Received: May 17, 2023. Revised: July 24, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

MEFNet-Micro Expression Fusion Network Based on Micro-Attention Mechanism and 3D-CNN Fusion Algorithms

Gomathi R[1] Logeswari S[2] Jothimani S[1]* Sangeethaa SN[1] Arul Sangeetha S[3] LathaJothi V[4]

[1]*Bannari Amman Institute of Technology, Sathyamangalam, India*
[2]*Karpagam College of Engineering, Coimbatore, India*
[3]*Rathinam Technical Campus, Coimbatore, India*
[4]*Velalar College of Engineering and Technology, Erode, India*
* Corresponding author's Email: jothimanis.phd@gmail.com

Abstract: Due to its impartiality in emotion identification, micro-expression can be used in emotional computing. Deep learning has proved successful at recognizing nuanced facial emotions. When the stakes are high, micro-expressions (MEs) reveal people's true sentiments. Early micro-expression recognition (MER) methods use traditional traits. MEs are delicate, rapid facial movements, making them harder to measure and annotate than macro-expressions. Recent deep learning (DL) methodologies aim to improve ME and MER performance. Micro-expression in small, localized areas of the face and a lack of large databases now hinder emotional facial movement recognition. To meet these issues, this study suggests a unique attention mechanism dubbed micro-attention that operates in tandem with residual networks called micro expression fusion network (MEFNet). Enhance the ME model by incorporating attention mechanisms to focus on the most informative regions of the face. This can help improve the accuracy of spatial and temporal data extraction, especially during subtle micro-expressions. Extend the two-stream CNN model to incorporate additional modalities, such as audio or textual cues, in addition to the eye and mouth regions. This multi-modal fusion will enable a more comprehensive understanding of emotions and increase the system's robustness to variations in facial expressions. The MEFNET achieved a specificity of 99.2%, sensitivity of 99.5%, and accuracy of 100% on the CAS(ME)2 dataset, and a specificity of 99.6%, sensitivity of 99.3%, and accuracy of 100% on the SMIC dataset. Rendering to the experimental results, the suggested framework compares favourably to the state-of-the-art methods.

Keywords: Attention, Deep Learning, Facial expression, Fusion, Micro-expression, CNN.

1. Introduction

Microexpression recognition is popular for good reason. Micro-expressions—short, involuntary facial gestures—can reveal a person's true feelings, even when they're trying to hide them. These looks last only a few milliseconds and are too delicate to see. Technology and deep learning algorithms can now detect and evaluate these expressions in real-time. Microexpressions, or fleeting facial expressions, reveal a person's true feelings. Microfacial expressions, which last half a second to four seconds [1], are the most common and effective way to express emotions. Meanwhile, much research has gone into teaching computers to identify human emotions by analyzing macro expressions ([2, 3]).

Psychologists [4, 5] have shown that humans aren't always good at reading emotions from appearances. Micro-expression, unlike macro-expression, is subconscious, revealing true emotion. Due to its impartiality, micro-expression recognition is used in affect monitoring [4], criminal detection [6], and home safety [7]. Deep learning has transformed AI research. Its ability to automatically learn and extract characteristics from photos and videos makes it useful for micro-expression identification. Deep learning algorithms can identify

emotional cues in facial photos and videos due to their massive data processing. It impacts psychology, policing, and HCI.

Deep learning for microexpression identification creates real-time microexpression recognition and categorization systems. We require deep learning architectures to extract fine-grained properties from huge face expression datasets. Deep learning microexpression recognition can improve human-robot interaction, VR experiences, fraud detection, and emotion detection. Attention was added to the micro-expression-recognizing deep CNN. This method can also extract certain face features.

Focusing on micro-expressions helps learning and acquisition. Deep learning solves most computer vision problems better than hand-crafted ones. Recently published CNN-based micro-expression recognition algorithms [22–25]. Face micro-expression recognition often uses CNNs, RNNs, and hybrids. These methods employ CNNs to extract spatial parameters from each expression video clip and feed them into recurrent neural networks (RNNs) to capture the temporal link between frames. The techniques cannot encode video information's spatial and temporal connection. We present two 3D convolutional neural network (CNN) models that simultaneously extract spatial and temporal information from a video to improve current approaches.

This paper's primary contributions are summed up as follows:

1. We present a spatial and temporal MECNN model to categorize facial expression movies. We achieved state-of-the-art performance on benchmark micro-expression datasets using the specified MECNN model. A two-stream MFCNN model merges eye and lip traits.

2. STAM, a modest yet powerful CAM module, adjusts convolutional feature and activation settings.

3. 3D CNNs trained on eye and mouth regions are evaluated in intermediate and late fusion settings. Salience maps can analyze facial attributes. Different 3D kernel sizes were investigated for micro-expression recognition.

