Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.032525
Article

A Multi-Modal Deep Learning Approach for Emotion Recognition

H. M. Shahzad[1,3], Sohail Masood Bhatti[1,3,*], Arfan Jaffar[1,3] and Muhammad Rashid[2]

[1]The Superior University, Lahore, Pakistan
[2]National University of Technology, Islamabad, Pakistan
[3]Intelligent Data Visual Computing Research (IDVCR), Lahore, Pakistan
*Corresponding Author: Sohail Masood Bhatti. Email: sohailmasood@superior.edu.pk
Received: 20 May 2022; Accepted: 24 June 2022

Abstract: In recent years, research on facial expression recognition (FER) under mask is trending. Wearing a mask for protection from Covid 19 has become a compulsion and it hides the facial expressions that is why FER under the mask is a difficult task. The prevailing unimodal techniques for facial recognition are not up to the mark in terms of good results for the masked face, however, a multi-modal technique can be employed to generate better results. We proposed a multi-modal methodology based on deep learning for facial recognition under a masked face using facial and vocal expressions. The multimodal has been trained on a facial and vocal dataset. We have used two standard datasets, M-LFW for the masked dataset and CREMA-D and TESS dataset for vocal expressions. The vocal expressions are in the form of audio while the faces data is in image form that is why the data is heterogenous. In order to make the data homogeneous, the voice data is converted into images by taking spectrogram. A spectrogram embeds important features of the voice and it converts the audio format into the images. Later, the dataset is passed to the multimodal for training. neural network and the experimental results demonstrate that the proposed multimodal algorithm outsets unimodal methods and other state-of-the-art deep neural network models.

Keywords: Deep learning; facial expression recognition; multi-model neural network; speech emotion recognition; spectrogram; covid-19

1 Introduction

Facial expressions of emotion are the most essential signs of the face because it reveals people's personalities, emotions, motivations, or intent. Wearing a mask for protection from Covid 19 has become a compulsion [1]. Nowadays, emotion recognition tools and methods for non-verbal communication are widely used in every field of life e.g., A teacher can get the mentality of his student by getting the facial expressions of that particular student. Similarly, Health care professionals can achieve the goal of better understanding with the patients and improving medical rehabilitation. In a business, a person can easily negotiate with the other person with the help of facial expressions. Most importantly, facial expression and recognition are playing an important role in national security and law enforcement [2,3].

A study has shown that people wearing masks are less likely to be accurately identified because with a hidden face certain emotions like happiness, sadness, anger, etc. Due to this, participants are less accurate in identifying emotions when the person was wearing a mask [4]. In addition, the use of a face mask has also impaired people's ability to understand the gestures of others, which can be important for understanding the speaker's intentions [5].

This finding suggests that people are better able to read the emotions of people whose faces they can see clearly than those whose faces are hidden. The present study aimed to investigate the impact of wearing a face mask on emotion recognition accuracy is less likely to accurately interpret others' emotions. Due to covering the nose and mouth, the accuracy to recognize expression has been reduced [6].

In the past few months, face masks have become an essential part of our lives. Although face masks are an effective means of preventing the spread of respiratory viruses. A major issue with face masks is that they obscure facial expressions, which makes it more difficult to detect emotions. In this paper, we Propose a multi-model approach using the datasets in which most of the facial expressions are hidden under the mask.

2 Related Works

The current study found that when people had to wear a mask, their performance on tests of facial recognition and expression declined. The author suggests that this may be because face masks obscure the lower half of the face, where most of the important facial cues are located. There was a decrease in the accuracy of emotion recognition in the masked target faces when compared to the unmasked target faces. Overall accuracy for unmasked is (69.9% *vs*. 89.5%) and for masked face (48.9% *vs*. ~71.7%) targets were considerably lower [7].

Convolution Neural Network (CNN) and Deep learning (DL) architectures are getting popular nowadays. DL has shown impressive results in various fields such as image recognition, natural language processing, and machine translation [8]. CNN architectures are composed of multiple layers of processing units, with each layer performing a specific task. The layers are usually connected in a feedforward manner, meaning that the output of one layer is the input to the next layer. CNNs can effectively learn complex spatial relationships between features in an image. This is because each layer in a CNN can learn to detect different kinds of patterns, and the layers can be stacked to form a hierarchy of patterns [9].

