Journal of Computer Science

Research Article

An Occlusion Aware Facial Expression Recognition Model Using Fitness Based Cheetah Optimizer and Adaptive Multi-Scale ViT-CNN With Attention Mechanism

A. Reddy Prasad and A. Rajesh

Department of Computer Science and Engineering Vels Institute of Science, Technology and Advanced Studies (VISTAS) Chennai, Tamil Nadu, India

Article history
Received: 23-10-2024
Revised: 18-03-2025
Accepted: 12-05-2025

Corresponding Author:
Reddy Prasad A.
CSE, Vels Institute of Science, Technology and Advanced Studies (VISTAS),India
Email: areddyprasad11@gmail.com

Abstract: As a highly nuanced aspect of human communication, facial expression recognition presents a computationally complex problem, making it a prominent area of research in computer vision and affective computing. Problems like poor image quality, occlusions, inconsistent illumination, and head attitude changes are frequently observed in images taken from unstructured sources such as the internet that affect the accuracy of facial expression performance. With the aim of resolving these issues, an innovative occluded Facial Expression Recognition (FER) using an advanced deep learning model is proposed. For recognizing facial expressions, images are gathered in benchmark sources. The Viola-Jones (VJ) facial detector model is processed using the collected images. The detected face images from the VJ are given to the Regions of Interest (ROI) extraction process. The extracted ROI is passed to the Adaptive and Multiscale Vision Transformer-Convolutional Neural Network with Attention Mechanism (AMViTCNN-AM) for recognizing facial expressions. AMViTCNN-AM accurately identifies the expression in the face images even in the presence of occlusion. To get better performance in the FER process, the parameters in the network are optimized by the Fitness-based Cheetah Optimizer (F-CO). Experiments are carried out to prove the efficiency of the designed framework. The outcomes show that the implemented approach attained an accuracy value of 98.43%, which proves the potential of a developed deep learning model in the FER.

Keywords: Occlusion Aware Facial Expression Recognition, Fitness-Based Cheetah Optimizer, Adaptive and Multiscale Vision Transformer-Convolutional Neural Network With Attention Mechanism

Introduction

Face recognition technology is rapidly growing because of the advancement and widespread use of mobile devices (Li *et al*., 2019). Recognizing facial expressions is an important process of face recognition technology and it has gained significant attention in Human-Computer Interaction (HCI), healthcare, and travel, making it a popular area of study in business and academia (Xia *et al*., 2022). FER plays a vital role in medical care due to its numerous important applications (Geetha *et al*., 2009). Human perception (Xie *et al*., 2022) and interaction mainly influence facial expression since it is one kind of non-verbal communication (Chen *et al*., 2023). The morphological alteration in the face is represented by the facial expression.

Humans communicate their emotions and interact with others through their facial expressions (Pantic and Patras, 2006), which are rich sources of emotional information. In a variety of human-computer interaction applications, including intelligent question answering (Zheng *et al*., 2006), intelligent healthcare, intelligent classroom instruction, criminal detection, and tiredness monitoring, facial expression recognition is essential (Li *et al*., 2013). It becomes an essential part of HCI since it is the most

straightforward and efficient method for computers to understand human emotions (Zhang *et al*., 2022). As a result, a lot of work is being done by advanced field of FER (Chang *et al*., 2006). The FER is adopted in numerous applications like interactive gaming, robotics, and virtual reality. Most FER methods mainly concentrate on interpreting static features of the face images (Ahlawat and Nehra, 2017). However, expressions on the face are intrinsic in nature and it is dynamically changed for a series of consecutive facial movements (Kotsia and Pitas, 2007).

Deep learning models are mostly utilized by the researchers for recognizing the emotions of the human (Eleftheriadis *et al*., 2015). The capability of these models is excellent unless the facial images have an occlusion effect. But, in real-world scenarios, the recognition of the unconstrained facial expression is a difficult process (Pantic and Rothkrantz, 2004). However, traditional techniques failed to fully leverage the facial features, resulting in unsatisfactory identification results (Liu *et al*., 2023b). In order to address occlusion and pose fluctuations in real-world scenarios, the Region Attention Network (RAN) adaptively learns significant regions in occluded and pose-transformed images (Selvakumar *et al*., 2015). The situation of occlusion is rarely taken into account in conventional techniques for facial emotion detection (Zhang *et al*., 2020). Various techniques utilize attention mechanisms to concentrate on the most distinctive face features, to enhance the precision of FER (Aleksic *et al*., 2006). These techniques mostly rely on spatial local features, and it doesn't fully capture facial features in the given images (Yang *et al*., 2018). The deep learning model overcomes the limitations of local receptive fields by continually stacking convolutional layers to capture the global characteristics of images (Huang *et al*., 2022). However, in real-world scenarios, particularly with limited resources, deepening the converging network to enhance recognition precision is impractical. Therefore, it is difficult to enhance FER performance with deep learning-based models (Kumar and Kumar, 2025).

Motivation of the Developed Methods

This study implements diverse FER prediction methods by utilizing conventional methods because the human face has diverse unique features. FER prediction plays a vital role in deciding human actions and it is highly validated based on the myriad emotions on people's faces. Yet, it needs more reliable datasets to accurately recognize facial emotions. Also, it is tightly restricted by illumination and the variability of poses (Devi and Preetha, 2025). It does not have the flexibility to categorize the captured images in an uncontrolled manner and provides poor generalizability outcomes (Haq *et al*., 2024). It has the ability to provide better results, yet the system performance is affected by various lighting, spatial alterations, occlusions, complex patterns, and computational resources which can significantly alter the appearance of expression detection (Colombo *et al*., 2011). With the aim of resolving these issues, a novel technique is implemented to identify facial expressions with the help of adaptive and multi-scale ViT-CNN with an attention mechanism to maximize the recognition framework.

Importance of FER

Facial communication is an important way to express people's feelings and convey their emotional states. Human angry, surprise, coldness, and truth can be significantly express their emotions without language. FER has the ability to easily recognize people's complex expressions by identifying the gathered features. It can effectively extract optimal features and differentiate the expressions and variations. This helps to prevent violence and enhance the safety and security measures to track people's mental health patterns and their behaviors. FER is a crucial part of communication to enhance empathy and enable the reliable social interactions in diverse fields such as marketing, education, psychology, and healthcare. For different surveillance systems, the FER helps to recognize suspicious behavioural activities to enhance public safety in emergency situations. By suggesting this context, deep learning techniques are incorporated in FER that tends to be highly employed in various applications.

Thus, we suggested an innovative occlusion-aware FER system in this paper to effectively capture the emotions from the occluded face images.

Key Contributions of the Proposed Occlusion-Aware FER System

We implement an efficient deep learning model for recognizing the expression of the face from the occluded images of the human to validate the rich emotional condition of the human. The developed FER system is being helped in security and healthcare applications to take the necessary actions based on the facial expression of the human

We recommend a VJ-facial detector model for detecting the face images hidden by the objects. The face detection process using the VJ is also used to lower the overall duration necessary for the emotion recognition framework since it helps to focus the relevant region in the face to get accurate results in the FER

We perform the ROI extraction on the face images to capture the relevant features from the specific area of the face images. The ROI extraction process is most importantly to capture the distinctive features of the face images that provide precise results in the FER

We design an AMViTCNN-AM with the help of ViT, CNN, and AM for recognizing the deep expression of the human. The overall face structure and the subtle features

of the face image are captured by AMViTCNN-AM so it attains high accuracy in the FER. The AMViTCNN-AM effectively determines the connection among different components of the face images so it provides efficient results in the FER

Literature Survey

Related Work

Facial Expression Recognition (FER) has become a pivotal area of research, particularly with the rise of deep learning. The application of these algorithms has significantly advanced the field by enabling more accurate and efficient analysis. A critical aspect of this improvement lies in feature extraction, which substantially reduces processing time and enhances overall system performance. This technological evolution has spurred extensive exploration, with numerous researchers proposing diverse frameworks to recognize human expressions. Key contributions in this area are reviewed in the following section.

Dapogny *et al.* (2018) proposed a novel multifaceted occlusion detection method for the identification of 3Dfaces that had been partially obscured by unanticipated, unrelated things. The reconstructive module used a suitable foundation for the region where the non-occluded features exist to reconstruct the entire face utilizing the data supplied from the non-occluded portion of one's face.

Hu *et al.* (2020) employed Random Forests (RF) for the emotion recognition process. The Local Expression Prediction (LEP) was used for categorizing the facial emotions. Such a system was taught to hierarchically and locally collect a variety of training information. Numerous tests demonstrated that the suggested LEP encoding produced strong results for the FER process.

Yang *et al.* (2022) developed an innovative system using Speeded Up Robust Feature (SURF) and heterogeneity soft partitioning networks to recognize faces. In such a structure, the horizontal asymmetric region of the occluded zone was located by using an obstruction recognition unit that could be used to identify the occluded component. Then, to quickly complete the process of face illustration within an unattended situation, a facial painting component according to reflection transitioning was presented.

Kuruvayil and Palaniswamy (2022) proposed a ViT to increase the precision of facial-mask-aware FER. Firstly, the facial-mask-aware face parser model was used to enhance the robustness of identifying the unimpeded facial area of the mask images. Secondly, a novel FER decoder model was presented to dynamically reweight the importance of the obscured and non-obscured faces to achieve the optimum recognition of facial expressions.

Wang *et al.* (2022) proposed Emotion Recognition with Meta-Learning across Occlusion, Pose, and Illuminations (ERMOPI) technique for recognizing emotions from static images. This approach was powerful for the FER even in the presence of biased occlusions, different head postures, and different lighting levels. The main advantage of the approach has adopted less training data than previous studies while recognizing emotions.

Wang *et al.* (2023) designed a Cascade Regression-assisted Facial Fractalization (CRFF) technique to rebuild a clear, frontally, and expression-aware image from naturally taken images. The bilateral spatial connection between the non-frontal shape of the face and the pre-frontal equivalent was initially carried out in predicting a pre-frontal facial form. The CRFF model progressively adjusted with the front perspective to get effective results in the FER. To improve prediction efficiency, a variety of coefficients was incorporated with the proposed model. Active appearing model parameterization was used to distort the face and provide an immaculate visage for face texturing restoration. Training the generative algorithm on deliberately chosen clean-face collections has eliminated occlusions from source faces. The suggested technique was less computationally expensive than the current face restoration techniques which were appropriate for the dynamical study of expressions on the face.

El Sayed *et al.* (2023) presented an innovative framework for FER in-the-wild statistics to address the potential risks. Initially, the system has collected the face images into two uncertainty categories based on their characteristics. The Graph Convolution Networks (GCN) structure was used to derive geometrical clues, such as the association of Action Units (AUs) which aid in predicting the likelihood of the underpinning emotion labeling. The projected hidden label frequency and the provided label were combined to create emotional label dispersion. According to findings from experiments, the suggested framework outperformed than other approaches.

Liu *et al.* (2023a) introduced a Hybrid-CNN (HCNN) assisted with regional binary sequence to precisely identify characteristics, notably for masking faces. It became crucial to understand the fundamental feelings of frustration, joy, sorrow, shock, disdain, dissatisfaction, and terror. Two information sets, CK and CK+, were utilized to apply the suggested approach. Results have indicated that the developed model facial mask and recognizing feelings was 70.76% accurate for three sentiments. Results demonstrated significant improvements than current techniques.

Bellamkonda *et al.* (2023) suggested a Patch Attention Convolutional Vision Transformer (PACVT) to address the occlusion issue. The backbone CNN was employed to extract face map features, which were subsequently divided into various patches to retrieve local as well as global characteristics. The face patches were converted into visual token patterns and the ViT was used to record the relationships and connections among these visual

symbols on a global level. Tests were carried out on three commonly utilized expression databases and associated occlusion groups, and the findings showed that the suggested PACVT surpassed cutting-edge approaches for occluded FER.

Liang *et al.* (2023) developed a Component-based Ensemble Stacked CNN (CES-CNN) to address the issue of partially obscured faces. CES-CNN was used on reaction modules of specific face parts such as the pupils, eyebrows, nose, face, referral, and glabella in an online sub-domain. To get the highest accuracy in recognition, a max-voting-based collective classifier was employed to combine the subnets' judgments. The suggested CES-CNN was verified by doing tests on standard datasets and comparing its results to conventional frameworks. The outcomes of the experiments proved that the developed approach improved recognition precision significantly in comparison with current models.

Naveen *et al.* (2023) presented a Convolution-Transformer Dual Branch Network (CT-DBN) that used local and global face data to address real-world occlusions and head-pose variation in FER. The CT-DBN has a pair of branches. Then, a local-global feature combination component was presented, which would adaptively combine both features into mixed characteristics and describe their connection. The network, using the feature combination section, not only fused global and local characteristics in an adaptive weighted way, but it might additionally acquire the matching distinctive characteristics on its own. Research findings from four main facial expression datasets showed that the suggested CT-DBN surpassed current state-of-the-art approaches and delivered resilient efficiency under real-world conditions.

Li *et al.* (2024b) suggested a technique that included Hopfield networks, Deep Belief Networks (DBN), and Lanczos correction. Lanczos interpolation improved the quality of images and lowered resizing times. The Hopfield networks were used to extract characteristics, including expression of the face, even when there were occlusions. The DBN was used for representational learning, and the network was fine-tuned using DenseNet to adjust occluded face expressions. Numerous experiments with diverse datasets have been carried out to assess the developed approach. The outcomes showed that the recommended framework outperformed alternatives of occlusion management and expression detection accuracy.

Zakieldin *et al.* (2024) recommended a hybrid ViT model for FER. The improved attention module in the model retrieves the most important multiscale features from the image. The patch-dropping approach is utilized for this network to enhance the FER. The developed network only focused on the relevant features from the images and obtained precise outcomes in the emotion recognition framework.

Xiong *et al.* (2024) designed a Hybrid Vision Transformer with Temporal Convolution (ViTCN) for interpreting the emotions of the human. The real-world emotion of the human was recognized by the ViTCN model. It only adopts the single frame for emotion recognition thus reducing the time consumed by the emotion recognition process. The facial expression during the interview process was effectively determined by the proposed model.

Zhang *et al*. (2024) have developed a Context Transformer (CoT) utilized with the ViT and CNN model for the FER. Here, the variation among local and global features was analyzed by the recommended approach. Here, the parameters of the ViT model were adjusted to enhance the accuracy of emotion recognition by removing the occlusion and noise effects. The implemented approach can significantly perform the FER even in a complex environment.

ViT-Based FER Model

Li *et al*. (2024a) have developed a novel Three-Stream Vision Transformer-based Network with Sparse Sampling and Relabeling (SSRLTS-ViT) technique. Initially, the network was studied with micro-expressions in the optical components. Then, the sparse sampling method was introduced to connect the optical flow components through the images. This could expand sample capacity and accurately generate data variations. Finally, the relabeling mechanism was developed in the training data to minimize the impact of annotations, which has enhanced the system's recognition accuracy.

Xiong *et al*. (2024) proposed a new Poker Face Vision Transformer (PF-ViT) method to recognize and differentiate the diverse emotions in a static facial image by producing an identical poker face without the help of paired images. The developed method used vanilla vision transformers and their components were pre-trained in auto encoders in a large facial expression dataset to capture the facial information.

Ngwe *et al*. (2024) developed the Context Transformer (CoT) technique among Convolutional Neural Network (CNN) and ViT approaches to enhance the learning performance among the local features to maintain the consistency between local and global area feature expression. Here, the interference of light and noises was reduced with the help of an adaptive learning approach and effectively handled the parameters that have attained better and more reliable performance than other traditional methods.

Comparative Analysis With Recent FER Approaches

Chen *et al*. (2024) developed a Lightweight Patch and Attention Network based on the MobileNetV1 (PAtt-Lite) technique to enhance the performance of FER in diverse situations. Here, a truncated ImageNet-pre-trained

MobileNetV1 was used to effectively extract the optimal features of facial images. The overall performance outcomes have shown that the suggested framework has outperformed than other methods.

