IEEE Open Journal of Engineering in Medicine and Biology

Technology

Multimodal Emotion Recognition based on Facial Expressions, Speech, and EEG

Jiahui Pan, *Member, IEEE,* Weijie Fang, Zhihang Zhang, Bingzhi Chen*, Zheng Zhang, *Senior Member, IEEE,* Shuihua Wang, *Senior Member, IEEE*

***Abstract*—*Goal:* As an essential human-machine interactive task, emotion recognition has become an emerging area over the decades. Although previous attempts to classify emotions have achieved high performance, several challenges remain open: 1) How to effectively recognize emotions using different modalities remains challenging. 2) Due to the increasing amount of computing power required for deep learning, how to provide real-time detection and improve the robustness of deep neural networks is important. *Method:* In this paper, we propose a deep learning-based multimodal emotion recognition (MER) called Deep-Emotion, which can adaptively integrate the most discriminating features from facial expressions, speech, and electroencephalogram (EEG) to improve the performance of the MER. Specifically, the proposed Deep-Emotion framework consists of three branches, i.e., the facial branch, speech branch, and EEG branch. Correspondingly, the facial branch uses the improved GhostNet neural network proposed in this paper for feature extraction, which effectively alleviates the overfitting phenomenon in the training process and improves the classification accuracy compared with the original GhostNet network. For work on the speech branch, this paper proposes a lightweight fully convolutional neural network (LFCNN) for the efficient extraction of speech emotion features. Regarding the study of EEG branches, we proposed a tree-like LSTM (tLSTM) model capable of fusing multi-stage features for EEG emotion feature extraction. Finally, we adopted the strategy of decision-level fusion to integrate the recognition results of the above three modes, resulting in more comprehensive and accurate performance. *Result and Conclusions:* Extensive experiments on the CK+, EMO-DB, and MAHNOB-HCI datasets have demonstrated the advanced nature of the Deep-Emotion method proposed in this paper, as well as the feasibility and superiority of the MER approach.**

***Index Terms*—Multimodal emotion recognition, Electroencephalogram, Facial expressions, Speech.**

***Impact Statement*—This study is the first attempt to combine the multiple modalities of facial expressions, speech, and EEG for emotion recognition. The accuracy and robustness of the emotion recognition method are improved using three improved deep learning models for each modality and an optimal weight distribution-based decision-level fusion method.**

This work was partially supported by the STI 2030—Major Projects under grant 2022ZD0208900, and the National Natural Science Foundation of China under grants 62076103 and 62271217. (*Corresponding author: Bingzhi Chen*)

J. Pan, W. Fang, Z. Zhang and B. Chen* are with the School of Software, South China Normal University, Guangzhou 510631, China (e-mail: panjiahui@m.scnu.edu.cn, chenbingzhi@m.scnu.edu.cn).

Z. Zhang is with the Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: zhengzhang@hit.edu.cn).

S. Wang is with the School of Computing and Mathematical Sciences, University of Leicester, UK (e-mail: sw546@leicester.ac.uk).

This article has supplementary downloadable material available at https://doi.org/***, provided by the authors

I. Introduction

HUMAN emotions can be understood as people's attitudes, experiences, and corresponding behavioral responses to the objective environment [1]–[3]. Emotions play an essential role in people's daily lives and work [4]. With the rapid development of multimedia and human-computer interaction applications, intelligent machines with emotion recognition have been widely used in medical assistance [5], driving safety [6], and other fields. The definition of emotion can be divided into two paradigms, i.e., discrete paradigm and multi-dimensional paradigm. The discrete paradigms refer to the categories of emotions that people describe in daily life, such as happiness, anger, depression, etc. In contrast, the most commonly used multi-dimensional paradigm is the arousal-valence 2D model proposed by Russell [7], arousal and valence are levels of excitement and positivity, respectively, and this definition method is conducive to our quantitative research on emotions. The ways of expressing emotions can be broadly classified into two categories: external representations, such as facial expressions and speech, and internal representations, such as electroencephalography (EEG) and heart rate [1].

Over the past decade, a majority of emotion recognition studies have been focused on unimodal emotion recognition (UER) using only one mode [8]–[10]. However, emotions are considered a complex representation that cannot be reliably captured with unimodal signals, since genuine emotion can be hidden by different facial expressions or tones [11]. However, even so, facial expressions and speech are still the dominant external channels for conveying emotion. One study [12] showed that these two modalities account for 93% of the emotional information in human communication, and they are critical for multimodal emotion recognition (MER) using external channels [13]. Some recent studies [14]–[16] have attempted to leverage multiple modalities to boost the performance of emotion recognition, which can demonstrate the complementarity of emotion among multiple modes. These fusion strategies still need internal representation modalities, and their reliability still needs to be improved. In a previous study [17], it was proposed that fusing facial expressions, speech, and EEG could be a promising direction for future research in emotion recognition. Inspired by this, on the basis

of MER that integrates facial expressions and speech, this paper introduces EEG, a signal that is not subject to the individual subjective will, to improve the reliability of the emotion recognition method [11].