From the existing literature, it has been found that a deep learning-based facial expression recognition method has achieved better accuracy as compared to the traditional Machine Learning algorithms. The proposed method uses a multi-layer neural network to learn the features of facial expressions from a database of facial images. The proposed method is evaluated on the FERET (without mask) database and the results show that the proposed method achieves higher recognition accuracy than the existing methods [10].

Facial expression is recognized easily and accuracy has been achieved but in occlusion faces or covering the face with a mask, the accuracy to recognize facial expression drastically declines because a major portion of the faces is hidden. There are many ways to improve the accuracy of a multi-model technique, among which, one way to do it is to add more features. If the features are hidden, then adding more features can help improve the accuracy of the technique. Many researchers nowadays are working on multimodal techniques to improve the accuracy by using CNN where data are complex [11].

The author has evaluated a multimodal approach for facial recognition. The proposed method uses a combination of the low-level facial key point feature and the high-level self-learning feature. The experimental results show that the proposed method can achieve better recognition results than the single modal features [12].

In a similar work, a method Multichannel Convolutional Neural Network (MCCNN) was proposed and it was evaluated on the FER dataset. The results show that the proposed MCCNN achieves better recognition accuracy than the traditional CNN-based architectures [13].

Many researchers are using multimodal [14] which means multiple sources of information are being used. In multimodal multiple types of data are being used to improve the accuracy of the model. This could include using data from multiple sensors, using data from multiple periods, or using data from multiple geographical locations. By using multiple sources of information, the model can better account for variability and improve its predictions.

In this paper, a multimodal technique is proposed to identify facial expressions while wearing a mask because it is less likely to accurately identify facial features due to the reason that certain emotions and their facial expressions may be hidden and difficult to be read because the masks obscure certain features of the face.

3 Datasets

We have used the M-LFW-FER dataset for the masked dataset, CREMA-D, and TESS dataset to create our multi-model neural network to tune our model.

3.1 M-LF-W Dataset

The M-LFW-FER [15] dataset is a collection of 9825 images of faces wearing masks (5479 positive, 799 negative, 3547 neutral) and the testing dataset contains 1213 images (676 positive, 96 negative, 441 neutral). Tab. 1 shows the statistics of the MLFW database.

Table 1: Statistic of mask count of MLFW database

MLF-W-FER	Positive	Negative	Neutral	Total sample sets
Training	5194	766	3347	9307
Testing	644	95	416	1155

Each image is labeled with the identity of the person wearing the mask and the expression on the face. The MLFW dataset was created by automatically detecting the face and the mouth in each image, and then cropping the image which includes only the face and the mask. The labels of the dataset include neutral, positive, and negative expressions as shown in the Fig. 1. We are using the same dataset of MLFW images which is publicly available.

Figure 1: Images from each emotion class in the M-LFW-FER dataset

3.2 Voice Datasets

We have used the TESS dataset and CREMA-D dataset to test with M-LFW dataset.

3.3 TESS Dataset

The Toronto Emotional Speech Set (TESS) [16] is a speech dataset consisting of 4048 audio files. There are 7 emotions of humans that have been considered for classification including happy, angry, sad, neutral, fearful, disgust, and surprised.

3.4 CREMA-D Dataset

The CREMA-D (Crowd-Sourced Emotional Multimodal Actors Database) dataset is a large-scale dataset that is used to train and test emotion recognition models. The dataset contains audio recordings of people expressing seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. CREMA-D includes 7,442 clips from 91 actors and actresses, of diverse ages and ethnicity [17].

4 Proposed Model

4.1 Pre-processing and Data Augmentation

In the TESS dataset, there are more categories (7 expressions such as happy, neutral, disgust, sad, surprise and fear) as compared to the M-LFW-FER dataset (only 3 expressions i.e., positive, neutral, and negative). For this purpose, we have selected only 3 categories of voice expression e.g., happy, angry, and neutral. The names of the categories in both the models must be the same, that's why we have changed the names of the two categories of the TESS dataset, happy as positive, and angry as negative and we did not change the name of the neutral category and its name is unchanged. Similarly, we have changed the name of the categories given in the CREMA-D voice dataset as we did in the TESS dataset.

The recordings of the TESS and CREMA-D datasets are available in WAV format. The final step in the preprocessing of the voice dataset is to convert the WAV files into spectrograms of images as mentioned in Figs. 2 and 3.

