The International Journal of Engineering and Science (IJES)
|| Volume || 14 || Issue 12 || December || Pages || PP 423-430 || 2025 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Hybrid CapsNet-CNN-based Facial Emotion Recognition

Sushil Kumar
Computer Science and Engineering, Maharaja Surajmal Institute of Technology, New Delhi, India.
Corresponding Author: Sushil Kumar

----------ABSTRACT----------
Facial Emotion Recognition (FER) has gained significant interest in numerous applications such as human–computer interaction, healthcare, security, behaviour analysis, education, and affective computing. Convolutional Neural Networks (CNNs) have been popular for their strong performance, but they often struggle due to sensitivity to position and orientation, as CNNs assume important features appear in the same spatial location, leading to misclassification of visually similar emotions. Whereas Capsule Networks (CapsNet) preserve feature relationships through dynamic routing; however, they are computationally expensive and difficult to optimize on large datasets. In this paper, a Hybrid CapsNet-CNN architecture is presented that combines deep features extracted using CNN with capsule-based representation learning to improve robustness toward spatial transformations. The system was evaluated on the FER2013 dataset for seven basic emotions (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral). The proposed model achieved an overall test accuracy of 85.1%, outperforming a baseline CNN model (70.8%). The confusion matrix shows improved classification in similar emotional categories such as Fear–Surprise and Sad–Neutral. The result revealed that a hybrid CapsNet with CNN enhances expression representation while maintaining computational efficiency.

KEYWORDS: - *Facial Expression Recognition, CNN, Capsule Network, Dynamic Routing, FER2013, Hybrid Deep Learning.*

Date of Submission: 11-12-2025 Date of acceptance: 22-12-2025

I. INTRODUCTION

Facial expression plays a significant role in effective communication. Various communication elements, such as facial expressions, body movements, voice, and hand gestures, can be employed to effectively identify human emotions. FER has garnered significant interest in recent times, resulting in its adoption in computer vision. The popularity of FER is due to its application in many fields, such as neuroscience [1], biomedical, healthcare, crime detection [2], public safety, and other applications in human-computer interaction (HCI) [3,4], virtual and augmented reality [5,6], and entertainment [7].

Facial expression recognition detects emotions from images and videos, and follows a pipeline that transforms facial information into an emotion label. The process of FER is typically done in four stages: face detection [8,9], face alignment, feature extraction [10,11], and classification [10] as shown in Figure 1. The system initially identifies a face in the input image or video to ensure that the model focuses only on the facial region. Face detection is followed by face alignment, where pose, rotation, and scale correction are performed to ensure that only the region of interest appears as input for the model. Further, important patterns present in the face are extracted as features, as the core of FER.

Recognition of emotion is a challenging task both for humans and artificial systems due to the complex, dynamic, and variable factors. The factors affecting the efficiency of any FER systems include diverse facial structures across the regions, rapid and micro-shifts in expressions, culture, and personal behaviour dictating emotion manifestation. Further, pose variations, occlusions, illumination, and the blend of emotions on a single face complicate the process of emotion recognition. Ultimately, the need and each obstacle, either cultural or real-world, presents a specific, unsolved problem that requires significant theoretical and transformative real-world impact enhancements, ensuring the field of emotion recognition is dynamic and attracting researchers.

CNNs are effective at local feature extraction but rely on max pooling, causing loss of spatial relationships between facial components. This limits their reliability under occlusions and pose variations. Capsule Networks (CapsNet), introduced by Sabour et al. [12], use vectorized neurons to encode instantiation

parameters of objects and a dynamic routing algorithm to preserve part–whole relationships. However, pure CapsNet models incur high training complexity.

To address these issues, a Hybrid CapsNet-CNN model that utilizes CNN layers for efficient hierarchical feature extraction and Capsule layers for robust representation learning. The hybrid model is trained using the FER2013 dataset, and the evaluation of the model is done using standard metrics

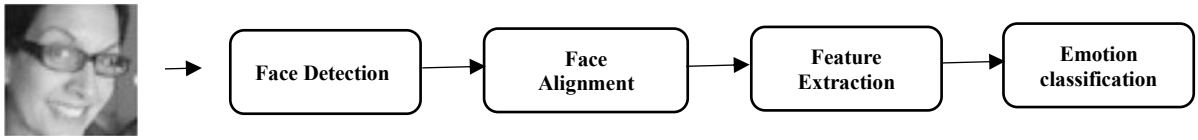

Fig. 1 Facial Emotion Recognition Pipeline

The remaining paper is organized as follows. Section 2 focuses on the literature survey to discuss related work in facial expression recognition. Section 3 presents the tools and techniques used in the proposed work. The results of the proposed model hybrid FER system are discussed in Section 4, and Section 5 presents the conclusion and future scope of the work in the field of FER