EEG signals extracted from the central nervous system can more accurately and objectively reflect changes in people's emotions than other signals [2], [11]. As mentioned above, emotions can be expressed from multiple dimensions, and facial expressions and speech, as the most critical external representations of emotions [12], should also be considered. This paper uses three modalities of facial expressions, speech, and EEG to study MER for the first time. Unlike previous MER methods, the proposed method considers the three most relevant external and internal representations of emotions, which has better accuracy and reliability.

For multimodal fusion, fusion methods can be divided into feature-level fusion and decision-level fusion [18]. We found that in previous studies [19]–[21], decision-level fusion methods are not only easy to implement but also exhibit better performance than feature-level fusion. For example, the winning methods of the EmotiW challenge were almost decision-level fusion [14]. However, the increase in patterns in decision-level fusion means that multiple models must be designed, leading to inefficient multi-pattern recognition algorithms that are difficult to port to mobile devices and provide real-time detection in daily use [22]. This implies that we need to pay attention to the model's size when designing sentiment recognition models rather than just striving for accuracy.

With the improvement of chip computing processing power and deep learning performance, many novel emotion recognition methods have emerged in recent years. Some mainstream neural network models have achieved good results in emotion recognition, such as CNN [4], LSTM [8], [20], DBN [23], and GCN [24]. These deep learning methods have gradually replaced traditional feature extraction methods as the primary research methods for emotion recognition. The proposed Deep-emotion recognition framework in this paper utilizes three deep learning models to extract emotional features from facial expressions, speech, and EEG, respectively. The decision-level fusion method is then applied to integrate the recognition results from each modality, resulting in a more comprehensive and accurate recognition rate. In addition, to prevent the final model from being too large due to excessive classification models, we reduced the number of model parameters as much as possible on the premise of ensuring the classification accuracy of each model. Our contributions in this paper can be summarized as follows:

- This study is the first attempt to combine the multiple modalities of facial expressions, speech, and EEG for emotion recognition. In the decision-level fusion stage, we propose an optimal weight distribution algorithm. Compared with traditional equal-weight fusion, this method can better judge the reliability of each mode and thus effectively enhance the fusion performance.
- In this paper, a carefully improved GhostNet [25] structure is proposed for facial expressions recognition (FER). This method can effectively alleviate the overfitting phenomenon of the original GhostNet in the training process, and effectively improve classification accuracy.
- For speech emotion recognition (SER), we design a lightweight full convolutional neural network (LFCNN), which has good feature learning performance with only a few parameters. Reducing model parameters as much as possible is also a factor to be considered in the model design process for decision-level fusion requiring multiple classifiers.
- In the work of EEG emotion recognition (EER), this paper designs a tree-like LSTM (tLSTM) model that can fuse multi-stage features. This model combines shallow and deep features in the feature extraction process and performs better.

The rest of this paper is organized as follows: Section II describes the proposed emotion recognition methods and related experiments. Next, the comprehensive experimental results and discussion are reported in Section III. Finally, Section IV presents the conclusion of this work.

II. Materials And Methods

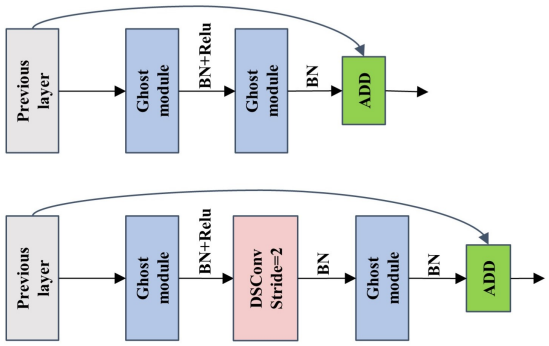
A. Data preprocessing

The method of data preprocessing can be found in the Supplementary Materials of this manuscript.