2. Related works

Microexpressions indicate emotions [39-41]. Neurology, criminology, and HCI employ microexpression identification. Deep learning increases facial emotion recognition. Recent deep learning microexpression detection studies are reviewed here. Researchers identified micro-expressions using public video data [8]. "Microexpression detection" finds a movie's beginning, peak, and end. Micro-expression movies must be labeled. This study addresses real-time microexpression recognition. Distinguishing micro-expression is hard. Broad expression feature generation allows micro-expression recognition.

CNNs [9] and GPUs advanced computer vision. After huge dataset training, deep learning model features exceeded baseline procedures. FER and MER have various CNN models. Sangeethaa et al. [11] promptly designed the effective visual geometry group (VGG) architecture [10] to overcome the FER issue with strong supervision in each layer. The authors in [13] improved MER accuracy using picture classification attention. CNN computed optical fluxes from the start and end frames (Khor et al., [27] and Liu, [12]). ME database shortage inhibits MER research. Pre-processing photos to identify ME characteristics may solve this issue. DNNs identify. Wang et al. [14] presented transfer-learning-based transferring long-term CNN model. Takalkar et al. [15] created a framework to merge deep CNN and handcrafted characteristics. The aforementioned image classification algorithms perform well with a few RGB pixels. This approach saves processing but ignores video motion and temporal information. Zhao et al. [16] produced a keyframe sequence. 3D-CNN calculates optical flow using keyframe sequences.

3D-CNN retrieved spatiotemporal properties for action recognition [17-19]. Haddad et al. found FER 3D convolution promising [20]. 3D-convoluted MER. [21]. 3D-CNN's biggest issue is computation expense. 3D convolution duplicate parameters overfit tiny datasets. Reddy et al. [22] built a shallow, strong 3D-CNN. Another model's cut-off facial areas. Full-face models exceed it. Research suggests hand-designed features are less precise and lasting. Accuracy makes deep learning approaches attractive. 3D CNN recognizes micro-expressions using optical flow data. This paper presents two integrated spatiotemporal training-based 3DCNN techniques. MFCNN implies the eyes and mouth, whereas MECNN suggests the face. The authors [22-31] used various deep learning techniques including transfer learning, and CNN. Micro-expressions are delicate and ephemeral. Advanced approaches are needed to identify and categorize micro-expressions due to their short duration and subtle motions. Hand-designed facial expression representations are less accurate and robust than deep learning-based approaches.

Received: May 17, 2023. Revised: July 24, 2023.

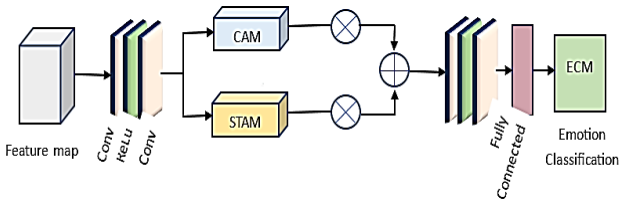


Figure. 1 MECNN model structure

3. Methods and materials

Inspired by deep networks' ability to extract spatio-temporal characteristics for microexpression identification, we present two 3D-CNN models, MEFNET. MECNN considers the whole face, while MFCNN concentrates on the eyes and mouth. Both suggested models use the 3DCNN to combine spatial and temporal context. 3D-CNNs with careful layer and filter size selection are presented for face micro-expression recognition. The proposed MECNN and MFCNN use 3D-CNN to identify face micro-expressions.

MECNN:

CAM, STAM, and ECM are the channel attention module, spatial-temporal attention module, and emotion classification module, respectively. Each convolutional block utilizes a CAM layer and a STAM layer following the convolutional layers. They improve feature-map-based inter-channel and intra-channel focus. To further improve categorization activities, ECM is implemented after completely linked layers. Channel attention module (CAM) and spatial-temporal attention module (STAM) are the two main components of our proposed model's overall design, as shown in Fig. 1. While ECM is placed after fully connected layers, CAM and STAM are placed after convolutional layers in convolutional blocks.

In order to quantify attention value, CAM exploits the connections between channels of convolutional maps. Some of the channels in the convolutional layer may be more concerned with discriminative characteristics, while others may be less so. Those global clues in forward propagation may be overlooked by common deep networks because they treat all channels the same. To make up for this, CAM was developed in an effort to clearly represent the channel interactions. As a result, CAM allows deep networks to capture cross-channel interactions and identify "what" is crucial early on.

The MECNN model includes activation functions, dropouts, fully-connected layers, 3D convolutional and pooling layers. 3D convolutional layers can extract spatial and temporal properties from 3D kernel convolutions. Unlike the 2D CNN, the 3D-CNN uses both spatial and temporal filters. The 3D pooling layer gradually downscales the 3D convolutional layer's output while preserving key characteristics. 3D pooling picks the most informative feature representation across a brief time and space frame. Dropout reduces model overfitting. Dropout improves regularization in the recommended network. The flatten layer is input dimension lengthening before the completely linked layer. Hierarchical feature extraction needs dense or fully-connected layers. The softmax layer grades dataset segments.