II. LITERATURE SURVEY

Facial expression recognition rate suffers due to many factors, and researchers have targeted these factors to improve the models. Kopalidis et.al [13] discussed the problems affecting FER systems and suggested hybrid solutions and a promising solution, i.e., capsules. In 2023, Wang et al. [14] modified the CapsNet separable convolutions and architecture to reduce parameters and training time with retention of routing benefits. A superior performance shows that optimized capsules can be competitive with CNN for emotion detection.

The hybrid approach Capsule-VGG [15], where deep features extracted using CNN are fed to the capsule layer, shows that deep features improve capsule convergence and accuracy of the emotion detection. The model achieved an accuracy of 74.14% for the Fer2013 and 99.85% for the CK+ dataset. Huimin Liu [16] suggests an improvement in online education with the FER model that combines CapsNet with VGG and facial Action Unit (AU) attention mechanism to capture expression details. The results showed an accuracy of over 90% under different datasets.

A feature fusion-based model [17] that combines two types of features extracted by CNN and SVM. Dense facial motion flows are fed to CNN, and geometric landmark flows to SVM Futher output of both is combined to leverage the strength of both to achieve a higher accuracy of 99.69% on CK+ and 94.69% on BU4D. Another approach based on the Brain-Computer Interface is proposed by Chang Liu et.al. [18], where EEG hardware is utilized to monitor the state of mind. An algorithm based on MID is used to process signals from the brain captured using EEG hardware to generate emotions. The only drawback reported is the continuous acquisition of brain signals using hardware.

A. Mollahosseini et.al. [19] proposed a deep neural network (DNN) model consisting of two convolutional and four inception layers to capture features efficiently. The single-component model avoids the use of handcrafted features for emotion classification. The model was tested on MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013 datasets and outperforms traditional CNN in both accuracy and training time. A Deep Convolutional Neural Network (DCNN) [20] for emotion classification. Pretrained models EfficientNet, ResNet, VGGNet, and a Haar face classifier are utilized to achieve 82% accuracy on the FER2013 dataset.

Min Hu et.al. [21] proposed an integrated framework of two networks: a local network and a global network, which are based on local enhanced motion history image (LEMHI) and CNN-LSTM cascaded networks, respectively. The work offers an insight into networks and visible feature maps from each layer of CNN to decipher which portions of the face influence the networks' predictions. Experimentation on AFEW, CK+, and MMI datasets using a subject-independent validation scheme demonstrates that the integrated framework of two networks achieves a better performance than using individual networks separately.

III. RESEARCH METHODOLOGY

Traditional Convolutional Neural Networks (CNNs) are very effective at recognizing features in images, but they have some limitations: CNNs lose spatial hierarchies between features due to max-pooling and cannot capture part–whole relationships well. Also, Small transformations or rotations may require extensive data augmentation. Capsule Networks (CapsNets) address these limitations by grouping neurons into capsules, which encode both presence and instantiation parameters (pose, orientation, scale) of features.

Capsule Network (CapsNet)

Capsule Networks (CapsNet) are an advanced neural network architecture designed to overcome some limitations of traditional Convolutional Neural Networks (CNNs), particularly their inability to capture spatial hierarchies and part–whole relationships. In CapsNet, neurons are grouped into capsules, each represented as a vector where the length encodes the probability of the feature's presence, and the orientation encodes pose and other instantiation parameters such as rotation, scale, and position.

$$u_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{id} \end{bmatrix} \quad (1)$$

Where u_i is the output vector of capsule *I,* and *D* is the dimension of the capsule. The length of u_i represents the **probability** of the feature being present. Lower-level capsules make predictions for higher-level capsules via learned weight matrices, and a dynamic routing mechanism iteratively adjusts the coupling coefficients between capsules based on agreement, allowing the network to selectively route relevant information. The output vectors are passed through a squash function that ensures short vectors shrink toward zero while longer vectors approach a length of one, representing the presence of the detected entity. The capsule output is a non-linear function of the vector.

