Multimedia Tools and Applications (2024) 83:28373–28394
https://doi.org/10.1007/s11042-023-16443-1

Interpretable multimodal emotion recognition using hybrid fusion of speech and image data

Puneet Kumar[1] · Sarthak Malik[2] · Balasubramanian Raman[1]

Received: 11 January 2023 / Revised: 26 July 2023 / Accepted: 2 August 2023 /
Published online: 5 September 2023

Abstract

This paper proposes a multimodal emotion recognition system based on hybrid fusion that classifies the emotions depicted by speech utterances and corresponding images into discrete classes. A new interpretability technique has been developed to identify the important speech and image features leading to the prediction of particular emotion classes. The proposed system's architecture has been determined through intensive ablation studies. It fuses the speech & image features and then combines speech, image, and intermediate fusion outputs. The proposed interpretability technique incorporates the divide and conquer approach to compute shapely values denoting each speech and image feature's importance. We have also constructed a large-scale dataset, IIT-R SIER dataset, consisting of speech utterances, corresponding images, and class labels, i.e., 'anger,' 'happy,' 'hate,' and 'sad.' The proposed system has achieved 83.29% accuracy for emotion recognition. The enhanced performance of the proposed system advocates the importance of utilizing complementary information from multiple modalities for emotion recognition.

Keywords Affective computing · Interpretable AI · Multimodal analysis · Information fusion · Speech and image processing

Puneet Kumar and Sarthak Malik are both contributed equally to this work.

✉ Puneet Kumar
pkumar99@cs.iitr.ac.in

Sarthak Malik
sarthak_m@mt.iitr.ac.in

Balasubramanian Raman
bala@cs.iitr.ac.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, India

[2] Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee, India

1 Introduction

The multimedia data has overgrown in the last few years, leading multimodal emotion analysis to emerge as an important research trend [1]. The need to develop multimodal emotion processing systems capable of recognizing various emotions from images and texts is rapidly increasing. Research in this direction aims to help machines become empathetic as emotion analysis is used in various applications such as cognitive psychology, automated identification, intelligent devices, and human-machine interface [35]. The speech and image modalities portray human emotions and intentions very effectively [16]. Combining complementary information from both modalities could increase emotion recognition accuracy [54].

Researchers have attempted to identify emotions by processing audio and visual information separately [28, 36, 49]. However, multimodal emotion recognition, where the emotional context from multiple modalities are analyzed together, performs better than unimodal emotion recognition [54]. In this context, multimodal emotion recognition from speech and text modalities and image and text modalities have been performed; however, emotion recognition from speech and image modalities has yet to be fully explored. The domain of interpretable multimodal emotion recognition using the hybrid fusion of speech and image data aims to create systems capable of accurately recognizing and understanding human emotions expressed through speech and images. By leveraging both speech and image modalities, multimodal emotion recognition systems can achieve higher accuracy in emotion detection [54]. This field has gained significant attention due to the ever-growing multimedia data and the increasing need for empathetic machines in applications such as cognitive psychology, automated identification, intelligent devices, and human-machine interfaces [35].

The domain of multimodal emotion recognition using the hybrid fusion of speech and image data faces several challenges that impact the development and effectiveness of the systems. The majority of deep learning-based multimodal emotion recognition systems lack interpretability. They act as a black box, making it difficult to interpret their internal mechanisms. Developing interpretable models that can provide insights into the importance of speech segments and visual features in emotion recognition is crucial for enhancing trust and reliability in these systems. It inspired us to develop a multimodal emotion recognition system capable of recognizing emotions portrayed by speech utterances and corresponding images and explaining the importance of each speech segment and visual feature towards emotion recognition.

Another major challenge is the unavailability of sufficient labeled datasets for training. The real-life multimodal data contains generic images with facial, human, and non-human objects, but most existing multimodal datasets contain only facial and human images [2]. A few multimodal datasets are available that contain generic images; however, they consist of positive, negative, and neutral sentiment labels and do not contain multi-class emotion labels [8, 46]. A new dataset, the IIT-R SIER[1] dataset, has been constructed to address this issue. It contains generic images, corresponding speech utterances, and discrete class labels, i.e., 'anger,' 'happy,' 'hate,' and 'sad.' We used the data instances with identical predicted emotion labels for image and text modalities to construct the dataset.

One more challenge for multimodal emotion recognition research is the fusion of information from speech and image modalities, which requires a delicate balance between intermediate and late fusion to ensure optimal performance. Designing effective architectures for multimodal systems that can capture complementary information from both modalities without losing crucial details remains a complex task. The proposed system, 'ParallelNet,'

[1] The abbreviations have been defined in Appendix A.

recognizes emotions in speech utterances and corresponding images. It implements two networks, $N1$ and $N2$, to fuse the information of speech and image modalities in a hybrid manner of intermediate and late fusion. The architectures for $N1$ and $N2$ are determined through extensive ablation studies. A technique for interpreting important input features and predictions has also been developed. The improvements in SER after combining the complementary information from corresponding images have been analyzed. The proposed system has performed with an accuracy of 83.29% on the IIT-R SIER dataset. The dataset and code for this paper are accessible at https://github.com/MIntelligence-Group/SpeechImg_EmoRec.
The major contributions of this paper are documented as follows.

- **IIT-R SIER dataset**: The first contribution of this paper is the construction of a new large-scale dataset, named the IIT-R SIER dataset. It contains speech utterances, corresponding images, and emotion labels, which can be used for multimodal emotion recognition. The unique feature of this dataset is that it includes generic (with human faces, non-facial and non-human components) images. This is different from many existing datasets, which typically only include facial images. This dataset is crucial for training the proposed system to recognize emotions based on both speech and images.
- **ParallelNet**: The second key contribution of this paper is the proposal of a novel system called 'ParallelNet.' This system is capable of classifying an input that contains a speech utterance and its corresponding image into discrete emotion classes. The system's architecture was determined after extensive ablation studies. ParallelNet works by combining information from both speech and image modalities, using a hybrid of intermediate and late fusion techniques. It first combines or 'fuses' information from both speech and image at an intermediate level, then further combines the information at a later stage in processing, leading to more accurate emotion recognition.
- **Interpretability**: Finally, the paper introduces a new interpretability technique to identify the important parts of the input speech and image that contribute the most to recognizing emotions. This technique uses a 'divide and conquer' approach to calculate the Shapley values (a concept from cooperative game theory often used in machine learning to measure feature importance) of each speech and image feature. This method enhances the transparency and understanding of the system's decision-making process, an area often critiqued in machine learning models due to their 'black box' nature.

The rest of the paper has been organized as follows. The related works have been reviewed in Section 2. The proposed dataset, system, and interpretability technique have been described in Section 3 along with the dataset compilation procedure. Sections 4 and 5 discuss the experiments and results. Finally, Section 6 concludes the paper and highlights the directions for future research.

2 Related works

This Section surveys the existing literature on speech and image emotion recognition and the interpretability of deep neural networks.

2.1 Speech emotion recognition

The deep learning-based approaches using spectrogram features and attention mechanisms have shown state-of-the-art results for SER [4, 17, 50]. In this context, Xu et al. [49] generated multiple attention maps, fused and used them for SER. They observed an increased performance as compared to non-fusion-based approaches. In the context of using speech features for emotion understanding, Mustaqeem and Soonil Kwon [20] used salient and discriminative speech features for SER using CNN and Signal processing techniques. In another work, Majumder et al. [28] implemented a deep neural network to track speakers' identities showing specific emotions.

