced to overcome the limitations imposed by the small size of convolutional kernel sizes in CNNs, which restricts each convolutional operation to covering a small area around the pixel. This module leverages non-local operations to calculate the correlation weight between each position of the feature map and the image signal representation, enabling the capture of long-distance features. Through various transformations and normalization functions, the self-attention layer transforms input feature maps into two embedding spaces, reducing both the computational complexity and the number of channels. The output of the attention layer is obtained through matrix multiplication with learning parameters, contributing to the propagation of gradients through the residual connection. During model training, binary cross-entropy loss is employed to measure the dissimilarity between predicted probabilities and true binary labels, providing a mechanism for optimizing model performance.

2.5.2. Learning parameter optimization using mountain gazelle optimizer

Inspired by the hierarchical structure and social dynamics observed in wild mountain gazelles, the mountain gazelle optimizer (MGO) optimizes learning rate parameters within the self-attention mechanism-integrated deep convolutional neural network (SMGCNN). Leveraging territorial behaviors, herd dynamics, and migration patterns of gazelles, this optimizer orchestrates an effective learning rate for the classifier. The algorithm of the MGO [34] is enumerated below. The Algorithm 1, described is a heuristic optimization approach inspired by the social behaviors of gazelles in nature. It begins by initializing a population of gazelles with 50 and total iterations to 100. Each gazelle's fitness is then evaluated based on mathematical representations of behaviors such as territorial solitary males (TSM), maternity herds (MH),

bachelor male herds (BMH), and migration to search for food (MSF). These behaviors are formulated as equations that assess aspects like territory establishment, reproduction, and resource searching. The fitness function is defined as the minimization of the loss function of the SMGCNN, considering predicted probabilities and true binary labels. The population is then arranged based on fitness, and the position of the best-performing gazelle is updated accordingly. This process iterates until a termination condition is met, typically involving a maximum number of iterations or satisfactory fitness levels. Ultimately, the algorithm aims to optimize learning rate parameters within the SMGCNN model, enhancing its predictive capabilities by leveraging principles inspired by gazelle social dynamics.

Algorithm 1. Mountain gazelle optimizer

```

1   Input  $\leftarrow$  TotalIteration=100, PopulationSize=50
2   Outputs  $\leftarrow$  OptimalPosition, BestFitness
3   Initialize Population randomly as  $X_i$  ( $i=1, 2, \dots, \text{PopulationSize}$ )
3   Calculate fitness levels for each individual in the population
4   While condition not met do:
5       For each Gazelle ( $X_i$ ) in Population do:
6           Calculate TSM, MH, BMH, and MSF behaviors
7           Calculate fitness values for each behavior:
8            $\text{TSMFitness} = |\text{male} - (\text{ri}_1 * \text{BH} - \text{ri}_2 * X(t)) * F| * \text{Cof}_r$ 
9            $\text{MHFitness} = |(\text{BH} + \text{Cof}_{(1,r)}) + (\text{ri}_3 * \text{male} - \text{ri}_4 * X_{\text{rand}}) * \text{Cof}_{(1,r)}|$ 
10           $\text{BMHFitness} = |(X(t) - D) + (\text{ri}_5 * \text{male} - \text{ri}_6 * \text{BH}) * \text{Cof}_r|$ 
11           $\text{MSFFitness} = |(\text{ub} - \text{lb}) * \text{ri}_7 + \text{lb}|$ 
12          Calculate fitness value as the minimization of the loss function of SMGCNN
13           $\text{Fitness} = \text{Min}\{\sum(y * \log(\bar{y}) + (1 - y) * \log(1 - \bar{y}))\}$ 
14          Arrange the gazelle population based on fitness values
14          Update the position of the best gazelle
15   End while
16   Return the best gazelle's position and its fitness value
17   End Procedure