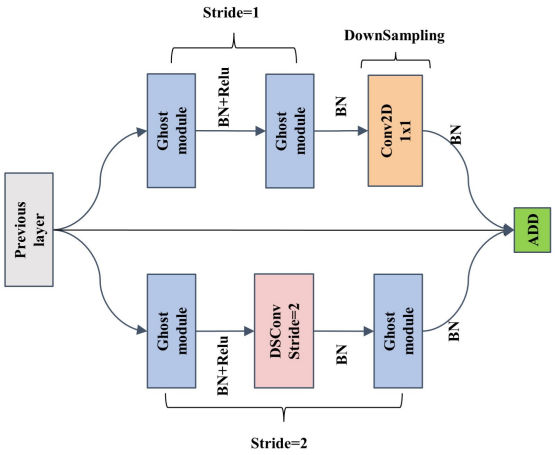
B. Deep learning model

1) The improved GhostNet for FER: GhostNet is mainly composed of multi-layer Ghost bottlenecks, among which Ghost bottlenecks are mainly composed of the Ghost module. The structures of the Ghost module and Ghost bottleneck are shown in Fig. 2 and Fig. 1(a), respectively. Our work focused on improving the Ghost bottleneck architecture, as shown in Fig. 1(b). The original Ghost bottleneck is divided into stride=1 and stride=2 modes, which perform feature extraction from different scales to obtain different feature map sizes. However, different sizes of feature maps have certain reference value for subsequent feature extraction. Inspired by this, the Ghost bottleneck proposed in this paper combines the characteristics of these two modes to provide more comprehensive characteristics. The specific implementation method introduces a 1×1 convolution for downsampling in the case of the original stride=1 to obtain the same shape when the stride=2. When the input shape is $48 \times 48 \times 1$, the specific structure of the improved GhostNet is shown in Table I.

2) Architecture of LFCNN for SER: The overall structure of our proposed LFCNN is shown in Fig. 3, which is mainly composed of three parts: parallel convolution structure, residual structure, and serial convolution structure. Depthwise separable convolution (DSC) has been found in past research to have a smaller number of parameters than traditional convolution [26]. The success of Xception proves the superiority of DSC over traditional convolution, and we will use it to design the LFCNN. The convolutional layers mentioned later in this

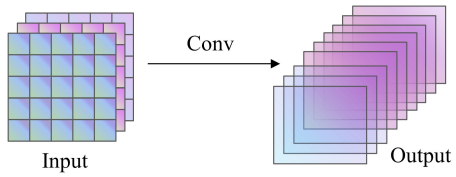


(a) The original Ghost bottleneck. Top: when stride=1; Bottom: when stride=2.

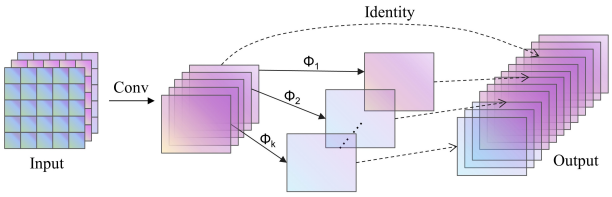


(b) Our improved Ghost bottleneck.

Fig. 1. Structure of Ghost bottleneck.



(a) The traditional convolution layer.



(b) The Ghost module.

Fig. 2. Structure of traditional convolution and Ghost module.

section are DSC. Further details on the structure of LFCNN can be found in the Supplementary Materials accompanying this manuscript.

3) Architecture of tLSTM for EER: Our proposed tLSTM structure is shown in Fig. 4. For the tree part, the LSTMs of the leaf nodes all have the same number of neurons to ensure

TABLE I. Structure of the proposed improved GhostNet model for FER. EXP: expansion size. OUT: the number of output channels. SE: whether using the SE module.

Operator	Output	Ghost bottleneck Setting		
		$\cdot EXP$	$\cdot OUT$	$\cdot SE$
Conv2d, 16, 3×3	(batch, 24, 24, 16)	-	-	-
Ghost_bottleneck	(batch, 12, 12, 40)	120	40	True
Dropout, 0.3	(batch, 12, 12, 40)	-	-	-
Ghost_bottleneck	(batch, 6, 6, 80)	240	80	False
Dropout, 0.3	(batch, 6, 6, 80)	-	-	-
Ghost_bottleneck	(batch, 3, 3, 160)	672	160	True
Dropout, 0.3	(batch, 3, 3, 160)	-	-	-
Ghost_bottleneck	(batch, 2, 2, 160)	960	160	False
Dropout, 0.3	(batch, 2, 2, 160)	-	-	-
Conv2d, 256, 1×1	(batch, 2, 2, 256)	-	-	-
Dropout, 0.3	(batch, 2, 2, 256)	-	-	-
GAVPool, Reshape	(batch, 1, 1, 256)	-	-	-
Conv2d, 512, 1×1	(batch, 1, 1, 512)	-	-	-
Dense, Softmax	(batch, 7)	-	-	-

that their output shapes are consistent since their outputs will be merged and fed into the sequence part. Our proposed tree structure consists of four levels, each representing a stage of features. The leaf nodes are located at different levels, which are used to fuse the features of each stage together to obtain a more comprehensive feature. It is worth noting that the output of the LSTM we use in the tree part is the output of the entire sequence, while the output of the LSTM in the sequence part is the last hidden layer. Finally, the network outputs the arousal and valence scores separately through the dense layer.

