IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1566~1577
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1566-1577

Multimodal facial expression recognition using residual mogrifier long short-term memory

Mamatha Kariyappa Rajanna[1], Thejaswini Shankar[1], Rashmi Narasimhamurthy[1], Nandhini Vannivedu Lakshmanan[2], Hariprasad S. Ananthapadmanabharao[3]
[1]Department of Electronics and Communication Engineering, BMS Institute of Technology and Management, Bengaluru, India
[2]Department of Electronics and Communication Engineering, Government Sri Krishnarajendra Silver Jubilee Technological Institute, Bengaluru, India
[3]Department of Electronics and Communication Engineering, Jain University, Bengaluru, India

Article Info

Article history:

Received Jul 12, 2025
Revised Jan 20, 2026
Accepted Feb 6, 2026

Keywords:

Feature interaction
Gradient flow
Human expressions
Multimodal facial expression recognition
Residual mogrifier long short-term memory

ABSTRACT

Multimodal facial expression recognition aims to improve emotion analysis by integrating visual, audio, and textual cues to achieve accuracy and robustness. However, effectively recognizing facial expressions across video, text, and audio presents challenges due to inconsistencies in how emotions are expressed among these modalities. To overcome this issue, this research proposes a residual mogrifier long short-term memory (RMLSTM) model to enhance robustness in multimodal facial expression recognition. By integrating residual connections into the long short-term memory (LSTM), the model improves its ability to capture complex dependencies among various modalities, including video, text, and audio. The residual connection overcomes the vanishing gradient problem and ensures stable training with better gradient flow in deeper networks. The mogrifier mechanism refines the input features dynamically, enhancing feature interaction and alignment across modalities. The RMLSTM achieves 99.57% and 97.83% accuracy on the SAVEE and YouTube datasets, respectively, outperforming both the mel-frequency cepstral coefficients time-domain feature with iterative dilated convolutional neural network (MFCCT-1DCNN) and attention-based multi-modal popularity prediction model of short-form videos (AMPS).

Corresponding Author:

Mamatha Kariyappa Rajanna
Department of Electronics and Communication Engineering, BMS Institute of Technology and Management
Yelahanka, Bengaluru, India
Email: mamathakr@bmsit.in

1. INTRODUCTION

Emotion recognition aims to identify the emotion of every utterance within a dialogue, where human emotions are expressed through different modalities, such as visual, acoustic, and textual [1]. The field of automatic emotion recognition has made significant progress using individual modalities; however, considerable challenges remain before this problem can be fully solved [2]. Human emotional expression is inherently complex, as people combine multiple emotional factors [3]. The same emotions can be represented with diverse signals, such as vocal expressions and nonverbal cues, including spoken words, facial expressions, and body movements [4]. The combination of various indicators from all potential emotion sources to build a multimodal recognition system holds great promise for improving the accuracy of single-source recognition [5]. Emotion recognition systems have practical applications across behavioral analysis, forensic work, and healthcare settings [6]. Research in psychology indicates that a driver's emotional state directly affects road safety, with negative emotions such as anger and sadness increasing the

likelihood of car accidents. As a result, transportation safety systems require real-time emotion monitoring due to the critical driving hazards posed by emotional instability [7], [8]. Detecting emotions is challenging because individual perceptions and the duration of emotional states impact how emotions are expressed [9]. The initial facial expression recognition techniques relied on handcrafted features, which resulted in limited accuracy and vulnerable system design [10]. Deep learning (DL) networks, particularly convolutional neural networks (CNN) and recurrent neural networks (RNN), have led to substantial advancements in feature extraction and classification, enhancing recognition results [11], [12]. Novel network architectures with attention mechanisms, as well as data augmentation methods, have created sophisticated approaches to enhance facial expression analysis and generalization. Research has primarily focused on detecting emotions in controlled settings involving individuals [13]. Multi-facial emotion recognition presents a significant challenge due to the multiple face detection models that impact real-time operation [14].

The development of end-to-end solutions is still in progress, as researchers have performed limited work to assess collective emotions in educational settings involving student groups. Research indicates that real-world multi-subject emotion recognition requires further development beyond current limitations [15]. Multimodal facial expression recognition utilizes video, audio, and text data to enhance accuracy. However, the main challenges include inconsistencies in emotion presentation across different modalities, model instability, and ineffective feature fusion. To address these challenges, this research proposes a residual mogrifier long short-term memory (RMLSTM) that improves cross-modal alignment through dynamic feature transformation, while residual connections ensure stable training. Furthermore, experiments on the SAVEE and YouTube datasets demonstrate that RMLSTM outperforms existing methods. This research examines facial expression recognition, utilizing various DL techniques along with their advantages and limitations. This analysis helps identify research gaps to enable the development of improved recognition techniques, leading to better precision. Singh *et al.* [16] suggested a 3D convolutional neural network (3DCNN)-convolutional long short-term memory (ConvLSTM) framework for emotional video facial recognition. By combining CNN and long short-term memory (LSTM), the model effectively captured spatial and temporal dependencies, resulting in dynamic emotion pattern recognition. However, the 3DCNN-ConvLSTM combination faced overfitting problems when dealing with high-dimensional spatiotemporal data. Singh *et al.* [17] developed an attention-based 2DCNN with LSTM to perform speech emotion recognition tasks. This architecture integrated four blocks of 2DCNN feature extractors with 2DCNN-LSTM dependency learners and incorporated an attention mechanism to filter LSTM-generated significant data, followed by a dropout operation to enhance emotion recognition. The proposed method exhibited sensitivity to various speech patterns and noise exposure, as the model failed to properly identify essential features.