The proposed network we present here consists of two fully-connected (dense) layers, a stacking layer, a 3D convolutional layer with 32 filters of size $3 \times 3 \times 15$, and a 3D pooling layer with a kernel size of $3 \times 3 \times 3$. The number of expression labels in a dataset determines the ultimate size of the fully connected layer. The convolutional block takes convolutional feature maps N as input, where N is in the form of $R^{H\times W\times C}$ for 2D networks or $R^{T\times H\times W\times C}$ for 3D networks. Through the use of channel attention map $A_C(N)$ and spatial-temporal attention map $A_s(N)$, CAM and STAM respectively obtain their respective attention maps. The performance of the STAM is observed to vary across the dimensions of the network input.

$$N_C = A_C(N) \otimes N \tag{1}$$

$$N_s = A_s(N) \otimes N \tag{2}$$

$$X = \text{concat}(N_C, N_s) \tag{3}$$

Here, $\otimes$ represents the element-wise multiplication. The channel attention map is denoted by $A_C(N) \in R^{C\times 1\times 1}$ and the spatial-temporal attention is signified by $A_s(N)) \in R^{1\times H\times W}$. Increasing the convolutional operation from 2 dimensions to 3 dimensions results in unmanageable parameter expansion and training expense, hence we present a (2+1) D convolutional layer for 3D STAM in 3D convolutional networks. Three-dimensional convolution has been decomposed into two-dimensional and one-dimensional convolutions in an effort to simplify the convolutional operation and minimize the number of parameters required. By employing a (2+1) D convolutional layer in our approach, we are able to bring the channel-wise mixed pooling layer into the third dimension. In addition, the dropout layer is only used during training since a consistent classification outcome is required for testing.

MFCNN

In the MFCNN model, two independent 3D

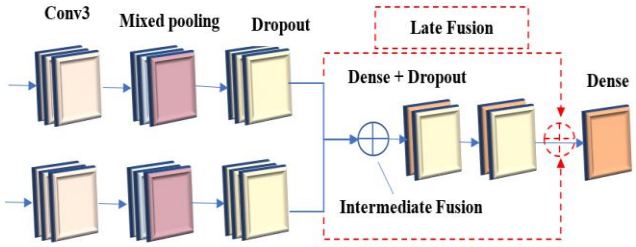


Figure 2. MFCNN model

spatiotemporal CNNs are fed data from only areas around the eyes and mouth. Later, the two CNNs are combined to form a single network. Following the identification of 68 facial landmarks using the DLib face detector, we do the preliminary processing on each frame of the expression video to determine the corresponding eye and mouth areas. The areas around the eyes and lips can be cropped using these points of reference. We present two variants of MFCNN models, which we call Intermediate MFCNN and Late MFCNN, founded on the distinct fusion procedures.

As its name implies, the Intermediate MFCNN Model combines the characteristics of two 3D-CNN as shown in the figure 2. One 3D-CNN receives input from the upper face (containing the eyes) and the lower face (including the lips). At this juncture, we combine the extracted characteristics from the eye and mouth areas. Comparable to the MECNN model, the proposed MFCNN employs two independent 3D-CNN, each of which is composed of stacking, one 3D convolutional layer with 32 filters of size 3×3 ×15, and one 3D pooling layer with a kernel size of 3 × 3 ×3. Activation maps may be flattened down into feature vectors using a flatten layer. Combining the two networks' normalized features results in a third vector. Class scores are generated using a combination of dense and dropout layers applied to the fused features in Intermediate MFCNN.

As its name implies, the characteristics of two 3DCNN are combined in late MFCNN model right before the last thick layer. In this model, one 3D convolutional neural network (CNN) receives input from the eye region of the face, while the second 3D CNN receives input from the mouth region. At the final fully connected layer, features extracted from the eye and mouth regions are combined. This model also includes two independent 3D CNNs, each of which is built from a stacking layer, a 3D convolutional layer with 32 filters of size 3×3×15, a 3D pooling layer with a kernel size of 3×3 ×3, and a flatten layer to produce a feature vector with a single dimension. Both networks make use of dropout, flatten, and dense layers. Both networks are combined into one prior to the final dense layer.