```

3. EXPERIMENTAL RESULTS

For experimentation, Windows 10 Pro, specifically on build 19045.2965, was employed as the operating system. The programming language Python version 3.10.9 and libraries such as NumPy, Librosa, Matplotlib, OpenCV, and TensorFlow are utilized to implement the tasks within the experiment. The datasets used for analysis are sourced from RAVDESS. The proposed OMSN-DER model is built and utilized for emotion recognition activity. The system is equipped with 16.0 GB of RAM, with 15.8 GB being usable for processing tasks throughout the experiments. To make a complete description of multimodal emotion recognition, this section discusses the RAVDESS dataset-based different scores to compute the similarities and differences between proposed and traditional methods. This section also takes a quantitative analysis in terms of some evaluation metrics implemented in PYTHON. The evaluation metrics such as accuracy, recall, F1-score, specificity, precision, and receiver operating characteristics (ROC) curve are used to compare the proposed model with the related methods.

3.1. Confusion matrix for multimodal features

The performance assessment, as depicted by the confusion matrix, highlights the model's accuracy in predicting emotional states. It also reveals potential areas of misclassification and confusion between certain emotions. According to Figure 3, the model demonstrated excellent performance for most samples, with a few exceptions. At 'neutral', some two classes are predicted as 'sad' and 'fearful'. Moreover, the two 'happy' classes are predicted as 'neutral' and 'surprised' and the 40 classes are accurately predicting the 'happy' class. The two classes in 'surprised' are confused and predicted as 'calm' and 'sad'. This confusion between 'sad' and 'fearful' can be attributed to the low arousal levels present in both emotions.

3.2. Analysis of runtime accuracy and loss

The method underwent training and evaluation over 200 epochs, dividing the RAVDESS dataset into 80% training and 20% validation subsets. The OMSN-DER approach is analyzed in Figure 4, where Figure 4(a) shows the high training accuracy attained, and Figure 4(b) demonstrates a decreased training loss for multimodal data. The results specify that the proposed model exhibits higher accuracy in emotion

identification for multimodal data (audio and video). Furthermore, it demonstrated superior performance in both the training and testing sets. Better accuracy and low loss value are achieved after crossing 100 iterations.