Middya *et al.* [18] introduced a DL-based multimodal emotion recognition system using audio-visual modalities and model-level fusion. Model-level fusion was performed to establish the best multimodal model for emotion recognition by combining audio and video modality data. The model developed an optimal multimodal emotion recognition system through the combination of audio and video features at the model level. However, when model-level fusion combined separate feature extractors, the approach became prone to overfitting when working with small datasets with limited emotional expression diversity. Alluhaidan *et al.* [19] introduced an mel-frequency cepstral coefficients time-domain feature with iterative dilated convolutional neural network (MFCCT-1DCNN) to recognize speech expressions. The CNN framework contained one-dimensional layers along with activation layers, max-pooling, dropout features, and fully connected (FC) components for speech expression classification. However, there was spectral information loss in MFCCT-1DCNN because it performed a conversion from the frequency domain to the time domain. Cho *et al*. [20] presented an AMPS for facial expression recognition. It incorporated bidirectional long short-term memory (BiLSTM) with self-attention and co-attention methods, which allowed it to achieve a better understanding of intra- and inter-modal relationships. The model struggled to capture fine details in temporal connections between audio and visual components due to the mixed utilization of recurrent and convolutional layers. The existing techniques suffered from multiple limitations, including poor capabilities to detect fine temporal connections between audio and visual modalities, noise sensitivity and diverse speech patterns because the model failed to discern proper features while handling multiple modalities, thereby resulting in complicated integration challenges. To overcome these issues, the RMLSTM is proposed in this research for facial expression recognition.

The main contributions are as follows:

i) The RMLSTM is proposed in this research for multimodal facial expression recognition. By integrating residual connections into LSTM, the model effectively captures complex dependencies among various modalities such as video, text, and audio.
ii) The residual connection provides stable training and better gradient flow in deeper networks while addressing the vanishing gradient issue. The mogrifier mechanism transforms input features dynamically, thereby enhancing feature interaction and alignment across modalities.

iii) EfficientNetB0 is used as an extractive feature mechanism for image/video samples through its effective visual pattern recognition capabilities.
iv) Text data is processed through term frequency-inverse document frequency (TFIDF), which transforms verbal information into numerical vectors by assigning significance to important words without reducing the standard terms.
v) MFCC are used to extract audio features through spectral and frequency-based characteristics, providing better audio signal representations.

This research paper is further organized as follows. Section 2 details the proposed methodology. Section 3 presents the results and discussion. The conclusion of this study is provided in section 4.

2. PROPOSED METHODOLOGY

In this research, the RMLSTM is proposed for multimodal facial expression recognition utilizing data from the SAVEE and YouTube sources, which include video/images, text, and audio. To provide high-quality inputs, preprocessing techniques are applied, such as image normalization and punctuation removal. Image normalization is applied to standardize pixel intensities, punctuation removal is used for text cleaning, and MFCC features are directly extracted from audio data to capture key frequency characteristics. The feature extraction step includes EfficientNetB0 for visual data, TFIDF for textual information, and MFCC for speech signals. The extracted features from all modalities are fused to create a unified multimodal representation. Then, this fused data is fed into the RMLSTM model, which uses residual connections to overcome the vanishing gradient issue and provide model stability, while the mogrifier approach transforms input features dynamically, thereby enhancing feature alignment and interaction among various modalities. Lastly, the model classifies facial expressions based on multimodal inputs. Figure 1 shows the flow diagram of this research.

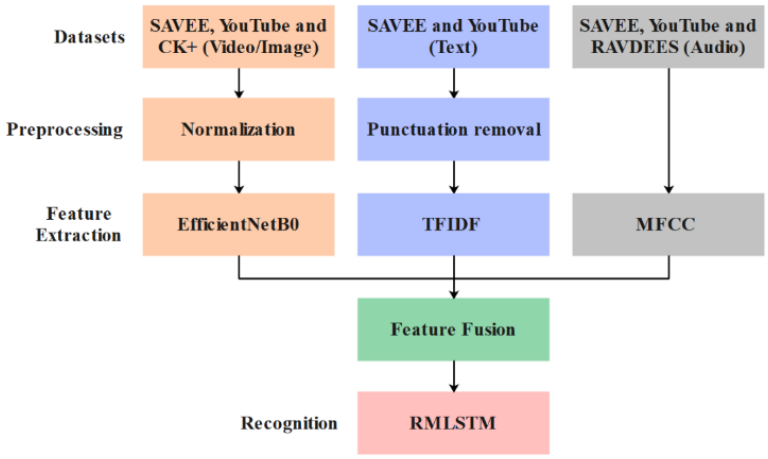

Figure 1. Flow diagram of the multimodal facial expression recognition

2.1. Dataset

The SAVEE [21], YouTube [7], CK+ [16], and RAVDEES [18] datasets are used to collect data for this research. These datasets are applied to create an automated emotion identification system that supports video, image, text, and audio. The SAVEE dataset includes 480 words, 10 general sentences, 3 standard sentences, and 2 emotion-specific sentences. The YouTube dataset includes seven distinct emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise, all of which are recorded and evaluated by 10 individuals. The CK+ dataset contains 593 sequences from 123 subjects, of which 327 sequences are labeled with emotions. Seven emotions are labeled, including surprise, fear, disgust, contempt, sadness, happiness, and anger. The RAVDEES dataset contains 1,440 audio files recorded by 12 female and 12 male actors, covering eight distinct emotions: anger, calmness, disgust, fear, neutrality, happiness, sadness, and surprise. All recorded audio files have a 48 kHz sample rate and 16-bit resolution.