C. Proposed decision-level fusion strategy

To find the reliability of each mode, we developed an optimal weight distribution algorithm. Taking arousal score decision fusion as an example, assume that there are n modes corresponding to n regression models and that a total of T trials are used for prediction. The predicted average arousal score for trial t in the kth model is A_{tk}, $k \in \{1,2,3\ldots,n\}, t \in \{1,2,3\ldots,T\}$. Let the weight set ϖ to be $\{0.00, 0.01, 0.02\ldots, 0.98, 0.99, 1.00\}$, an array that starts at 0.00 and ends at 1.00 with a step size of 0.01. The root mean square error (RMSE) is used as a measure to evaluate the performance of the current weight distribution. When it is in the best performing weight distribution, RMSE should be the smallest, denoted as RS_{min}. In accordance with relevant provisions of the above, the optimal weight distribution algorithm steps are as follows:

Step 1: The weights of n modes are enumerated in ϖ. Let the weight of the kth mode be ω_k; then, go to ***Step 2*** when (1) is satisfied. The algorithm ends when the enumeration is finished and the optimal weight distribution is saved.

$$\sum_{k=1}^{n} \omega_k = 1 \tag{1}$$

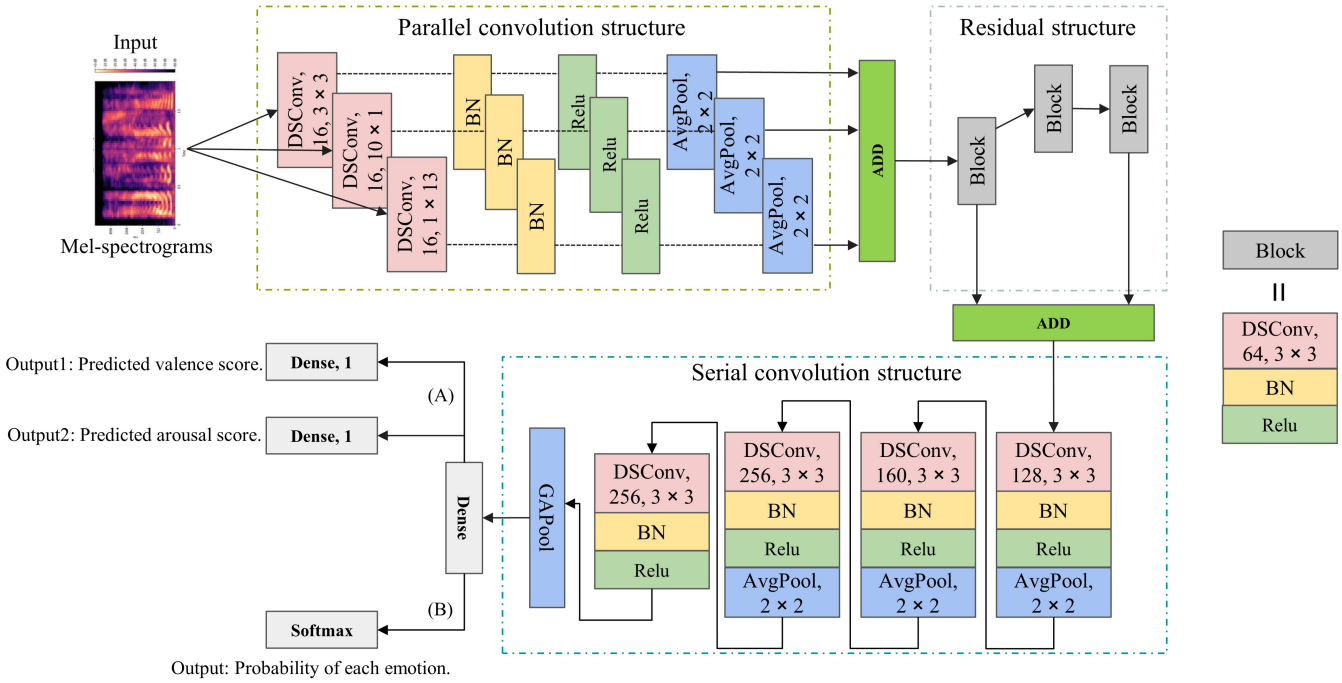


Fig. 3. Structure of the proposed LFCNN model for SER. (A): When the task is expressed with a discrete paradigm. (B): When the task is expressed with a arousal-valence 2D model.

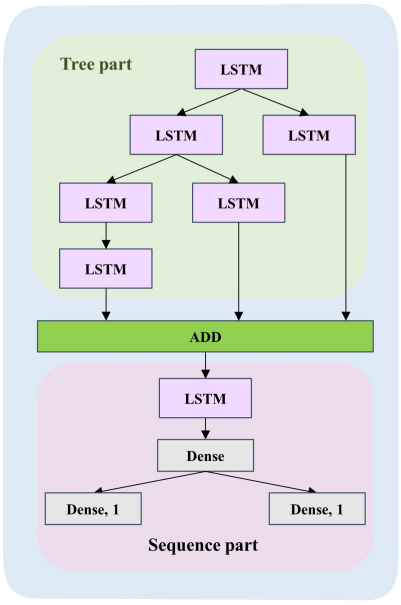


Fig. 4. Structure of the proposed tLSTM model for EER.