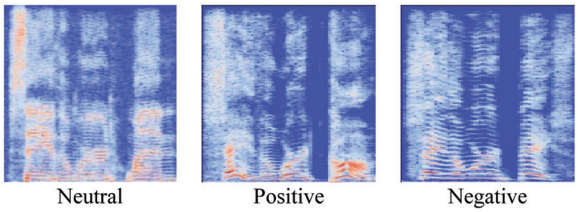


Figure 2: TESS dataset (Converted .wav file into spectrogram)

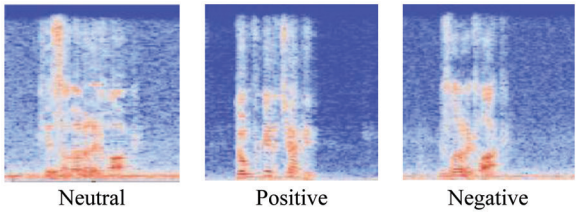


Figure 3: CREMA-D dataset (Converted .wav file into spectrogram)

These spectrogram files are generated by using the LIBROSA library and then cropped the important features of the spectrogram. In the multimodal technique, both dataset M-LFW-FER (Masked) and TESS (voice spectrogram) and M-LFW-FER and CREAM-D categories of expressions must be equalized. For this purpose, we have applied an augmentation rule in the TESS dataset and equalized the number of voice expressions as the number of expressions available in the M-LFW-FER dataset.

A spectrogram can be used to identify the different frequencies present in a sound wave and can be used to analyze the sound. The signal strength is represented by the amplitude of the waveform. The different frequencies are represented by different colors. It can also be a signal, speaker recognition and speech recognition [18,19].

The spectrogram is defined in Eq. (1)

$$y(t) = A\sin(2\pi fct) + B\cos(2\pi fct) \quad (1)$$

where y(t) is the amplitude of the sound at time t, A is the amplitude of the sound at time 0, f is the frequency of the sound, and C is the speed of sound.

Our proposed model consists of two types of layers. The First layers consist of convolution layers and the second one is fully connected layers for both inputs of visual and audio expression. The entire architecture of the proposed model is depicted in Fig. 4.

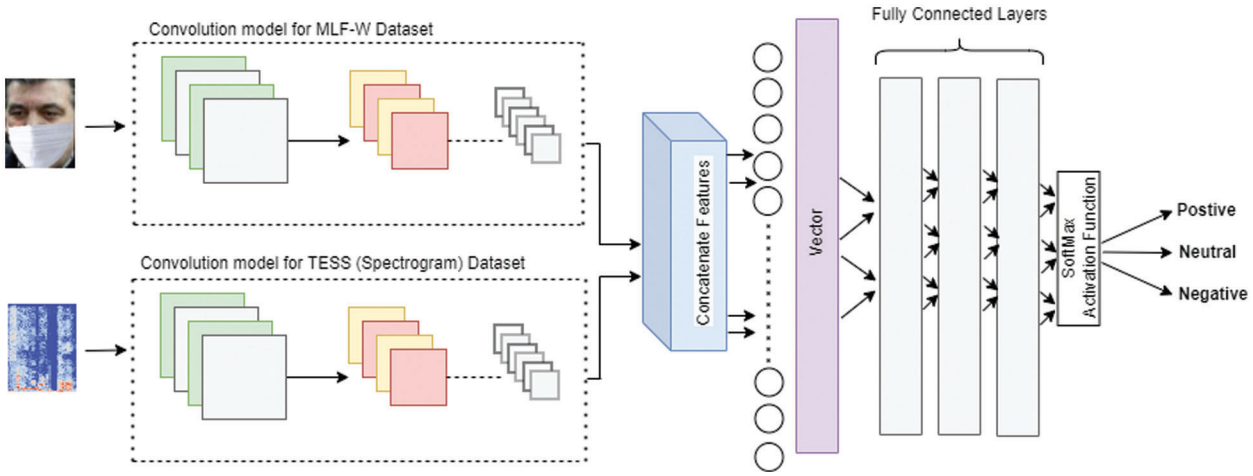


Figure 4: Flowchart of the proposed multimodal method

4.2 Convolution and Pooling Filters

Fig. 5 shows block 1 of the first CNN architecture for the MLF-W-FER dataset, the first layer is a convolution layer with 64 filters of size 3 × 3 by applying the regularization technique (ridge regularization l2). The second layer is another convolution layer with 64 filters of size 3 × 3 and activation is done using the RELU activation function. Then, the Max pooling is performed. The approach of Block 2, Block 3, and Block are similar to that of Block 1 except for the change of filter size. The filter size used in block 1 is 64, 128 in the second block, 256 in the third block, and 512 filters is used in the fourth block. The same process is done with the second CNN for CREMA-D and the TESS dataset.