2.2. Preprocessing

After data collection, normalization, and punctuation removal are applied to preprocess the data. Normalization is used for image/video data, while punctuation removal is applied to text data. Normalization is a preprocessing method applied to reduce the differences in face images, such as variations in lighting, to achieve better image quality. It enhances image intensity, resulting in improved clarity and higher recognition performance. The mathematical expression for this normalization is presented in (1). Where x_n denotes the normalized image, x_{min} and x_{max} denote minimum and maximum image intensities.

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Punctuation removal helps to standardize the textual data by eliminating unnecessary symbols that do not contribute to semantic meaning, thus enhancing model efficiency. By removing punctuation, noise is reduced, leading to better tokenization and feature extraction. The formula for removing punctuation is presented in (2). Where P denotes the set of punctuation marks and $T_{punc_{removed}}$ denotes the tokenized text after punctuation removal. After performing the preprocessing, feature extraction is applied to image/video, text, and audio data to extract meaningful and relevant information from each of these modalities.

$$T_{punc_{removed}} = \{t_i \in T_{stop_{removed}} : t_i \notin P\} \tag{2}$$

2.3. Feature extraction

EfficientNetB0 is used as an extractive feature mechanism for image/video samples through its efficient computation of visual patterns. Text data is processed through TFIDF, which transforms verbal information into numerical vectors by emphasizing significant words without diminishing standard terms. MFCC analyzes audio signals through spectral and frequency-based characteristics to create effective audio signal representations. A detailed description of these methods is provided in the subsequent sections.

2.3.1. EfficientNetB0 for image/video data

EfficientNet is a scaled-up model that optimizes both accuracy and efficiency. EfficientNetB0 is a key layer in the mobile inverted bottleneck MBConv, with compound scaling applied to three components: depth, width, and resolution. The constants α, β, and γ denote depth, width and resolution, respectively, and are derived for better results [22]. EfficientNetB0 searches for these scaling coefficients with less system capacity in smaller models.

2.3.2. Term frequency-inverse document frequency for text data

TFIDF is applied to extract features from raw text data, where weights are assigned to each term in a document based on term frequency (TF) and inverse document frequency (IDF). Higher-weight terms have greater significance compared to lower-weight terms [23]. The weight for each term is calculated using (3).

$$W_{i,j} = TF_{t,d}\left(\frac{N}{D_t}\right) \tag{3}$$

Where $TF_{t,d}$ is the number of occurrences of term t in document d, N is the total number of documents and D_t denotes the documents with term t. TFIDF is a type of scoring measurement approach widely used in summarization and data retrieval. TF estimates the frequency of token and gives more significance to common tokens in a given document. However, IDF estimates the rarity of tokens in the corpus. In this manner, if uncommon words appear in more than one document, they are considered significant. In a group of documents D, the IDF weights a token x using (4). Where, $n(x)$ is the frequency of x in D and $\frac{N}{n(x)}$ is the inverse frequency. The TF-IDF is estimated by combining TF and IDF, as represented in (5).

$$IDF(X) = \frac{N}{n(x)} \tag{4}$$

$$TF - IDF = TF \times IDF \tag{5}$$

TFIDF is applied to estimate the significant term weights, and the final output of TFIDF is in the form of a weight matrix. The scores increase gradually with the TFIDF count, but are balanced by the frequency of the word. It transforms variable-length text into feature vector of fixed-length thereby providing simpler integration with recognition models.

2.3.3. MFCC for audio data

The acoustic features of an audio signal represent the physical properties of speech in terms of amplitude, frequency, and loudness. The acoustic feature set includes distinct spectral features, voice quality features, and time-domain features to describe facial expressions. MFCC provides spectral information of speech and models human auditory perception. In MFCC, the cepstrum is obtained by applying a discrete cosine transform to the logarithm of the short-time power spectrum of the signal. The coefficient of the mel-scale spectrum is an improved technique adapted from the cepstrum. The mel-scale consists of a uniform space of triangular filter banks [24].

Hence, the bandwidth of individual filters enhances the logarithmic scale and normalizes the frequency. Consider the magnitude response of a sequence of the triangular filter, given as mel frequency. The mel-scale form is expressed in (6). Where, $mel(f)$ denotes the mel frequency scale and f denotes the actual frequency. The mel frequency converts the actual frequency of power spectrum magnitudes of the input speech signal through a mel-scale filter bank, as given in (7).

$$mel(f) = 2595log_{10}\left(1 + \frac{f}{700}\right) \quad (6)$$

$$M_{(c,n)} = mel\left(P_{(m,n)}\right) \quad (7)$$

Where $M_{(c,n)}$ represents the mel frequency coefficients and c is a sequence of coefficients obtained through the mel-scale filter bank. MFCC represents the tone aspects of speech that are significant for fetection facial emotions by pitch, stress patterns, and vocal intonation. Furthermore, it reduces high-dimensional audio into low-dimensional feature set while preserving discriminative features thereby enhancing classification performance.

2.4. Feature fusion

Feature fusion is performed after extracting features from EfficientNetB0 for video/image data, TFIDF for text, and MFCC coefficients for audio. Fusion is carried out by concatenating the extracted feature vectors into a single vector for multimodal representation. This fusion establishes essential connections between different modalities, allowing them to leverage each other's data to maximize recognition performance.

The integrated technique helps address variations in expression representations across different modalities. The interconnected feature vector provides a better understanding of facial expressions, leading to improved recognition results. Recognition performance depends on the concatenated feature vector, as it serves as input to the recognition process, which benefits from the additional information provided by combining multimodal data.