Notation	Meaning
$K_l(l = 1,2,3.\ldots\ldots n)$	Network's parameter
n	Number of categories
$\frac{1}{\sum_{l=1}^{n} e^{K_l^T a^{(l)}}}$	Normaliztion
1	Total no.of probabilities
P	Probability
K	Sample
V	Arbitrary value
L_f	Loss function
N	Number of classes

Loss functions generalize logistic functions, which are widely used to solve multi-classification problems. Softmax returns an n-dimensional vector (the sum of its elements is 1) that indicates the likelihood that the image used as source was assigned to one of the N classes. Labels and training data for k samples,

$$\{(a^{(1)}, b^{(1)}), (a^{(2)}, b^{(2)}), \ldots\ldots\ldots (a^{(k)}, b^{(k)})\} \quad (4)$$

Then the probability of each sample, $f_\theta(a^{(l)})$

$$\begin{bmatrix} P(b^{(l)} = 1|a^{(l)}; K) \\ P(b^{(l)} = 2|a^{(l)}; K) \\ \cdot \\ \cdot \\ \cdot \\ P(b^{(l)} = n|a^{(l)}; K) \end{bmatrix} = \frac{1}{\sum_{l=1}^{n} e^{V_l^T a^{(l)}}} \begin{bmatrix} e^{K_1^T a^{(l)}} \\ e^{K_2^T a^{(l)}} \\ \cdot \\ \vdots \\ e^{K_n^T a^{(l)}} \end{bmatrix} \quad (5)$$

In order to ensure convergence during training, the loss function uses the gradient descent technique.

The loss function L_f,

$$L_f = -\frac{1}{k} \sum_{m=1}^{n} \sum_{p=1}^{q} l\{b^{(i)} = p\} \, log \frac{e^{K_p^T a^{(i)}}}{\sum_{i=1}^{n} e^{K_i^T a^{(i)}}} \quad (6)$$

The value of the function is zero when the condition is false and one otherwise. The equation 6 is simplified as,

$$L_f = -\frac{1}{k} \left[\sum_{m=1}^{n} log \frac{e^{K_{b^{(i)}}^T a^{(i)}}}{\sum_{i=1}^{n} e^{K_i^T a^{(i)}}} \right] \quad (7)$$

In practice, we often augment the loss function with a weight attenuation to avoid producing arbitrary V when the parameter was 0. The likelihood that the classifier is an actual label decreases as the loss function increases. To get the best possible outcome, we iteratively determined the

loss function's minimal value.

4. Experimental setup

The experimental conditions used in this study are described here. Before diving into a detailed explanation of the hyper-parameter settings used to train the proposed networks, we present a description of each micro expression dataset utilized in the studies. The experiment employed Python 3.6 within the PyCharm IDE, and was based on the keras structure with a backend of the TensorFlow platform. Tests on 64-bit versions of Windows 10 and other platforms. The GPU was an NVIDIA 2080 Ti, and there was 11 GB of dedicated graphics memory on the system. Adam was used to optimize the loss with the following experimental parameters: 0.0001 learning rate, 1e-5 decay, 100 epochs, 0.5 dropout, 64 batches. In this study, 80% of the data is utilized for training and 20% for testing.

4.1 Datasets

This study's micro-expression recognition datasets are summarized below. Deep learning requires a huge data collection. To meet study requirements, we employed CAS(ME)2 and SMIC micro expression video datasets. Eighty percent of each dataset was trained and twenty percent validated. The training and validation split is done once, and all trials use the same sets. Face expression recognition researchers can utilize CAS(ME)2. 247 movies from 123 people—68 women and 55 men—represent a wide age and racial range. The movies were shot in English, Mandarin Chinese, Cantonese, and ASL using natural and manufactured face expressions. These films have great lighting and settings. Participants had to produce angry, scornful, terrified, pleased, sad, astonished, and neutral looks. Each video is labeled with emotion start and finish times, strength, and facial action unit presence. Because it incorporates spontaneous expressions, CAS(ME)2 is harder to identify. HCI, emotion analysis, and artificial facial expression detection have used the dataset extensively.

The SMIC dataset is a facial expression analysis benchmark. SJTU took it. There are 282 images and 164 640x480-pixel videos. 16 actors—8 men and 8 women—were instructed to express various moods. Neutral to intense delight, sadness, surprise, rage, disgust, and terror. Actors focused, thought, and bewildered. SMIC contains human-annotated images of humans with different facial expressions and mental states. Expressions and conditions have start and finish times and 0–5 intensity ratings.

5. Result and discussion

We compare the experimental results to the gold standard in this section. F1-score, weighted average recall (WAR) or accuracy, and unweighted average recall (UAR) measure experiment success. UAR is like "balanced" accuracy, which accounts for class size but averages accuracy results. The micro-averaged F1-Score is for extremely unbalanced data. K-fold cross-validation divides the sample into k equal-sized subsamples. Each subset is tested, while the rest is the training set. Cross-validation tests total K. Setting an appropriate K may reduce practice evaluation time. K=6 is typical. The recall rate is the percentage of positive cases correctly predicted. Sensitivity is this. It assesses a model's favourable data selection. Thus, it is the percentage of correct diagnoses.

$$\text{Recall(r)} = \frac{Tr_p}{Tr_p + Fa_n} \tag{8}$$

Accuracy is the proportion of correct diagnoses relative to all diagnoses.