Step 2: Calculate the predicted arousal score under the current weight distribution. Assuming that the predicted arousal score of trial t is $\hat{y}_t$, then the calculation formula can be expressed as:

$$\hat{y}_t = \sum_{k=1}^{n} \omega_k A_{tk} \tag{2}$$

Step 3: The RMSE of T trials under the current weight distribution, denoted as RS_{cut}, is calculated as (3), where y_t is the actual arousal score of trial t. By comparing the size relationship between RS_{cut} and RS_{min}, when $RS_{cut} < RS_{min}$, the current weight distribution is considered to have better performance. Thus, RS_{min} is updated to RS_{cut}, and the current weight distribution is saved. When $RS_{cut} \geq RS_{min}$, it is considered that the current weight distribution does not exhibit better performance. Regardless of the size relationship, ***Step 1*** is performed again.

$$RS_{cut} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2} \tag{3}$$

To provide a clear illustration of the algorithm's implementation, the flow chart of the algorithm execution as well as the pseudo-code for the case of fusion of three modes is presented in the Supplementary Materials of this manuscript.

D. Experiment

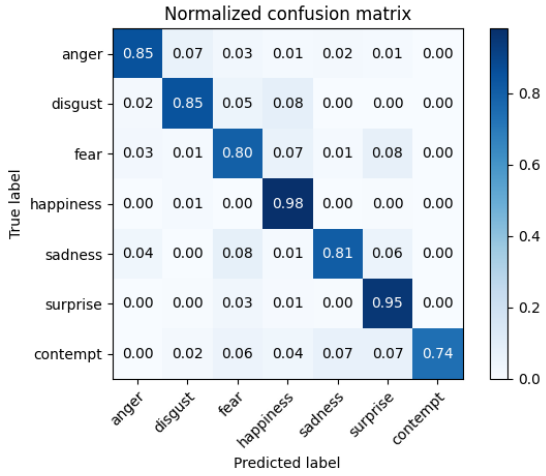
The setup and various details of the experiment can be found in the Supplementary Materials of this manuscript.

III. Results and Discussion

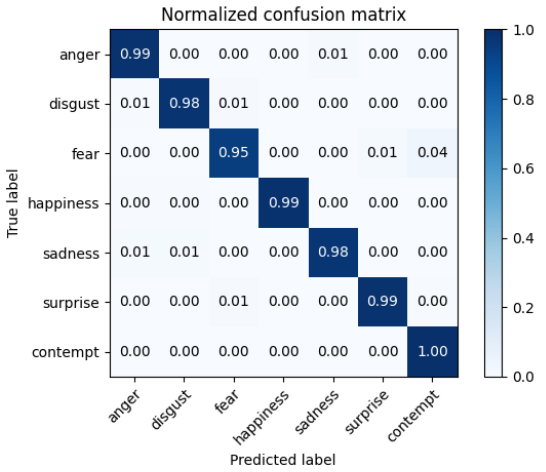
A. Results on CK+

Tenfold cross-validation was conducted on CK+ using both GhostNet and our improved GhostNet. The confusion matrix is shown in Fig. 5(a) and Fig. 5(b). The results show that our improved GhostNet achieved an average accuracy of 98.27%, outperforming the traditional GhostNet (90.21%).

Furthermore, we found that the overfitting phenomenon of the GhostNet model appears in the training process. In this regard, we introduced well-designed multiple dropout layers while modifying the Ghost bottleneck to alleviate the overfitting phenomenon. The curves of the accuracy and loss obtained during training of the GhostNet model on CK+ before and after improvement, as a function of epoch, can be found in the Supplementary Materials of this manuscript. The improved GhostNet achieved an average accuracy of 98.27%, but the accuracy of fear expression was only 95%, which may be because fear and contempt have similar features. Nevertheless, our proposed method has achieved advanced results in recent research. Table II shows the comparison with some recent studies. It can be seen from the table that the improved GhostNet proposed by us performs better than other classical classification models, which fully proves the superiority of our proposed method.



(a) GhostNet.



(b) Our improved GhostNet.

Fig. 5. The confusion matrix obtained by Ten-fold cross-validation on the CK+ dataset.

TABLE II. Comparison with recent studies on CK+ datasets. Val:Validation method, Acc: Accuracy.

Literatures	Model	Val	Acc(%)
Nasri et al. 2020 [27]	Xception	10-fold	98.20
Chowdary et al. 2021 [28]	Vgg19	-	96.00
	Inception-v3	-	94.20
Priya et al. 2021 [29]	MobileNet	10-fold	96.00
Mishra et al. 2022 [30]	ResNet50	5-fold	89.80
Shaik et al. 2022 [31]	CNN-Attention	10-fold	97.67
Ours	GhostNet	10-fold	90.21
	Imporved GhostNet	10-fold	**98.27**

TABLE III. Comparison with recent studies on EMO-DB datasets. Val: Validation method, Acc: Accuracy.