2.5. Recognition

The fused features are provided as input to the multimodal facial expression recognition process. While LSTM networks are widely used for sequential data processing, they struggle to capture complex dependencies among multiple modalities in facial expression recognition. Traditional LSTM [25] processes input sequences independently, limiting its ability to capture intricate relationships among visual, audio and text data. LSTM has several limitations, such as ineffective handling of multimodal dependencies, difficulty in capturing fine-grained interactions among various input modalities, and slower convergence due to unidirectional processing. These issues reduce its efficiency in multimodal facial expression recognition, where both spatial and temporal dependencies must be modeled effectively.

Standard LSTM struggles with handling complex interactions, particularly in multimodal scenarios. The mogrifier long short-term memory (MLSTM) addresses this by introducing the mogrifier mechanism, which dynamically transforms input features multiple times before feeding them into the LSTM cell. This enhances feature interaction and data alignment across modalities, improving model flexibility and feature representation in facial expression recognition.

The MLSTM improves multimodal information fusion by dynamically adapting feature importance, enhancing expressive power and generalization. Its alternative update mechanism allows deeper cross-modal feature refinement, leading to more accurate emotion recognition. By improving multimodal feature integration, minimizing information loss and enabling quicker convergence, the MLSTM becomes highly suitable for facial expression recognition. It extends the standard LSTM by adding two gating units on top of the LSTM, enhancing the interaction space between network inputs and outputs. Figure 2 shows the MLSTM structure.

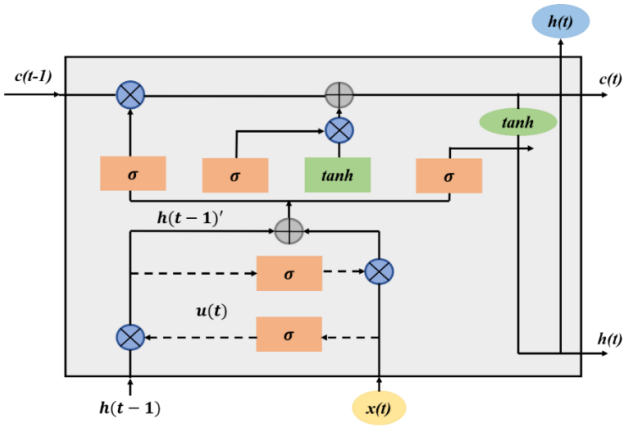

Figure 2. Cell structure of MLSTM

In MLSTM, the input $x(t)$ and previous timestamp output $h(t-1)$ are alternatively screened before entering the LSTM. The input $x(t)$ is passed through a sigmoid threshold unit to attain the control state $u(t)$. The sigmoid function ensures that the values of $u(t)$ lie between [0, 1]. The $u(t)$ and $h(t-1)$ transform every element in $h(t-1)$ to various degrees. If $u(t)$ has a value of 1 for a given element, and the corresponding element of $h(t-1)$ is passed into the network with its actual value. If $u(t)$ matches an element value of 0.5 in $h(t-1)$, the respective element in $h(t-1)$ flows into the neural network after splitting every element value into $u(t)$. The value of every element in $x(t)$ is a derivative from the $x(t)$ weight at each layer, updated continuously during training to reduce network loss. The update process for $u(t)$ and $h(t-1)'$ are mathematically formulated in (8) and (9).

$$u(t) = \sigma(W_u \cdot x(t) + b_u) \tag{8}$$

$$h(t-1)' = 2h(t-1) \cdot u(t-1) + h(t-1) \tag{9}$$

Where W_u is the weight of the input $x(t)$ that controls $h(t-1)$, and b_u is a bias. The $h(t-1)$ is converted into $u(t)$ by $h(t-1)'$. The final output of the RNN is obtained from $c(t)$. In the training phase, the neural network is updated using the time series $c(t)$. If a particular element within the time series of input shows a dramatic change, then the network needs to be better tuned. The MLSTM addresses this by introducing $h(t-1)'$ as a threshold unit of $x(t)$ control. The $h(t-1)'$ is processed through a sigmoid threshold structure to generate the control state $v(t)$. By using $v(t)$ to transform each element $x(t)$, the network uses the obtained loss value to fine-tune the weights of the layer for gradient updates. The update process for $v(t)$ and $x(t-1)'$ are given in (10) and (11).

$$v(t) = \sigma(W_v \cdot h(t-1)' + b_v) \tag{10}$$

$$x(t)' = 2x(t) \cdot v(t) + x(t) \tag{11}$$

Where W_v is the input weight $h(t-1)'$ on $x(t)$, providing adequate control, and b_v is the bias matrix, $x(t)$ is converted into $v(t)$ using $x(t)'$. Utilizing residual connections within MLSTM for multimodal facial expression recognition enhances the system's ability to process diverse input modalities including audio, text and image/video. Residual connections ensure that the original input features remain intact as they pass through untouched transformations, preserving essential data from all modalities across various processing layers. This significantly reduces the chance of gradient vanishing due to improved gradient flow within the network, allowing deeper architectures to perform more effectively. The MLSTM leverages these residual connections to maintain both transformed and raw input data, enabling the system to detect subtle emotional cues across different modalities. Residual connections in MLSTM improve precision in detecting complex facial expressions by promoting rapid model learning, while maintaining stable feature extraction and combination. They help preserve features from excessive distortion, facilitating effective modeling of micro-expressions and vocal tone variations. Consequently, the system becomes more adept at recognizing emotions in context, adapting to changes in speaker identity and environmental conditions, while successfully processing visual information loss.