Gong *et al*. (2024) developed a dual subspace manifold learning approach with the help of a Graph Convolutional Network (GCN) to differentiate FER tasks. It has the ability to significantly treat node categorization issues and study the manifold representation by utilizing both Locality Preserving Projection (LPP) and Peak-piloted Locality Preserving Projection (PLPP) methods. Here, the LPP method has the ability to handle local similarity between data and the PLPP method, thus it has significantly maintained the locality among the expressions of non-peak and peak to maximize the performance. The combination of LPP and PLPP methods has enhanced the FER performance.

Liu *et al*. (2024) implemented an Enhanced Spatial-Temporal Learning Network (ESTLNet) to enhance the performance of facial expression recognition with the help of a Spatial Fusion Learning Module (SFLM) and a Temporal Transformer Enhancement Module (TTEM). Initially, the SFLM has attained a better spatial feature characterization via a deep learning module. Further, the TTEM method has accurately extract the optimal contextual expression features by an encoder. Finally, the experimental outcome of the implemented method has attained a better performance.

Tao and Duan (2024) developed a novel Graph Neural Network (GNN) method in the systematic process of FER in human visual perception. Initially, the facial region was divided into diverse parts with the help of region division method to generate the efficient facial features. Then, a human visual cognition strategy was developed to recognize the relationship among various parts of facial expression and it utilized six regions to select the optimal key features. Finally, this developed method was examined and it has achieved best characterization and identification ability.

Vick *et al*. (2007) developed a Hierarchical Attention Module (HAM) with progressive feature fusion for characterizing the FER. Here, diverse feature aggregation blocks were developed to handle gradient, global, local, high and low-level features. A HAM method was developed to significantly fuse the features and improve the discriminative features from the facial images.

Methodological Gaps With Recent FER Approaches

Focusing on the existing FER method, it consumes more duration for training and testing phases and it is difficult to capture various face movements and micro-expressions while considering the lightweight PAtt-Lite technique (Chen *et al*., 2024). In the GCN model, it does not have the ability to manage large variations in pose and facial behaviors. It is restricted for recognizing limited facial expressions in the training data (Gong *et al*., 2024). Inaccurate performance happens while extracting the features of temporal and spatial in the SFLM model. In this context, it consumes more time during the process of feature extraction (Liu *et al*., 2024). Also, it is difficult to obtain accurate emotions in the dataset images. It does not have the ability to handle low-quality facial images and is complex for measuring the input of large images to generate poor outcomes (Tao and Duan, 2024). The existing HAM method has a more critical and sensitive task for extracting features from one face to another (Vick *et al*., 2007).

Problem Statement

Conventional techniques for recognizing face characteristics did not fully utilize facial features, resulting in unsatisfactory identification outcomes due to issues such as partial blockage, significant facial distortion processing, and differences in occlusion between areas. Table 1 presents a comparison of features and challenges between conventional techniques and the occlusion-aware FER model.

The conventional approaches do not provide accurate results in the FER process due to the presence of occlusion like hair, scarf mask, etc., which causes certain difficulties in fully analyzing the facial features from the images. To fill this gap, this research adopts the face detection algorithm that separately detects the face from the obscured portion and also it provides details on individual portions of the face images, which helps to get precise outcomes in the FER.

The subtle movement in the human face does not reflect on the face images of the human so the expression in the face images is not captured by the conventional approaches. To solve this difficulty, an ROI extraction is performed in this research work that captures the specific area of the face images that is helpful to analyze the expression caused by subtle changes in the human face.

The feature of the face images is greatly hindered by the illumination effects, which greatly reduces the overall precision of the FER process. Here, a multiscale deep approach is designed that processes the features of the images in multiple scales to accurately identify the expression caused by the human face images.

The traditional frameworks still face difficulties during the dynamic face movements so it does not provide sufficient results in the FER. This difficulty is solved by the utilization of the parameter-tuned hybrid deep model as it continuously monitors the visible features in the dynamic environment and generates precise outcomes in the FER.

Table 1: Features and challenges of the traditional FER-based methods

Reference	Methodology	Features	Challenges
Yanga *et al.* (2022)	VTC-FER	• It can considerably recognize a larger variety of gestures in the human with minimum time	• It became more difficult and easier to identify occluded faces • It needs more datasets to perform the FER
Wang *et al.* (2020)	GCN	• It offers the highest recognition of face reliability in the visual landmarking phase	• It can estimate only the 3D facial contour and it is not suited for 2D marker identification
Kuruvayil and Palaniswamy (2022)	ERMOPI	• It can recognize different feelings being shown by many individuals in just one image	• It requires high duration of training performance
Hu *et al.* (2020)	SURF	• The execution period of the model is 2.38s quicker than that of more advanced ones	• It does not explore the issue of face occluded and expression identification over continual sceneries due to the border brightness shift not being sufficiently smooth
Dapogny *et al.* (2018)	RF	• It produces strong description value for expressing emotions and it offers precise results in the occluded FER.	• It does not detect the facial expression of the human if it is obscured by the paint
El Sayed *et al.* (2023)	HCNN	• It can easily recognize the fundamental emotions of anger, joy, sorrow, surprise, disdain, dissatisfaction, and fear	• It did not accurately recognize feelings and movements of the face from the top of one's head when someone wears a face mask
Dapogny *et al.* (2018	ICP	• It is not dependent on any detection method and might increase the 3D recognition system's occluded accuracy despite limited processing power	• It has a lack of identifying emotions for unskilled classes • The calculated aspect conversion has the potential to distort the observed information and thus lead to poor accuracy
Wang *et al.* (2022)	CRFF	• It fills the gap left by the absence of dynamical FER methods for geographical alignment	• Integrating time-aligning techniques with the suggested spatial aligning method fails to boost the flexible FER effectiveness • It cannot be used in any other healthcare detection process

Materials and Methods

Designed Occlusion Aware Facial Expression Recognition Approach

The occluded FER process becomes very difficult because of unconstrained reasons like illumination, pose and scale variation, etc. The conventional FER techniques are excellent for recognizing face expressions, but they cannot perform better on partly occluded faces. The GAN model is implemented to offer effective performance. But it is prone to fail and complicated to congregate. The Facial Actions Coding System (FACS) (Ali *et al*., 2001) is designed. It effectively detects the occluded faces and helps to enhance the accuracy. But it takes more time to process. It requires a greater quantity of trained data. The FER task is still challenging as it has large variations in the expressions of every person. Then, the Local Salient Independents of the Component Analysis (LS-ICA) (El Maghraby *et al*., 2014) technique is suggested. It automatically detects the occluded face without any difficulties. However, the frontal-shaped face evaluation cannot be processed by this model. However, they depend on the significant features that could be exploited by the tracking models and cannot detect 2D and 3D faces. The proficiency of the conventional approaches in detecting the expression of the human is varied and particularly the lower resolution images are not handled by the traditional model. Thus, an effective FER model is designed based on deep learning to tackle the conventional issues. Figure 1 deployed the structural view of the suggested occlusion-aware FER model.

The developed occlusion-aware FER model is used to effectively detect the expression of faces from occluded images. Face occlusions like a scarf, sunglasses, mask, etc., make it most challenging to detect the expression of an individual. The suggested framework can accurately detect facial expressions even in occluded faces. The developed model is also used in various fields since it effectively identifies the expression of the human, which helps to fulfill the necessity of the user.

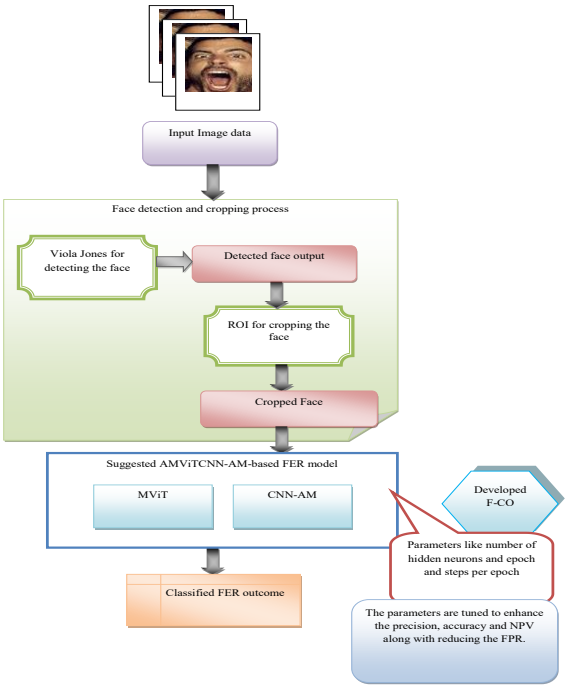


Fig. 1: Pictorial Depiction for the Suggested Occlusion Aware FER Model

The developed model is also capable of identifying suspicious behavior in individuals whose faces are partially obscured by masks. Furthermore, in the educational domain, the occlusion-aware FER system can be applied to monitor student engagement. By interpreting facial expressions in real-time, it enables educators to adapt their teaching strategies and provide more responsive instruction. The occluded objects are gathered from the standard sources. Then the gathered images are taken for further processing. In this case, the VJ approach is suggested, which helps to detect face of occluded images. The VJ face detection approach detects the face in the occluded images for capturing the most relevant information from the face, nose, and eyes of the images. Further, the non-face region from the occluded images is easily retrieved by the VJ and it reduces the overall computational burden in the FER. Because of the rapid processing capability, the VJ is used for the FER process. After the detection process, the detected images are passed to the ROI model. The ROI is utilized in this model to crop the detected images for accurate recognition. The ROI extraction process extracts the relevant region from the face images that provide more details about the facial expression. In addition, the ROI extraction is used to analyze the expression caused by the subtle changes in the human face. The ROI extraction process mainly focuses on the specific area of the face images, which reduces the burden and enhances the accuracy of the FER process. Further, the output provided by the ROI is given to the recommended AMViTCNN-AM-based FER model. The AMViTCNN-AM is developed using the ViT, CNN, and AM. The complex pattern in ROI is analyzed by the ViT and it also determines the connection among the individual components of the facial images. The visible parts of the face images are apparently focused by the CNN model and it also analyzes the temporal as well as spatial features of the gathered images to provide better outcomes in the FER. In AMViTCNN-AM, the input is provided in a multiscale manner so that it can effectively capture the variation in the expression of the human so it greatly maximizes the developed method's accuracy. During the recognition process, the suggested F-CO algorithm is used for parameter tuning. The suggested F-CO helps to resolve the complex tuning problems over several dimensions. The parameters such as epoch, steps per epoch, and hidden neuron number are accurately tuned by the recommended F-CO to maximize the precision, Negative Predictive Value (NPV), and accuracy along with minimizing the False Positive Rate (FPR). Finally, the suggested AMViTCNN-AM scheme provides FER outcomes.

Collection of Facial Expression Images

The experimental analysis is done by fetching data from a benchmark database to predict the accurate facial expression. Considering a small dataset in existing techniques, it does not have the ability to generate accurate details in the prediction framework. Also, the consideration of a smaller dataset is not effective for statistical inference and causesdata imbalance issues. Here, the developed method utilizes standard datasets for an effective facial expression recognition approach. However, both dataset includes a high volume of data to generate efficient results during the training process. The aggregation of images related to the FER approach is achieved by utilizing the below-mentioned datasets.

Occluded FER Dataset

The Occluded FER dataset is taken from the link of https://github.com/savya08/Occluded-Facial-Expression-Recognition. The required images are taken from the two standard databases available in this dataset. And they are RAF-DB and AffectNet.

RAF-DB Dataset

This dataset has 29672 real-world face images, 12 compound feeling classes, and 5 numbers of precise landmarks. The images are categorized according to age, gender, boundary box, location, and race. The height, width, and size of the images in this dataset are 388, 512, and 512, respectively.

AffectNet Dataset

This dataset contains 1250 facial emotions of the people. It is classified into 11 sections and they are non-face, none, anger, fear, sad, neutral, uncertain, contempt, disgust, surprise, and happy, and each has several images

DOI: 10.3844/jcssp.2025.3051.3080

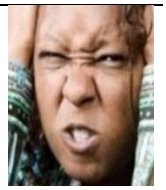
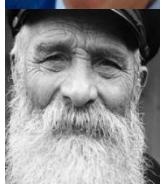
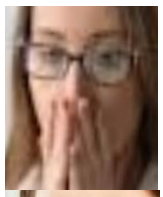
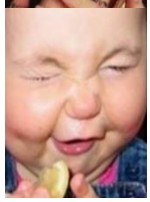
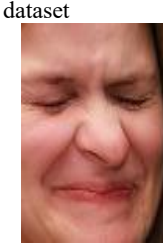
based on the emotions. A total of 420299 images are available here. The collected images are indicated by FS_g^{FR}. The size, height, and width of the images in this dataset are 15509, 512, and 512, respectively.

Advantages

In this research work, the occluded FER dataset is selected because it provides high-dimensional human face data for processing with large volume of facial expression images. It enables the facial expression identification and emotions using the occluded FER dataset, which is crucial for providing the accurate predictive models. This dataset ensures that the model train with large number of samples to make the model more robust and generalize on the unseen data in an effective manner. RAF-DB dataset contains a substantial number of images like basic and compound emotions, producing sufficient data for training process to improve the generalization of developed performance. It has the ability to capture images from natural settings, generating a more sensible characterization of different populations under different challenging conditions. Also, it can provide detailed annotations while the variation occurs in image size and imbalanced scenarios. Affect Net datas*et al*lows the researchers to validate and train the FER model with diverse facial expressions are captured in various people to generate with more reliable outcomes. It can effectively maximize the classification process in FER and easily recognize the facial expression.

Table 2 provides the sample images for the FER framework and here various emotion states caused by humans are visually presented. Here, neutral anger, surprise, fear, disgust, sadness, and happy are the most common emotions in human in their daily activity. Both of the datasets consist of expressions of various nationalities. The ages of the people available in the images range from 0 to 75 years and the skin color of the user varies from person to person.

Table 2: Sample frames collected from the online resources

Image Description	Image-1	Image-2	Image-3
RAF-DB dataset			
Anger			
Happy			
Sadness			
Fear			
Surprise			
Disgust			
Neutral			
AffectNet dataset			
Disgust			

Occlusion Aware FER Model using Facial Detector and Face ROI Cropping Process

Facial Detector With Viola Jones

The raw images FS_g^{FR} are passed to the input of the VJ. The VJ (Lin *et al*., 2007) system, developed in 2001 is the first object detection system to perform object recognition in an instantaneous manner.

While it may be programmed to recognize a range of object categories, face identification was the primary motivation for the development of the technology. It is typical to use this great effectiveness of this area and precision to detect the area around the face within an image to identify a substance of uncertain size. Various procedures in the VJ technique are:

Integral Images for Extracting the Features

Haar-like features are extracted from the entire image. These features, which are based on rectangular regions, are computed efficiently using an integral image representation.

Cascade Classifier

Several characteristics are effectively combined using a cascade algorithm. The term cascade suggests that its ultimate stage is formed of numerous smaller ones, which are used successively to achieve an ROI until the applicant fails to qualify for each of the phases that have been completed. After the cascade, the model may ultimately extract the non-face area and facial region as the input. The output is termed by VJ_l^{Dec}.

ROI-Based Face Region Cropping

VJ detection outcome VJ_l^{Dec} is undergone for the ROI cropping (Akbari *et al*., 2022). More effectively than scaling an image directly, a classifier is used to identify the items that are intriguing at various sizes. Several image scans ought to be carried out using various sizes to locate a substance with an uncertain dimension in the image itself. The ROI extraction on the face images is performed to capture the relevant features from the specific area of the face images. The ROI extraction process is most importantly used to capture the distinctive features of the face images that provide precise results in the FER. Lastly, the cropped image CI_l^{Crp} is obtained as the outcome.