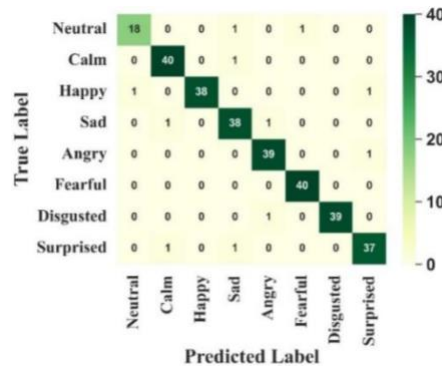


Figure 3. Confusion matrix of RAVDESS dataset

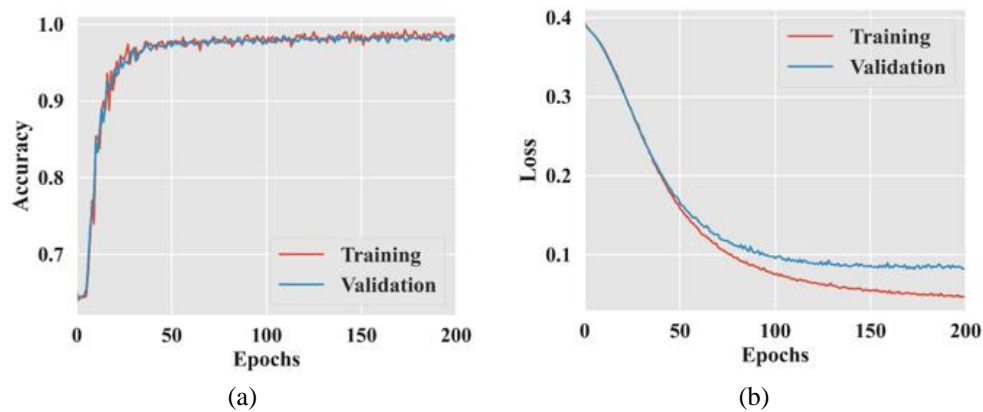


Figure 4. Analysis for different epochs of (a) accuracy and (b) loss

3.3. Receiver operating characteristics curve

Figure 5 shows the ROC curve for the minimum number of samples needed to train an SMGCNN, which is highly dependent on the specific classification task. The ROC curve uses the calculated false positive rate (FPR) and true positive rate (TPR) values. It is a plot of TPR against FPR, with different points on the curve corresponding to different classification thresholds. The ROC curve starts at (0.0, 0.0) and ends at (1.0, 1.0), indicating that it successfully classifies all positive samples.

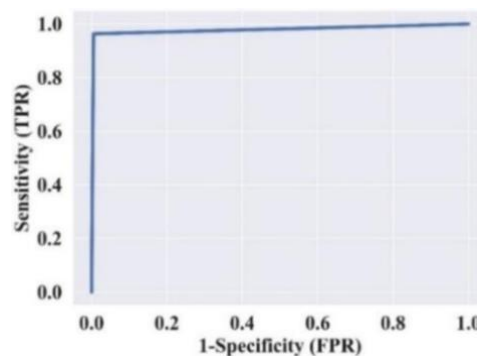


Figure 5. ROC curve of SMGCN

3.4. Comparison of OMSN-DER with existing emotion recognition methods

Table 1 shows the comparison of OMSN-DER with existing emotional recognition methods where “-” indicates that the corresponding evaluation metrics are not available in that research. All evaluation shows that the proposed method achieves high accuracy and other scores which concludes that the OMSN-DER attains better performance in the classification of all classes. The fusion approach provides the best and most relevant features of all classes where all the individual features are concatenated and given to classification. This reduces the computational complexity and time.

Table 1. Comparison of OMSN-DER with existing methods

Techniques	Precision	F1-score	Accuracy	Recall	Specificity
SER and FER [13]	0.9399	-	0.9805	0.7237	-
Optimal multimodal emotion recognition [14]	0.86	0.86	0.86	0.86	0.98
Bidirectional parallel ESN [15]	0.8656	0.8519	0.7539	0.8854	-
MFA [16]	0.97	0.97	0.9507	0.98	-
MCCA [17]	0.9098	0.9014	0.9525	0.9061	-
Self-supervised learning algorithm [28]	0.8950	0.8168	0.876	0.8550	-
3DCNN-LSTM [29]	-	0.9609	0.90	-	99.45
OMSN-DER (Proposed)	0.9633	0.9633	0.9908	0.9633	0.9947

3.5. Comparison of OMSN-DER with existing emotion recognition methods

Table 2 illustrates the comparison of the accuracy of OMSN-DER with existing methods based on 8 emotional classes (happy, fearful, calm, angry, sad, disgusted, neutral, and surprised) where “-” indicates that the corresponding class was not classified by the existing methods. From all the observations, it is obvious that the OMSN-DER attains high accuracy in all 8 classes. It is concluded that the OMSN-DER attains better performance in the classification of all classes. Classification results are based on factors such as size, and quality of the dataset, the architecture and parameters of the model, the preprocessing techniques applied, and the fusion task at hand.

Table 2. Comparison of OMSN-DER accuracy with existing methods based on 8 emotional classes

Techniques	Happy	Fearful	Calm	Angry	Sad	Disgusted	Neutral	Surprised
Bidirectional parallel ESN [11]	0.73	0.82	0.89	0.84	0.58	0.89	0.57	0.72
MFA [12]	0.97	0.98	-	0.96	0.99	0.99	0.96	-
Self-supervised learning algorithm [13]	0.79	0.806	0.816	0.830	0.782	0.851	0.773	0.883
Multi-head self-attention mechanism [14]	0.87	0.85	-	0.97	0.98	0.97	-	0.95
Deep ESN model [16]	0.75	0.97	0.91	0.95	0.87	0.89	0.94	-
OMSN-DER (Proposed)	0.993	0.996	0.99	0.99	0.983	0.996	0.99	0.986