$$\text{Precision(p)} = \frac{\text{Tr}_\text{p}}{\text{Tr}_\text{p} + \text{Fa}_\text{p}} \tag{9}$$

$$Accuracy(Ac) = \frac{Tr_p + Tr_n}{Tr_p + Tr_n + Fa_p + Fa_n} \tag{10}$$

Taking into account the relative relevance of each class in the dataset, weighted average recall (WA_R) measures the average recall across numerous classes:

$$\text{WA}_\text{R} = \sum_{\text{l}=1}^{\text{P}} \text{w}_\text{l} * \text{r}_\text{l} \tag{11}$$

where M is the number of classes, w_l is the weight of the lth class, and r_l is the recall of the kth class. Unweighted average recall (UAR) is a measure of the average recall across multiple classes, where each class is given equal importance:

$$\text{UAR} = \frac{1}{\text{M}} \sum_{\text{k}=1}^{\text{M}} \text{r}_\text{k} \tag{12}$$

The F1 score combines a model's accuracy and recall measures.

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision(p)} \times \text{Recall(r)}}{\text{Precision(p)} + \text{Recall(r)}} \tag{13}$$

Here, we additionally ran each algorithm on CAS(ME)2 and SMIC for verification. The findings are presented in Table 1. Its recognition

Table 1. The performance analysis of the MFCNN, MECNN and base model

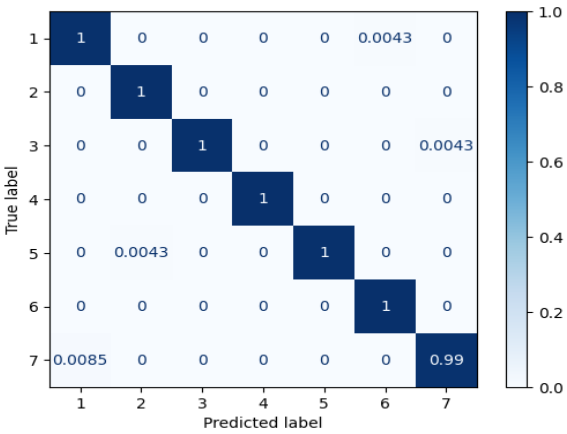
Data set	Model	Accuracy	Recall	F1-Score	WAR	UAR
CAS (ME)^2	Base model	90.23%	91	90	91	90
	MECNN	98.9%	98	99	98	99
	MFCNN	100%	99	100	99	100
SMIC	Base model	89.9%	89	89	90	89
	MECNN	98.2%	98	98	98	98
	MFCNN	99.8%	99	99	99	100

performance for the SMIC database is inferior to that of the CAS(ME)2. Low camera frame rate (100/fps) and environmental factors like lighting and shadows may be the reason for this performance. There was statistically significant variation in the recognition outcomes between MFCNN (99.8%), MECNN (98.9%), and Basic Network (90.23%). However, the suggested MEFNET outperformed the competition in terms of recognition rate.

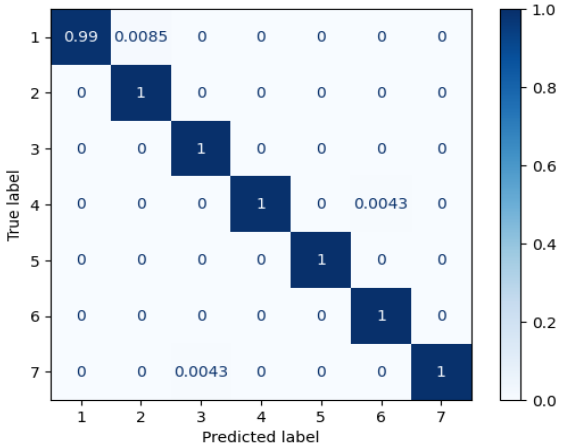
The total efficacy of all algorithms is shown in Table 1. MFCNN, an algorithm based on conventional techniques, has the highest mean accuracy of any micro-expression recognition system at 100%. The MECNN method has the best accuracy (98%) of any deep learning technique. In addition, we use the confusion matrix to assess the algorithm's effectiveness in recognizing each emotion class. The algorithm's ability to distinguish between emotions may be evaluated by seeing the total number of TPs, FPs, TNs, and FNs it has learned for each category. Since feature learning is so challenging, it stands to reason that a class with more examples will have better recognition performance when it comes to emotions. This is because additional data allows the model to learn the associated characteristics.

On CASME II and SMIC, we also calculated the confusion matrices for each method. Fig. 3 shows the MEFNET's confusion matrix. The results of several approaches on CASME II and SMIC are shown in Table 1.