Implemented F-CO-Based Parameter Tuning

Numerous algorithms are developed to generate a better performance of FER prediction framework. The consideration of conventional models faces several challenges. The existing optimization technique does not have the capability to classify the captured images in an uncontrolled manner. The existing optimization model needs large computational resources, and it takes maximum time for the detection and prediction process. Due to large and diverse facial expression datasets, the conventional algorithms are not suitable for handling these datasets. To overcome these issues, the CO algorithm is utilized in this research work. It has the ability to solve large-scale problems and provides timely optimal outcomes. Also, it can significantly reduce the computational resources and training duration. The accuracy of the CO model is also higher. However, there is a slight downfall in the traditional CO algorithm. CO generates the imbalance between the exploitation and exploration phases. Also, it cannot be applied to complex tuning issues and it cannot provide excellent results for the hybrid models. In order to surmount such problems, the conventional CO algorithm is enhanced by integrating a fitness-based mechanism that intends to propose a novel model as F-CO algorithm.

Novelty of F-CO Methods

F-CO is developed for tuning the attributes to enhance the deep approach. The suggested F-CO algorithm tends to optimize the essential hyper parameters such as steps per epoch, epoch, and hidden neurons to enhance the NPV, accuracy, and precision along with lowering the FPR. It supports avoiding premature convergences and entrapment in a local optimum. It offers a powerful, fast, and easy mechanism. The random number is upgraded with the support of the adaptive concept as shown in Eqs. 1 and 2. Here, the upgraded random number is calculated based on mean fitness values. If the mean fitness value $F_m > 1$, then the new random number is calculated using

Eq. (1), or else the new arbitrary number is determined using Eq. (2):

$$q = \frac{F_b}{F_w * F_m * Fitness(h)} \tag{1}$$

$$q = \frac{F_w * F_m * Fitness(h)}{F_b} \tag{2}$$

Here, the best-fit, mean-fit, and worst-fit are mentioned by the terms F_b, F_m and F_w. The term $Fitness(h)$ represents the fitness solution of the current iteration. The random value is termed by q. The stagnation in the local optimal condition is banned using upgraded random number and it also maximize the convergence speed, which helps to provide optimal solution.

Parameter Tuning Process Using F-CO Algorithm

Parameter tuning is the process of choosing optimal values using deep learning. The tuning of parameters is carried out through the use of developed F-CO optimization algorithm. The goal of parameter tuning tends to capture the better values that lead to provide reliable performance of the developed method. Here, the trial and error methods are utilized to find the optimal solution. The optimal values are achieved based on the iterations and this process is carried through until the optimal solution is attained. An iterative process is initiated to manage the model parameter for initiating the optimization process. Also, it permits the prediction accuracy through reducing the error rate and provides reliable generalization in the unseen data. The developed method's parameters are optimized for maximizing the value of accuracy thus, it helps to enhance the system in an effective manner. In this research work, it can control the learning process of the method, like the number of steps per epoch values, epoch count, and hidden neurons. Thus, the developed algorithm generates flexibility and robust performance and provides a better convergence speed.

CO (Wang *et al*., 2020): The CO's theoretical models are outlined below.

Searching Strategy

Cheetahs pursue food in two distinct manners: Forcefully patrolling the region when resting or walking. On the contrary, when the target is dispersed and engaged, choosing a mode of activity that consumes more power over scanning is preferable. Because of this, the cheetah can pick from a combination of both of those searching modes throughout hunting, according to the state of victims, health, and the protection of area. The term $W_{h,i}^{s}$ represents the nearby location of cheetah $h\ (h = 1,2,..,m)$ in configuration $i\ (i = 1,2,..,C)$, where m the total population is size of Cheetah and C denotes optimization issue size.

Next, utilizing the present location of every cheetah with a predetermined size of steps, an additional randomly generated Eq. (3) is recommended for upgrading the inventive location of every cheetah throughout the array:

$$W_{h,i}^{s+1} = W_{h,i}^{s} + q_{h,i}^{-1} \cdot \gamma_{h,i}^{s} \tag{3}$$

Here, the location of cheetah h in configuration i is represented by $W_{h,i}^{s+1}$ and $W_{h,i}^{s}$, correspondingly. The hunting time represents S, while the index s denotes the present quantity of the hunting period. The randomizing variable and several steps for Cheetahs h in the configuration i are represented by $\gamma_{h,i}^{s}$.

The commander as well as the victim is separated by some distance. By adjusting a few parameters in the best possible way, the dominant location is therefore chosen according to the target's position. If a hunting period is not interrupted, it is commonly anticipated that the commander and the target will get closer gradually, resulting in an upgraded position of authority. Thus, with any randomized variable and arbitrary size of steps, i.e., $q_{h,i}^{-1}$ and $\gamma_{h,i}^{s}$, the CO may successfully address the optimization issues in the right manner.

Sit and Wait Strategy

A cheetah could select to attack to get adequate proximity to its target to alleviate this process. Because of that, the cheetah stays in place with this posture and watches for its target to get closer. These steps can be used to model this behavior as shown in Eq. (4):

$$W_{h,i}^{s+1} = W_{h,i}^{s} \tag{4}$$

Here, the upgraded and present places of cheetah h in the configuration i are represented by the terms $W_{h,i}^{s+1}$ and $W_{h,i}^{s}$. This approach demands the CO to resist many cheetahs in an ensemble at once in order to improve hunting (find an improved alternative), which can help it prevent excessive convergence.

Attacking Strategy

Velocity and flexibility are two essential elements that cheetahs employ to assault victims. Every cheetah within the group has the ability to change positions depending on the positioning of the commander or nearby cheetah and the location of the target. Simply, every one of the cheetahs' striking strategies is defined statistically by Eq. (5):

$$W_{h,i}^{s+1} = W_{A,i}^{s} + q_{h,i} \cdot \alpha_{h,i}^{s} \tag{5}$$

DOI: 10.3844/jcssp.2025.3051.3080

Here, the term $W_{A,i}^{s}$ represents the prey's present spot in configuration i. The deciding element and interacting factor related to the cheetah in the i^{th} configuration is $q_{h,i}$ and $\alpha_{h,i}^{s}$. The term q can be upgraded by the adaptive concept as in Eq. (1,2). The term $W_{A,i}^{s}$ is utilized in Eq. (5) since when in hunting manner, cheetahs hurry to reach as near as feasible to their prey's location quickly using their optimum speed. Therefore, relying upon the prey's present location, it determines the new spot of h^{th} cheetah in a striking posture. The connection between two cheetahs or between a cheetah and a commander in the capture form is affected by the rotation variable $\alpha_{h,i}^{s}$ in the subsequent phase. The pseudo-code for the suggested F-CO algorithm and its flowchart are shown in Algorithm 1 and Table 2.

Algorithm 1: Developed F-CO algorithm

Input: Number of epochs, hidden neuron, and steps per epoch
Output: Optimal solution
Initialize the population
Initialize the iteration and population values
While $i < MaxIteration$ do
  Determine γ and α
    If $F_m > 1$
      Upgrade the random value q **by Eq. (1)**
    else
      Upgrade the random value q **by Eq. (2)**
      Perform the Searching phase in Eq. (3)
      Perform the Sit and Wait phase in Eq. (4)
      Perform the Attacking phase in Eq. (5)
    end
  end
end

Adaptive and Multiscale Vision Transformer-Convolutional Neural Network with Attention Mechanism for Facial Expression Recognition.

CNN

CNN (Wu *et al*., 2022) is a type of artificially intelligent neural network that specializes in generalization and employs convolutional operations to derive a high degree of results. The input, convolutional layer (C_L), fully-connected layer (F_L), pooled layer (P_L), and output are the basic components of the CNN framework.

Convolutional Layer (C_L)

A fundamental component of the CNN approach is C_L that uses kernels of convolution of various shapes and sizes to derive various local conceptual characteristics from the source image.

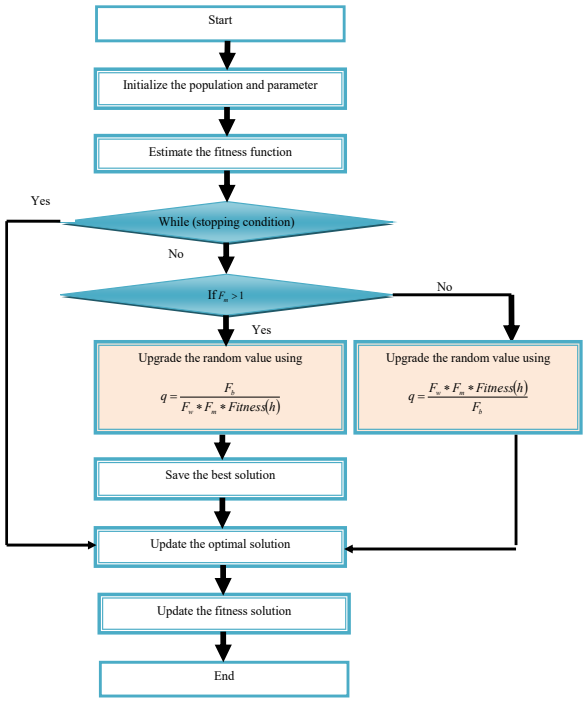


Fig. 2: Flowchart of the Developed F-CO

Assuming that, represents the set of input maps of features, Eq. (6) can be used to explain this convolution procedure:

$$W = e\left(\sum w * v_{h,i} + a\right) \tag{6}$$

Here, the symbol $*$ stands for the convolution calculation, $v_{h,i}$ is the kernels of convolution, a is the cumulative bias, and $e(\ )$ is the activation function, which is often the activated ReLU function.

Pooling Layers (P_L)

To decrease the dimension of the feature convolution maps and avoid the effects of dimension, then P_L is frequently used after the C_L. It can be expressed in writing as in Eq. (7).

$$W = e\left(Tf\left(w\right) \times v + a\right) \tag{7}$$

Here, the term $Tf(w)$ stands for the pooled function and w stands for the multiplicative distortions. Pooling with Max can maintain localized translation consistency as well as minimize the size of features, which enhances the classifier's resilience effectiveness.

Fully-Connected Layer (F_L)

The F_L is needed to transfer dispersed representations of features on the sample's tag field in order to realize the characteristic class once the pooling and convolution

processes have mapped with the image input to a hidden-layered space of features. Eq. (8) can be used to define F_L.

$$g(w) = e(vw + a) \tag{8}$$

Here the term v denotes the weighting level, w and $g(w)$ represents both outputs and inputs for the FC layers accordingly. The SoftMax, a function of activation for task classification, is indicated by e ( ).

Description of the Developed MViTCNN-AM

The cropped images are produced as the input in the MViT approach which supports to extract the input features. MViT (Thakur *et al*., 2023) enhances the performance of ViT according to two significant strategies. First, MViT acquires multiscale features in phases, beginning with smooth modeling of tiny regions and progressing to a high degree modeling of bigger areas in subsequent phases. This is done by maintaining a single resolution M throughout the entire network. Second, to significantly lower the computing expenditure of attention layers, MViT employs pooling attention O, which aggregates the temporal elements P J, and U.

Then attention mechanism is connected before the fully connected layer of the CNN for processing the significant data.

Attention Layer

To enhance the effectiveness and precision of FER, the attention mechanism is used with the CNN framework. A query string and an array of value-key pairs are the main components of the attention operation, which transforms these to an outcome that represents the total calculated of all entries. The resultant vector of the attention described in Eq. (9) for the input variable with key, value, and query value of size fm:

$$A(S, M, X) = Sftm\left(\frac{SM^{V}}{\sqrt{fm}}\right)X \tag{9}$$

Here, the terms M and X stand for the key and value matrix, and S stand for the query matrices.

Developed AMViTCNN-AM for Facial Expression Recognition

The cropped images CI_l^{Crp} are inputted to the designed AMViTCNN-AM model, which is effectively processed to provide better results. The input images are provided in the ViT network. The ViT model extracted the features and put them to the CNN for further processing; multiscale and attention are added in the implemented method to improve the accuracy. Finally, we obtained the classified FER image as the outcome. The CNNs employ convolutional operations to derive a high degree of conceptual characteristic from series or visuals and perform approaches like detecting the object, classification process as well as image recognition. Thus, we choose CNN to perform the process of the suggested occlusion-aware FER approach. The AMViTCNN-AM model integrates the strengths of MViT, CNN, and attention mechanism to maximize the robustness and accuracy of FER analysis. The AMViTCNN-AM model begins with the MViT component, where facial images transform sequences of tokens. This transformation allows the model to process facial features as structured sequences, enabling a more profound analysis of spatial relationships and dependencies within the face. By breaking down the images into tokens, the model can capture fine details and subtle nuances crucial for significant FER. Then, CNN architecture is crucial for extracting hierarchical features in the tokenized sequences obtained from MViT. CNN excels in feature extraction by applying convolutional filters across the data, enabling the implemented approach to detect features and patterns at diverse levels of abstraction. By leveraging the capabilities of CNN, the model can extract essential features from the tokenized sequences, enhancing its ability to recognize and differentiate various facial expressions. The fusion of MViT and CNN creates a synergistic relationship where MViT provides a structured understanding of facial features through tokenization, while CNN excels in feature extraction and hierarchical representation learning. This collaboration allows the model to benefit from the strengths of both architectures, leading to a more complete and detailed analysis of facial expressions.

Integration of Multi-Scale and Attention Mechanism

The integration of a multi-scale mechanism can significantly improve the FER performance. The multi-scale mechanism is highly utilized to provide patches of various sizes for FER. It has the ability to easily recognize the blur and illumination conditions of facial images. Also, it can effectively extract the optimal features from the different scales with limited duration to provide better and more reliable FER outcomes. To enhance the performance, the attention mechanism is further integrated to the developed method by enabling crucial parts of the input sequences. By assigning attention weights to diverse tokens, the model can prioritize significant features and filter out irrelevant information, improving its ability to capture essential facial expressions for accurate FER. This attention mechanism empowers the technique to adaptively attend to relevant details, enhancing its overall performance in recognizing different facial expressions with precision and efficiency. The adaptability of the AMViTCNN-AM model is a key feature that sets it apart, allowing it to vigorously adjust its internal representations based on the input data it receives. This adaptability enables the model to refine its understanding of facial expressions in real-time,

DOI: 10.3844/jcssp.2025.3051.3080

making it more resilient to variations, and making it more resilient to variations in lighting conditions, facial orientations, and other factors that can impact expression recognition accuracy. The graphical depiction of the proposed AMViTCNN-AM-assisted FER system is depicted in Fig. 3.

Steps per epoch, hidden neurons and epoch are the parameters tuned by the recommended F-CO to increase the NPV, accuracy, and precision and reduce the FPR. The objective function for the recommended approach is illustrated in Eq. (13):

$$FB_{Nf} = \underset{\left\{Cn_b^t, CnE_k^t, Eh_l^t,\right\}}{\arg\min} \left(\frac{1}{Yc + Sn + V} + FR \right) \tag{13}$$

Here, the terms Cn_b^t , CnE_k^t , and Eh_l^t define the number of hidden neurons of AMVCNN-AM lies in between $[5,255]$, the number of epochs of AMVCNN-AM stuck in between $[5,250]$, and the step per epoch of AMVCNN-AM in between $[5,255]$.

Choosing Specific Parameters and Impact on the Model's Performance

The parameters are highly important to generate optimal solutions in FER. Several works fail to manage large datasets with insufficient parameters and it needs large computational demand for the training phase. The ineffective parameters are not efficient for handling complex issues in optimization.