3.6. Discussions

The OMSN-DER model marks a significant advancement in the field of multimodal emotion recognition, showcasing remarkable accuracy and efficiency in classifying a wide range of emotional states from audio-visual data. It is innovative use of optimized multi-layer self-attention mechanisms and dynamic emotion recognition techniques sets a new benchmark and outperforms existing models in precision, recall, and specificity. This model’s ability to accurately interpret complex emotional expressions holds substantial promise for enhancing human-computer interaction and emotional analysis applications. The research, though advanced in multimodal emotion recognition, has limitations. It primarily focuses on audio and video, neglecting the richness of textual and physiological data. Future research should aim to integrate text modality to enrich emotion recognition capabilities.

4. CONCLUSION

This research emphasizes emotion representation and feature fusion using OMSN-DER for accurate emotional assessment from audio and video modalities. Employing MPDRF-Net and SMGCNN, the method achieves superior emotion classification (fearful, angry, neutral, surprised, happy, calm, disgusted, and sad) on the RAVDESS dataset. Evaluation against existing methods reveals outstanding performance with an accuracy of 0.9908, precision of 0.9633, recall of 0.9633, F1-score of 0.9633, and specificity of 0.9947. These results demonstrate the effectiveness of OMSN-DER in multimodal emotion recognition, contributing a robust approach to address complexity in emotion assessment. The research primarily focuses on audio and video modalities, potentially overlooking the benefits of incorporating other modalities. This includes text, which is

crucial for comprehensive emotion recognition systems. Additionally, the study's reliance on specific datasets, like RAVDESS, raises concerns about generalizability, warranting exploration across diverse datasets for a more robust evaluation. In the future, more evaluation parameters will be focused on finding the performance of the research work and will improve the multimodal evaluation with text modality too. The current model is validated by employing various advanced fusion models. Additionally, optimization of crucial hyperparameters is performed, potentially influencing the classification approach significantly.