The RMLSTM introduces key innovations that enhance multimodal facial expression recognition by improving feature integration and temporal modeling. The mogrifier mechanism dynamically transforms and refines input interactions among modalities, ensuring effective contextual understanding and data fusion. The RMLSTM addresses vanishing gradient issues, enabling deeper network training while retaining long-term dependencies. Additionally, its iterative feature refinement optimizes hidden state representations, leading to accurate, discriminative embeddings for emotion recognition. Through these advancements, the RMLSTM significantly improves recognition accuracy in multimodal facial expression analysis.

3. RESULTS AND DISCUSSION

The RMLSTM is simulated in MATLAB (R2020b) with system requirements of Windows 10 OS, 16 GB RAM, and an i7 processor. The recognition performance is evaluated using metrics of accuracy, positive predictive value (PPV), specificity, sensitivity, and Matthews correlation coefficient (MCC). The mathematical equations for these metrics are given in (12) to (16). Where, TP, FP, FN, and FP are true positives, false positives, false negatives, and false positives, respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

$$PPV = \frac{TP}{TP+FP} \tag{13}$$

$$Specificity = \frac{TN}{TN+FP} \tag{14}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{15}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{16}$$

3.1. Performance of multimodal features for SAVEE and YouTube datasets

In this subsection, the performance of multimodal features for the SAVEE and YouTube datasets is analyzed using various metrics. Table 1 shows the performance results for the SAVEE dataset, while Table 2 presents the performance results for the YouTube dataset. The hybrid features leverage complete modality complementarity between audio, video/image, and text to achieve optimal performance. The hybrid model integrates audio signals, visual signals and linguistic information, resulting in stronger feature representations and improved decision-making. By combining the participating modalities, the performance metrics show superior accuracy and precision, alongside enhanced specificity and sensitivity, as well as a higher MCC score compared to individual or dual modalities. The hybrid features achieve 99.57%, 99.31%, 99.16%, 98.65%, and 98.83% for accuracy, PPV, specificity, sensitivity, and MCC, respectively, on the SAVEE dataset. Similarly, hybrid features achieve 97.83%, 97.56%, 97.24%, 96.24%, and 96.75% for accuracy, PPV, specificity, sensitivity, and MCC, respectively, on the YouTube dataset.

Table 1. Performance of different feature extraction on SAVEE dataset

Multimodal features	Accuracy (%)	PPV (%)	Specificity (%)	Sensitivity (%)	MCC (%)
Audio	97.52	97.28	97.14	96.67	96.81
Video/image	95.43	95.19	95.07	94.58	94.72
Text	93.61	93.37	93.22	92.79	92.93
Audio+video	99.21	98.89	98.76	98.34	98.49
Video+text	97.33	97.11	97.05	96.59	96.74
Text+audio	95.48	95.26	95.13	94.62	94.79
Hybrid features (audio, video, text)	99.57	99.31	99.16	98.65	98.83

Table 2. Performance of different feature extraction on YouTube dataset

Multimodal features	Accuracy (%)	PPV (%)	Specificity (%)	Sensitivity (%)	MCC (%)
Audio	95.72	95.48	95.31	94.12	94.36
Video	93.61	93.42	93.25	92.07	92.29
Text	91.55	91.33	91.17	90.05	90.28
Audio+video	97.21	96.89	96.66	95.41	95.63
Video+text	95.33	95.12	94.98	93.81	94.05
Text+audio	93.47	93.26	93.11	91.89	92.12
Hybrid features (audio, video, text)	97.83	97.56	97.24	96.49	96.75

3.2. Performance of recognition for SAVEE and YouTube datasets

In this subsection, the performance of recognition for the SAVEE and YouTube datasets is analyzed using various metrics. Table 3 presents the performance for the SAVEE dataset, while Table 4 shows the performance for the YouTube dataset. The RMLSTM outperforms vision transformer (ViT), swin transformers, bidirectional encoder representations from transformer (BERT), multi-layer perceptron (MLP), RNN, LSTM, and MLSTM by incorporating advanced mechanisms for detecting temporal patterns and transforming dynamic feature associations. The memory and gating mechanisms in RMLSTM enable it to handle sequential data more effectively compared to MLP and RNN. Additional features in RMLSTM improve upon standard LSTM and MLSTM by utilizing attention-like mechanisms and gradient update methods to address non-linear data patterns and long-term dependencies. The RMLSTM achieves 99.57%, 99.31%, 99.16%, 98.65%, and 98.83% for accuracy, PPV, specificity, sensitivity, and MCC, respectively, on the SAVEE dataset. Similarly, RMLSTM achieves 97.83%, 97.56%, 97.24%, 96.24%, and 96.75% for accuracy, PPV, specificity, sensitivity, and MCC, respectively, on the YouTube dataset. RMLSTM outperforms transformer-based structures due to its superior ability to capture long-term dependencies. In contrast to transformers, RMLSTM performs better on moderate datasets by effectively maintaining sequential dependencies. Furthermore, as a time-series and sequential model, RMLSTM demonstrates a clear advantage over transformer models in this context.