2.2 Image emotion recognition

The IER research is also an active domain. For instance, Kim et al. [16] built a deep feed-forward neural network to combine different levels of emotion features obtained by using the semantic information of the image. In another work, Rao et al. [36] prepared hierarchical notations for emotion recognition in the visual domain. Human emotions can be expressed in various modalities, out of which speech and image express the emotional intentions most effectively [16]. Analysis in a single modality may not recognize the emotional context completely, which leads to the need for multimodal emotion recognition approaches that analyze multimodal audio-visual emotional context [54].

2.3 Multimodal emotion recognition

Multimodal emotion analysis from audio-visual data has started getting researchers' attention lately [9, 12, 19]. For instance, Siriwardhana et al. [42] fine-tuned Transformers-based models to improve the performance of multimodal speech emotion recognition. Multimodal emotion recognition has been carried out for text and speech modalities [18, 29], and text and image modalities [8, 19, 46]. However, it has yet to be fully explored for speech & image modalities. Moreover, most deep learning-based multimodal emotion recognition systems work as a black box where it is difficult to interpret their inside mechanism. It inspired us to develop an interpretable multimodal emotion recognition system for speech and image modalities.

In the context of multimodal feature fusion, Teng et al. [45] proposed a regularization-based loss function to enhance weakly supervised video moment retrieval. In another work, Lu et al. [24] presented an unsupervised approach for video object segmentation using co-attention siamese networks. Lu et al. [25] performed object segmentation in relational visual data, focusing on scenes with complex object interactions and Maji et al. [27] proposed an advanced fusion-based system for speech emotion recognition, utilizing a dual attention mechanism with Conv-Caps and Bi-GRU features. In the context of multimodal sentiment analysis, Zeng et al. [53] developed a framework that combines heterogeneous graph convolution with in-domain self-supervised learning, whereas Han et al. [10] introduced a system that maximizes mutual information in multimodal fusion, effectively preserving task-related information and offering efficient solutions to the MI bounds issue. Researchers have used semantic heterogeneous graph convolutional networks [52] and pre-training at the word and sentence level [6] to improve the performance of multimodal sentiment analysis.

2.4 Interpretability of deep neural networks

The existing interpretability approaches compute each input feature's importance by back-propagating the network or observing the changes in output on changing the input [23]. In this direction, Riberio et al. [37] explained a network based on each input's importance.

Researchers have explained the layer-by-layer learning of deep neural networks and the output based on all the neurons' contributions [18, 40]. There are interpretability methods for visual analysis to compute input pixels' importance [23, 30, 37]. However, such methods still need to be sufficiently explored for speech modality. It inspired us to develop an interpretability technique for multimodal emotion recognition to explain the importance of each speech segment and each visual feature of the input.

3 Proposed methodology

3.1 Dataset construction

The IIT-R SIER dataset has been constructed using the B-T4SA dataset [46]. It contains generic (with human faces, non-facial and non-human components) images as opposed to only facial images contained by the existing datasets. The impact of universal images of non-human objects on emotional analysis in the article is twofold: a) *Dataset Diversity*: The inclusion of universal images containing non-human objects in the IIT-R SIER dataset enriches the diversity of emotional expressions [7]. This diversification allows the proposed multimodal emotion recognition system to handle a wider spectrum of emotional expressions, making it more versatile and applicable to real-world scenarios where emotions can be elicited by a multitude of factors beyond just human interactions. b) *Interpretable AI Research*: The presence of non-human objects in the images introduces complexities in emotion recognition, requiring an innovative interpretability technique [22]. The proposed interpretability technique explains the contributions of speech and visual features in recognizing emotions influenced by these objects. By understanding the impact of non-human objects on emotion recognition, it contributes to the advancement of interpretable AI in multimodal emotion recognition, enabling the development of more robust and reliable emotion analysis systems. Overall, universal images of non-human objects advance multimodal emotion analysis, dataset diversity, and interpretable AI research, fostering empathetic AI systems for understanding human emotions in diverse contexts.

The schematic diagram of the data construction process is shown in Fig. 1 and the details of the proposed method have been discussed in Section 3.2. The recent TTS models generate high-quality audio that can be used as a valid approximation of natural audio signals [5, 33,

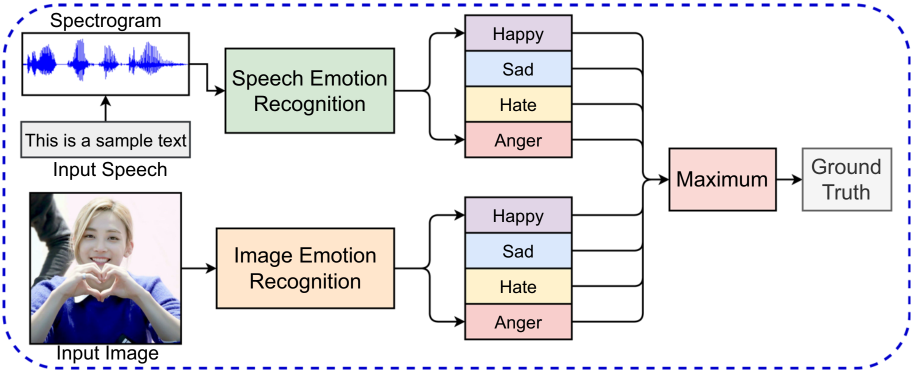


Fig. 1 Schematic diagram of data preparation process

47]. A pre-trained state-of-the-art TTS model, DeepSpeech3 [33], has been used to convert the text from the B-T4SA dataset to speech. The samples are manually cleaned by removing the corrupt and duplicate samples. Further, the following procedure has been followed to generate the ground-truth labels according to the overall emotional context represented by both modalities in combination.

Various parameters of the SIER dataset have been summarised in Fig. 2 whereas the procedure to construct the same has been described as follows. For emotion understanding, humans concentrate on the modality which reflects the emotions more strongly and use the complementary information from other modalities. To mimic this behavior, we have used the max operation to sufficiently combine the emotional context from different modalities.

The speech component of each data sample is passed through the SER model trained on the IEMOCAP [2] dataset; classification probabilities for each emotion class are obtained, and the maximum among the probabilities for all emotion classes is noted as max_1. Likewise, each sample's image component is passed through the IER model trained on FI [51] dataset, and the maximum classification probabilities for all emotion classes, i.e., max_2 is noted. The higher among max_1 and max_2 is observed, and the corresponding emotion label is assigned as the ground-truth label to the data sample. For example, if the IER model returned probabilities 0.1, 0.8, 0.05, and 0.05 for four emotion classes while the SER model gave 0.1, 0.1, 0.7 and 0.1 then we assigned the second emotion class to the sample considering $max(0.8$ and $0.7)$. The samples with $max(max_1, max_2)$ less than a threshold of 0.5 are discarded as the predicted class label must be at least double confident than the random prediction (probability 0.25). The samples labeled as 'excitement' and 'disgust' have been re-labeled as 'happy' and 'hate' as per Plutchik's wheel of emotions [34]. The final dataset contains $80,893$ samples, with $42,958$ labeled as 'happy,' $13,621$ as 'sad,' and $4,401$ and $19,913$ as hate and 'anger,' respectively. We did not take the samples with the same predicted labels by SER and IER, as speech and image modalities might favor different emotion classes in isolation. In contrast, we are interested in the emotion class denoted by both modalities, so the maximum of both probabilities is taken. Furthermore, the data samples with a high probability greater than 0.5 are retained in the IIT-R SIER dataset, denoting high confidence in the ground-truth label.