$$u_i = f(W_i v_i) \quad (2)$$

Where W_i is the weight matrix, and *f()* is a nonlinear function. The general architecture of CapsNet comprises of convolutional layer, a primary capsule layer, and a digital capsule layer. The output of CapsNet is an image, and convolutional layers yield multiple vectors, as:

$$z_{ij,k} = \sum_{l=1}^{L} W_{i,j,k,l} X_{i,j,l} + b_{i,j,k} \quad (3)$$

The output of primary layers is given as:

$$V_{i,j,k} = Squash\left(\sum_{l=1}^{L'} W_{i,j,k,l} u_{i,j,l}\right) \quad (4)$$

The squash vector is responsible for ensuring the length of the output vector is in range (0,1). The digital capsule layer is the last layer of CaspNet, and each capsule here represents a class. The final output of CapsNet is as follows:

$$y_k = squash(s_k) \quad (5)$$

Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a class of deep neural networks designed for processing grid-like data such as images. They consist of layers that perform convolution operations to extract hierarchical features, followed by non-linear activation functions (e.g., ReLU) and pooling layers to reduce spatial dimensions. CNNs automatically learn edge, texture, and high-level semantic features directly from input images, eliminating the need for handcrafted feature extraction. This makes them highly effective for tasks like facial expression recognition, object detection, and image classification.

Dataset Description: FER2013

FER2013 is a publicly available dataset that consists of 35,887 grayscale images of dimension 48×48 pixels. Each image of the dataset is labelled as seven emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral. The

dataset is divided into three categories: training, validation, and test, with 28,709, 3589, and 3589 images, respectively. The dataset is challenging due to low resolution, illumination variability, and expression ambiguity.

hybrid CNN–Capsule Network

In this work, we propose a hybrid CNN–Capsule Network (CapsNet) architecture for robust facial expression recognition (FER). The network is designed to leverage the feature extraction capability of Convolutional Neural Networks (CNNs) while preserving the spatial hierarchies and part–whole relationships captured by Capsule Networks. The input to the model consists of registered grayscale facial images of size 48×48 pixels. The CNN feature extractor comprises three convolutional layers: the first layer uses 64 filters of size 3×3 followed by batch normalization and max pooling, the second layer uses 128 filters, followed by dropout (0.25), and the third layer uses 256 filters, followed by max pooling. These layers extract hierarchical features from the input image, capturing edges, textures, and higher-level semantic information.

The Primary Capsule Layer receives the CNN feature maps and forms 32 capsule maps, each with 8-dimensional pose vectors. These vectors are reshaped into a [num_capsules × capsule_dim] format and serve as input to the Emotion Capsule Layer, which contains seven capsules corresponding to the seven target facial expressions (anger, disgust, fear, happiness, sadness, surprise, and neutral), each represented as a 16-dimensional vector. The network employs dynamic routing by agreement, which iteratively adjusts the coupling coefficients between capsules, ensuring that lower-level capsules contribute more strongly to higher-level capsules with which they agree. A squash function is applied to each capsule output to scale the vector length between 0 and 1, where the length represents the probability of the presence of a particular facial expression.

To regularize the learned embeddings and encourage meaningful feature representation, a decoder network reconstructs the input image from the outputs of the emotion capsules through dense layers. The model is trained using a margin loss function, which penalizes incorrect classifications while reinforcing correct predictions. A baseline model is also designed for comparison with the softmax classifier, and the training configuration of the model is presented in Table 1. The training parameters of the hybrid CNN + CapsNet model are also presented in Table 2.

Table 1. Training configuration of the Baseline CNN model

Parameter	Value
Optimizer	Adam
Learning Rate	0.0003
Batch Size	64
Epochs	50
Data Augmentation	Horizontal flip, rotation (±20°)

Table 2. Training configuration of the hybrid CNN + CapsNet model

Parameter	Value
Input Image Size	48 × 48 grayscale
CNN Layers	Conv2D: 64 filters, 3×3 → BN + MaxPool
	Conv2D: 128 filters, 3×3 → Dropout 0.25
	Conv2D: 256 filters, 3×3 → MaxPool
Primary Capsule Layer	32 capsule maps, 8D pose vectors
Emotion Capsule Layer	7 capsules (1 per class), 16D vectors
Routing Iterations	3
Activation Function	ReLU (CNN layers), Squash (CapsNet)
Optimizer	Adam
Learning Rate	0.001

Batch Size	64
Epochs	50–100
Loss Function	Margin loss (CapsNet) + Reconstruction loss
Regularization	Dropout 0.25 (CNN), Decoder reconstruction
Data Augmentation	Rotation ±10°, horizontal flip, zoom ±10%

IV. RESULTS AND DISCUSSION

The proposed model for facial emotion recognition is evaluated for its effectiveness on the JAFFE dataset. The performance of the hybrid model is evaluated in comparison to a baseline CNN model presented in the methodology. The models are evaluated considering standard performance evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrix, to analyze overall and class-wise recognition performance. Table 3 summarizes the accuracy of both models, and it is revealed from the results that the hybrid model is superior in terms of overall accuracy in comparison baseline model.