Table 3. Performance of different classifier on SAVEE dataset

Classifier	Accuracy (%)	PPV (%)	Specificity (%)	Sensitivity (%)	MCC (%)
ViT	87.64	87.46	87.28	86.52	86.41
Swin transformers	89.27	89.14	89.04	88.76	88.59
BERT	90.93	90.57	90.36	89.64	89.44
MLP	91.52	91.38	91.19	90.05	90.24
RNN	93.67	93.49	93.31	92.12	92.34
LSTM	95.73	95.58	95.42	94.19	94.45
MLSTM	97.49	97.29	97.08	96.07	96.25
RMLSTM	99.57	99.31	99.16	98.65	98.83

Table 4. Performance of different classifier on YouTube dataset

Classifier	Accuracy (%)	PPV (%)	Specificity (%)	Sensitivity (%)	MCC (%)
ViT	85.73	85.35	84.59	84.31	84.09
Swin transformers	87.27	87.08	86.73	86.47	86.23
BERT	89.62	89.27	88.85	88.52	88.18
MLP	90.83	90.55	90.27	89.46	89.71
RNN	91.15	91.71	91.46	90.62	90.06
LSTM	93.41	93.14	93.63	92.04	92.33
MLSTM	95.06	95.38	95.09	94.60	94.17
RMLSTM	97.83	97.56	97.24	96.49	96.75

The integration of residual connections in RMLSTM improves gradient flow and mitigates vanishing gradient issues in deeper networks. By incorporating these residual connections, RMLSTM ensures better propagation across layers, leading to enhanced learning stability and efficiency during training. Specifically, RMLSTM excels due to its ability to capture long-term dependencies in sequential data while handling less computational time compared to transformer models.

Table 5 shows the complexity analysis of recognition models for the SAVEE and YouTube datasets. The ViT, Swin Transformer, BERT, MLP, RNN, LSTM, and MLSTM are considered as existing methods, and the proposed RMLSTM is compared with these models. The complexity is analyzed in terms of training time and testing time for both the SAVEE and YouTube datasets. The RMLSTM achieves a training time of 174 minutes and a testing time of 0.03 seconds for the SAVEE dataset. Similarly, the RMLSTM achieves a training time of 176 minutes and a testing time of 0.04 seconds for the YouTube dataset. This is due to its superior residual connections and gating mechanisms compared to ViT, swin transformer, BERT, MLP, RNN, LSTM, and MLSTM. In contrast to transformer models such as ViT, swin transformer, and BERT, the RMLSTM maintains a sequential processing structure with linear time complexity. Compared to LSTM and MLSTM, the residual connection in RMLSTM provides better convergence through enhanced gradient flow, thereby reducing the vanishing gradient issue. Furthermore, it requires fewer parameters than MLP and RNN, thus reducing computation time. The proposed RMLSTM achieves p-values of 0.008 and 0.0012 for the SAVEE and YouTube datasets, respectively.

Table 5. Complexity analysis of recognition models for SAVEE and YouTube datasets

Classifier	SAVEE		YouTube	
	Training time (min)	Testing time (s)	Training time (min)	Testing time (s)
ViT	198	0.2	195	0.21
Swin transformers	194	0.16	192	0.19
BERT	189	0.13	189	0.17
MLP	186	0.11	187	0.15
RNN	184	0.9	184	0.12
LSTM	180	0.07	181	0.09
MLSTM	177	0.05	178	0.07
RMLSTM	174	0.03	176	0.04

3.3. Comparative analysis

This section presents a comparative analysis of the proposed RMLSTM with existing methods is provided for the SAVEE, YouTube, CK+, and RAVDEES datasets. The 3D-CNN-ConvLSTM [16], LSTM+attention+CNN-2D [17], DL-based feature extractor networks [18], MFCCT-CNN [19], and AMPS [20] are considered as existing methods to demonstrate the performance of the proposed RMLSTM. The RMLSTM achieves accuracy rates of 99.57%, 97.83%, 97.28%, and 94.63% for the SAVEE, YouTube, CK+, and RAVDEES datasets, respectively. Table 6 shows the comparative analysis.

Table 6. Comparative analysis of proposed model on four different datasets

Dataset	Method	Accuracy (%)
SAVEE	3D-CNN-ConvLSTM [16]	98.83
	LSTM+attention+CNN-2D [17]	57.50
	DL-based feature extractor networks [18]	99
	MFCCT-CNN [19]	93
	RMLSTM	99.57
YouTube	AMPS [20]	79.7
	RMLSTM	97.83
CK+	3D-CNN-ConvLSTM [16]	95.10
	RMLSTM	97.28
RAVDEES	LSTM+attention+CNN-2D [17]	74.44
	DL-based feature extractor networks [18]	86
	MFCCT-CNN [19]	92
	RMLSTM	94.63

3.4. Discussion

The advantages of RMLSTM and the limitations of existing methods are discussed in this section. The existing techniques suffer from multiple limitations, including poor capabilities in detecting fine temporal connections between audio and visual modalities, noise sensitivity and diverse speech patterns. These models fail to discern the proper features while handling multiple modalities, leading to complex integration challenges. To overcome these issues, the RMLSTM is proposed for multimodal facial expression recognition because it improves both feature representation and sequential learning methods. Each residual element helps reduce gradient vanishing during backpropagation, enabling effective training in DL models. In MLSTM, the input representations undergo continuous transformations that improve the relationship between temporal and visual features. This mechanism improves the model's efficiency in detecting subtle facial movements and temporal relationships. The use of RMLSTM strengthens the system's performance against facial expression noise, resulting in higher recognition accuracy. The results are analyzed in terms of quantitative and comparative analysis for both the SAVEE and YouTube datasets. The RMLSTM achieves 99.57%, 99.31%, 99.16%, 98.65%, and 98.83% for accuracy, PPV, specificity, sensitivity, and MCC for the SAVEE dataset. Similarly, RMLSTM achieves 97.83%, 97.56%, 97.24%, 96.24%, and 96.75% for accuracy, PPV, specificity, sensitivity, and MCC for the YouTube dataset.