3.1.1 Human evaluation

We had two human readers (one male and one female) who spoke out and recorded the text components of the data samples. The evaluators listened to the machine-synthesized and human speech recorded by the human readers and labeled the emotion classes portrayed by them. The samples have been picked randomly, and the average of the evaluators' scores

Emotion	Samples
Happy	42,958
Sad	13,621
Hate	4,401
Anger	19,913
Total	**80,893**

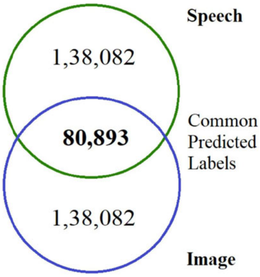


Fig. 2 Summary of IIT-R SIER dataset. Left: Class-wise data samples distribution. Right: Modality-wise data distribution

has been reported in Table 1. Here, A_i denotes the emotion classification accuracy when the human evaluators predicted the emotions considering the image components. Likewise, A_{ss} and A_{hs} are the accuracy values on considering the synthetic and human speech components, and A_{ss-i} and A_{hs-i} are the accuracies on considering both speech and image modalities.

The following two major observations can be drawn from Table 1: **i)** The similar values of 74.49% for synthetic speech and 78.91% for human speech advocate that the speech component of the data generated through TTS is mature enough and embodies the appropriate emotional context. **ii)** Considering complementary information from speech and image modalities led to higher emotion recognition performance. The evaluators also reported that 78.93% of the samples considering machine-synthesized speech along with the corresponding image was in line with the determined emotion label, whereas this is comparable to the value of 80.46% on considering human speech along with the corresponding image with is significantly higher than the accuracies on considering only image or only speech components.

3.2 Proposed multimodal emotion recognition system

Figure 3 depicts the architecture of the proposed multimodal emotion recognition system, which is determined in Section 4.4 through the ablation studies. A hybrid of intermediate and late fusion is implemented where intermediate fusion combines various modalities' information before classifying, while late fusion fuses the results after classification. The input image is in the space domain. The speech has been converted from the time domain to a log-mel spectrogram, i.e., the space domain. The proposed system contains networks $N1$ and $N2$ and dense, multiply, weighted addition, and softmax layers. $N1$ uses convolution and max-pool layers while $N2$ uses pre-trained networks VGG16 and VGG19 [41]. Both of these networks contain batch-normalization, flattened, and dense layers.

The intuition behind our architecture was to include a mechanism somehow So that each modality affects the other while making predictions. Here the two modalities are combined in two ways:- intermediate fusion and late fusion. First, to bring both modalities in the same domain audio signal is converted to a log-mel spectrogram to convert it from the time to space domain. Now, let us consider two networks, N1 and N2. N1 consists of a pre-trained network, a batch normalization layer, a flattening layer, and a dense layer of 512 neurons. While N2 has the following architecture: First, two convolution layers have 64 filters, then a max-pooling layer, then two more convolution layers of 128 filters, then a max-pooling

Table 1 Human evaluation of SIER dataset

Class	A_i	A_{ss}	A_{hs}	A_{ss-i}	A_{hs-i}
Happy	63.89%	66.67%	69.44%	73.48%	75.87%
Sad	75.00%	77.08%	78.13%	82.43%	83.27%
Hate	67.86%	71.43%	72.32%	77.64%	81.32%
Anger	70.31%	82.81%	85.94%	82.17%	84.19%
Overall	69.26%	74.49%	76.46%	78.93%	80.46%

Where A_m denotes the emotion classification accuracy for modality m, i: image modality, ss: synthetic speech, hs: human speech, $ss-i$: multimodal context combining synthetic speech and image modalities and $hs-i$: multimodal context combining human speech and image modalities

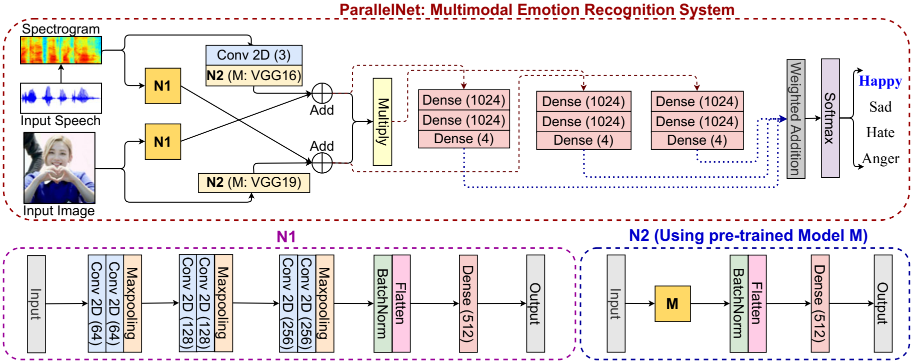


Fig. 3 Architecture of the proposed system (top), $N1$ (bottom left) and $N2$ (bottom right) where M is a pre-trained model

layer again, comes two more convolution layers of 256 filters, and then a max-pooling layer. Then it consists of a batch normalization layer, a flattening layer, and a dense layer of 512 neurons.

3.2.1 Intermediate fusion phase

Consider the networks fed with the image input as $N1_i$ and $N2_i$ while the networks $N1_s$ and $N2_s$ process the speech input. The speech is expressed as a spectrogram of sizes (128, 128, 1) and passed first to a convolution layer having three convolutional filters of size (1, 1) each and then to $N2_s$ where a pre-trained VGG16 network is used. The image with sizes (128, 128, 3) is passed to $N2_i$ which uses a pre-trained VGG19 network. As shown in (1), the output of $N1_s$ is added to the output of $N2_i$ to get F_s. Likewise, the outputs of $N1_i$ and $N2_s$ are added to obtain F_i. Then F_s and F_i are element-wise multiplied to obtain F_{mul}.

$$\begin{aligned} F_i &= Add(output(N1_i),\ output(N2_s)) \\ F_s &= Add(output(N1_s),\ output(N1_i)) \\ F_{mul} &= Multiply(F_i,\ F_s) \end{aligned} \tag{1}$$

The choice of using multiplication instead of weighted addition in (1) to combine F_s and F_i in the low-level fusion has been determined experimentally. Moreover, theoretically, if the speech and image modalities predict the same emotion class, they should support each other. However, let us consider a case where one modality predicts i^{th} emotion very strongly while another predicts another emotion j^{th} weakly. We expect the i^{th} emotion to be predicted weakly. It would not have been the case in the case of using addition, and the i^{th} emotion would have the upper hand. In comparison, the multiplication of both modalities would dilute the assertive behavior of the i^{th} emotion and give us the expected prediction.

3.2.2 Late fusion phase

The intermediate outputs $F_i,\ F_s,\ and\ F_{mul}$ are passed from three dense layers of size 1024, 1024, and 4 to obtain O_{sp} for speech, O_{img} for image, and O_{mul} for multiplied. These outputs are combined using the weighted addition layer as per (2) in a late fusion manner and passed from a softmax layer to get the final predicted label, $\hat{y}$. The weights $w_1,\ w_2,$ and w_3

are randomly initialized and passed to a softmax layer to normalize them to non-negative values. Their final values are learned using the Gradient Descent algorithm. It combines the information from speech and image modalities and the output of intermediate fusion in a hybrid manner.

$$\begin{aligned} &O = w_1 \times O_{sp} + w_2 \times O_{img} + w_3 \times O_{mul} \\ &\hat{y} = Softmax(O) \end{aligned} \tag{2}$$

3.3 Proposed interpretability technique

While making predictions, a deep learning-based classifier is expected to consider the input features that a human would consider. However, it is challenging to look into it and understand what input features it is considering [37]. To work on this challenge, we have developed an interpretability technique based on 'shapely values' [23] that denotes each input feature's importance. Theoretically, shapely values' computation takes exponential time. The computation has been approximated using the divide and conquer approach, as shown in (3). For a model with two features f_1 and f_2, shapely value $\mathscr{S}_{\{f_1\}}$ for feature f_1 denoting its importance is computed as follows.