It is clear that MEFNET can outperform the Basic Network in all facets of recognition, with a specificity of 99.2%, a sensitivity of 99.5%, and an accuracy of 100% on CAS(ME)2 dataset and specificity of 99.6%, a sensitivity of 99.3%, and an accuracy of 100% on SMIC dataset. A model's



(a)



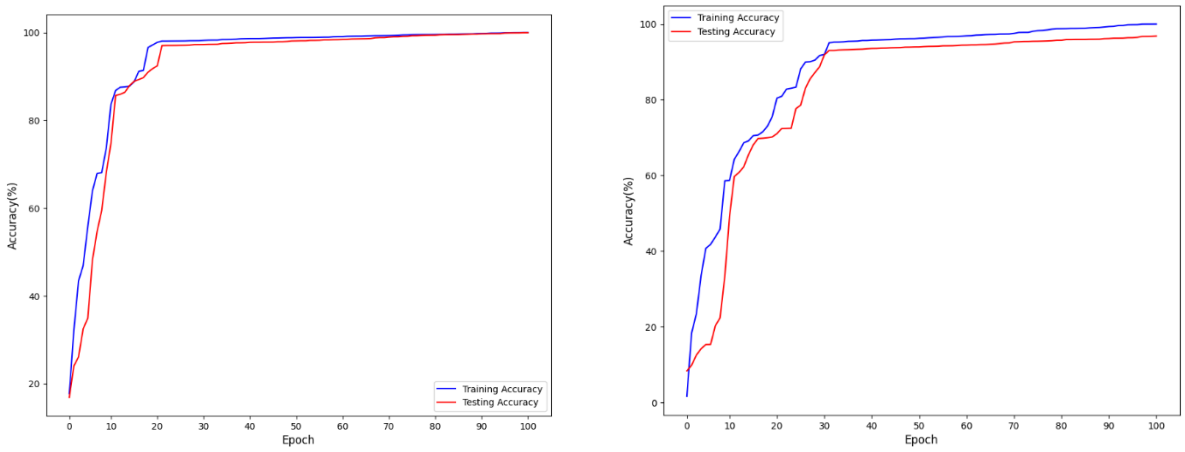
(b)

Figure. 3 confusion matrix on the dataset CAS(ME)2 and SMIC (a) MFCNN on CAS(ME)2 dataset (b) MFCNN on SMIC dataset

dependability depends on its test data predictions. Fig. 4 shows a sample accuracy graph for a dataset-trained model. The model's training data accuracy may initially be 0 to 10.

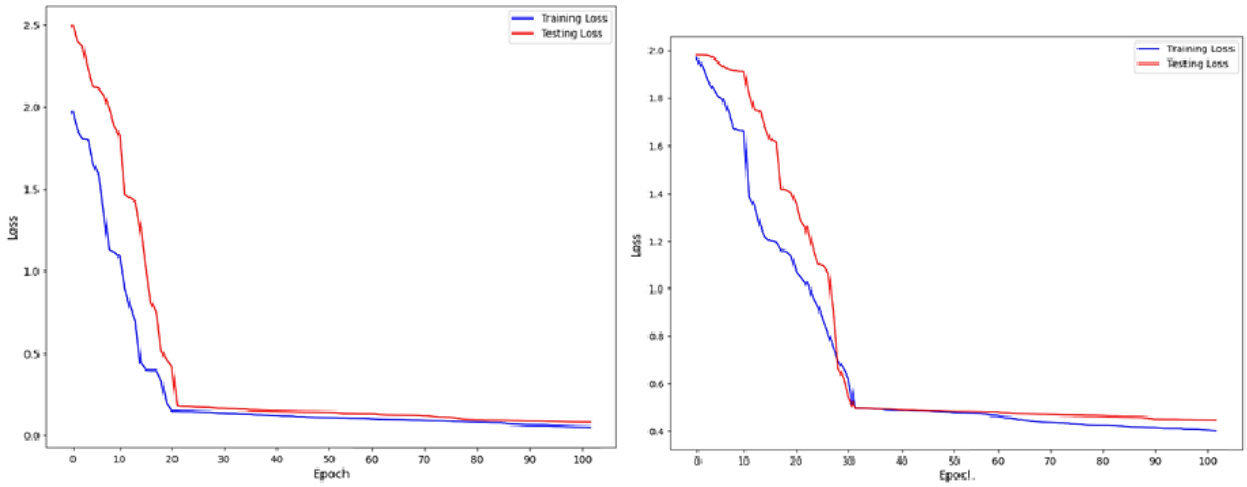
Each training cycle often results in improved accuracy on the training data, as the model's weights and biases are fine-tuned. If the model is overfitting the training data, nevertheless, it may start to underperform on the validation data even while it continues to do well on the training data. In this situation, the validation accuracy will drop while the training accuracy keeps rising, creating an imbalance in the sample accuracy graph. The accuracy of the model MFCNN is static after the epoch 20 and reaches 100% on both datasets. The model MECNN reaches 98% and static after the epoch 30.