4. CONCLUSION

The RMLSTM is proposed in this research for multimodal facial expression recognition. The SAVEE and YouTube datasets are used, which include video/images, text, and audio modalities. Three preliminary operations are performed: normalizing images, removing punctuation from text data, and directly extracting MFCC features from audio inputs. In the feature extraction step, EfficientNetB0 is used for processing visual data, while TFIDF and MFCC are employed for textual data and speech signals,

respectively. EfficientNetB0 serves as an extractive feature mechanism for image/video samples due to its efficient computation of visual patterns. The TFIDF processes text data by converting verbal information into numerical vectors, enhancing the importance of meaningful words without diminishing standard terms. The MFCC analyzes audio signals through spectral and frequency-based characteristics to create effective audio signal representations. Subsequently, the features from each modality are fused, and the combined features are input into the RMLSTM. The gating and memory mechanisms in RMLSTM enhance its capabilities for handling sequential data, thereby enhancing multimodal facial expression recognition performance. The RMLSTM achieves 99.57% and 97.83% accuracy for the SAVEE and YouTube datasets, respectively. In the future, various DL techniques can be applied to further enhance recognition performance. Additionally, the developed model could be improved to enhance robustness against noisy environments and adversarial attacks, exploring preventive measures.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support and facilities provided by BMS Institute of Technology and Management, Yelahanka, Government Sri Krishnarajendra Silver Jubilee Technological Institute, Bangalore, and Jain University, Kanakapur Campus, Bengaluru, for carrying out this research work.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mamatha Kariyappa Rajanna	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Thejaswini Shankar	✓			✓			✓		✓			✓		✓
Rashmi Narasimhamurthy		✓				✓		✓	✓	✓	✓	✓		
Nandhini Vannivedu	✓		✓	✓			✓			✓	✓		✓	✓
Lakshmanan Hariprasad S.					✓		✓	✓		✓		✓		✓
Ananthapadmanabharao														

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dataset generated during the current study are available in the SAVEE, CK+, and RAVDEES repository at http://kahlan.eps.surrey.ac.uk/savee/, https://gts.ai/dataset-download/ck-dataset-ai-data-collection/, and https://www.innovatiana.com/en/datasets/ravdess.

REFERENCES

[1] Y. Dai, J. Li, Y. Li, and G. Lu, "Multi-modal graph context extraction and consensus-aware learning for emotion recognition in conversation," *Knowledge-Based Systems*, vol. 298, 2024, doi: 10.1016/j.knosys.2024.111954.
[2] Y. Lei and H. Cao, "Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2954–2969, 2023, doi: 10.1109/TAFFC.2023.3234777.
[3] H. M. Shahzad, S. M. Bhatti, A. Jaffar, and M. Rashid, "A multi-modal deep learning approach for emotion recognition," *Intelligent Automation and Soft Computing*, vol. 36, no. 2, pp. 1561–1570, 2023, doi: 10.32604/iasc.2023.032525.

[4] C. Dixit and S. M. Satapathy, "A customizable framework for multimodal emotion recognition using ensemble of deep neural network models," *Multimedia Systems*, vol. 29, no. 6, pp. 3151–3168, 2023, doi: 10.1007/s00530-023-01188-6.
[5] K. Ali and C. E. Hughes, "A unified biosensor–vision multi-modal transformer network for emotion recognition," *Biomedical Signal Processing and Control*, vol. 102, 2025, doi: 10.1016/j.bspc.2024.107232.
[6] S. B. Abdullahi, Z. A. Bature, L. A. Gabralla, and H. Chiroma, "Lie recognition with multi-modal spatial–temporal state transition patterns based on hybrid convolutional neural network–bidirectional long short-term memory," *Brain Sciences*, vol. 13, no. 4, 2023, doi: 10.3390/brainsci13040555.
[7] R. Gummula, V. Arumugam, and A. Aranganathan, "Facial emotion recognition using enhanced multi-verse optimizer method," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 2, pp. 1519–1529, 2024, doi: 10.11591/ijece.v14i2.pp1519-1529.
[8] J. Chen, S. Dey, L. Wang, N. Bi, and P. Liu, "Attention-based multi-modal multi-view fusion approach for driver facial expression recognition," *IEEE Access*, vol. 12, pp. 137203–137221, 2024, doi: 10.1109/ACCESS.2024.3462352.
[9] S. Chopparapu and J. B. Seventline, "An efficient multi-modal facial gesture-based ensemble classification and reaction to sound framework for large video sequences," *Engineering, Technology and Applied Science Research*, vol. 13, no. 4, pp. 11263–11270, 2023, doi: 10.48084/etasr.6087.
[10] G. Tang, Y. Xie, K. Li, R. Liang, and L. Zhao, "Multimodal emotion recognition from facial expression and speech based on feature fusion," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 16359–16373, 2023, doi: 10.1007/s11042-022-14185-0.
[11] 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, 2023, doi: 10.3390/electronics12020288.
[12] N. Sun, C. You, W. Zheng, J. Liu, L. Chai, and H. Sun, "Multimodal sentimental privileged information embedding for improving facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 16, no. 1, pp. 133–144, 2024, doi: 10.1109/taffc.2024.3415625.
[13] 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, doi: 10.1016/j.ins.2022.11.076.
[14] S. Liu, S. Huang, W. Fu, and J. C. W. Lin, "A descriptive human visual cognitive strategy using graph neural network for facial expression recognition," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 1, pp. 19–35, 2024, doi: 10.1007/s13042-022-01681-w.
[15] L. Wang, J. Zhao, H. Song, and X. Xu, "E2E-MFERC: a multi-face expression recognition model for group emotion assessment," *Computers, Materials and Continua*, vol. 79, no. 1, pp. 1105–1135, 2024, doi: 10.32604/cmc.2024.048688.
[16] R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra, and S. Singh, "Facial expression recognition in videos using hybrid CNN & ConvLSTM," *International Journal of Information Technology*, vol. 15, no. 4, pp. 1819–1830, 2023, doi: 10.1007/s41870-023-01183-0.
[17] J. Singh, L. B. Saheer, and O. Faust, "Speech emotion recognition using attention model," *International Journal of Environmental Research and Public Health*, vol. 20, no. 6, 2023, doi: 10.3390/ijerph20065140.
[18] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, 2022, doi: 10.1016/j.knosys.2022.108580.
[19] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech emotion recognition through hybrid features and convolutional neural network," *Applied Sciences*, vol. 13, no. 8, 2023, doi: 10.3390/app13084750.
[20] M. Cho, D. Jeong, and E. Park, "AMPS: predicting popularity of short-form videos using multi-modal attention mechanisms in social media marketing environments," *Journal of Retailing and Consumer Services*, vol. 78, 2024, doi: 10.1016/j.jretconser.2024.103778.
[21] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (SAVEE) database," *SAVEE*. [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/
[22] M. S. Alam, M. M. Rashid, R. Roy, A. R. Faizabadi, K. D. Gupta, and M. M. Ahsan, "Empirical study of autism spectrum disorder diagnosis using facial images by improved transfer learning approach," *Bioengineering*, vol. 9, no. 11, 2022, doi: 10.3390/bioengineering9110710.
[23] M. Z. Naeem, F. Rustam, A. Mehmood, Mui-zzud-din, I. Ashraf, and G. S. Choi, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," *PeerJ Computer Science*, vol. 8, 2022, doi: 10.7717/PEERJ-CS.914.
[24] A. Nosan and S. Sitjongsataporn, "Enhanced feature extraction based on absolute sort delta mean algorithm and MFCC for noise robustness speech recognition," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 4, pp. 422–436, 2021, doi: 10.22266/ijies2021.0831.37.
[25] D. R. I. Moses Setiadi *et al.*, "Integrating hybrid statistical and unsupervised LSTM-guided feature extraction for breast cancer detection," *Journal of Computing Theories and Applications,* vol. 2, no. 4, pp. 536–550, 2025, doi: 10.62411/jcta.12698.