Literatures	Val	Acc(%)	Model size(MB)
Chen et al. 2018 [32]	10-fold	82.82	323.46
Sajjad et al. 2020 [8]	5-fold	85.57	128.00
Muppidi et al. 2021 [33]	5-fold	88.70	31.20
Kwon et al. 2021 [34]	5-fold	93.00	14.40
Andayani et al. 2022 [35]	10-fold	85.55	-
Ours	10-fold	**94.36**	**2.28**

B. Results on EMO-DB

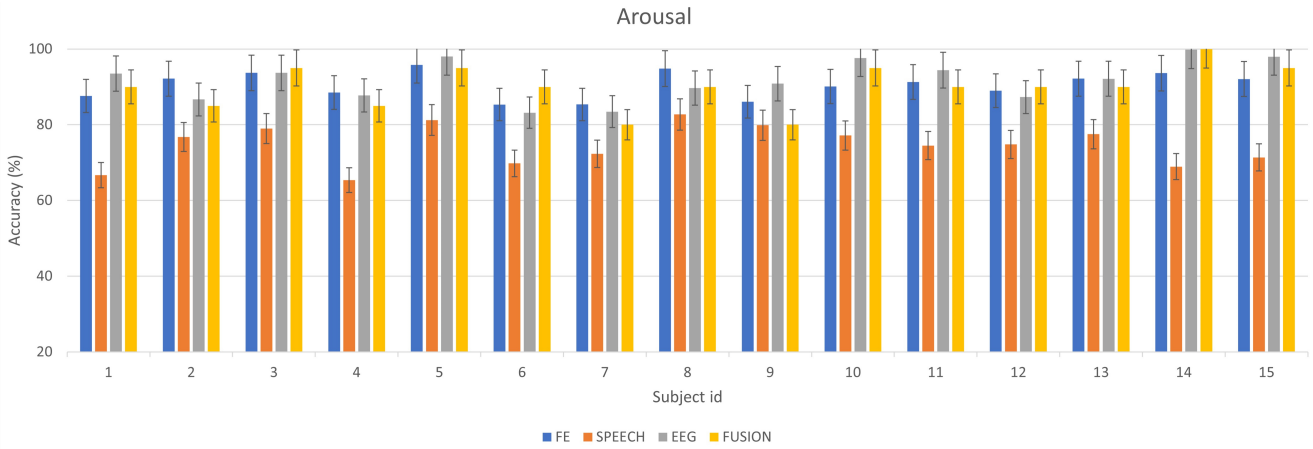
For experiments on EMO-DB, we achieve an average accuracy of 94.36% with an F1-scores of 94.38%, which almost surpasses most recent studies. The size of our proposed model is only 2.28 MB, which is much smaller than other models and is more likely to be applied to future mobile devices. The number of parameters for each component of the LFCNN structure can be found in the Supplementary Materials accompanying the manuscript. Table III shows the comparison between our work and previous work. The table shows the prediction accuracy and the size of the model. According to the comparison table data, we can see that the superiority of our method was thus validated.

C. Results on MAHNOB-HCI

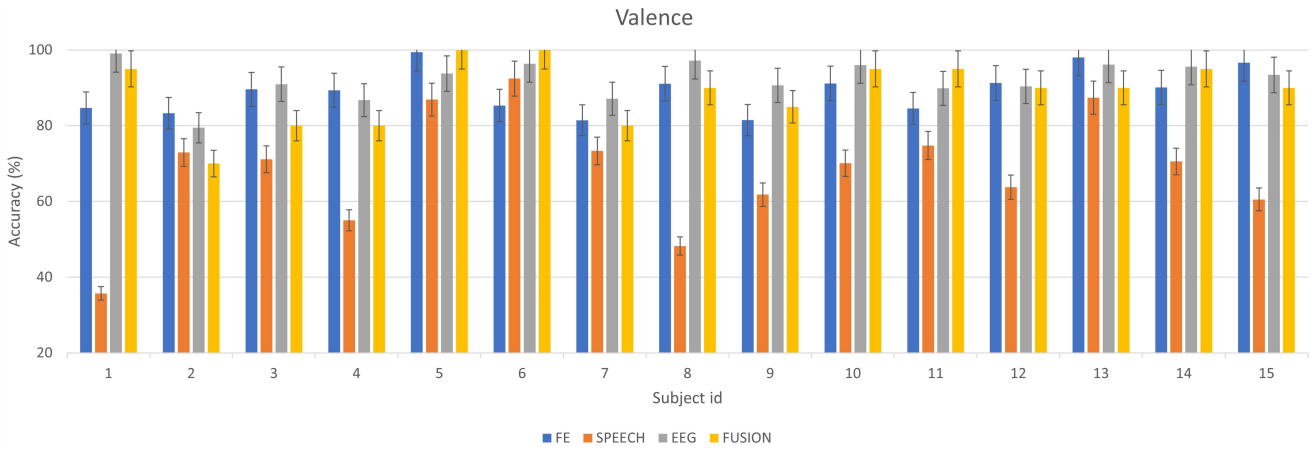
The experimental results obtained in the MAHNOB-HCI dataset show that the tLSTM model has obtained advanced results in EER and the feasibility of the decision-level fusion method. Fig. 6 shows the verification results of subjects 1 to 15. Table IV shows the average RMSE of some subjects after fusion. We found that the fusion method of Deep-Emotion proposed in this paper has improved performance in both arousal and valence dimensions. However, the fusion method is only significantly higher than that of the speech mode ($p < 0.05$, paired t test). We did not find that the fusion method outperformed the facial expressions or EEG mode ($p > 0.05$, paired t test). In emotion recognition experiments, facial expressions are often associated with high volatility since