Loss graph shape may indicate network performance. The loss number is often high early in training because the network cannot reliably predict the output and its parameters were initially set randomly. As training continues, the loss value



(a) (b)

Figure. 4 Accuracy of the MEFNET (a) MFCNN on CAS(ME)2 dataset (b) MECNN on CAS(ME)2 dataset



(a) (b)

Figure. 5 Loss of the MEFNET (a) MFCNN on CAS(ME)2 dataset (b) MECNN on CAS(ME)2 dataset

Table 3. Performance comparison

Methodology	Dataset	Model	Accuracy
Reddy, S.P.T et.al [22]	CASME 2, SMIC	CNN	51%
He.K et.al [23]	CIFAR-10	ResNet	50%
Peng, M. et.al [24]	CASME 2	TL	46%
Xia, Z et.al [26]	CASME 2, SMIC	RCNN	80%
Huang, X. et.al [14]	CASME 2, SMIC	LSTM	65%
Khor, H. et.al [27]	CASME 2, SAMM	DSN	71%
Zhi, R. et.al [28]	CASME 2, SMIC	3D-CNN	66%
Proposed methodology	CASME 2, SMIC	MEFNet	100%

decreases to 0.015 and 0.25 on CAS(ME)2 and SMIC datasets, indicating the network is improving prediction accuracy.

6. Performance comparison with existing methodology

On this particular dataset CASME 2 and SMIC, deep learning models were utilized for the purpose of micro-expression recognition. The Table 4 shows a list of some of the deep learning models that have been applied to CASME 2 and SMIC along with the stated accuracy of those models.

Table 3 provides a summary of the comparison of the results obtained from the CASME2 and SMIC database. The authors of the research papers [22-28] have made significant progress in improving the accuracy of micro-expression recognition. Their work has resulted in accuracy improvements ranging from 51% to 80%. The initial accuracy reported in these studies ranged from 51% to 80%. This indicates that the existing methods and models had limitations in effectively capturing the subtle and fleeting nature of micro-expressions. These

relatively lower accuracies can be attributed to the challenges associated with distinguishing micro-expressions, limited dataset availability, and the difficulty in extracting discriminative features. To address these challenges and improve the accuracy of micro-expression recognition, the authors proposed a novel approach, which achieved a remarkable accuracy of 100%. As can be shown in Table 3, our technique is capable of achieving a recognition accuracy that is 19% greater than that which was acquired from reference [37].

7. Conclusion and future work

Video ME recognition using 3DCNN deep learning. MEs seldom produce symptoms. We propose a two-stream 3D CNN architecture to overcome this challenge. First stream generates spatiotemporal characteristic from source picture series optical data. A second 3D CNN stream receives op tical flow vectors to detect motion changes. Our 3DCNN model beats cutting-edge models on SMIC and CASME2 ME datasets. New action recognition module MEFNet employs convolution and activation to improve processing. We present a parallel fusion strategy for the convolutional attention module and a highlight and dropout layer in STAM to boost attention in space and time to address order. MEFNet may possibly enhance any CNN without training samples or modalities. MEFNet improves CASME2 and SMIC deep networks. MEFNET surpasses the basic network in all recognition parameters, with 99.2% specificity, 99.5% sensitivity, and 100% accuracy on the CAS(ME)2 dataset and 99.6%, 99.3%, and 100% accuracy on SMIC. MEFNet's attention recalibration may enhance action recognition.

Our future focus will be improving attention modules. We will investigate a computationally efficient MER framework. We wish to examine temporal simulation of video clips using a pooling method to reduce parameters and processing cost.

Conflicts of interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Author contributions

Conceptualization, A, B, C and D; methodology, A; software, B; validation, D, E and F; formal analysis, A; investigation, B; resources, C; data curation, D; writing—original draft preparation, E; writing—review and editing, F; visualization, A; supervision, A; project administration, D;

References

[1] P. Ekman, *Emotions revealed: recognizing faces and feelings to improve communication and emotional life*, NY: OWL Books, 2007.

[2] Corneanu, C. Adrian, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 8, pp. 1548-1568, 2016.

[3] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 6, pp. 1113-1133, 2018.

[4] Porter, Stephen, and L. T. Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions", *Psychological Science*, Vol. 19, No. 5, pp. 508-514, 2008.

[5] T. A. Russell, E. Chu, and M. L. Phillips, "A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool", *British. Journal of Clinical Psychology*, Vol. 45, pp. 579– 583, 2006.

[6] F. Xu and J. P. Zhang, "Facial microexpression recognition: a survey", *Zidonghua Xuebao/Acta Automatica Sinica*, Vol. 43, No. 3, pp. 333–348, 2017.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks", In: *Proc. of 25th International Conf. on Neural Information Processing Systems*, Lake Tahoe, *USA*, Vol. 1, pp. 1097–1105, 2012.