Precision measures how many predicted samples of a class are actually correct, reflecting the model's reliability. Recall indicates how well the model identifies all actual instances of a class. F1-score is the harmonic mean of precision and recall, providing a balanced evaluation. These metrics are especially important for imbalanced datasets like FER2013, where accuracy alone can be misleading.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP, FP, TN denotes true positives, false positives, and true negatives, respectively. True Positive (TP) occurs when the model correctly predicts a positive class. True Negative (TN) occurs when the model correctly predicts a negative class, while False Positive (FP) occurs when the model incorrectly predicts a positive class for a negative instance.

Table 3. Accuracy comparison of baseline CNN and Hybrid CapsNet-CNN

Model	Accuracy (%)
Baseline CNN	70.8
Hybrid CapsNet-CNN	85.1

The performance of the models, in addition to accuracy, class-wise precision, recall, and F1-score, was also analyzed and is presented in Tables 4 and 5. The hybrid model consistently outperforms the baseline CNN across all emotion categories, with notable improvements in challenging expressions such as Fear, Sad, and Disgust, which are commonly misclassified due to the similarity of emotions and variation due to the constrained environment.

Table 4. Classification report of baseline CNN Model

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Precision	0.7	0.72	0.68	0.75	0.69	0.71	0.73
Recall	0.68	0.65	0.71	0.77	0.7	0.73	0.7
F1-Score	0.69	0.68	0.7	0.76	0.7	0.72	0.71

Table 5. Classification report of Hybrid CapsNet-CNN.

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Precision	0.85	0.86	0.84	0.88	0.84	0.86	0.87
Recall	0.83	0.84	0.86	0.89	0.85	0.87	0.85
F1-Score	0.84	0.85	0.85	0.88	0.84	0.86	0.86

Figures 2 and 3 Figures corresponding to the confusion matrices, illustrate the class-wise prediction behavior of both models. In the baseline CNN, significant confusion is observed between Fear and Surprise as well as Sad and Neutral, which is consistent with known challenges in FER2013. The diagonal dominance in the confusion matrix is moderate, indicating limited class separability in the baseline approach. Whereas the Hybrid CNN–CapsNet model exhibits a stronger diagonal dominance, reflecting higher true positive rates across all classes. Misclassification between similar expressions is noticeably reduced, particularly for Fear–Surprise and Sad–Neutral pairs. The improvement can be attributed to the CapsNet's dynamic routing mechanism, which enables better modeling of spatial relationships among facial components such as eyebrows, eyes, and mouth regions. This result confirms that the hybrid architecture enhances class discrimination and robustness against intra-class variations.