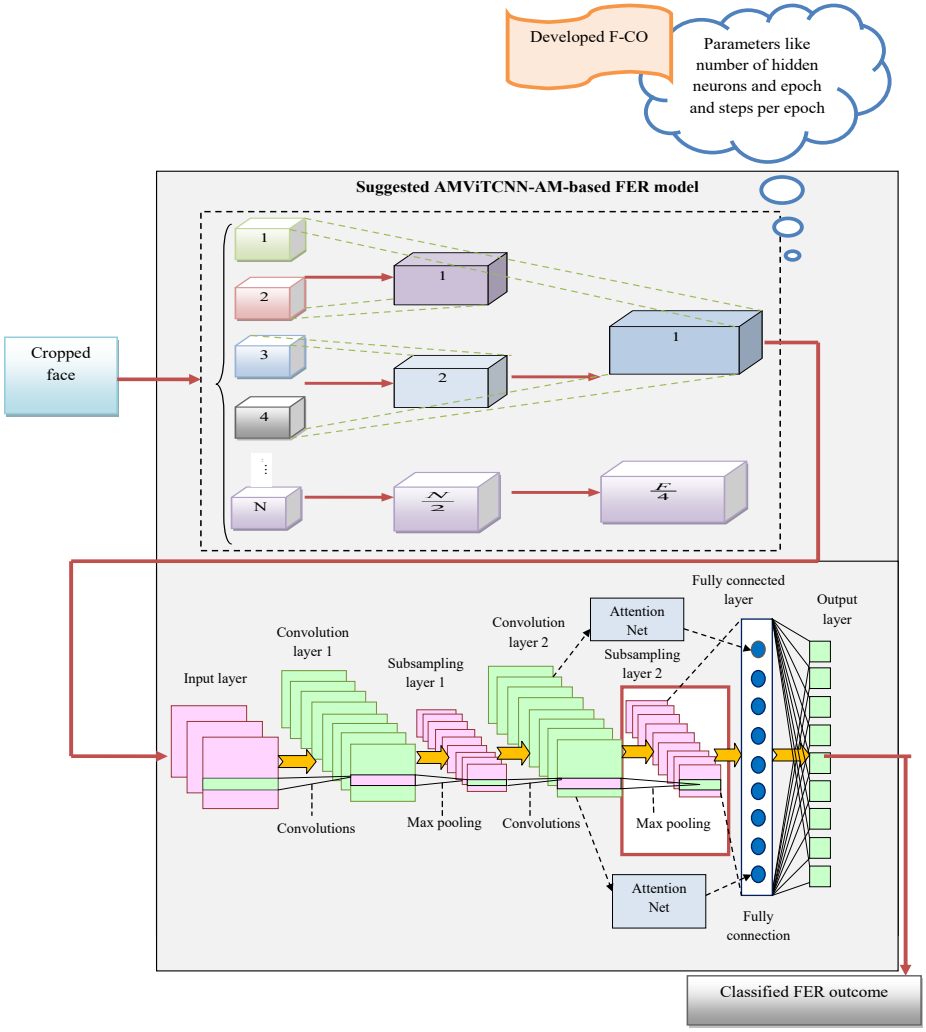


Fig. 3: Graphical Depiction of the Proposed AMViTCNN-AM-Assisted FER System

The process of tuning is significantly done by choosing the optimal parameters in this research work. In this implemented methodology, the process of tuning is highly reliant when considering the parameters like epoch, steps per epoch, and hidden neurons are optimized to choose better solutions. The hidden neuron count parameter can significantly generate the information about the reused data between

the layers and it has the ability to learn and handle complex problems to provide optimal solutions. The epoch parameter is used to strictly monitor the training process by tracking the performance, which helps to prevent overfitting issues in the developed framework. It can easily manage with large datasets to minimize the computational issues. The step per epoch parameter enhances the accuracy in a consistent manner during each training process. It has the ability to determine the training length and batch size. Focusing on these parameters efficiently improves the performance utilizing the F-CO algorithm. Also, the implemented model's convergence is enhanced through selecting the prescribed parameters and attains better performance.

Enhance Robustness and Accuracy of the Framework by Considering the Specific Parameters

Optimizing parameters can efficiently maximize the robustness and accuracy of the developed method. Higher accuracy enhances the developed FER and ensures reliable and better outcomes. It helps to reliable recognition performance in facial expression-related input data. Robust performance can manage the system in various lighting conditions, and variations. Utilizing optimization techniques helps to fine-tune these hyperparameters and it can accurately improve the robustness of the implemented method.

Eqs. 14-17 elaborate the expression for precision (Sn), accuracy $(\mathrm{Y}c)$, NPV (V), and FPR (FR).

Precision: Sn is employed to calculate the approach's efficiency and it is denoted in Eq. (14):

$$Sn = \frac{f}{f+g} \tag{14}$$

Accuracy: Y_c is determined based on the positive and negative detection experimental rates and it is represented in Eq. (15):

$$\mathrm{Y}c = \frac{(f+l)}{(f+l+g+k)} \tag{15}$$

NPV: V is calculated based on the total of every individual without detected facial expression in the observation and it is mentioned in Eq. (16):

$$V = \frac{g}{k+g} \tag{16}$$

FPR: FR is the measurement to denote the proportion of false detection to the overall negative detection outcomes and it is mentioned in Eq. (17):

$$FR = \frac{g}{g+l} \tag{17}$$

Here, the true and false positives are termed by f and k. The true and false negatives are termed as g and l, respectively.

Results and Discussion

Experimental Setup

The performance of the developed model was validated with diverse algorithms and approaches. The classifiers were CNN (Wu *et al*., 2022), ViT-CNN (Li *et al*., 2022), ResNet (Ye *et al*., 2021), Visual Geometry Group-16 (VGG-16) (Liu *et al*., 2022), and MobileNet (Rao *et al*., 2022) were taken to find the performance of the developed model. The algorithms were Mexican Axolotl Optimization (MAO) (Xie *et al*., 2021), Tuna Swarm Optimization (TSO) (Zhong *et al*., 2022), Beluga Whale Optimization (BWO) (Kotsia and Pitas, 2007) and CO (Wang *et al*., 2020) were also adopted for the comparison. The population used by the suggested scheme was 10, and the iteration was 50. The chromosome length was 3. The implemented model was implemented in Python. Here, the experimental details of the implemented FER technique are given in Table 3.

Performance Measures

Performance measures are adopted for validating the developed occlusion-aware FER model are as follows.

False Negative Rate (FNR): Rn is the percentage of positives that yield negative detection results with the process and it is mentioned in Eq. (18):

Table 3: Experimental details of the implemented FER method

Software Requirements	
Software	Pycharm
Version	3.11 and anaconda; version 3
Hardware Requirements	
RAM	8GB
Machine	Windows
Processor	i3
ROM	500GB
Version	11

$$Rn = \frac{k}{k+f} \tag{18}$$

False Discovery Rate (FDR): Dr is the percentage of false positives identified through the developed model among overall positive detection is computed and it is mentioned in Eq. (19):

$$Dr = \frac{g}{g+f} \tag{19}$$

Recall: CR calculates the accurate positive detection values in the total positive values and it is denoted in Eq. (20):

$$CR = \frac{f}{f+k} \tag{20}$$

Matthews's correlation coefficient (MCC): DC is the quality evaluation of binary classification in testing and it is mentioned in Eq. (21):

$$DC = \frac{f \times l - g \times k}{\sqrt{(f+g)(f+k)(l+g)(l+k)}} \tag{21}$$

Specificity: YsP is the proportion of negatives that are accurately detected and it is evaluated in Eq. (22):

$$YsP = \frac{l}{l+g} \tag{22}$$

F1-Score: $E1S$ is the quantification of accuracy in the specific examination and it is mentioned in Eq. (23):

$$E1S = \frac{2*f}{2*(f+g+k)} \tag{23}$$

Sensitivity: Sy is the proportion of positive rates that are accurately identified and it is mentioned in Eq. (24):

$$Sy = \frac{f}{f+k} \tag{24}$$

Cost Function Analysis

Figure 4 provides the cost function analysis of the suggested model. The developed F-CO-AMViTCNN-AM framework provided the cost function is 10.31, 23.07, 20, and 33.33% enhanced than CO-AMViTCNN-AM, MAO-AMViTCNN-AM, BWO-AMViTCNN-AM, and TSO-AMViTCNN-AM for dataset 1 and dataset 2. The outcomes showed that the recommended approach converges rapidly towards the best optimal solution with minimum iteration values. The rapid convergence of the recommended model also eliminates the overfitting issues in the FER process. In addition, recommended algorithms do not struggle in the local optimal because of the rapid convergence rate. The outcomes denote that the developed approach effectively understands the facial expression of the human even in the presence of the occlusion effects. The final outcome confirmed the recommended model's effectiveness over other conventional techniques.

Algorithmic Performance Analysis

The analysis of Dataset-1 and Dataset-2 for the suggested occlusion-aware FER system is depicted in Figs. 5 and 6. The dataset-1-based FDR analysis of the suggested F-CO-AMViTCNN-AM technique is 60.89, 78.57, 67.85, and 74.64% enhanced than CO-AMViTCNN-AM, MAO-AMViTCNN-AM, BWO-AMViTCNN-AM, and TSO-AMViTCNN-AM when the K-fold at 1. Here, the implemented method's accuracy rate is maintained in the same value at the end of the K fold value, which indicates that the performance of the F-CO-AMViTCNN-AM is not compromised at any K fold value and provides excellent results than the conventional model. From the accuracy analysis, it is observed that the implemented model effectively captures various emotional expressions of the human like happy, anger and sad with high accuracy. Moreover, the subtle variation in the face features is also determined by the developed model and generates effective outcomes in the expression recognition framework. The result proved that the efficacy of the designed occlusion-aware FER technique.

Classifiers-Based Performance Analysis

Analysis of the suggested occlusion-aware FER system is depicted in Figs. 7 and 8. The recommended F-CO-AMViTCNN-AM scheme gives the FDR analysis for dataset-2 is 82.22, 80, 65.21, 77.14, and 55.55% more than CNN, ViT-CNN, ResNet, VGG-16 and MobileNet when analyzing the K-fold at 3. Here, the FPR value of the F-CO-AMViTCNN-AM maintained fewer than 3 in all K fold values.

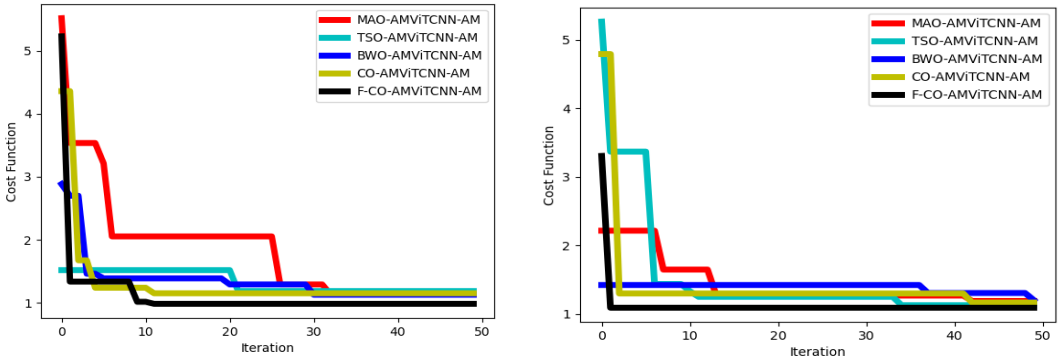
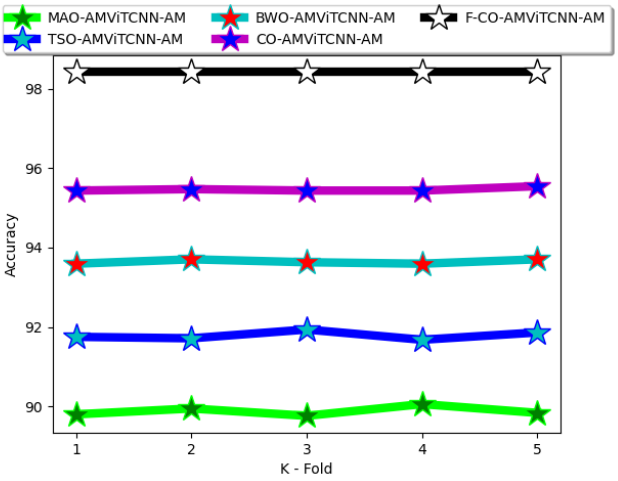
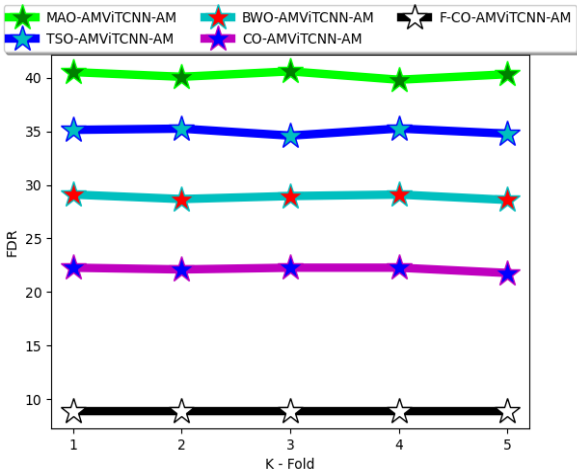


Fig. 4: Cost function analysis for the Suggested Approach regarding (a) Dataset-1 and (b) Dataset-2

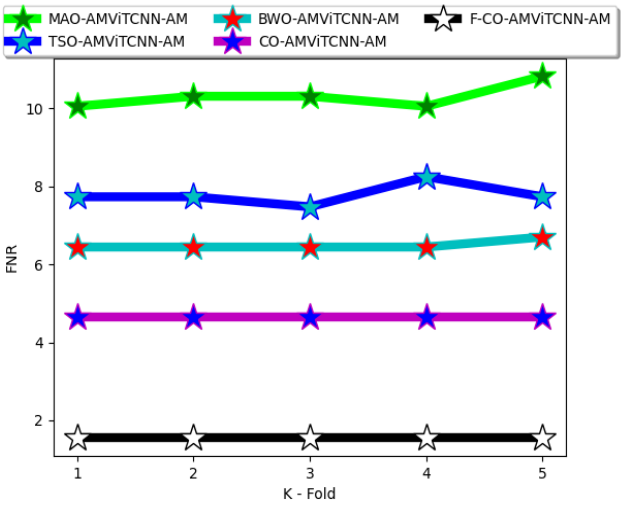
DOI: 10.3844/jcssp.2025.3051.3080



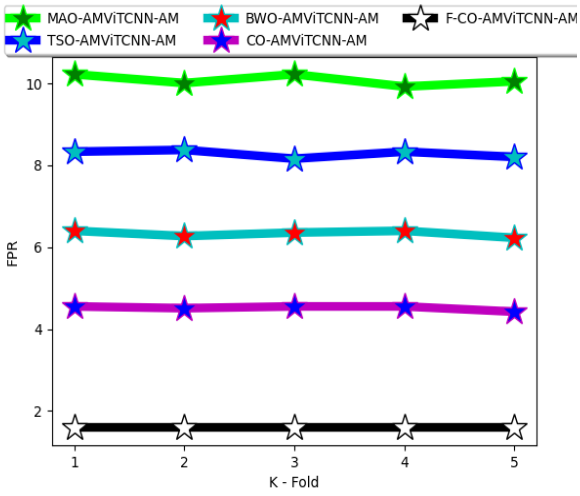
(a)