$$\mathscr{S}_{\{f_1\}} = (1/2) \times MC_{f_1,\{f_1\}} + (1/2) \times MC_{f_1,\{f_1,f_2\}} \tag{3}$$

Here, $MC_{f_1,\{f_1\}}$ is feature f_1's marginal contribution to the model containing only f_1 and given by (4) where $score_{\{f_1\}}$ denotes the prediction for the ground-truth label using the model with feature f_1.

$$MC_{f_1,\{f_1\}} = score_{\{f_1\}} - score_{\{\phi\}} \tag{4}$$

The respective speech and image inputs are segregated and fed into the model while keeping the other as zero to compute the individual contribution of each modality. As depicted in Fig. 4, each modality's input is divided into two parts for a specified number of times, and the importance of each part towards the model's prediction is computed as per (2). Moreover, the calculation of the importance score follows the basic requirement of shapely values given by (5).

$$\mathscr{S}_{\{f_1\}} + \mathscr{S}_{\{f_2\}} = \mathscr{S}_{\{f_1,\ f_2\}} - \mathscr{S}_{\{null\}} \tag{5}$$

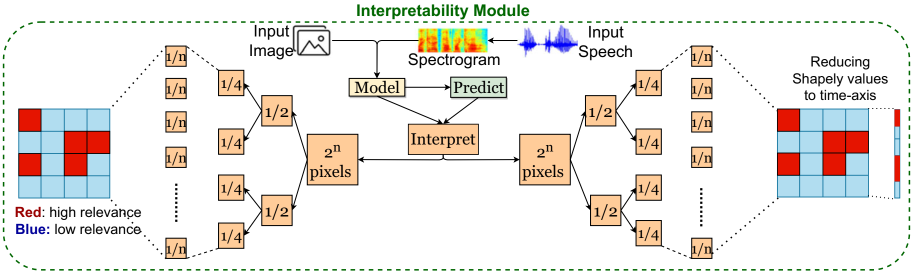


Fig. 4 Proposed interpretability technique's illustration. Here, each part's importance is computed using Divide & Conquer

The important image features for the predictions can be directly observed through the shapely values. In contrast, the important speech features are analyzed after transforming them to wave, i.e., time-domain representation. We first applied the shapely values directly and converted the spectrogram to speech; however, the speech reconstructed by this method was not meaningful. Then, we used the method of averaging the shapely values along the frequency axis and reducing them to the time axis to find the features' importance at a given time. The speech segments below a threshold shapely values of 30 percentile have been reduced to zero. The leftover segment with high importance is converted to text using a pre-trained STT model [3] and interpreted to understand how the model classifies each instance. The proposed interpretability technique has been summarised in Algorithms 1 and 2 describes various functions used by it.

Algorithm 1 Proposed interpretability technique.

```
function DNCSHAP_MM($model, data\_img, data\_speech, wd, ht, times$)
    ▷ Initialize data using 'numpy' library
    $data\_1 \leftarrow$ np.random.rand($wd, ht, 3$)
    $data\_2 \leftarrow$ np.random.rand($wd, ht, 1$)
    $data\_1 \leftarrow$ data_1.reshape($1, wd, ht, 3$)
    $data\_2 \leftarrow$ data_2.reshape($1, wd, ht, 1$)
    ▷ Reshape the data to be fed into the model
    $data\_f\_img \leftarrow data\_img$.reshape($1, wd, ht, 3$)
    $data\_f\_speech \leftarrow data\_speech$.reshape($1, wd, ht, 1$)

    ▷ Predict the label using the multimodal deep neural network
    $pred \leftarrow model$.predict($[data\_f\_img, data\_f\_speech]$)
    $arg\_max \leftarrow$ np.argmax($pred$)

    ▷ Calculate the predicted probability with original data
    $pred\_f \leftarrow pred[0][arg\_max]$

    ▷ Calculate the predicted probability with blank data
    $pred\_b \leftarrow model$.predict($[data\_1, data\_2])[0][arg\_max]$

    ▷ Calculate the predicted probability with only the image modality
    $pred\_1 \leftarrow model$.predict($[data\_f\_img, data\_2])[0][arg\_max]$

    ▷ Calculate the predicted probability with only the speech modality
    $pred\_2 \leftarrow model$.predict($[data\_1, data\_f\_speech])[0][arg\_max]$

    ▷ Compute the importance scores of image and speech modalities
    $score\_1 \leftarrow ((pred\_1 - pred\_b) + (pred\_f - pred\_2))/2$
    $score\_2 \leftarrow ((pred\_2 - pred\_b) + (pred\_f - pred\_1))/2$

    ▷ Initialize placeholders for SHAP values
    $SHAP\_value\_img \leftarrow$ np.zeros($[wd, ht]$)
    $SHAP\_value\_speech \leftarrow$ np.zeros($[wd, ht]$)
    $times \leftarrow times - 1$

    ▷ Compute $SHAP\_value\_img$ and $SHAP\_value\_speech$
    Compute_SHAP_value_img($SHAP\_value\_img, score\_1, times$)
    Compute_SHAP_value_speech($SHAP\_value\_speech, score\_2, times$)
    return $SHAP\_value\_img, SHAP\_value\_speech$
end function
```

Algorithm 2 Compute_SHAP_value_img and Compute_SHAP_value_speech functions.

```
function COMP_SHAP_VAL_IMG($SHAP\_value\_img$, $score\_1$, $times$)
    if $times \leq 0$ then
        ▷ Base case: assign the current score to the SHAP value
        $SHAP\_value\_img \leftarrow SHAP\_value\_img + score\_1$
    else
        ▷ Recursive case: divide image into parts and compute SHAP values
        for each division in image do
            ▷ Simulate removing the division (setting it to blank data)
            $blank\_data \leftarrow$ np.zeros($div$.shape)
            ▷ Calculate predicted probability with div set to blank data
            $pred\_blank \leftarrow model$.predict([$blank\_data$, $speech\_data$])[0][$arg\_max$]
            ▷ Calculate the score for the current div
            $div\_score \leftarrow (score\_1 - (pred\_blank - pred\_2))/2$
            ▷ Recursively compute SHAP values for the current div
            Comp_SHAP_val_img($SHAP\_value\_img$, $div\_score$, $times - 1$)
        end for
    end if
end function

function COMPUTE_SHAP_VAL_SP($SHAP\_value\_speech$, $score\_2$, $times$)
    if $times \leq 0$ then
        ▷ Base case: assign the current score to the SHAP value
        $SHAP\_value\_speech \leftarrow SHAP\_value\_speech + score\_2$
    else
        ▷ Recursive case: divide spectrogram into parts and compute SHAP values
        for each division in speech_spectrogram do
            ▷ Simulate removing the div (setting it to blank data)
            $blank\_data \leftarrow$ np.zeros($div$.shape)
            ▷ Calculate the predicted probability with division set to blank data
            $pred\_blank \leftarrow model$.predict([$image\_data$, $blank\_data$])[0][$arg\_max$]
            ▷ Calculate the score for the current div
            $div\_score \leftarrow (score\_2 - (pred\_blank - pred\_1))/2$
            ▷ Recursively compute SHAP values for the current div
            Comp_SHAP_val_sp($SHAP\_value\_speech$, $div\_score$, $times - 1$)
        end for
    end if
end function
```

4 Experiments

4.1 Experimental setup

The proposed system's network has been trained using Nvidia Quadro P5000 Graphics Card, whereas 64 bit Core(TM) i7-8700 Ubuntu system with 3.70 GHz 16GB RAM has been used for model evaluation.