[8] Y. Fan, V. O. Li, and J. C. Lam, "Facial expression recognition with deeply-supervised attention network", *IEEE Transactions on Affective Computing*, Vol. 13, pp. 1057–1071, 2020.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", In: *Proc. of International Conf. on Learning Representations*, San Diego, *USA*, pp. 1–14, 2015.

[10] C. Wang, M. Peng, M, T. Bi, and T. Chen,

"Micro-attention for micro-expression recognition", *Neurocomputing*, Vol. 410, No. 3, pp. 54–362, 2020.

[11] S. N. Sangeethaa and P. U. Maheswari, "An Intelligent Model for Blood Vessel Segmentation in Diagnosing DR Using CNN", *Journal of Medical Systems*, Vol. 42, No. 10, 2018.

[12] Y. Liu, H. Du, and L. G. Zheng, "A neural micro-expression recognizer", In: *Pro. of 14th IEEE International Conf. on Automatic Face and Gesture Recognition*, Lille, France, pp. 14–18, 2019.

[13] A. T. Lopes, E. D. Aguiar, A.F. D. Souza, and T. O. Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order", *Pattern Recognition*, Vol. 61, pp. 610–628, 2017.

[14] S. J. Wang, B. J. Li, Y. J. Liu, W. J. Yan, X. Ou, X. Huang, X, F. Xu, and X. Fu, "Micro-expression recognition with small sample size by transferring long-term convolutional neural network", *Neurocomputing*, Vol. 312, pp. 251–262, 2018.

[15] M. A. Takalkar, M. Xu, and Z. Chaczko, "Manifold feature integration for micro-expression recognition", *Multimedia Systems*, Vol. 26, pp. 535–551, 2020.

[16] Zhao, S.; Tao, H.; Zhang, Y.; Xu, T.; Zhang, K.; Hao, Z.; and Chen, E. "A two-stage 3D CNN-based learning method for spontaneous micro-expressionrecognition", *Neurocomputing*, Vol. 448, pp. 276–289, 2021.

[17] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, pp. 221–231, 2013.

[18] J. Haddad, O. Lézoray, and P. Hamel, "3D-CNN for facial emotion recognition in videos", In: *Proc. of International Symp. on Visual Computing*, San Diego, CA, USA, pp. 5–7, 2020.

[19] S. N. Sangeethaa, "Presumptive discerning of the severity level of glaucoma through clinical fundus images using hybrid PolyNet", *Biomedical Signal Processing and Control*, Vol. 81, No. 104347, 2023.

[20] J. Guo, S. Zhou, J. Wu, J. Wan, X. Zhu, Z. Lei, and S. Z. Li, "Multimodality network with visual and geometrical information for micro emotion recognition in Automatic Face & Gesture Recognition", In: *Proc. of International Conf. on IEEE*, pp. 814–819, 2017.

[21] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro expression recognition in Pattern Recognition (ICPR)", In: *Proc. of 23rd International Conf. on IEEE*, pp. 2258–2263, 2016.

[22] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks", In: *Proc. of International Joint Conf. on Neural Networks*, Budapest, Hungary, pp. 1–8, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016.

[24] M. Peng, Z. Wu, Z. Zhang, and T. Chen, "From Macro to Micro Expression Recognition: Deep Learning on Small Datasets Using Transfer Learning", In: *Proc. of 13th IEEE International Conf. on in Automatic Face & Gesture Recognition*, pp. 657-661, 2018.

[25] M. Peng, C. Wang, and C. T. Chen, "Attention Based Residual Network for Micro-Gesture Recognition", In: *Proc. of 13th IEEE International Conf. on in Automatic Face & Gesture Recognition*, pp. 790-794, 2018.

[26] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions", *IEEE Trans. Multimedia*, Vol. 22, No. 3, pp. 626–640, 2020.

[27] H. Khor, J. See, S. Liong, R. C. W. Phan, and W. Lin, "Dual-stream Shallow Networks for Facial Micro-expression Recognition", In: *2019 IEEE International Conference on Image Processing*, pp. 36–40, 2019.

[28] R. Zhi, H. Xu, M. Wan, and T. Li, "Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition", *IEICE Transactions on Information Systems*, Vol. 102, pp. 1054–1064, 2019.

[29] S. Jothimani and K. Premalatha, "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network", *Chaos, Solitons & Fractals*, Vol. 162, No. 112512, 2022.

[30] S. N. Sangeethaa and S. Jothimani, "Detection of exudates from clinical fundus images using machine learning algorithms in diabetic maculopathy", *International Journal of Diabetes in Developing Countries*, Vol. 42, pp.

1-11, 2022.

[31] S. Jothimani, S. N. Sangeethaa, and K. Premalatha, "Advanced Deep Learning Techniques with Attention Mechanisms for Acoustic Emotion Classification", In: *Proc. of 2022 International Conf. on Inventive Computation Technologies,* pp. 1235-1240, 2022.