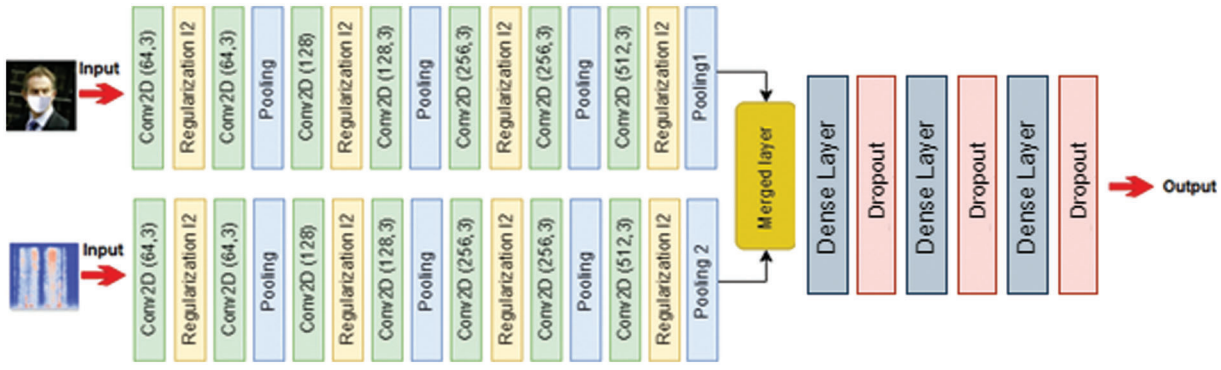


Figure 5: Flowchart of the proposed multimodal method

We have added regularization techniques in the proposed architecture like dropout, ridge regression, and batch normalization. Dropout helps [20] to prevent overfitting by randomly dropping out (setting to zero) some features (usually hidden units) during training. This forces the model to learn multiple independent representations of the same data, which reduces overfitting and improves generalization. Ridge classification [21] is another regularization technique that can be used to reduce overfitting. Ridge classification adds a penalty term to the loss function that encourages the weights to be small. This can help to prevent overfitting by reducing the complexity of the model. Batch normalization [22] is a process that helps stabilize training by normalizing the inputs to each layer. This can help improve the accuracy of the model by reducing the amount of variance in the data.

Convolution is applied on an image "I" to filter and extract information from it. The filter "k" determines which information is extracted is defined in Eq. (2)

$$f_m = 1 * k(a, b) = \sum_{i=0}^{\text{colum}} \sum_{j=0}^{\text{rows}} I(i, j) k(i - a, j - b) \tag{2}$$

The model uses ReLU activation function and it is the most commonly used activation function in neural networks. The main advantage of using the ReLU activation function is that it does not have a vanishing gradient, which is a problem with other activation functions such as the sigmoid function [23]. ReLU also has a very simple mathematical form, which makes it easy to compute and it is shown in Eq. (3):

$$y = \max(0, x) \tag{3}$$

4.3 Features Combination

In the proposed framework, our model takes two separate inputs of visual and audio features. Visual and audio features are learned in two separate streams of convolution neural network (CNN), which then combine both features of visual and audio images.

4.4 Flattening

After combining the features of both inputs of images, then we flattened them into a 1D of an array for feeding into the fully connected layer.

4.5 Fully Connected (FC) Layers / Neural Network

The fully connected layer is the main learning unit of CNNs.

4.6 SoftMax and Cross Entropy

The final step is to interpret the output of the FC layer through a SoftMax function. These outputs the probabilities of each class [24]. The cross entropy is used to calculate the loss.