4.2 Training strategy

The model has been trained using a batch size of 64, a train-test split of 70-30, 5-fold cross-validation, *Adam* optimizer, and *ReLU* activation function with a learning rate of 8×10^{-6}. The baselines and proposed models converged regarding validation loss in 18-23 epochs. The models have been trained for 30 epochs as a safe upper bound. A weighted combination of categorical cross entropy with weights 1 and 0.5 and categorical focal loss [21] has been used as the loss function. *EarlyStopping* and *ReduceLROnPlateau* have been incorporated with patience values 5 and 2. Accuracy, macro f1 [31], and *CohenKappa* [48] have been analyzed for evaluation.

4.3 Evaluation metric

To evaluate the performance of the baselines and proposed model, the accuracy metric has been used. As depicted in (6), it is defined as the ratio of the number of examples predicted with correct emotion or sentiment to the total number of data points.

$$Accuracy = \frac{\text{No. of correct predictions}}{\text{Total number of predictions}} \tag{6}$$

We have used the unweighted accuracy, which gives the same weight to each emotion class, regardless of how many samples of that class the dataset contains.

4.4 Ablation studies and models

The following studies analyze the Effect of using multimodal information and various network configurations.

4.4.1 Effect of multiple modalities

We first worked on SER and IER alone, using only speech samples and images from the IIT-R SIER dataset. Then we combined the information from speech and image modalities and performed multimodal emotion recognition. The IER-only experiments demonstrated high training but low validation accuracy. The convergence of accuracy and f1 score was not in line, and the *CohenKappa* metric's value was low, denoting over-fitting for a particular class. The accuracy and f1 score converged in line for SER-only experiments, though the accuracy was less.

4.4.2 Effect of various network configurations

As depicted in Fig. 3, ParallelNet consists of a family of networks where $N1$ and $N2$ can be varied in different situations. We first keep $N2$ fixed as EfficientNet [44] and evaluate three configurations for $N1$ – Configuration 1 uses two criss-crosses before and after $N2$. A *criss-cross* is a position combining two different modalities' networks. Configurations 2 and 3 implement single criss-cross before and after $N2$. Three baseline models have been implemented in line with these configurations. Configuration 3 was chosen for final implementation, showing in-line convergence and improved performance.

Further, keeping Configuration 3 fixed for $N1$'s configuration, following choices have been evaluated for $N2$ – VGG [41] (VGG-16, VGG-19), ResNet [11] (ResNet-34, ResNet-50, ResNet-101), InceptionNet [43] (Inception 3a, Inception 4a), MobileNet [13] and DenseNet [14]. The best performance has been observed with VGG16 as $N2_s$ and VGG19 as $N1_i$, which have finally been implemented by the 'ParallelNet.' The baseline and proposed models determined through the aforementioned studies are listed below, and their performance in terms of validation accuracies has been summarized in Table 2.

Table 2 Ablation studies' summary

Model	Accuracy	Avg. Time per epoch
SER Only	60.17%	10.6 min
IER Only	66.93%	12.4 min
Baseline 1	63.93%	23.3 min
Baseline 2	61.81%	23.2 min
Baseline 3	67.70%	23.4 min
Proposed ('ParallelNet')	83.29%	26.5 min

- **Baseline 1** – $N1$: Two criss-cross, $N2$: EfficientNet. It divides $N1$ into two parts and uses two crisscrosses before and after $N2$. A *criss-cross* is a position combining two different modalities' networks.
- **Baseline 2** – $N1$: Criss-cross before $N2$; $N2$: EfficientNet.
- **Baseline 3** – $N1$: Criss-cross after $N2$; $N2$: EfficientNet.
- **Proposed** – $N1$: Criss-cross after $N2$; $N2$: VGG.

The SER is likely affected by variations in speech information such as tone, speed, loudness, and pitch which can sometimes lead to inaccuracies in emotion recognition. The information from the visual modality can help improve emotion recognition accuracy by combining the emotional context not available in the speech modality. Likewise, the emotional context from speech can also help the visual modality improve emotion recognition accuracy. Hence, combining the complementary information from image and speech modalities performs better, advocating the importance of including complementary information from multiple modalities for emotion recognition. The observations from baseline 3 and the proposed method's performance also align with the abovementioned theory.

5 Results and discussion

The emotion classification results have been discussed in this Section, along with their interpretation and a comparison of sentiment classification results with existing methods. Results and technical observations for various phases of the proposed models are summarized as follows, whereas the detailed quantitative and qualitative results have been presented in the upcoming sections.

- *Data construction phase*: The results for the constructed dataset's evaluation have been presented in Section 3.1.1.
- *Feature extraction phase*: The proposed system is based on deep-learning-based architecture that performs the feature extraction automatically without human intervention.
- *Model selection phase*: Section 4.4 demonstrates the experimental studies performed to select the appropriate model and its architecture.
- *Hyperparameter tuning phase*: We have the grid search to tune the hyperparameters for the proposed system's architecture. The values 10^{-5}, 4×10^{-6}, 8×10^{-6}, and 10^{-6} were considered for the learning rate, whereas the options for the weights for categorical cross-entropy and focal loss were 0.5, 1, and 2. The most suitable hyperparameters resulting after evaluating the aforementioned options (described in Section 4.2) are used for the model training.

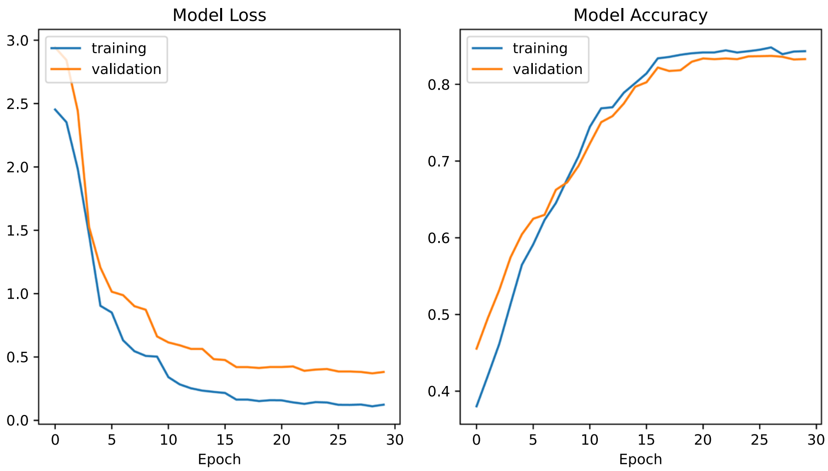


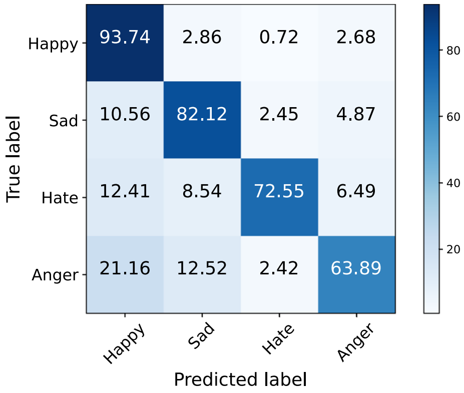
Fig. 5 Results for the model training phase