TABLE IV. The average RMSE report of some subjects under our proposed decision-level fusion method.

Subject Id			1	4	9	15	18	Average
Arousal	Scale	[1-9]	0.6235 ± 0.2153	0.7412 ± 0.3162	0.8526 ± 0.2431	0.7812 ± 0.2162	0.8512 ± 0.1425	0.7628 ± 0.2913
		[0-1]	0.0642 ± 0.0246	0.0758 ± 0.0186	0.0813 ± 0.0283	0.0784 ± 0.0213	0.0902 ± 0.0182	0.0813 ± 0.0241
Valence	Scale	[1-9]	0.9842 ± 0.3120	1.0143 ± 0.3124	0.8975 ± 0.1962	1.1321 ± 0.2548	0.8458 ± 0.3465	0.9450 ± 0.2627
		[0-1]	0.1031 ± 0.0326	0.0968 ± 0.0274	0.0862 ± 0.0164	0.1063 ± 0.0205	0.0872 ± 0.0242	0.0952 ± 0.0176



(a) Arousal dimension.



(b) Valence dimension.

Fig. 6. The accuracy of UER and MER for each subject in the MAHNOB-HCI dataset. The horizontal axis represents the subject ID, and the vertical axis represents the accuracy rate (%). FE: facial expressions. (a) Arousal dimension accuracy. (b) Valence dimension accuracy.

subjects may deceive the machine by mimicking certain facial expressions. In this case, the gap between the error associated with facial expressions and the error of accurate emotion detection can be filled by adding information sources (e.g., EEG and speech). Furthermore, in the MAHNOB-HCI dataset, the subjects were asked to behave normally rather than mimic certain facial expressions, which may be the main reason we could not find solid statistical evidence indicating significant improvement after fusion. For example, in this study, the FER accuracies of subjects 2 and 13 are higher than the result of the fusion method because the fusion result combines multiple modes for comprehensive consideration.

In addition to having a relatively high accuracy rate, DeepEmotion can also show relatively good robustness. For example, when subjects express facial expressions that are different from their real emotions, the results obtained by MER fusion will not deviate significantly from the real ones. This is because the subjects' EEG still represents their real emotion. A comparison of this study with other research on emotion recognition in the MAHNOB-HCI dataset can be found in the Supplementary Materials. Moreover, SER is a very challenging task in MAHNOB-HCI. This is because the speech signal in this dataset includes not only the voice of the subject but also the voice of the stimulus material, which makes it challenging

to extract the voice of the subject. This may be the reason for our relatively average recognition rate in SER.

Currently, our work is primarily based on the analysis of open source datasets, and we have not conducted independent data collection to further verify our findings. In future research, we plan to design a standardized experimental paradigm to collect additional data from subjects to more thoroughly evaluate the capabilities of Deep-Emotion.

IV. Conclusion

In our work, we propose a new MER method named Deep-Emotion, based on deep learning techniques to develop emotion recognition models for facial expressions, speech, and EEG. An improved GhostNet is proposed for facial expressions, which effectively alleviates the overfitting phenomenon and dramatically improves the model's performance. An LFCNN model is developed for speech signals, which can greatly reduce the model size on the premise of ensuring recognition accuracy. For EEG signals, a tLSTM model that can better learn the emotional characteristics of each stage was designed. Furthermore, we designed an optimal weight distribution search algorithm to find the reliability of each mode and achieve decision-level fusion. Our proposed methods are tested with open source datasets in MER experiments. To the best of our knowledge, this study is the first attempt to combine facial expressions, speech, and EEG for MER. The experimental results obtained in multiple public datasets validate the feasibility of the proposed method. In future work, we can further improve the fusion method that can dynamically assign weight to each mode to enhance the overall robustness of the algorithm. Dynamic weight allocation is a method of assigning different weights to different modalities, or sources of information, in a multimodal learning system. This allows the model to assign higher weights to the modalities that contain more relevant and useful information and lower weights to the modalities that contain less relevant or noisy information. This can improve the performance of the model by focusing on the most useful information and filtering out the noise. Overall, we believe that dynamic weight allocation is a promising approach for multimodal learning, and we will explore its potential in future research.