5 Results and Discussion

In this paper, we presented a new deep architecture for facial emotion recognition. The proposed architecture comprises two streams for extracting features from the facial and audio emotions, that are then combined into their features before passing it to the dense layer. The visual stream can be used to identify facial expressions, while the auditory stream can be used to identify the emotions associated with those expressions. The combination of these two streams leads to more accurate results than using each stream separately.

Table 2: Accuracies for VGG-16, ResNET-50 and efficent NEtV2M on MLF-W-FER dataset

Model	Accuracy
VGG-16	55.6
ResNet-50	56.8
EfficientNetV2M	49

We examine MLF-W-FER dataset on VGG-16, ResNet-50, and EfficentNetV2M architecture and achieve testing accuracy 55.6%, 56.8%, and 49% respectively, as shown in Tab. 2. The testing accuracy on mask data set is not good because a data set of wearing a mask covers most of the features used in emotions and this led to a drop in performance on CNN and other machine learning methods.

As far as the CREMA-D and TESS datasets are concerned, some authors have evaluated results on them by applying neural network and they achieved testing accuracy 55.01% [25] for CREMA-D and 97.15% [26] for TESS respectively as reflected in Tab. 3 which shows that the results achieved by single model technique in masked or complex are not good enough. Moreover, CREMA-D dataset has not achieved better results except for TESS dataset.

Table 3: Accuracies of TESS and CREMA-D dataset

Model	Datasets	Accuracy
VGG16	TESS	97.15
CNN	CREMA-D	55.01%

In the proposed model, a multi-model technique has been evaluated in different experiments. In the first experiment, MLF-W-FER with TESS dataset, we achieved an accuracy of 99.92%, while in the second experiment of MLF-W-FER with CREMA-D dataset 75.67% accuracy has been achieved, as shown in Tab. 4. The accuracy achieved with multimodal is higher than the single model technique used for the masked dataset.

Table 4: Performance evaluation of Multimodal CNN (Proposed method)

Dataset	Accuracy
MLF-W-FER & TESS	99.92%
MLF-W-FER & CREMA-D	75.67%

In our purposed multimodal technique, we have changed the architecture of VGG-16 by adding regularization, and dropout techniques to overcome the overfitting problem. We have added the ridge regularization technique (l2 = 0.0001) in the convolution of each layer. The sizes for hidden layers were kept at 4,096, 2,048, and 1024 respectively along with a SoftMax output layer of 3 emotion classes. Furthermore, we have added a 0.3 dropout technique in each layer as shown in Fig. 4. The proposed model trained using Adam optimizer with a batch size of 32 and a learning rate of 0.0003. The number of epochs was set to 50.

As multiple authors achieved better results with the TESS dataset and not with the CREMA-D dataset that is why the focus of this article is mainly on getting better results with CREMA-D. A series of

experimental analyses is performed using the multimodal architecture to evaluate the performance of the proposed framework using the MLF-W-FER and CREMA-D dataset. In different experiments. Dropout and regularization (L2) techniques are used to prevent overfitting problems in MLF-W-FER and CREMA-D datasets. Fig. 6 shows the results of experiments carried out in this research.

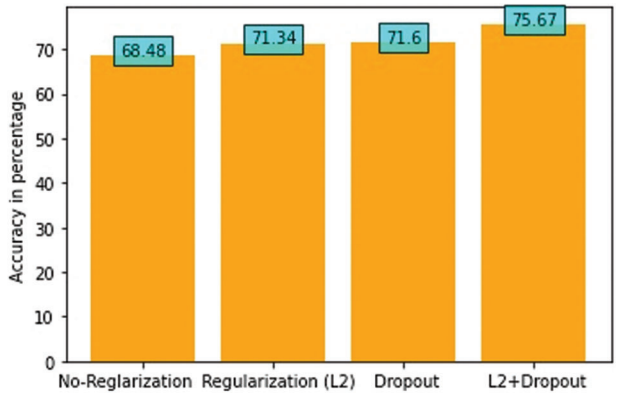
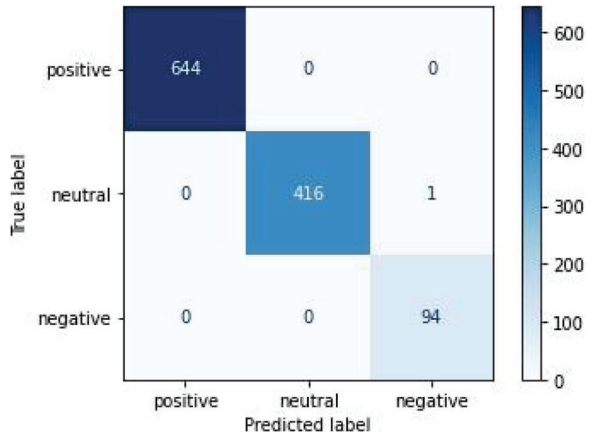
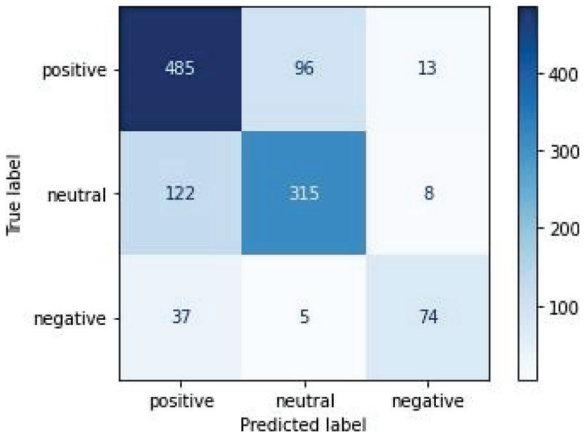