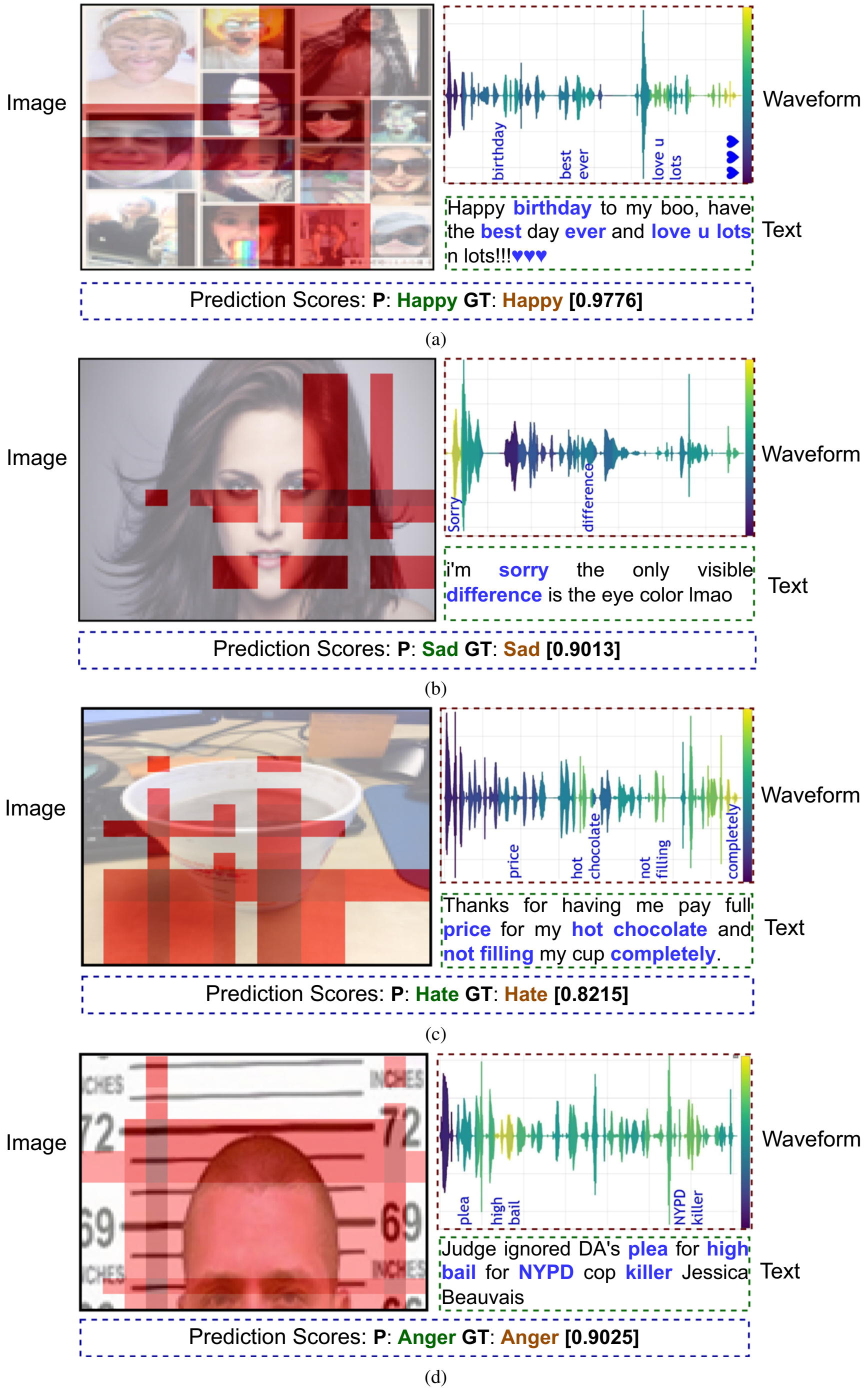
- *Model training phase*: The results for the model training phase in terms of loss and accuracy have been depicted in Fig. 5. The proposed system achieved an emotion classification accuracy of 89.29%.
- *Model evaluation phase*: The evaluation results for the proposed model have been included in Figs. 6 and 7 and Table 3.
- *Iterative refinement phase*: As discussed in Section 4.4.2, we first developed baseline 1 and then iteratively refined it to construct baseline 2, followed by baseline 3 and finally the proposed model. The results for the baselines and the proposed model have been included in Table 2.

5.1 Quantitative results

The ‘ParallelNet’ has achieved the emotion recognition accuracy of 83.29%. Its class-wise accuracies are shown in Fig. 6.

Fig. 6 Confusion matrix showing class-wise accuracies





Happy birthday to my boo, have the best day ever and love u lots n lots!!!♥♥♥

Prediction Scores: **P**: **Happy** **GT**: **Happy** **[0.9776]**

(a)

i'm sorry the only visible difference is the eye color lmao

Prediction Scores: **P**: **Sad** **GT**: **Sad** **[0.9013]**

(b)

Thanks for having me pay full price for my hot chocolate and not filling my cup completely.

Prediction Scores: **P**: **Hate** **GT**: **Hate** **[0.8215]**

(c)

Judge ignored DA's plea for high bail for NYPD cop killer Jessica Beauvais

Prediction Scores: **P**: **Anger** **GT**: **Anger** **[0.9025]**

(d)

Fig. 7 Sample results; here, 'P,' 'GT,' and 'Score' denote the predicted label, ground-truth label, and softmax score

Table 3 Comparing existing sentiment analysis methods

Approach	Author	Accuracy
Hybrid-T4SA FT-A	Vadicamo et al. [46]	49.10%
Hybrid-T4SA FT-F	Vadicamo et al. [46]	49.90%
VGG-T4SA FT-F	Vadicamo et al. [46]	50.60%
VGG-T4SA FT-A	Vadicamo et al. [46]	51.30%
Multimodal Sentiment Analysis	Gaspar et al. [8]	60.42%
Hybrid Fusion	Kumar et al. [19]	86.70%
ParallelNet (Proposed)		89.68%

The proposed method, ParallelNet, has been deployed for the B-T4SA dataset, and its sentiment classification performance has been compared against the existing results on the B-T4SA dataset

5.2 Qualitative results

Figure 7 shows sample emotion classification and interpretation results. The important speech and image features contributing to emotion classification are obtained, and corresponding words are highlighted. In the waveform, yellow and blue correspond to the most and least important features.

5.3 Results comparison

Comparison with existing Sentiment Analysis methods The emotion recognition results have been reported in Section 5.1. The IIT-R SIER dataset has been constructed from the B-T4SA dataset in this paper; hence, there are no existing emotion recognition results. However, sentiment classification (into neutral, negative, and positive classes) results on the B-T4SA dataset are available in the literature. This has been compared with the proposed method's sentiment classification results in Table 3.

As the same model (ParallelNet) has performed better than the existing methods for 2-class classification, it is empirical proof that it will perform better for 4-class classification as well. We have included the available sentiment analysis results for the B-T4SA dataset. Some more results are available for the T4SA dataset, which were not included because T4SA is a different dataset. We chose to work on the B-T4SA dataset than T4SA because B-T4SA is a balanced subset of the T4SA dataset, obtained by removing corrupted and near-duplicate images. It is an improvement over the T4SA dataset as it is less noisy and has a balanced distribution of samples. This makes it more suitable for training and evaluating machine learning models.

Comparison with human evaluation On considering the multimodal context from image and speech modalities, the human evaluation (See Table 1) and automatic evaluation (using ParallelNet. See Table 2) resulted in emotion classification accuracies of 80.46% and 89.68% respectively. In both cases, the emotion classification performance improved on considering the multimodal context compared to considering only image or speech modality. It establishes the importance of considering complementary information from multiple modalities for emotion recognition.