REFERENCES

- [1] M. Li *et al.*, "Multimodal emotion recognition and state analysis of classroom video and audio based on deep neural network," *Journal of Interconnection Networks*, vol. 22, Jun. 2022, doi: 10.1142/S0219265921460117.
- [2] K. Zhang, Y. Li, J. Wang, Z. Wang, and X. Li, "Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis," *IEEE Signal Processing Letters*, vol. 28, pp. 1898–1902, 2021, doi: 10.1109/LSP.2021.3112314.
- [3] L.-N. Do, H.-J. Yang, H.-D. Nguyen, S.-H. Kim, G.-S. Lee, and I.-S. Na, "Deep neural network-based fusion model for emotion recognition using visual data," *The Journal of Supercomputing*, vol. 77, no. 10, pp. 10773–10790, Oct. 2021, doi: 10.1007/s11227-021-03690-y.
- [4] C. A. Devi and D. K. Renuka, "Multimodal emotion recognition framework using a decision-level fusion and feature-level fusion approach," *IETE Journal of Research*, vol. 69, no. 12, pp. 8909–8920, Dec. 2023, doi: 10.1080/03772063.2023.2173668.
- [5] D. Dresvyanskiy, E. Ryumina, H. Kaya, M. Markitantov, A. Karpov, and W. Minker, "End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild," *Multimodal Technologies and Interaction*, vol. 6, no. 2, Jan. 2022, doi: 10.3390/mti6020011.
- [6] 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, Jan. 2023, doi: 10.1016/j.ins.2022.11.076.
- [7] S. Lee, D. K. Han, and H. Ko, "Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification," *IEEE Access*, vol. 9, pp. 94557–94572, 2021, doi: 10.1109/ACCESS.2021.3092735.
- [8] J. Zheng, S. Zhang, Z. Wang, X. Wang, and Z. Zeng, "Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 2213–2225, 2023, doi: 10.1109/TMM.2022.3144885.
- [9] M. Li *et al.*, "Multimodal emotion recognition model based on a deep neural network with multiobjective optimization," *Wireless Communications and Mobile Computing*, pp. 1–10, Aug. 2021, doi: 10.1155/2021/6971100.
- [10] D. Wen, S. Zheng, J. Chen, Z. Zheng, C. Ding, and L. Zhang, "Hyperparameter-optimization-inspired long short-term memory network for air quality grade prediction," *Information*, vol. 14, no. 4, Apr. 2023, doi: 10.3390/info14040243.
- [11] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887–2905, Jan. 2021, doi: 10.1007/s11042-020-08836-3.
- [12] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowledge-Based Systems*, vol. 229, Oct. 2021, doi: 10.1016/j.knosys.2021.107316.
- [13] C. L. -Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. F. -Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset," *Applied Sciences*, vol. 12, no. 1, Dec. 2021, doi: 10.3390/app12010327.
- [14] A. I. Mridha, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities," *Knowledge-Based Systems*, vol. 244, May 2022, doi: 10.1016/j.knosys.2022.108580.
- [15] H. Ibrahim, C. K. Loo, and F. Alnajjar, "Bidirectional parallel echo state network for speech emotion recognition," *Neural Computing and Applications*, vol. 34, no. 20, pp. 17581–17599, Oct. 2022, doi: 10.1007/s00521-022-07410-2.
- [16] P. V. V. S. Srinivas and P. Mishra, "Human emotion recognition by integrating facial and speech features: an implementation of multimodal framework using CNN," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130172.
- [17] F. M. Alamgir and M. S. Alam, "Hybrid multi-modal emotion recognition framework based on InceptionV3DenseNet," *Multimedia Tools and Applications*, vol. 82, no. 26, pp. 40375–40402, Nov. 2023, doi: 10.1007/s11042-023-15066-w.
- [18] M. McTear, Z. Callejas, and D. Griol, "The conversational interface: talking to smart devices," *The Conversational Interface: Talking to Smart Devices*, pp. 1–422, 2016, doi: 10.1007/978-3-319-32967-3.
- [19] B. W. Schuller and A. M. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. Hoboken, United States: John Wiley and Sons, 2013, doi: 10.1002/9781118706664.
- [20] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-Net: a lightweight CNN-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, Sep. 2020, doi: 10.3390/s20185212.
- [21] V. Franzoni, G. Biondi, D. Perri, and O. Gervasi, "Enhancing mouth-based emotion recognition using transfer learning," *Sensors*, vol. 20, no. 18, Sep. 2020, doi: 10.3390/s20185222.
- [22] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, USA: Association for Computational Linguistics, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [24] V. Santucci, S. Spina, A. Milani, G. Biondi, and G. D. Bari, "Detecting hate speech for Italian language in social media," in *EVALITA Evaluation of NLP and Speech Tools for Italian*, Torino: Accademia University Press, 2018, pp. 239–243, doi: 10.4000/books.aaccademia.4799.
- [25] V. Franzoni, A. Milani, and G. Biondi, "SEMO: a semantic model for emotion recognition in web objects," in *Proceedings of the International Conference on Web Intelligence*, New York, USA: ACM, Aug. 2017, pp. 953–958, doi: 10.1145/3106426.3109417.
- [26] C. L. -Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. F. -Martínez, "Multimodal emotion recognition on RAVDESS dataset using transfer learning," *Sensors*, vol. 21, no. 22, Nov. 2021, doi: 10.3390/s21227665.
- [27] A. Radoi, A. Birhala, N.-C. Ristea, and L.-C. Dutu, "An end-to-end emotion recognition framework based on temporal aggregation of multimodal information," *IEEE Access*, vol. 9, pp. 135559–135570, 2021, doi: 10.1109/ACCESS.2021.3116530.
- [28] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, May 2018, doi: 10.1371/journal.pone.0196391.