Figure 6: Comparison of different experiments in multimodal techniques

In the first experiment of the multimodal technique, where an accuracy of 68.48% has been achieved which is the least among all the experiments because no regularization technique was applied in the first place. In the second experiment, while using the regularization (L2) technique, an accuracy of 71.34% has been achieved which is better than the first experiment.

The dropout technique was applied in the experiment followed, where an accuracy of 71.60% was achieved which is quite similar to the previous one. In the final experiment, both the techniques (regularization (L2) and Dropout) were used to overcome the overfitting problem in the multimodal architecture, where accuracy of 75.67% was achieved, which is better than all the previous experiments carried out in this research. Confusion matrix for this experiment is shown in Tab. 5.

Table 5: Comparison of confusion matrix of multimodal and single model technique

Confusion matrix of M-LFW-FER & TESS multimodal CNN (proposed framework)	Confusion matrix of M-LFW-FER & CREMA-D multimodal CNN (proposed framework)
	

Looking at the bigger picture, it can be seen that the accuracy of a multimodal has been improved up to 8%, from 68.48% to 75.67%, by applying dropout and regularization technique, while the accuracy of a single model to multimodal approach enhances the overall test accuracy from 55.6% to 75.67%, that is a 20% increase in MLF-W-FER and CREMA-D dataset.

It has been observed that the prediction values of all the three classes (positive, negative, and neutral) are between 67 to 80. The accuracy for individual classes is depicted in the confusion matrix as shown in Fig. 2. Positive, neutral, and negative classification accuracy recorded in a single model technique that is 56%, 57% and 54% whereas VGG16 has shown better results. In the multi-modal technique, the accuracy of 81.64%, 70.78%, and 63.79% achieved respectively on the testing dataset for each class, which is higher than the single model CNN architecture.

6 Conclusion

In this work, a multimodal technique has been proposed to achieve better accuracy on a challenging dataset like MLF-W-FER. In the multimodal approach, voice emotions datasets of TESS and CREMA-D were considered for evaluation of masked facial expression of the MLF-W-FER dataset. The voice emotion datasets, TESS, and CREMA-D datasets were parallelly used for the evaluation of the proposed multimodal architecture with the MLF-W-FER faces dataset. Accuracy of 75.67% was achieved on M-LFW-FER and CREMA-D datasets, while 99.92% on MLF-W-FER and TESS datasets.

Furthermore, in our experiments, we applied different techniques to overcome the problem of overfitting and increased the accuracy of our multimodal approach. The experimental results showed that the multimodal technique enhances the accuracy up to 8% by applying the right combination of regularization techniques. Moreover, the overall accuracy increased by 20%.

Funding Statement: The authors are pleased to announce that this research is sponsored by The Superior University, Lahore.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Carbon and M. Serrano, "Wearing face masks strongly confuses counterparts in reading emotions," *Frontiers in Psychology*, vol. 11, no. 566886, pp. 1–8, 2020.