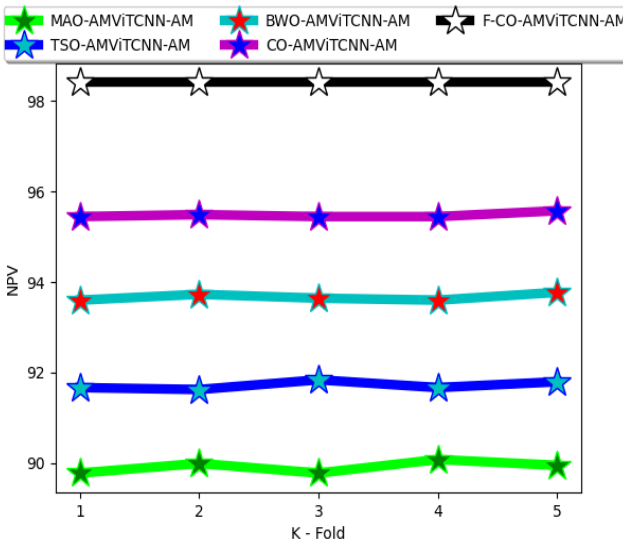
(b)



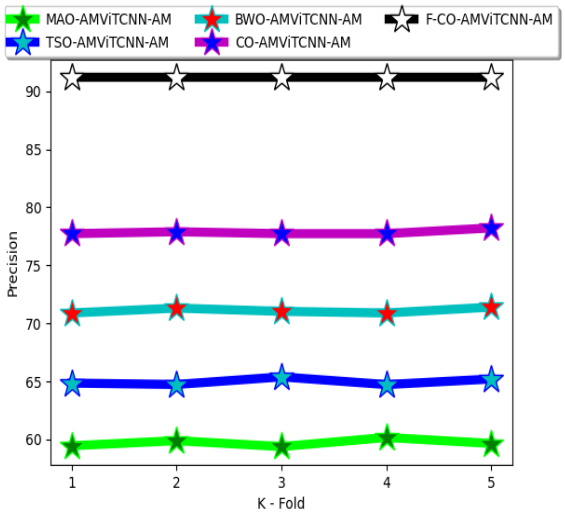
(c)



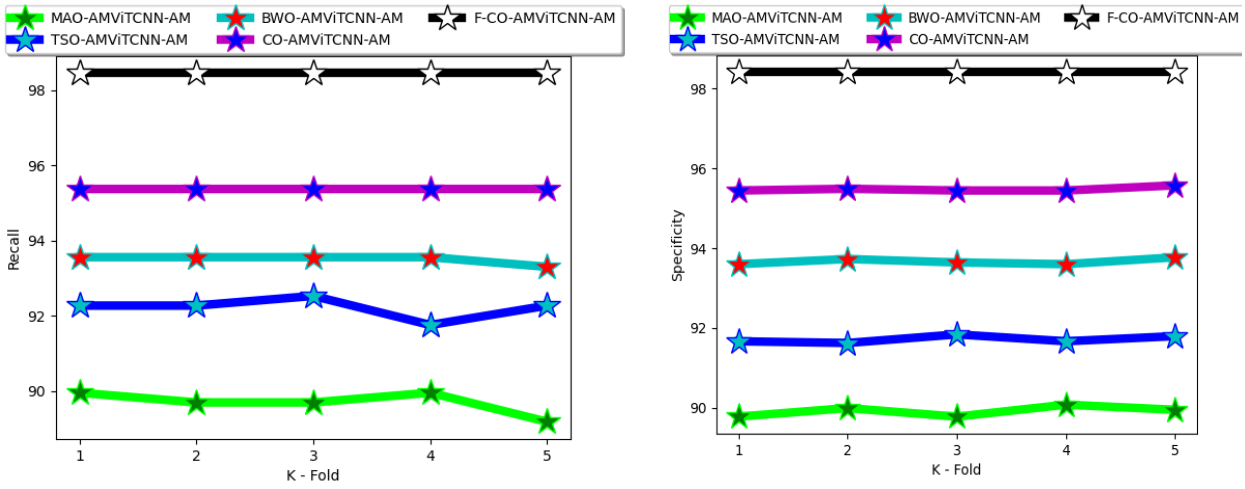
(d)



(e)

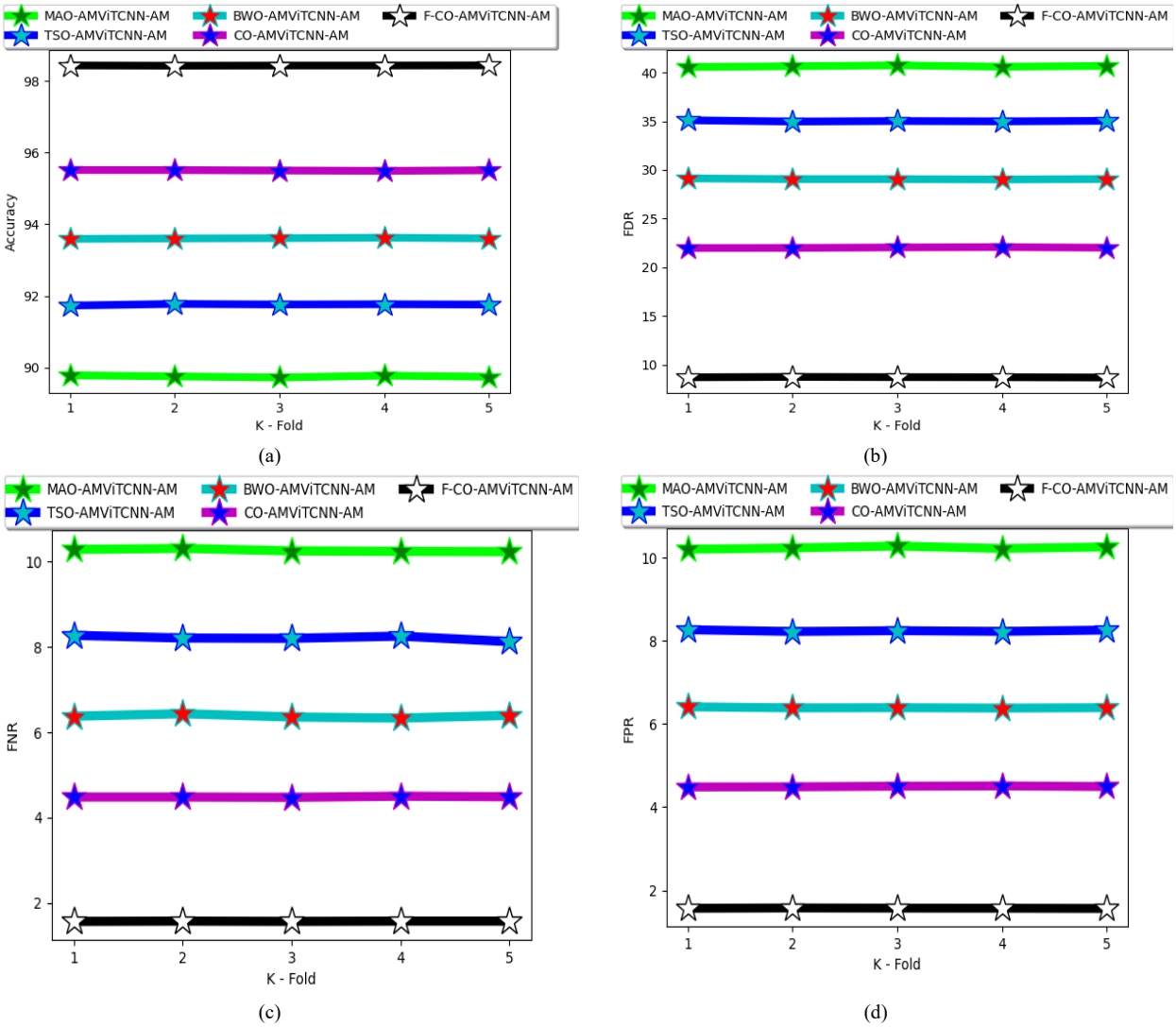


(f)



(g) (h)

Fig. 5: Analysis of the recommended model for dataset 1 in terms of (a) Accuracy, (b) FDR, (c) FNR, (d) FDR, (e) NPV, (f) Precision, (g) recall, and (h) Specificity



(a) (b)

(c) (d)

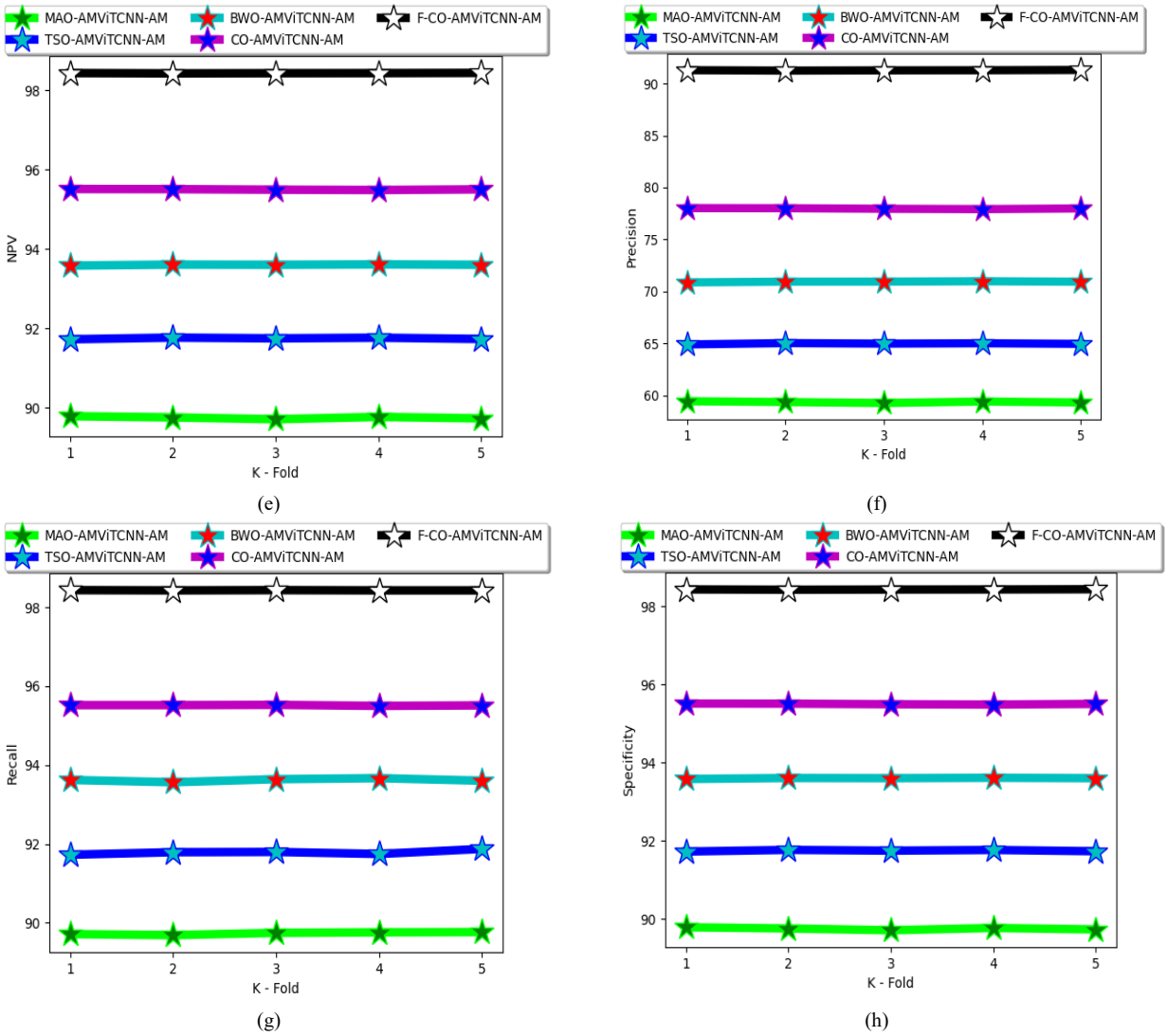
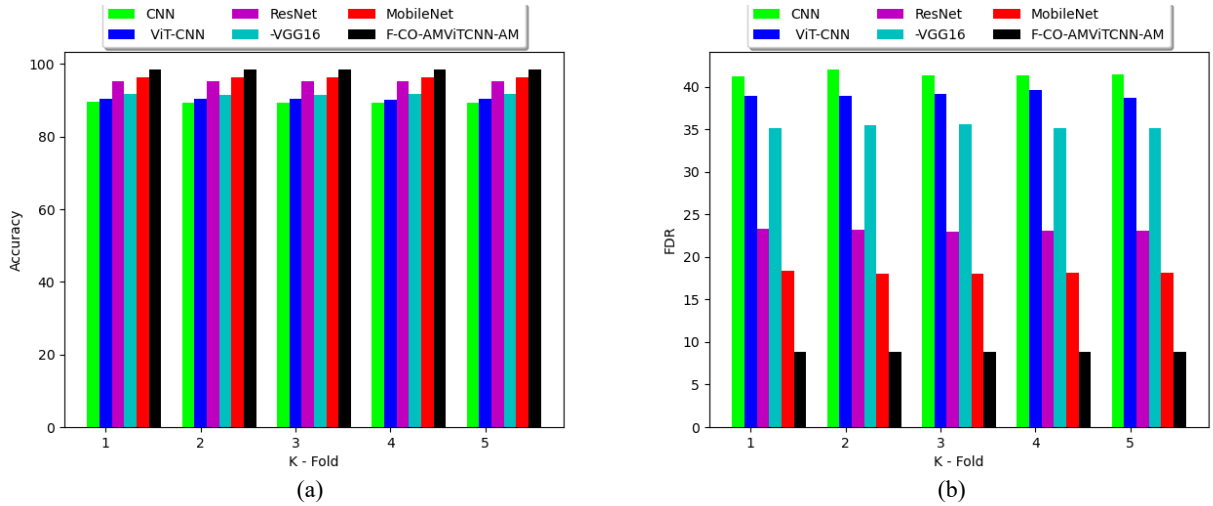


Fig. 6: Analysis of the recommended model for dataset 2 in terms of (a) Accuracy, (b) FDR, (c) FNR, (d) FDR, (e) NPV, (f) Precision, (g) recall, and (h) Specificity