- [29] A. Chaudhari, C. Bhatt, A. Krishna, and C. M. T. -González, "Facial emotion recognition with inter-modality-attention-transformer-based self-supervised learning," *Electronics*, vol. 12, no. 2, Jan. 2023, doi: 10.3390/electronics12020288.
- [30] O. Atila and A. Şengür, "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition," *Applied Acoustics*, vol. 182, Nov. 2021, doi: 10.1016/j.apacoust.2021.108260.
- [31] M. Khan, A. El Saddik, F. S. Alotaibi, and N. T. Pham, "AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network," *Knowledge-Based Systems*, vol. 270, Jun. 2023, doi: 10.1016/j.knsys.2023.110525.
- [32] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal emotion detection via attention-based fusion of extracted facial and speech features," *Sensors*, vol. 23, no. 12, Jun. 2023, doi: 10.3390/s23125475.
- [33] S. Singkul and K. Woraratpanya, "Vector learning representation for generalized speech emotion recognition," *Heliyon*, vol. 8, no. 3, Mar. 2022, doi: 10.1016/j.heliyon.2022.e09196.
- [34] B. Abdollahzadeh, F. S. Gharehchopogh, N. Khodadadi, and S. Mirjalili, "Mountain gazelle optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems," *Advances in Engineering Software*, vol. 174, Dec. 2022, doi: 10.1016/j.advengsoft.2022.103282.

BIOGRAPHIES OF AUTHORS






Basamma Umesh Patil    working as an Assistant Professor in the Department of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru. She received her M.Tech. from VTU in the year 2014 and her B.E. from VTU in the year 2009. She is pursuing her Ph.D. in Computer Science and Engineering in the Department of Computer Science and Engineering, JSSATE, Bengaluru. She has 12 years of academic teaching experience and years of research experience. Her research field of interests includes data science, artificial intelligence, and machine learning. She has published around 16 research papers in international journals and conferences like IEEE, Springer, and Elsevier. She can be contacted at email: bupatil25@gmail.com or patilbu@jssateb.ac.in.



Ashoka Davanageri Virupakshappa    working as a Professor (ISE) at JSS Academy of Technical Education Bengaluru. He served as Professor and Head, Dean at various reputed engineering colleges of Karnataka. He received his Ph.D. in Computer Science and Engineering from Dr. MGR University in 2009-10. He is a member of IEEE, MCSI, MISTE, Fellow Institute of Engineers. Six students have been awarded Ph.D. degrees. in Computer Science and Engineering under his guidance. He was one of the National Award winners in "Rashtriya Ekta Samman-2013". He is an editorial board member of reputed international journals. His biography is included in Who's Who in the World 2011-12. He has over 100 publications in different international and national journals and conferences. His areas of expertise include software engineering, software architecture, requirement engineering, data science, knowledge engineering, and operating system virtualization. He can be contacted at email: dr.dvashoka@gmail.com.



Ajay Prakash Basappa Vijaya    currently working as a Research Professor in the Department of Artificial Intelligence and Machine Learning at Ambedkar Institute of Technology, Bengaluru. He worked as an Associate Professor in the Department of Computer Science and Engineering, SJBIT Bengaluru. He received his Ph.D. in Computer Science and Engineering from VTU in 2018. He received his M.Tech. from VTU in the year 2008 and B.E. from VTU in the year 2005. He has 15 years of academic teaching experience and 8 years of research experience. He is a Wipro Certified Faculty (WCF) for project based learning framework in Java-J2EE and imparting the learning to students. His research field of interest includes knowledge engineering, software architecture, data science, and machine learning. He has published around 22 research papers in international journals and conference like IEEE, Springer, Elsevier, and Inderscience. Recently received Rs. 5,00,000/-Lahks grants from VGST, the government of Karnataka for the data science project titled "intelligent and interactive plant disease supervision using deep learning". He can be contacted at email: ajayprakas@gmail.com.