5.4 Discussion

The proposed solution in the research paper aims to enhance emotion recognition by leveraging both speech and image data. A cornerstone of the solution is the IIT-R SIER dataset, a large-scale, realistic dataset comprising speech utterances, corresponding images, and emotion labels. This dataset stands out for its inclusivity, capturing a range of images that include both human faces and non-human elements, as well as diverse emotion labels.

The proposed system classifies a given multimodal input containing speech and the corresponding image into 'anger,' 'happy,' 'hate,' and 'sad' classes. The proposed interpretability technique identifies the important speech and image features contributing to emotion recognition. An alternate procedure to construct the IITR-SIER dataset was to retain only those samples from the BT4SA dataset, for which SER and IER models predicted the same label and discarded the remaining samples. However, it would have caused a bias towards the models used in the first place for creating these labels. The SER and IER models have been retrained on the IITR-SIER dataset instead of using the pre-trained weights of the models used to construct the IITR-SIER dataset. However, suppose somebody uses the pre-trained models of either of the two modalities (trained on IEMOCAP and Flickr & Instagram datasets, respectively) used during dataset construction. In that case, they will get 100% accuracy. The closest to them any other evaluated machine learning model is, the more favorable its evaluation would be. The proposed procedure of considering the prediction probabilities for all emotion classes is more effective in capturing the overall emotional context represented by both modalities in combination. It leads to generating more accurate ground-truth labels.

Many existing datasets lack multimodal labels. To bridge that gap, we have proposed the IIT-R SIER dataset containing the emotion labels for image and speech modalities. At the same time, we have proposed a multimodal emotion recognition architecture that fuses the complementary information from both modalities. We first improved the unimodal SER and IER models that use the limited information from a single modality and then used them (unimodal models) to construct the multimodal ParallelNet. In this way, we have focused on preparing a strong dataset with multimodal labels and compiling a strong model for multimodal emotion recognition to conceal the weakness of unimodality.

The proposed system is a fusion-based emotion recognition architecture that consists of intermediate and late fusion phases, followed by an interpretability technique. In the intermediate fusion phase, the system uses two neural networks each for image and speech inputs, incorporating pre-trained VGG16 for speech and VGG19 for image processing. The outputs of these networks are then combined using an element-wise multiplication method, the efficacy of which has been determined through experimental evaluation. In the late fusion phase, the intermediate outputs undergo transformations through three dense layers before being combined using a weighted addition layer. The final predicted label is obtained after passing these combined outputs through a softmax layer, with the weights normalized to non-negative values using another softmax layer. The system then employs a Gradient Descent algorithm to learn the final values of these weights.

In addition to this fusion-based architecture, the system incorporates a novel interpretability technique based on 'shapely values'. This method allows it to determine the importance of each input feature, addressing the challenge of understanding what features a deep learning-based classifier is considering for its predictions. The technique uses a divide and conquer approach to compute the ‘shapely values’, making it more efficient. For speech inputs, this process involves segregation, division into parts, and importance computation for each part. Meanwhile, for image inputs, the important features for predictions can be observed directly. These computed values are then used to determine the most significant features for the model’s predictions.

The ParallelNet’s architecture has been determined through extensive ablation studies. It consists of a family of networks where $N1$ and $N2$ can be varied in different situations. We have first determined the optimal configuration for $N1$ to combine speech and image modalities’ information. Further, VGG, ResNet, InceptionNet, MobileNet, and DenseNet have been evaluated for $N2$. The best performance has been observed with VGG. The ResNet depicted very slow learning for a lower learning rate, while the learning fluctuated significantly for a higher learning rate. The model converged faster for the Inception Net and Efficient Net; however, the accuracy is lower. MobileNet and DenseNet have also resulted in low performance.

Apart from the experimental validation in Table 2, Fig. 7 qualitatively re-affirms the importance of combining complementary information from multiple modalities for more accurate emotion recognition. In the first and second cases, the image and speech features (shown by yellow parts of the waveform and denoted by corresponding words in blue) contribute to predicting the emotion class ‘sad.’ In the third and fourth cases, the image features have not been precisely captured, and the images seem neutral. However, the corresponding speech features *not filling* and *killer* contribute towards hatred and anger intent, which leads to recognizing the ‘hate’ and ‘anger’ classes. Although, as demonstrated in Fig. 5 (Sample Results), all the modalities contribute significantly towards emotion classification. However, as per the experimental results presented in Table 3, the image modality contributes slightly more than other modalities.

This work advances multimodal emotion recognition through hybrid fusion techniques and an interpretability framework. Integrating speech and image modalities in ParallelNet improves emotion classification by capturing complementary information. The interpretability technique sheds light on feature importance. The IIT-R SIER dataset contributes a large-scale resource for multimodal emotion recognition. Overall, this work highlights the significance of multiple modalities and interpretability for real and explainable emotion classification.

While the proposed multimodal emotion recognition system based on ParallelNet has achieved impressive accuracy, it has a few limitations. The interpretability technique provides insights into feature importance but needs more detailed explanations. To address this, future research could focus on advanced interpretability techniques for a more comprehensive understanding of emotion recognition. Additionally, the reliance on the specific IIT-R SIER dataset may introduce biases. To overcome this, researchers can explore diverse datasets for better generalizability.

6 Conclusions and future work

The importance of utilizing information from multiple modalities has been established for emotion recognition. The proposed system, *ParallelNet*, has resulted in better performance than SER alone, IER alone, and baseline models. The proposed interpretability technique identifies the important image and speech features contributing to emotion recognition. Future research plans include working on emotion recognition in other modalities such as text, videos, and emotion signal data. It is also planned to explore the interpretability of emotion recognition in the aforementioned modalities.

Appendix A

Table 4 defines various abbreviations that have appeared in this paper.

Table 4 List of Abbreviations

Abbreviation	Definitions
SER	Speech Emotion Recognition
IER	Image Emotion Recognition
CNN	Convolutional Neural Network
TTS	Text-to-speech
STT	Speech-to-text
IIT-R	Indian Institute of Technology Roorkee
SIER	Speech & Image Emotion Recognition
B-T4SA	Balanced Twitter for Sentiment Analysis
IEMOCAP	Interactive Emotional Dyadic Motion Capture
FI	Flickr & Instagram
VGG	Visual Geometry Group
N1	Network 1
N2	Network 2

Acknowledgements Ministry of Education, India, has supported this work with grant no. 1-3146198040. It has been carried out at the Machine Intelligence Lab, IIT Roorkee, India.

Author Contributions *Puneet Kumar*: Methodology, Implementation, Experiments, Result Analysis, Writing - original draft and editing. *Sarthak Malik*: Data Curation, Implementation, Conceptualization, Validation, Writing - review. *Balasubramanian Raman*: Conceptualization, Writing - review, Supervision, Project administration.

Funding This research was supported by the Ministry of Human Resource Development (MHRD) INDIA with reference grant number: 1-3146198040.

Availability of data and material available at https://github.com/MIntelligence-Group/SpeechImg_EmoRec.

Code Availability available at https://github.com/MIntelligence-Group/SpeechImg_EmoRec.