[2] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinex and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[3] M. Sajjad, M. Nasir, K. Muhammad, S. Khan, Z. Jan *et al.,* "Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities," *Future Generation Computer Systems*, vol. 108, pp. 995–1007, 2020.

[4] M. Gori, L. Schiatti and M. B. Amadeo, "Masking emotions: Face masks impair how we read emotions," *Frontiers in Psychology*, vol. 11, no. 1541, pp. 669432, 2021.

[5] N. Mheidly, M. Y. Fares, H. Zalzale and J. Fares, "Effect of face masks on interpersonal communication during the COVID-19 pandemic," *Frontiers in Public Health*, vol. 8, no. 898, pp. 582191, 2020.

[6] M. Grahlow, C. I. Rupp and B. Derntl, "The impact of face masks on emotion recognition performance and perception of threat," *PLoS One*, vol. 17, no. 2, pp. e0262840, 2022.

[7] F. Grundmann, K. Epstude and S. Scheibe, "Face masks reduce emotion-recognition accuracy and perceived closeness," *Plos One*, vol. 16, no. 4, pp. e0249792, 2021.

[8] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, pp. 1–20, 2021.

[9] H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Colorado, USA, pp. 1–9, 2016.

[10] Y. Han, X. Wang and Z. Lu, "Research on facial expression recognition based on multimodal data fusion and neural network," *Arxiv:2109.12724*, 2021.

[11] Al-Waisy, R. Qahwaji, S. Ipson and Al-Fahdawi, "A multimodal deep learning framework using local feature representations for face recognition," *Machine Vision and Applications*, vol. 29, no. 1, pp. 35–54, 2018.

[12] W. Wei, Q. Jia, Y. Feng, G. Chen and M. Chu, "Multi-modal facial expression feature based on deep-neural networks," *Journal on Multimodal User Interfaces*, vol. 14, no. 1, pp. 17–23, 2020.

[13] D. Hamester, P. Barros and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Int. Joint Conf. on Neural Networks (IJCNN)*, Killarney, Ireland, 2015.

[14] W. Sun, X. Chen, X. R. Zhang, G. Z. Dai, P. S. Chang *et al.,* "A multi-feature learning model with enhanced local attention for vehicle re-identification," *Computers, Materials & Continua*, vol. 69, no. 3, pp. 3549–3560, 2021.

[15] B. Yang, W. Jianming, G. Hattori, "Facial Expression Recognition with the advent of human beings all behind face masks," *Association for Computing Machinery*, 2020.

[16] P. Fuller, M. Kathleen and K. Dupuis, "Toronto emotional speech set (TESS)," *Scholars Portal Dataverse*, vol. 1, 2020.

[17] R. Pappagari, T. Wang, J. Villalba, N. Chen and N. Dehak, "X-Vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *45th Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 7169–7173, 2020.

[18] A. M. Badshah, J. Ahmad, N. Rahim and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Int. Conf. on Platform Technology and Service (PlatCon-17)*, Busan, Korea, pp. 1–5, 2017.

[19] Z. Arshad, S. M. Bhatti, H. Tauseef and A. Jaffar, "Heart sound analysis for abnormality detection," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1195–1205, 2022.

[20] L. Qian, L. Hu, L. Zhao, T. Wang and R. Jiang, "Sequence-dropout block for reducing overfitting problem in image classification," *IEEE Access*, vol. 8, pp. 62830–62840, 2020.

[21] L. Chen, M. Li, X. Lai, K. Hirota and W. Pedrycz, "CNN-based broad learning with efficient incremental reconstruction model for facial emotion recognition," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 10236–10241, 2020.

[22] S. H. Gao, Q. Han, D. Li, M. M. Cheng and Pai Peng, "Representative batch normalization with feature calibration," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Tennessee, USA, pp. 8669–8679, 2021.

[23] C. Nwankpa, W. Ijomah, A. Gachagan and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.

[24] H. Sun and R. Grishman, "Lexicalized dependency paths based supervised learning for relation extraction," *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 861–870, 2022.

[25] Aggarwal, Apeksha, A. Srivastava, A. Agarwal, N. Chahal *et al.,* "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, pp. 2378, 2022.

[26] A. Shukla, K. Vougioukas, P. Ma, S. Petridis and M. Pantic, "Visually guided self supervised learning of speech representations," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASS)*, Barcelona, Spain, pp. 6299–6303, 2020.