Supplementary Materials

The Supplementary Materials of this manuscript include relate work on emotion recognition. The data preprocessing is described, and the full structure of LFCNN is detailed. The flow chart of the algorithm execution as well as the pseudo-code of the optimal weight distribution algorithm is also provided. The experimental process of this study is described in detail. Finally, supplementary explanations of the experimental results are provided. This document can be accessed in the "Media" section of IEEE Xplore.

References

[1] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 02, pp. 52–58, 2021.

[2] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.

[3] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, and D. Zhang, "Multimodal emotion recognition with temporal and semantic consistency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3592–3603, 2021.

[4] M. Wu, W. Su, L. Chen, W. Pedrycz, and K. Hirota, "Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 805–817, 2022.

[5] W. Meng, Y. Cai, L. T. Yang, and W.-Y. Chiu, "Hybrid emotion-aware monitoring system based on brainwaves for internet of medical things," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 16014–16022, 2021.

[6] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, A. El-Sayed, and A. Elleithy, "A hybrid model for driver emotion detection using feature fusion approach," *International journal of environmental research and public health*, vol. 19, no. 5, p. 3085, 2022.

[7] J. A. Russell, "Affective space is bipolar.," *Journal of personality and social psychology*, vol. 37, no. 3, p. 345, 1979.

[8] M. Sajjad, S. Kwon, *et al.*, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.

[9] A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, "LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6912–6916, 2022.

[10] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn," *Speech Communication*, vol. 120, pp. 11–19, 2020.

[11] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen, "EEG based emotion recognition: A tutorial and review," *ACM Computing Surveys (CSUR)*, 2022.

[12] A. Mehrabian, "Communication without words," in *Communication theory*, pp. 193–200, Routledge, 2017.

[13] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, p. 108580, 2022.

[14] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2617–2629, 2021.

[15] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (avef): A deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2019.

[16] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and eeg for multimodal emotion recognition," *Computational intelligence and neuroscience*, vol. 2017, 2017.

[17] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, P. D. Barua, M. Murugappan, Y. Chakole, and U. R. Acharya, "Automated emotion recognition: Current trends and future perspectives," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106646, 2022.

[18] Z. He, Z. Li, F. Yang, L. Wang, J. Li, C. Zhou, and J. Pan, "Advances in multimodal emotion recognition based on brain–computer interfaces," *Brain sciences*, vol. 10, no. 10, p. 687, 2020.

[19] Y. Fang, R. Rong, and J. Huang, "Hierarchical fusion of visual and physiological signals for emotion recognition," *Multidimensional Systems and Signal Processing*, vol. 32, no. 4, pp. 1103–1121, 2021.

[20] R. Li, Y. Liang, X. Liu, B. Wang, W. Huang, Z. Cai, Y. Ye, L. Qiu, and J. Pan, "Mindlink-eumpy: An open-source python toolbox for multimodal emotion recognition," *Frontiers in Human Neuroscience*, vol. 15, p. 44, 2021.

[21] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomedical Signal Processing and Control*, vol. 70, p. 103029, 2021.

[22] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1034–1047, 2022.

[23] D. Liu, L. Chen, Z. Wang, and G. Diao, "Speech expression multimodal emotion recognition based on deep belief network," *Journal of Grid Computing*, vol. 19, no. 2, pp. 1–13, 2021.

[24] J. Li, S. Li, J. Pan, and F. Wang, "Cross-subject EEG emotion recognition with self-organized graph neural network," *Frontiers in Neuroscience*, p. 689, 2021.

[25] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1580–1589, 2020.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[27] M. Nasri, M. A. Hmani, A. Mtibaa, D. Petrovska-Delacretaz, M. B. Slima, and A. B. Hamida, "Face emotion recognition from static image based on convolution neural networks," in *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 1–6, 2020.

[28] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Computing and Applications*, pp. 1–18, 2021.

[29] R. B. Priya, M. Hanmandlu, and S. Vasikarla, "Emotion recognition using deep learning," in *2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–5, 2021.

[30] S. Mishra, B. Joshi, R. Paudyal, D. Chaulagain, and S. Shakya, "Deep residual learning for facial emotion recognition," in *Mobile Computing and Sustainable Informatics*, pp. 301–313, 2022.

[31] N. S. Shaik and T. K. Cherukuri, "Visual attention based composite dense neural network for facial expression recognition," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–14, 2022.

[32] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[33] A. Muppidi and M. Radfar, "Speech emotion recognition using quaternion convolutional neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6309–6313, 2021.

[34] S. Kwon *et al.*, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, p. 107101, 2021.

[35] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-Transformer model for emotion recognition from speech audio files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022.