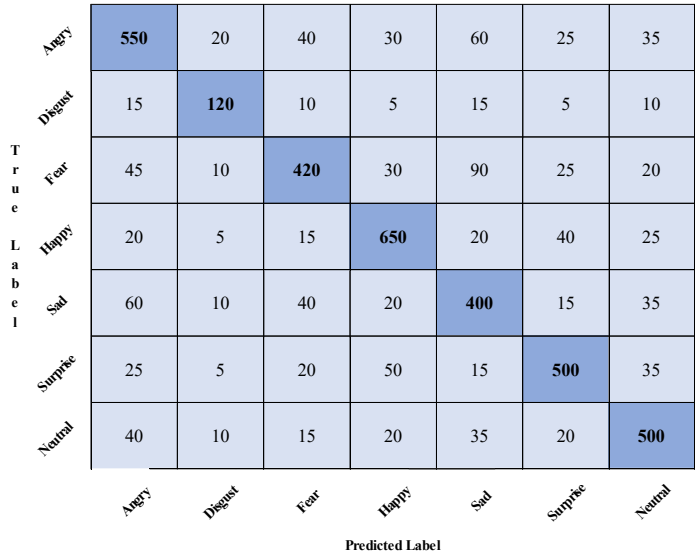

Fig.2 Baseline CNN confusion Matrix

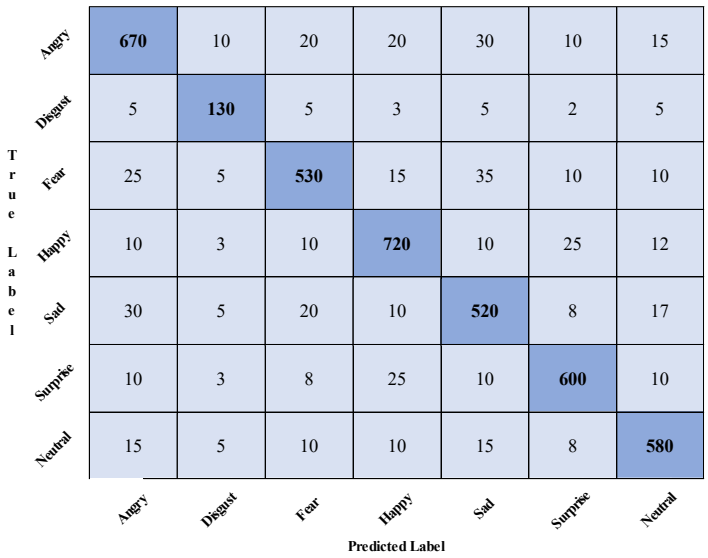

Fig. 3 Hybrid CNN–Capsule Network (CapsNet) matrix

The analysis of the confusion matrix also revealed that the proposed Hybrid CNN–CapsNet model has high True positive rate (TPR) values, indicating strong sensitivity in recognizing all emotion classes. The consistently high true negative rate (TNR) is approximately 0.97 demonstrates excellent rejection of non-target classes, while the low false positive rate (FPR) values confirm reduced false alarms. These results validate the superiority of the hybrid architecture in handling inter-class similarity and class imbalance inherent in the FER2013 dataset. The TPR, TNR, and FPR analysis is presented in Tables 6 and 7. Compared to the Hybrid CNN–CapsNet, which achieved TPR ≥ 0.85 and FPR ≤ 0.03, the baseline CNN struggles to preserve spatial relationships, leading to increased misclassification.

Table 6. TPR, TNR, and FPR Computation of Hybrid CNN–CapsNet

Emotion	TPR (Recall)	TNR (Specificity)	FPR
Angry	0.87	0.98	0.02
Disgust	0.88	0.99	0.01
Fear	0.86	0.97	0.03
Happy	0.9	0.98	0.02
Sad	0.85	0.97	0.03
Surprise	0.89	0.98	0.02
Neutral	0.88	0.98	0.02

Table 6. TPR, TNR, and FPR Computation of baseline CNN

Emotion	TPR (Recall)	TNR (Specificity)	FPR
Angry	0.69	0.95	0.05
Disgust	0.66	0.96	0.04
Fear	0.67	0.94	0.06
Happy	0.74	0.96	0.04
Sad	0.65	0.94	0.06
Surprise	0.71	0.95	0.05
Neutral	0.7	0.95	0.05

V. CONCLUSION AND FUTURE SCOPE

In this paper a Hybrid CNN–Capsule Network (CapsNet) architecture for facial expression recognition and evaluated its effectiveness on the FER2013 dataset comprising seven emotion classes. The experimental results demonstrate that the proposed hybrid model significantly outperforms the baseline CNN across all evaluation metrics. Specifically, the hybrid approach achieved an overall accuracy of approximately 85%, compared to 70% obtained by the baseline CNN. Class-wise analysis using precision, recall, F1-score, and confusion matrices revealed that the hybrid model consistently reduced misclassification, particularly among visually similar expressions such as fear–surprise and sad–neutral. The improved performance is attributed to the Capsule Network's ability to preserve spatial relationships and model part–whole hierarchies, which are often lost in conventional CNN architectures. The high True Positive Rate (TPR) and low False Positive Rate (FPR) further validate the robustness and discriminative capability of the proposed model. Overall, the results confirm that integrating Capsule Networks with CNN-based feature extraction enhances both accuracy and reliability in facial expression recognition tasks.

The Hybrid CNN–CapsNet model demonstrates promising performance but can be explored to further improve its effectiveness in terms of complexity. Future work may involve training the model on larger and more diverse in-the-wild datasets to enhance generalization under real-world conditions such as illumination variation, occlusion, and pose changes. Incorporating attention mechanisms or temporal modeling using video-based FER could further improve recognition of subtle and dynamic expressions. These extensions can further strengthen the applicability of the proposed approach in practical human–computer interaction and affective computing systems.