DOI: 10.3844/jcssp.2025.3051.3080

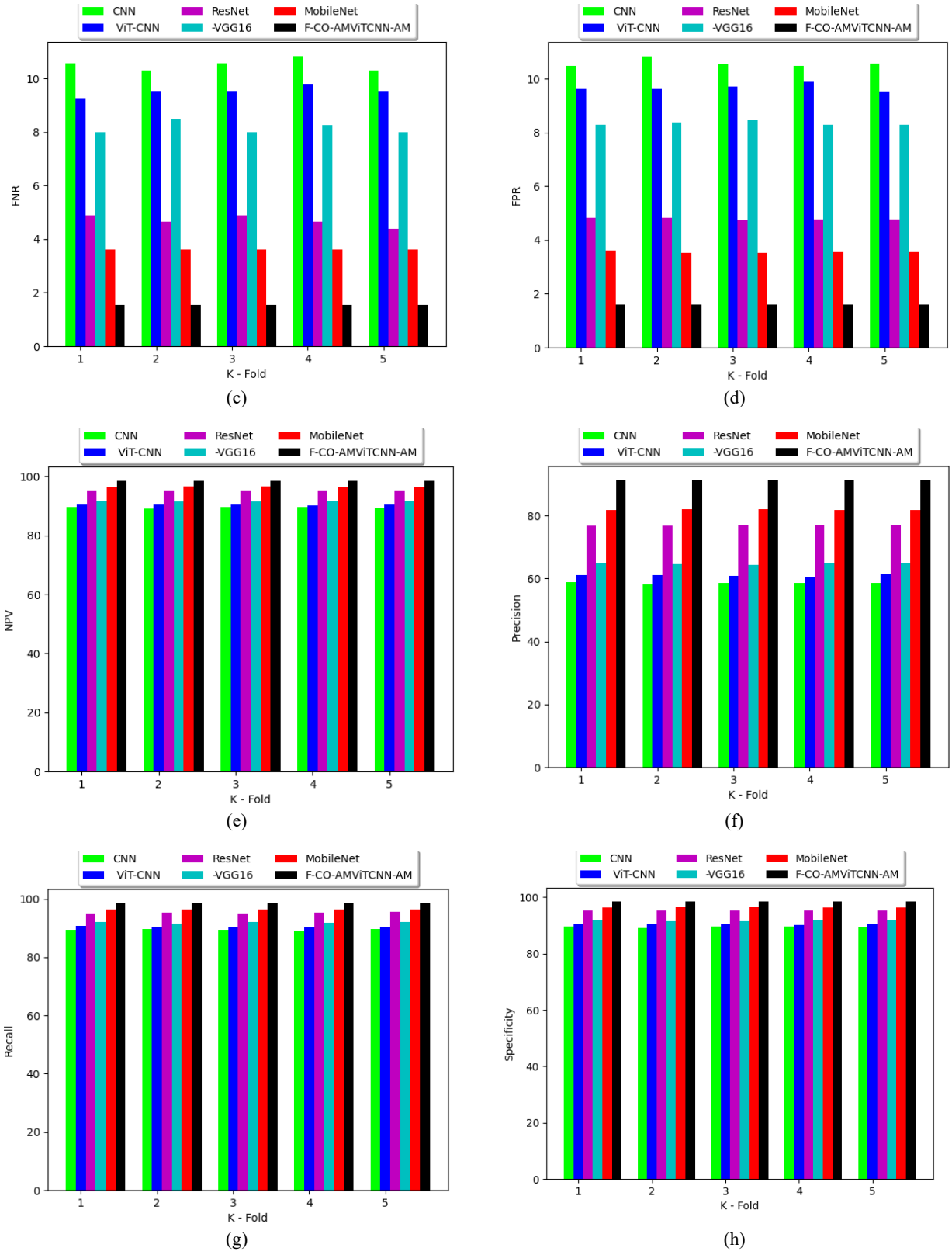


Fig. 7: Analysis of the proposed model for dataset 1 in terms of (a) Accuracy, (b) FDR, (c) FNR, (d) FDR, (e) NPV, (f) Precision, (g) Recall, and (h) Specificity

DOI: 10.3844/jcssp.2025.3051.3080

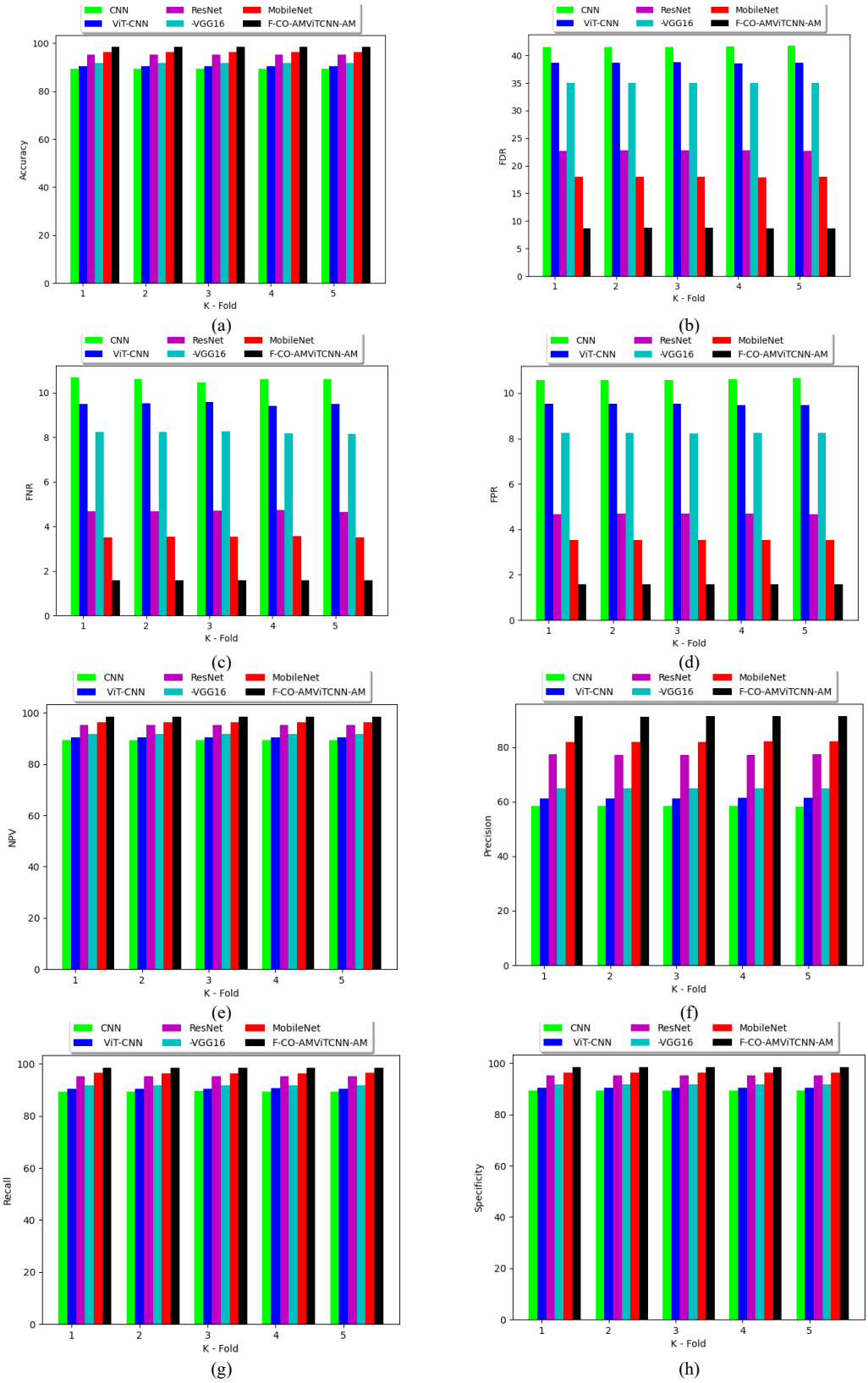


Fig. 8: Analysis of the proposed model for dataset 2 in terms of (a) Accuracy, (b) FDR, (c) FNR, (d) FDR, (e) NPV, (f) Precision, (g) Recall, and (h) Specificity

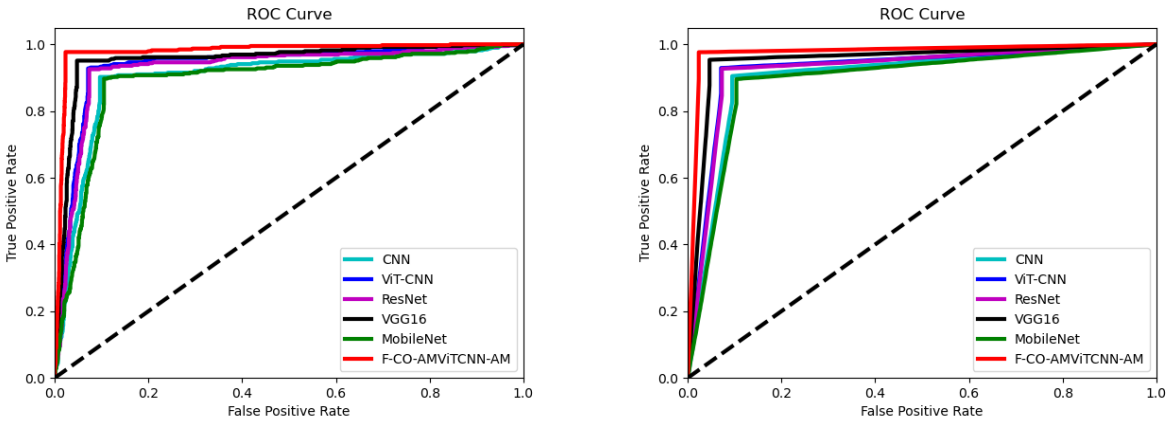


Fig. 9: ROC Performance for the Suggested Occlusion aware FER Model based on (a) Dataset-1 and (b) Dataset-2

The lower value of FER denotes that the proposed approach yielded maximum accuracy in the FER. The F-CO-AMViTCNN-AM can effectively handle the complexity of both datasets and provide the same results in the same dataset and it also verifies the ability of the designed framework in the FER process. Considering accuracy values in dataset-1, the conventional CNN method shows minimal accuracy value, this may significantly reduce the recognition performance. However, the developed F-CO-AMViTCNN-AM method attains a maximum accuracy rate than other existing methods. The maximum accuracy helps to maximize the entire FER performance framework.

ROC Analysis

The ROC-based performance for the developed occlusion-aware FER model is shown in Fig. 9. This helps to compare the effectiveness of diverse recognition techniques. Investigators may use ROC analysis to validate which statistical methods are the most efficient for recognizing the FER in humans.

The suggested F-CO-AMViTCNN-AM scheme provides the ROC is 11.11%, 5.26, 8.69, 4.16, and 12.35 more than CNN, ViT-CNN, ResNet, VGG-16, and MobileNet when analyzing the false positive rate at 0.4. Here, the ROC experiment is conducted using CNN, ViT-CNN, ResNet, and VGG-16 to demonstrate the effectiveness of the FER process. Here, the recommended model is situated above the threshold line, which denotes that the designed framework surpassed the existing methods in the FER. The developed model captures the multiscale facial features to capture efficient outcomes in the emotion recognition framework.

Classifiers-Based Comparison Analysis

Table 4 illustrates the comparison analysis of the suggested occlusion-aware FER model. The F1-Score value of the designed F-CO-AMViTCNN-AM is 33.35%, 29.6%, 11, 24.37 and 6.94 increased than CNN, ViT-CNN, ResNet, VGG-16, and MobileNet based on dataset-1. From the experimental outcome, the recall of the proposed model on the 1st dataset is 98.45%, which is maximum than the classical models. The outcomes showed that the developed approach is practically applicable to the FER as it effectively analyzes the useful features in the images using the deep model. Considering Table 4 (Dataset 1), the conventional CNN method shows a minimal 89.47% accuracy rate. This minimized accuracy value significantly reduces the recognition performance of FER. Also, it reduces the decision-making performance when dealing with large datasets. However, the developed F-CO-AMViTCNN-AM method attains a higher 98.41% accuracy value than several traditional frame works. This enhanced accuracy rate can efficiently enhance the decision-making process with the possibility of handling large datasets and neglecting the overfitting issues. Specificity evaluation executed in the developed method gained more reliable performance than the conventional approaches. Enhancing the specificity in the suggested approach helps to maximize the FER prediction process. In the specificity evaluation, the developed method shows higher FER prediction performance than 10 CNN, 8.7 ViT-CNN, 3.3 ResNet, 7.3 VGG16, and 2.05% MobileNet. In Table 4 (Dataset 2), the traditional CNN method shows a high FPR rate of 10.659. The conventional technique needs additional computational resources for predicting the FER performance. However, the implemented F-CO-AMViTCNN-AM approach can effectively reduce the FPR than the existing methods. Minimizing FPR has the ability to easily predict the FER in a restricted duration. This denoted the maximized performance of the developed F-CO-AMViTCNN-AM scheme.

Algorithms-Based Comparison Analysis

Table 5 illustrates the comparison analysis for the algorithms based on the suggested occlusion-aware FER

model. The FDR of the recommended F-CO-AMViTCNN-AM is 60.49, 78.6, 70.08, and 75.05% enhanced than CO-AMViTCNN-AM, MAO-AMViTCNN-AM, BWO-AMViTCNN-AM, and TSO-AMViTCNN-AM according to the dataset-2. This provides a negative outcome and optimally represents the absence of irrelevant emotions. Here, the developed model's recognition accuracy is high for the second dataset. The outcomes showed that the recommended approach provides robust performance in the FER as it effectively captures the complex features in the images. In addition, the proposed model is also useful to analyze the expression of humans who are undergone the FER. The developed model considers the sequence of the images to get efficient outcomes in the FER. Consider Table 5, the conventional MAO-AMViTCNN-AM method shows a low precision rate of 59.253. This may significantly reduce the system performance and enhance the computational overhead. However, the developed F-CO-AMViTCNN-AM method attains a maximum precision rate of 91.16% than different existing approaches. Thus, the enhanced precision value helps to maximize the prediction process of FER. It can maximize the decision-making and robustness of the system. Also, the NPV rate of the existing MAO-AMViTCNN-AM method shows a low 89.69% value This can maximize the duration of the detection and prediction process. It does not have the ability to handle large-size datasets. However, the developed method achieves a better value of 98.41% than other conventional methods. Consider Table 5 (Dataset 2), the developed method has attained a better MCC than the conventional 12% CO-AMViTCNN-AM, 38% MAO-AMViTCNN-AM, 20% BWO-AMViTCNN-AM, and 28% TSO-AMViTCNN-AM. The result findings proved that the designed framework attained the best performance.

Statistical Analysis

Table 6 provides the statistical analysis for the developed occlusion-aware FER approach. The mean of the proposed F-CO-AMViTCNN-AM is 14.16, 38.79, 16.54, and 15.02% enhanced than CO-AMViTCNN-AM, MAO-AMViTCNN-AM, BWO-AMViTCNN-AM, and TSO-AMViTCNN-AM when considering the dataset-1. The statistical analysis is used to find the effect of the F-CO in the AMViTCNN-AM.

The statistical outcomes attained that the developed approach has certain advantages in the FER and it classifies the diverse emotions especially anger and fear caused in humans. In addition, the designed model consumes a very minimum amount of training data to detect similar features in the human face images. It shows that the suggested scheme can offer a more accurate result in all measures.