Declarations

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Conflicts of interest Authors have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

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

References

1. Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell (T-PAMI) 41(2):423–443
2. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang J-N, Lee S, Narayanan S-S (2008) IEMOCAP: Interactive Emotional dyadic MOtion CAPture data. Lang Resour Eval 42(4)
3. Chan W, Jaitly N, Le Q, Vinyals O (2016) Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 4960–4964
4. Dai D, Wu Z, Li R, Wu X, Jia J, Meng H (2019) Learning discriminative features from spectrograms using center loss for SER. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 7405–7409
5. Deep Mind (2016) Wavenet: a generative model for raw audio. http://deepmind.com/blog/article/wavenet-generative-model-raw-audio. Accessed on 20 Feb 2022
6. Fan S, Lin C, Li H, Lin Z, Su J, Zhang H, Gong Y, Guo J, Duan N (2022) Sentiment aware word and sentence level pre-training for sentiment analysis. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp 4984–4994
7. Finka L-R, Luna S-P, Brondani J-T, Tzimiropoulos Y, McDonagh J, Farnworth M-J, Ruta M, Mills D-S (2019) Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar. Sci Rep 9(1):9883
8. Gaspar A, Alexandre L-A (2019) A multimodal approach to image sentiment analysis. In Springer International Conference on Intelligent Data Engineering and Automated Learning (IDEAL). pp 302–309
9. Guanghui C, Xiaoping Z (2021) Multimodal emotion recognition by fusing correlation features of speech-visual. IEEE Signal Process Lett 28:533–537

10. Han W, Chen H, Poria S (2021) Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp 9180–9192
11. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 770–778
12. Hossain M-S, Muhammad G (2019) Emotion recognition using deep learning approach from audio visual emotional big data. Inf Fusion 49:69–78
13. Howard A-G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. Accessed 06 Jan 2023
14. Huang G, Liu Z, Van Der Maaten L, Weinberger K-Q (2017) Densely Connected Convolutional Networks. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). pp 4700–4708
15. Hu A, Flaxman S (2018) Multimodal sentiment analysis to explore the structure of emotions. In ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). pp 350–358
16. Kim H-R, Kim Y-S, Kim S-J, Lee I-K (2018) Building emotional machines: recognizing image emotions through deep neural networks. IEEE Trans Multimed (T-MM) 20(11):2980–2992
17. Kumar P, Jain S, Raman B, Roy P-P, Iwamura M (2021) End-to-end triplet loss based emotion embedding system for speech emotion recognition. In IEEE International Conference on Pattern Recognition (ICPR). pp 8766–8773
18. Kumar P, Kaushik V, Raman B (2021) Towards the explainability of multimodal speech emotion recognition. In INTERSPEECH. pp 1748–1752
19. Kumar P, Khokher V, Gupta Y, Raman B (2021) Hybrid fusion based approach for multimodal emotion recognition with insufficient labeled data. In 2021 IEEE International Conference on Image Processing (ICIP). IEEE, pp 314–318
20. Kwon S (2019) A CNN assisted enhanced audio signal processing for speech emotion recognition. Sensors 20(1):183
21. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV). pp 2980–2988
22. Lu X, Adams R-B, Li J, Newman M-G, Wang J-Z (2017) An investigation into three visual characteristics of complex scenes that evoke human emotion. In The Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp 440–447
23. Lundberg S-M, Lee S-I (2017) A unified approach to interpreting model predictions. In The 31st International Conference on Neural Information Processing Systems (NeurIPS). pp 4768–4777
24. Lu X, Wang W, Ma C, Shen J, Shao L, Porikli F (2019) See more, know more: unsupervised video object segmentation with co-attention siamese networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 3623–3632
25. Lu X, Wang W, Shen J, Crandall D-J, Gool L-V (2021) Segmenting objects from relational visual data. IEEE Trans Pattern Anal Mach Intell (T-PAMI) 44(11):7885–7897
26. Lu X, Wang W, Shen J, Crandall D, Luo J (2020) Zero shot video object segmentation with co-attention siamese networks. IEEE Trans Pattern Anal Mach Intell (T-PAMI) 44(4):2228–2242
27. Maji B, Swain M (2022) Advanced fusion-based speech emotion recognition system using a dual attention mechanism with conv-caps and Bi-GRU features. Electron 11(9):1328
28. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E (2019) DialogueRNN: an attentive RNN for emotion detection in conversations. In Conference on Artificial Intelligence (AAAI) 33:6818–6825
29. Makiuchi M-R, Uto K, Shinoda K (2021) Multimodal emotion recognition with high-level speech and text features. In IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)
30. Malik S, Kumar P, Raman B (2021) Towards interpretable facial emotion recognition. In The 12th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP). pp 1–9
31. Opitz J, Burst S (2019) Macro f1 and Macro f1. arXiv:1911.03347
32. Pagé Fortin M, Chaib-draa B (2019) Multimodal multitask emotion recognition using images, texts and tags. In ACM workshop on cross-modal learning and application. pp 3–10
33. Ping W, Peng K, Gibiansky A, Arik S-O, Kannan A, Narang S, Raiman J, Miller J (2018) DeepVoice 3: scaling text-to-speech with convolutional sequence learning. In The 6th Int. Conference on Learning Representations (ICLR)
34. Plutchik R (2001) The nature of emotions. J Stor Digit Lib Am Sci J 89(4):344–350
35. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion. Elsevier Inf Fus J 37:98–125

36. Rao T, Li X, Xu M (2019) Learning multi-level deep representations for image emotion classification. Neural Process Lett 1–19
37. Ribeiro M-T, Singh S, Guestrin C (2016) Why should i trust you? Explaining predictions of any classifier. In International Conference on Knowledge Discovery & Data mining (KDD). pp 1135–1144
38. Salamon J, Bello J-P (2017) Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett 24(3):279–283
39. Selvaraju R-R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In The IEEE/CVF International Conference on Computer Vision (ICCV). pp 618–626
40. Shrikumar A, Greenside P, Kundaje A (2017) Learning Important Features Through Propagating Activation Differences. In International Conference on Machine Learning (ICML). pp 3145–3153
41. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Accessed 06 Jan 2023
42. Siriwardhana S, Reis A, Weerasekera R (2020) Jointly fine tuning 'BERT-Like' Self supervised models to improve multimodal speech emotion recognition. INTERSPEECH pp 3755–3759
43. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). pp 1–9
44. Tan M, Le Q (2019) EfficientNet: rethinking model scaling for CNN. In International Conference on Machine Learning (ICML). pp 6105–6114
45. Teng J, Lu X, Gong Y, Liu X, Nie X, Yin Y (2021) Regularized two granularity loss function for weakly supervised video moment retrieval. IEEE Trans Multimed (T-MM) 24:1141–1151
46. Vadicamo L, Carrara F, Cimino A, Cresci S, Dell'Orletta F, Falchi F, Tesconi M (2017) Cross-media learning for image sentiment analysis in the wild. In IEEE International Conference on Computer Vision Workshops (ICCV-W). pp 308–317
47. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: a generative model for raw audio. arXiv:1609.03499
48. Vieira S-M, Kaymak U, Sousa J-MC (2010) Cohen's kappa coefficient as a performance measure for feature selection. In International Conference on Fuzzy Systems. IEEE, pp 1–8
49. Xu M, Zhang F, Khan S-U (2020) Improve accuracy of speech emotion recognition with attention head fusion. In IEEE Annual Computing and Communication Workshop and Conference (CCWC). pp 1058–1064
50. Yenigalla P, Kumar A, Tripathi S, Singh C, Kar S, Vepa J (2018) Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In INTERSPEECH pp 3688–3692
51. You Q, Luo J, Jin H, Yang J (2016) Building a large scale dataset for image emotion recognition: the fine print and the benchmark. In The 30th AAAI Conference on Artificial Intelligence (AAAI). pp 308–314
52. Zeng Y, Li Z, Chen Z, Ma H (2023) Aspect-level sentiment analysis based on semantic heterogeneous graph convolutional network. Front Comput Sci 17(6):176340
53. Zeng Y, Li Z, Tang Z, Chen Z, Ma H (2023) Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis. Exp Syst Appl 213:119240
54. Zeng Z, Pantic M, Roisman G-I, Huang T-S (2009) A survey of affect recognition: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell (T-PAMI) 31(1):39–58

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.