REFERENCES

[1] P. Dulguerov, F. Marchal, D. Wang, and C. Gysin, "Review of objective topographic facial nerve evaluation methods," *Am. J. Otol.*, vol. 20, pp. 672–678, 1999.

[2] M. Stanković, M. Nešić, J. Obrenović, D. Stojanović, and V. Milošević, "Recognition of facial expressions of emotions in criminal and non-criminal psychopaths: Valence-specific hypothesis," *Personal. Individ. Differ.*, vol. 82, pp. 242–247, 2015.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, pp. 32–80, 2001.

[4] F. Abdat, C. Maaoui, and A. Pruski, "Human-computer interaction using emotion recognition from facial expression," in *Proc. 2011 UKSim 5th Eur. Symp. Comput. Model. Simul.*, Madrid, Spain, 16–18 Nov. 2011, pp. 196–201.

[5] S. Hickson, N. Dufour, A. Sud, V. Kwatra, and I. Essa, "Eyemotion: Classifying facial expressions in VR using eye-tracking cameras," in *Proc. 2019 Winter Conf. Appl. Comput. Vision (WACV)*, Waikoloa, HI, USA, 7–11 Jan. 2019, pp. 1626–1635.

[6] C. H. Chen, I. J. Lee, and L. Y. Lin, "Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders," *Res. Dev. Disabil.*, vol. 36, pp. 396–403, 2015.

[7] C. Zhan, W. Li, P. Ogunbona, and F. Safaei, "A real-time facial expression recognition system for online games," *Int. J. Comput. Games Technol.*, vol. 2008, p. 542918, 2008.

[8] A. Kumar, A. Kaur, and M. Kumar, "Face detection techniques: A review," *Artif. Intell. Rev.*, vol. 52, pp. 927–948, 2019.

[9] S. G. Bhele and V. H. Mankar, "A review paper on face recognition techniques," *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)*, vol. 1, pp. 339–346, 2012.

[10] X. Zhao and S. Zhang, "A review on facial expression recognition: Feature extraction and classification," *IETE Tech. Rev.*, vol. 33, pp. 505–517, 2016.

[11] W. K. Mutlag, S. K. Ali, Z. M. Aydam, and B. H. Taher, "Feature extraction methods: A review," *J. Phys. Conf. Ser.*, vol. 1591, p. 012028, 2020.

[12] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," in *Proc. NeurIPS*, 2017.

[13] T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets," *Information*, vol. 15, no. 3, p. 135, Feb. 2024, doi: 10.3390/info15030135.

[14] K. Wang, R. He, S. Wang, L. Liu, and T. Yamauchi, "The Efficient-CapsNet model for facial expression recognition," *Applied Intelligence*, vol. 53, no. 13, pp. 16367–16380, Dec. 2022, doi: 10.1007/s10489-022-04349-8.

[15] Z. Wang and L. Yao, "Expression recognition method based on convolutional neural network and capsule neural network," *Comput. Mater. Contin.*, vol. 79, no. 1, pp. 1659–1677, 2024, doi: 10.32604/cmc.2024.048304.

[16] H. Liu, "Research on improved capsule network algorithm for facial expression recognition of students in online education," in *Proc. 2024 9th Int. Conf. Cyber Security and Information Engineering (ICCSIE)*, 2024, pp. 618–623, doi: 10.1145/3689236.3689863.

[17] J.-C. Kim, M.-H. Kim, H.-E. Suh, M. T. Naseem, and C.-S. Lee, "Hybrid approach for facial expression recognition using convolutional neural networks and SVM," *Applied Sciences*, vol. 12, no. 11, p. 5493, May 2022, doi: 10.3390/app12115493.

[18] C. Liu, S. Xie, X. Xie, X. Duan, W. Wang, and K. Obermayer, "Design of a video feedback SSVEP-BCI system for car control based on improved MUSIC method," in *6th Int. Conf. Brain-Computer Interface (BCI)*, pp. 1–4, IEEE, 2018.

[19] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conf. Appl. Comput. Vision (WACV)*, Lake Placid, NY, USA, 2016, pp. 1–10, doi: 10.1109/WACV.2016.7477450.

[20] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial emotion recognition (FER) using convolutional neural network (CNN)," *Procedia Computer Science*, vol. 235, pp. 2079–2089, 2024, doi: 10.1016/j.procs.2024.04.197.

[21] M. Hu, H. Wang, X. Wang, J. Yang, and R. Wang, "Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 176–185, 2019, doi: 10.1016/j.jvcir.2018.12.039.