Table 4: Overall performance validation on the recommended model over various models

Terms	CNN (Wu *et al*., 2022)	ViT-CNN (Thakur *et al*., 2023)	ResNet (Li *et al*., 2022)	VGG16 (Ye *et al*., 2021)	MobileNet (Liu *et al*., 2022)	F-CO-AMViTCNN-AM
Dataset-1						
Accuracy	89.470	90.464	95.287	91.753	96.429	98.417
Recall	89.691	90.464	95.619	92.010	96.392	98.454
Specificity	89.433	90.464	95.232	91.710	96.435	98.411
Precision	58.586	61.257	76.971	64.909	81.838	91.169
FPR	10.567	9.536	4.768	8.290	3.565	1.589
FNR	10.309	9.536	4.381	7.990	3.608	1.546
NPV	89.433	90.464	95.232	91.710	96.435	98.411
FDR	41.414	38.743	23.029	35.091	18.162	8.831
F1-Score	70.876	73.049	85.287	76.119	88.521	94.672
MCC	67.70	69.84	83.52	72.49	86.88	93.98
Dataset-2						
Accuracy	89.347	90.512	95.339	91.760	96.479	98.437
Recall	89.387	90.502	95.351	91.850	96.499	98.427
Specificity	89.341	90.514	95.337	91.745	96.475	98.439
Precision	58.292	61.392	77.315	64.966	82.023	91.309
FPR	10.659	9.486	4.663	8.255	3.525	1.561
FNR	10.613	9.498	4.649	8.150	3.501	1.573
NPV	89.341	90.514	95.337	91.745	96.475	98.439
FDR	41.708	38.608	22.685	35.034	17.977	8.691
F1-Score	70.566	73.158	85.391	76.103	88.674	94.734
MCC	66.26	69.85	83.83	72.79	87.70	93.99

Table 5: Algorithms-based Overall Performance Evaluation on the Recommended Model

Terms	MAO-AMViTCNN-AM (Rao *et al*., 2022)	TSO-AMViTCNN-AM (Xie *et al*., 2021)	BWO-AMViTCNN-AM (Zhong *et al*., 2022)	CO-AMViTCNN-AM (Akbari *et al*., 2022)	F-CO-AMViTCNN-AM
Dataset-1					
Accuracy	89.728	91.716	93.778	95.471	98.417
Recall	89.948	92.268	94.072	95.619	98.454
Specificity	89.691	91.624	93.729	95.447	98.411
Precision	59.253	64.738	71.429	77.778	91.169
FPR	10.309	8.376	6.271	4.553	1.589
FNR	10.052	7.732	5.928	4.381	1.546
NPV	89.691	91.624	93.729	95.447	98.411
FDR	40.747	35.262	28.571	22.222	8.831
F1-Score	71.443	76.089	81.201	85.780	94.672
MCC	67.36	72.69	78.56	83.97	93.78
Dataset-2					
Accuracy	89.770	91.800	93.608	95.512	98.439
Recall	89.767	91.734	93.565	95.525	98.440
Specificity	89.770	91.811	93.616	95.510	98.439
Precision	59.392	65.121	70.951	78.002	91.310
FPR	10.230	8.189	6.384	4.490	1.561
FNR	10.233	8.266	6.435	4.475	1.560
NPV	89.770	91.811	93.616	95.510	98.439
FDR	40.608	34.879	29.049	21.998	8.690
F1-Score	71.487	76.170	80.704	85.879	94.741
MCC	67.36	72.99	78.70	83.18	93.99

Table 6: Statistical Analysis of the Suggested Occlusion-Aware FER Model

Terms	MAO-AMViTCNN-AM (Rao *et al*., 2022)	TSO-AMViTCNN-AM (Xie *et al*., 2021)	BWO-AMViTCNN-AM (Zhong *et al*., 2022)	CO-AMViTCNN-AM (Akbari *et al*., 2022)	F-CO-AMViTCNN-AM
Dataset-1					
Median	2.054	1.187	1.292	1.152	0.985
Best	1.134	1.187	1.129	1.152	0.985
Standard deviation	0.878	0.164	0.376	0.630	0.600
Worst	5.515	1.520	2.884	4.359	5.228
Mean	1.842	1.327	1.351	1.314	1.127
Dataset-2					
Best	1.191	1.129	1.202	1.168	1.092
Standard deviation	0.341	0.840	0.057	0.690	0.310
Worst	2.215	5.263	1.423	4.790	3.305
Mean	1.435	1.521	1.391	1.418	1.136
Median	1.270	1.251	1.423	1.300	1.092

MCC and F1 Score Analysis

Figure 10 provides the MCC and F1 score analysis of the recommended F-CO-AMViTCNN-AM for the two datasets. The MCC and F1score metrics are used to validate the FER performance of the implemented method. Here, we compare the results of both metrics with the conventional methods. The developed model attained acceptable results in both metrics at the FER process since the developed model processes the input data at different scales so it generates optimal outcomes in the FER. Thus, the designed approach is suitable to find the expression of the people. Further, in the educational domain, the expression of the student is determined to find the perplex lesson of the human.

Computational Time Analysis of the Implemented Method Over Traditional Techniques

Table 7 deploys the analysis of the computational time of the implemented framework with traditional methods. Consider Table 7, the traditional TSO-AMViTCNN-AM model shows a maximum computation time of 21.36 (secs). It does not have the ability to easily detect the facial expression in large datasets. It takes more duration to process the performance. However, the developed F-CO-AMViTCNN-AM algorithm achieves a minimum time of 16.78 (secs) than other conventional algorithms. It can significantly enhance the detection and facial expression prediction framework without any interference.

DOI: 10.3844/jcssp.2025.3051.3080

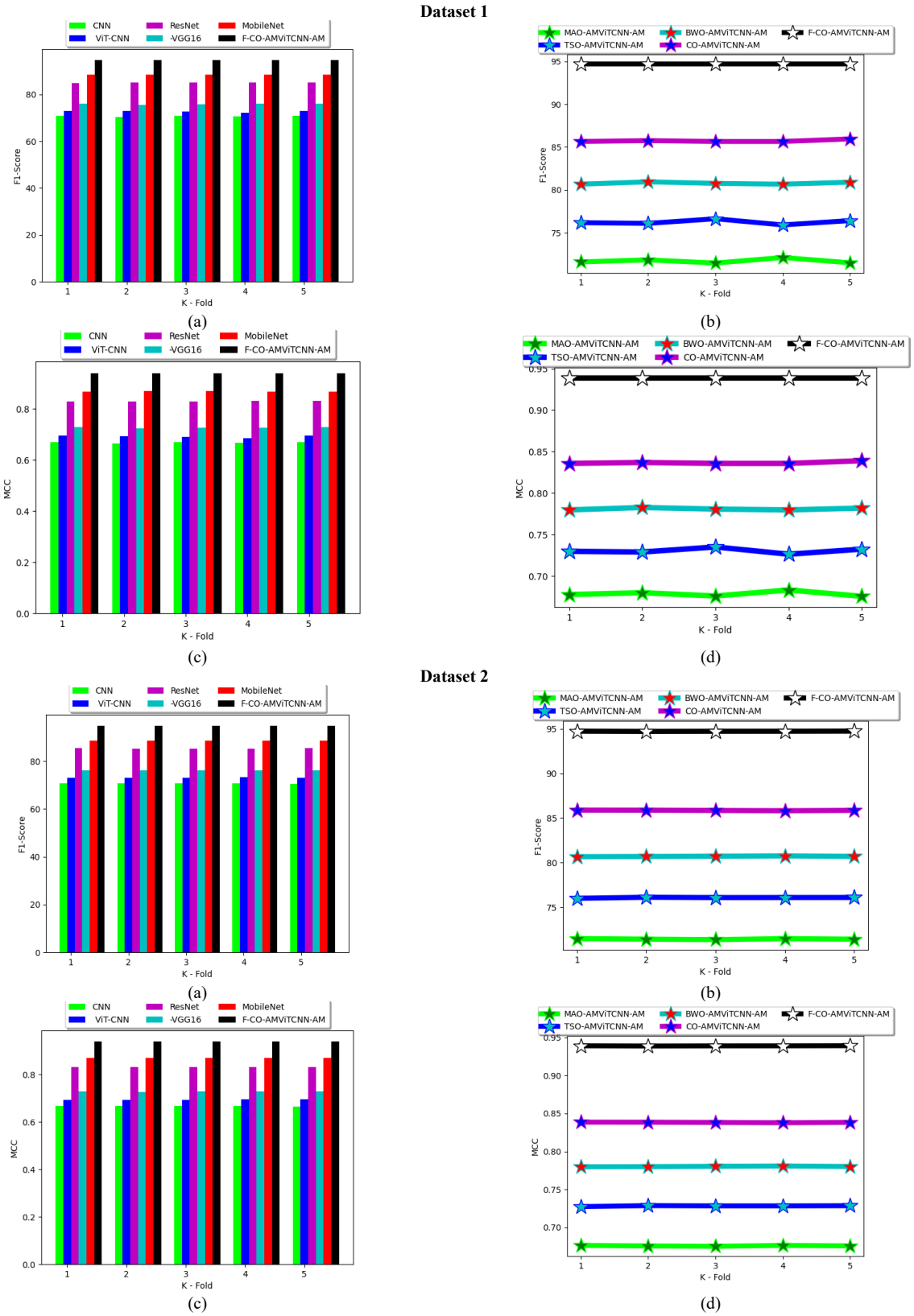


Fig. 10: MCC and F1 score analysis of the recommended F-CO-AMViTCNN-AM for the two datasets among various (a) Algorithms and (b) Techniques

Table 7: The analysis of computational time of the designed model and classifier

Algorithms	Dataset Time (seconds)
MAO-AMVITCNN-AM (Rao *et al*., 2022)	20.65
TSO-AMVITCNN-AM (Xie *et al*., 2021)	21.36
BWO-AMVITCNN-AM (Zhong *et al*., 2022)	17.85
CO-AMVITCNN-AM (Akbari *et al*., 2022)	18.75
Developed F-CO-AMViTCNN-AM	16.78
Classifier analysis	
CNN (Wang *et al*., 2020)	22.63
VIT-CNN (Thakur *et al*., 2023)	20.32
ResNet (Li *et al*., 2022)	19.54
VGG16 (Ye *et al*., 2021)	17.52
Developed F-CO-AMViTCNN-AM	16.78

Therefore, the efficiency of the developed method is more reliable than diverse FER prediction frameworks.

Analysis of Computational Complexity of the Implemented Approach With Conventional Methods

Table 8 shows the developed method's computational complexity analysis over traditional methods. In Table 8, the population number is represented as N_{pop}, and the maximum iterations is mentioned as *Iter*. Also, the chromosome length is denoted as *Chlen*. Here, these mentioned hyperparameters help to optimize the fitness rate in the optimization method to improve the implemented approach's performance.

Table 8: Computational complexity of the designed framework

Algorithms	Computational Complexity
MAO-AMViTCNN-AM (Rao *et al*., 2022)	$O[Iter+2+N_{pop}+3+Chlen+2]$
TSO-AMViTCNN-AM (Xie *et al*., 2021)	$O[Iter+3+N_{pop}+2+Chlen+1]$
BWO-AMViTCNN-AM (Zhong *et al*., 2022)	$O[Iter+2+N_{pop}+2+Chlen+1]$
CO-AMViTCNN-AM (Akbari *et al*., 2022)	$O[Iter+2+N_{pop}+1+Chlen]$
Developed F-CO-AMViTCNN-AM	$O[Iter+N_{pop}+Chlen]$

Statistical Significance Testing

Table 9 displays the statistical significance of Friedman aligned ranks at 0.005 significance level of testing. The Friedman-aligned rank test is a highly utilized nonparametric method for the comparison of classifiers with diverse datasets.

It can significantly validate the equality of the joint distribution and examine the rank sums. When dealing with different groups of data, this test can efficiently provide better robust performance and outcomes. It can detect significant facial emotion differences with higher sensitivity. Also, it does not need any assumption in the distribution of data to generate reliable outcomes.

Table 9: Statistical significance testing of the developed method

Comparison	Statistic	Adjusted p-value	Result
MAO-AMVITCNN-AM (Rao *et al*., 2022) vs Developed F-CO-AMVITCNN-AM	1.78885	0.73638	H0 is accepted
BWO-AMViTCNN-AM (Zhong *et al*., 2021) vs CO-AMViTCNN-AM (Akbari *et al*., 2022)	0.44721	1	H0 is accepted
BWO-AMViTCNN-AM (Zhong *et al*., 2022) vs MAO-AMViTCNN-AM (Rao *et al*., 2022)	1.34164	1	H0 is accepted
TSO-AMViTCNN-AM (Xie *et al*., 2021) vs CO-AMViTCNN-AM (Akbari *et al*., 2022)	0.44721	1	H0 is accepted
TSO-AMViTCNN-AM (Xie *et al*., 2021) vs Developed F-CO-AMVITCNN-AM	1.34164	1	H0 is accepted
TSO-AMViTCNN-AM (Xie *et al*., 2021) vs MAO-AMViTCNN-AM (Rao *et al*., 2022)	0.44721	1	H0 is accepted
BWO-AMViTCNN-AM (Zhong *et al*., 2022) vs TSO-AMViTCNN-AM (Xie *et al*., 2021)	0.89443	1	H0 is accepted
BWO-AMViTCNN-AM (Zhong *et al*., 2022) vs Developed F-CO-AMVITCNN-AM	0.44721	1	H0 is accepted
MAO-AMViTCNN-AM (Rao *et al*., 2022) vs CO-AMViTCNN-AM (Akbari *et al*., 2022)	0.89443	1	H0 is accepted
Developed F-CO-AMVITCNN-AMvs CO-AMViTCNN-AM (Akbari *et al*., 2022)	0.89443	1	H0 is accepted

Comparison With Recent Methods

Table 10 shows the comparative analysis of the implemented framework over recent state-of-the-art frameworks. Here, different performance measures are used to validate the effectiveness of the designed FER prediction performance. In Table 10 (Dataset 1), the conventional GRO method shows a minimal accuracy rate of 89.64%, which may gradually reduce the FER performance at different expressions, and pose

variations. However, the developed F-CO-AMViTCNN-AM method attains a maximum accuracy rate of 98.42% than other conventional methods which can effectively extract the relevant features and differentiate the emotions with large datasets at a limited duration. It has the ability to capture dynamic facial movements without any interference. Considering.

Table 10: Comparative analysis of the implemented framework with recent state-of-the-art-approaches

Measures/Methods	SSRLTS-ViT (Zhang *et al*., 2024)	PF-ViT (Li *et al*., 2024a)	CoT- CNN based ViT (Xiong *et al*., 2024)	Developed F-CO-AMViTCNN-AM
Dataset 1				
Accuracy	89.64	92.33	91.82	98.42
Recall	81.22	85.76	84.88	98.45
Specificity	94.54	96.01	95.73	98.41
Precision	89.65	92.33	91.82	91.17
FPR	5.46	3.99	4.27	1.59
FNR	18.78	14.24	15.12	1.55
NPV	89.64	92.33	91.82	98.41
FDR	10.35	7.67	8.18	8.83
F1-Score	85.23	88.92	88.21	94.67
MCC	77.50	83.21	82.11	93.98
Dataset 2				
Accuracy	88.62	93.20	90.54	98.44
Recall	98.77	99.30	99.00	98.43
Specificity	42.84	56.87	47.95	98.44
Precision	88.63	93.19	90.54	91.31
FPR	57.16	43.13	52.05	1.56
FNR	1.23	0.70	1.00	1.57
NPV	88.50	93.23	90.46	98.44
FDR	11.37	6.81	9.46	8.69
F1-Score	93.42	96.15	94.58	98.73
MCC	56.65	69.67	61.67	93.99

Table 10 (Dataset 2), the performance validation of the implemented F-CO-AMViTCNN-AM method is improved by 97% GRO, 96% HCO, 97 CFO, and 95% POA in terms of FPR analysis. The designed framework has a minimum FPR rate than other traditional approaches thus it generates better and more reliable recognition. It efficiently enhances the FER performance. In the F1-score evaluation (Dataset 2), the developed mechanism attained maximum FE recognition performance than 5.68% GRO, 2.68% HCO, 4.38% CFO, and 1.44% POA respectively. Attaining maximum performance in the developed approach helps to minimize the error in the suggested method to enhance the FER performance. Thus, the analysis showed that the recommended method accomplished superior performance for recognizing facial expressions than the classical methods.

Conclusion

The developed occlusion-aware FER model was used for detecting the expression of faces from occluded images. Initially, the raw images were gathered in the standard databases. The collected images were given to VJ for face detection, and then the detected images of the faces were put to the ROI to crop the images. Next, the cropped face images were passed to AMViTCNN-AM for FER. The parameters, such as epoch, steps per epoch, and the hidden neurons number were tuned by the recommended F-CO to enhance the NPV precision, and accuracy and also it reduces the FPR value. Finally, the implemented AMViTCNN-AM-based FER model provided the classified results. The accuracy of the suggested F-CO-AMViTCNN-AM model was 98.43%, which was higher than the traditional models. The efficacy of the implemented method was explored with existing techniques to validate the best result. The outcome showed that the developed occlusion-aware FER method successfully recognized the expressions from the occluded faces more than cutting-edge models. In future, real-time facial emotion recognition will be implemented using advanced approaches and more classes of emotions will be also determined in future research work with the help of an advanced model.

Handling Occlusions in FER

In generally, occlusions are presented in objects. Occlusion refers to the circumstance where the person's face parts are blocked and obscured through hair, glasses, masks, and hands; handling these occlusion objects is complex for significantly recognizing their facial expression while missing visual details in the occluded places. This occlusion image can highly impact the robustness and accuracy of the FER system. However, the developed F-CO-AMViTCNN-AM method can efficiently reduce the occlusion in an object with the help of the attention mechanism.

Acknowledgment

I would like to express my very great appreciation to the co-authors of this manuscript for their valuable and constructive suggestions during the planning and development of this research work.

Funding Information

This research did not receive any specific funding.