BIOGRAPHIES OF AUTHORS

Mamatha Kariyappa Rajanna has around 20 years of academic experience as faculty. She is interested in signal and image processing and has published several papers in international/national conferences and Journals. She is guiding projects in the field of signal processing, communication, and cryptography. She got funding from Karnataka State Council for Science and Technology for several student projects and got a patent grant from Indian Patent office. She can be contacted at email: mamathakr@bmsit.in.

Thejaswini Shankar completed her master's in Digital Communication and Networking and Ph.D. in Signal Processing (computational neuroscience) from Visvesvaraya Technological University. With more than 20 years of teaching experience she is expertise in biomedical signal processing, communication, cryptography, and actively engaged in the field of computational neuroscience. Additionally, she has successfully managed various projects in the field of signal processing, and has received a funding of 40 Lakhs as CO-PI for the project titled "An adaptive motor imagery-based brain-computer interface" under GRE scheme from VGST. She can be contacted at email: thejaswini.s@bmsit.in.

Rashmi Narasimhamurthy has Ph.D. degree in wireless communication from Visveswaraya Technological University (2020). She is currently an assistant professor of BMS Institute of Technology and Management working in the Department of Electronics and Communication Engineering, she has experience in wireless communication, emphasis on physical layer design of trans receiver, embedded system, block chain technology, and optimization techniques. She can be contacted at email: rashmiswamy@bmsit.in.

Nandhini Vannivedu Lakshmanan is working as an associate professor in the Department of Electronics and Communication Engineering at Government Sri Krishnarajendra Silver Jubilee Technological Institute (Govt. SKSJTI), K. R. Circle, Bangalore, under Visveswaraya Technological University (VTU). She obtained her Doctorate in the field of photonics at University Visvesvaraya College of Engineering, Department of Electronics and Computer Science Engineering, Bangalore University, Bengaluru. She has more than 24 years of enriched teaching experience in both private and government organizations. She has published and presented research papers in both international and national conferences. Many of her research works were published in reputed international journals and few of them are Indian Patented. She has presented her academic talents in few of the well-known conclaves. She can be contacted at email: sunandi7276@gmail.com.

Hariprasad S. Ananthapadmanabharao a distinguished academician with over three decades of teaching expertise, complemented by 16 years of dedicated research and administrative experience. Held pivotal roles such as pro chancellor, vice principal, and head of the department. Successfully spearheaded government-funded projects and contributed significantly to the academic community with over 150 publications in esteemed national and international journals and conferences, including six Indian patents. Authored a textbook on advanced microprocessor and mentored nine research scholars towards their doctoral degrees, with eight currently under supervision. Pioneered the establishment of advanced laboratories in collaboration with leading industries, adept at managing intricate courses at both undergraduate and postgraduate levels. Renowned for training industry professionals, organizing various workshops, conferences, and faculty development programs (FDPs). Delivered numerous keynotes and invited talks, serving as a visiting professor at prestigious engineering institutions and featured as a guest on television programs discussing educational topics. Recognized with accolades such as best teacher and best leader awards. Currently working as director, Faculty of Engineering and Technology, JAIN (Deemed-to-be University). He can be contacted at email: sa.hariprasad@jainuniversity.ac.in.