Author's Contributions

All authors have made substantial contributions to conception and design, revising the manuscript, and the final approval of the version to be published. Also, all authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Ahlawat, D., & Nehra, V. (2017). An efficient hybrid PC-SIFT-based feature extraction technique for face recognition. *International Journal of Signal and Imaging Systems Engineering*, *10*(5), 237. https://doi.org/10.1504/ijsise.2017.087766

Akbari, M. A., Zare, M., Azizipanah-abarghooee, R., Mirjalili, S., & Deriche, M. (2022). The cheetah optimizer: a nature-inspired metaheuristic algorithm for large-scale optimization problems. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-14338-z

Aleksic, P. S., & Katsaggelos, A. K. (2006). Automatic Facial Expression Recognition Using Facial Animation Parameters and Multistream HMMs. *IEEE Transactions on Information Forensics and Security*, *1*(1), 3–11. https://doi.org/10.1109/tifs.2005.863510

Ali, H. B., Powers, D. M. W., Leibbrandt, R., & Lewis, T. (2011). *Comparison of Region Based and Weighted Principal Component Analysis and Locally Salient ICA in Terms of Facial Expression Recognition*. 81–89. https://doi.org/10.1007/978-3-642-22288-7_7

Bellamkonda, S., Gopalan, N. P., Mala, C., & Settipalli, L. (2023). Facial expression recognition on partially occluded faces using component based ensemble stacked CNN. *Cognitive Neurodynamics*, *17*(4), 985–1008. https://doi.org/10.1007/s11571-022-09879-y

Chang, K. I., Bowyer,K. W., & Flynn, P. J. (2006). Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1695–1700. https://doi.org/10.1109/tpami.2006.210

Chen, D., Wen, G., Li, H., Chen, R., & Li, C. (2023). Multi-Relations Aware Network for In-the-Wild Facial Expression Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(8), 3848–3859. https://doi.org/10.1109/tcsvt.2023.3234312

Chen, J., Shi, J., & Xu, R. (2024). Dual subspace manifold learning based on GCN for intensity-invariant facial expression recognition. *Pattern Recognition*, *148*, 110157. https://doi.org/10.1016/j.patcog.2023.110157

Colombo, A., Cusano, C., & Schettini, R. (2011). Three-Dimensional Occlusion Detection and Restoration of Partially Occluded Faces. *Journal of Mathematical Imaging and Vision*, *40*(1), 105–119. https://doi.org/10.1007/s10851-010-0252-0

Dapogny, A., Bailly, K., & Dubuisson, S. (2018). Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection. *International Journal of Computer Vision*, *126*(2–4), 255–271. https://doi.org/10.1007/s11263-017-1010-1

Devi, B., & Preetha, M. M. S. J. (2025). Facial emotion recognition using convolutional neural network based krill head optimisation. *Expert Systems*, *42*(1). https://doi.org/10.1111/exsy.13376

Eleftheriadis, S., Rudovic, O., & Pantic, M. (2015). Discriminative Shared Gaussian Processes for Multiview and View-Invariant Facial Expression Recognition. *IEEE Transactions on Image Processing*, *24*(1), 189–204. https://doi.org/10.1109/tip.2014.2375634

El Maghraby, A., Abdalla, M., Enany, O., & Y. El Nahas, M. (2014). Detect and Analyze Face Parts Information using Viola- Jones and Geometric Approaches. *International Journal of Computer Applications*, *101*(3), 23–28. https://doi.org/10.5120/17667-8494

El Sayed, Y., ELSayed, A., & Abdou, M. A. (2023). An automatic improved facial expression recognition for masked faces. *Neural Computing and Applications*, *35*(20), 14963–14972. https://doi.org/10.1007/s00521-023-08498-w

Geetha, K. P., Sundaravadivelu, S., & Singh, N. A. (2009). Facial expression recognition using intelligent optical neural networks. *International Journal of Signal and Imaging Systems Engineering*, *2*(3), 141. https://doi.org/10.1504/ijsise.2009.033726

Gong, W., Qian, Y., Zhou, W., & Leng, H. (2024). Enhanced spatial-temporal learning network for dynamic facial expression recognition. *Biomedical Signal Processing and Control*, *88*, 105316. https://doi.org/10.1016/j.bspc.2023.105316

Haq, H. B. U., Akram, W., Irshad, M. N., Kosar, A., & Abid, M. (2024). Enhanced Real-Time Facial Expression Recognition Using Deep Learning. *Acadlore Transactions on AI and Machine Learning*, *3*(1), 24–35. https://doi.org/10.56578/ataiml030103

Hu, K., Huang, G., Yang, Y., Pun, C.-M., Ling, W.-K., & Cheng, L. (2020). Rapid facial expression recognition under part occlusion based on symmetric SURF and heterogeneous soft partition network. *Multimedia Tools and Applications*, *79*(41–42), 30861–30881. https://doi.org/10.1007/s11042-020-09566-2

Huang, W., Zhang, S., Zhang, P., Zha, Y., Fang, Y., & Zhang, Y. (2022). Identity-Aware Facial Expression Recognition Via Deep Metric Learning Based on Synthesized Images. *IEEE Transactions on Multimedia*, *24*, 3327–3339. https://doi.org/10.1109/tmm.2021.3096068Kotsia, I., & Pitas, I. (2007). Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE Transactions on Image Processing*, *16*(1), 172–187. https://doi.org/10.1109/tip.2006.884954

Kumar, A., & Kumar, A. (2025). Human emotion recognition using Machine learning techniques based on the physiological signal. *Biomedical Signal Processing and Control*, *100*, 107039. https://doi.org/10.1016/j.bspc.2024.107039

Kuruvayil, S., & Palaniswamy, S. (2022). Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 7271–7282. https://doi.org/10.1016/j.jksuci.2021.06.012

Li, H., Yang, W., Zhang, X., Wei, X., & Xu, X. (2022). A ResNet-Based Method for Complex Channel Interpretation in Seismic Volumes. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/lgrs.2022.3223422

Li, J., Nie, J., Guo, D., Hong, R., & Wang, M. (2024a). Emotion Separation and Recognition from a Facial Expression by Generating the Poker Face With Vision Transformers. *IEEE Transactions on Computational Social Systems*, 1–15. https://doi.org/10.1109/tcss.2024.3478839

Li, N., Huang, Y., Wang, Z., Fan, Z., Li, X., & Xiao, Z. (2024b). Enhanced Hybrid Vision Transformer with Multi-Scale Feature Integration and Patch Dropping for Facial Expression Recognition. *Sensors*, *24*(13), 4153. https://doi.org/10.3390/s24134153

Li, Y., Wang, S., Zhao, Y., & Ji, Q. (2013). Simultaneous Facial Feature Tracking and Facial Expression Recognition. *IEEE Transactions on Image Processing*, *22*(7), 2559–2573. https://doi.org/10.1109/tip.2013.2253477

Li, Y., Zeng, J., Shan, S., & Chen, X. (2019). Occlusion Aware Facial Expression Recognition Using CNN with Attention Mechanism. *IEEE Transactions on Image Processing*, *28*(5), 2439–2450. https://doi.org/10.1109/tip.2018.2886767

Liang, X., Xu, L., Zhang, W., Zhang, Y., Liu, J., & Liu, Z. (2023). A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition. *The Visual Computer*, *39*(6), 2277–2290. https://doi.org/10.1007/s00371-022-02413-5

Lin, H., Si, J., & Abousleman, G. P. (2007). Region-of-Interest Detection and its Application to Image Segmentation and Compression. *2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems*. 2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems, Waltham, MA, USA. https://doi.org/10.1109/kimas.2007.369827

Liu, C., Hirota, K., & Dai, Y. (2023a). Patch attention convolutional vision transformer for facial expression recognition with occlusion. *Information Sciences*, *619*, 781–794. https://doi.org/10.1016/j.ins.2022.11.068

Liu, S., Huang, S., Fu, W., & Lin, J. C.-W. (2024). A descriptive human visual cognitive strategy using graph neural network for facial expression recognition. *International Journal of Machine Learning and Cybernetics*, *15*(1), 19–35. https://doi.org/10.1007/s13042-022-01681-w

Liu, T., Li, J., Wu, J., Du, B., Chang, J., & Liu, Y. (2023b). Facial Expression Recognition on the High Aggregation Subgraphs. *IEEE Transactions on Image Processing*, *32*, 3732–3745. https://doi.org/10.1109/tip.2023.3290520

Liu, W., Li, C., Xu, N., Jiang, T., Rahaman, M. M., Sun, H., Wu, X., Hu, W., Chen, H., Sun, C., Yao, Y., & Grzegorzek, M. (2022). CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. *Pattern Recognition*, *130*, 108829. https://doi.org/10.1016/j.patcog.2022.108829

Naveen, P. (2023). Occlusion-aware facial expression recognition: A deep learning approach. *Multimedia Tools and Applications*, *83*(11), 32895–32921. https://doi.org/10.1007/s11042-023-17013-1

Ngwe, J. L., Lim, K. M., Lee, C. P., Ong, T. S., & Alqahtani, A. (2024). PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition. *IEEE Access*, *12*, 79327–79341. https://doi.org/10.1109/access.2024.3407108

Pantic, M., & Patras, I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *36*(2), 433–449. https://doi.org/10.1109/tsmcb.2005.859075

Pantic, M., & Rothkrantz, L. J. M. (2004). Facial Action Recognition for Facial Expression Analysis From Static Face Images. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *34*(3), 1449–1461. https://doi.org/10.1109/tsmcb.2004.825931

Rao, Ch. S. V. P., Pandian, A., Reddy, Ch. R., Aymen, F., Alqarni, M., & Alharthi, M. M. (2022). Location Determination of Electric Vehicles Parking Lot With Distribution System by Mexican AXOLOTL Optimization and Wild Horse Optimizer. *IEEE Access*, *10*, 55408–55427. https://doi.org/10.1109/access.2022.3176370

Saaidia, M., Zermi, N., & Ramdani, M. (2016). Fuzzy linear projection on combined multi-feature characterisation vectors for facial expression recognition enhancement. *International Journal of Signal and Imaging Systems Engineering*, *9*(4/5), 252. https://doi.org/10.1504/ijsise.2016.078266

Selvakumar, K., Jerome, J., Shankar, N., & Sarathkumar, T. (2015). Robust embedded vision system for face detection and identification in smart surveillance. *International Journal of Signal and Imaging Systems Engineering*, *8*(6), 356. https://doi.org/10.1504/ijsise.2015.072928

Tao, H., & Duan, Q. (2024). Hierarchical attention network with progressive feature fusion for facial expression recognition. *Neural Networks*, *170*, 337–348. https://doi.org/10.1016/j.neunet.2023.11.033

Thakur, P. S., Chaturvedi, S., Khanna, P., Sheorey, T., & Ojha, A. (2023). Vision transformer meets convolutional neural network for plant disease classification. *Ecological Informatics*, *77*, 102245. https://doi.org/10.1016/j.ecoinf.2023.102245

Vick, S.-J., Waller, B. M., Parr, L. A., Smith Pasqualini, M. C., & Bard, K. A. (2007). A Cross-species Comparison of Facial Morphology and Movement in Humans and Chimpanzees Using the Facial Action Coding System (FACS). *Journal of Nonverbal Behavior*, *31*(1), 1–20. https://doi.org/10.1007/s10919-006-0017-z

Wang, H., Xu, J., Yan, R., Sun, C., & Chen, X. (2020). Intelligent Bearing Fault Diagnosis Using Multi-Head Attention-Based CNN. *Procedia Manufacturing*, *49*, 112–118. https://doi.org/10.1016/j.promfg.2020.07.005

Wang, S., Zhao, A., Lai, C., Zhang, Q., Li, D., Gao, Y., Dong, L., & Wang, X. (2023). GCANet: Geometry cues-aware facial expression recognition based on graph convolutional networks. *Journal of King Saud University - Computer and Information Sciences*, *35*(7), 101605. https://doi.org/10.1016/j.jksuci.2023.101605

Wang, Y., Dong, X., Li, G., Dong, J., & Yu, H. (2022). Cascade Regression-Based Face Frontalization for Dynamic Facial Expression Analysis. *Cognitive Computation*, *14*(5), 1571–1584. https://doi.org/10.1007/s12559-021-09843-8

Wu, C.-Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., & Feichtenhofer, C. (2022). MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA. https://doi.org/10.1109/cvpr52688.2022.01322

Xia, Y., Yu, H., Wang, X., Jian, M., & Wang, F.-Y. (2022). Relation-Aware Facial Expression Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(3), 1143–1154. https://doi.org/10.1109/tcds.2021.3100131

Xie, L., Han, T., Zhou, H., Zhang, Z.-R., Han, B., & Tang, A. (2021). Tuna Swarm Optimization: A Novel Swarm-Based Metaheuristic Algorithm for Global Optimization. *Computational Intelligence and Neuroscience*, *2021*(1). https://doi.org/10.1155/2021/9210050

Xie, W., Wu, H., Tian, Y., Bai, M., & Shen, L. (2022). Triplet Loss With Multistage Outlier Suppression and Class-Pair Margins for Facial Expression Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(2), 690–703. https://doi.org/10.1109/tcsvt.2021.3063052

Xiong, L., Zhang, J., Zheng, X., & Wang, Y. (2024). Context Transformer and Adaptive Method with Visual Transformer for Robust Facial Expression Recognition. *Applied Sciences*, *14*(4), 1535. https://doi.org/10.3390/app14041535

Yang, B., Cao, J., Ni, R., & Zhang, Y. (2018). Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images. *IEEE Access*, *6*, 4630–4640. https://doi.org/10.1109/access.2017.2784096

Yang, B., Wu, J., Ikeda, K., Hattori, G., Sugano, M., Iwasawa, Y., & Matsuo, Y. (2022). Face-mask-aware Facial Expression Recognition based on Face Parsing and Vision Transformer. *Pattern Recognition Letters*, *164*, 173–182. https://doi.org/10.1016/j.patrec.2022.11.004

Ye, M., Ruiwen, N., Chang, Z., He, G., Tianli, H., Shijun, L., Yu, S., Tong, Z., & Ying, G. (2021). A Lightweight Model of VGG-16 for Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 6916–6922. https://doi.org/10.1109/jstars.2021.3090085

Zakieldin, K., Khattab, R., Ibrahim, E., Arafat, E., Ahmed, N., & Hemayed, E. (2024). ViTCN: Hybrid Vision Transformer with Temporal Convolution for Multi-Emotion Recognition. *International Journal of Computational Intelligence Systems*, *17*(1), 64. https://doi.org/10.1007/s44196-024-00436-5

Zhang, F., Zhang, T., Mao, Q., & Xu, C. (2020). Geometry Guided Pose-Invariant Facial Expression Recognition. *IEEE Transactions on Image Processing*, *29*, 4445–4460. https://doi.org/10.1109/tip.2020.2972114

Zhang, H., Yin, L., Zhang, H., & Wu, X. (2024). Facial micro-expression recognition using three-stream vision transformer network with sparse sampling and relabeling. *Signal, Image and Video Processing*, *18*(4), 3761–3771. https://doi.org/10.1007/s11760-024-03039-x

Zhang, X., Zhang, F., & Xu, C. (2022). Joint Expression Synthesis and Representation Learning for Facial Expression Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1681–1695. https://doi.org/10.1109/tcsvt.2021.3056098

Zheng, W., Zhou, X., Zou, C., & Zhao, L. (2006). Facial Expression Recognition Using Kernel Canonical Correlation Analysis (KCCA). *IEEE Transactions on Neural Networks*, *17*(1), 233–238. https://doi.org/10.1109/tnn.2005.860849

Zhong, C., Li, G., & Meng, Z. (2022). Beluga whale optimization: A novel nature-inspired metaheuristic algorithm. *Knowledge-Based Systems*, *251*, 109215. https://doi.org/10.1016/j.knosys.2022.109215