Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 March 2026
Revised 20 April 2026
Accepted 25 May 2026

Natural Resources for Human Health 2026, Vol. 6 (1s): 360–375

KEYWORDS:
Autism Spectrum Disorder; Facial Emotion Recognition; Harris Hawk Optimization; Explainable Artificial Intelligence

An Explainable ConvNeXt-V2 Architecture for Facial Emotion Recognition from Children with Autism under Real-Time Visual Environments

Suresh Chanamala[1], S. Anu H Nair[2], K. P. Sanal Kumar[3]

[1] Research Scholar, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamil Nadu, India.
E-mail : chsureshscet@gmail.com
[2] Assistant Professor, Department of Computer Science and Engineering, Annamalai University (Deputed To WPT, Chennai), India.
E-mail : anu_jul@yahoo.co.in
[3]Assistant Professor, Department of Computer Science, RV Government Arts College, Chengalpattu, India.
E-mail : sanalprabha@yahoo.co.in

ABSTRACT: Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder caused by differences in brain development that affect a child's ability to communicate, interact socially, and interpret emotional cues. Individuals with ASD often experience difficulties in recognizing facial expressions, making emotion recognition a significant component for early screening and diagnosis. Moreover, conventional machine learning and deep learning methodologies face limitations such as poor generalization, high computational complexity, and difficulty in capturing subtle facial features. In addition, explainable artificial intelligence (XAI) approaches are incorporated to improve the transparency of model decisions. Consequently, there is a need for more robust and effective computational models that improve emotion recognition accuracy and support early ASD detection, ultimately helping to reduce social and emotional difficulties. This manuscript presents an Interpretable Explainable Deep Convolutional Framework-based Autism Facial Emotion Recognition (IDCF-AFER) framework. The proposed model aims to design an effective approach for the accurate identification and classification of facial emotions such as neutral, anger, fear, joy, sadness, and surprise. The proposed model employs a ConvNeXt-V2 framework to effectively extract deep spatial features from facial images. Moreover, the stacked sparse autoencoder with a softmax layer is used for classification, which enables the model to effectively learn discriminative feature representations. Furthermore, Harris Hawks Optimization (HHO) is applied to fine-tune the model parameters and improve classification performance. Eventually, Score-CAM is utilized to highlight significant regions in facial images. Extensive experimental evaluation of the IDCF-AFER model was conducted on the benchmark Autism Facial Emotion Recognition dataset, and the results demonstrate its superior performance with an accuracy of 95.39% compared to existing methods.

1. INTRODUCTION

Autism spectrum disorder (ASD) is a long-term neurological condition that affects a person's activities, such as social interaction, communication skills, and behavioral responses, as well as daily routine activities [1]. Early diagnosis and intervention are the first and foremost requirements for proper treatment with special attention, which enhances the quality of an ASD person's life. However, the subtle nature and difficulties of diagnosing ASD symptoms make diagnosis challenging and often lead to misdiagnosis or incorrect diagnosis [2]. As the root cause of autism disorder is unclear, there is no available evidence that autism can be cured permanently [3]. However, it is essential to identify ASD prior to the age of 3 or 4; typical early symptoms can be observed before the age of 2, such as difficulties in social interaction, limited response during assessment, and reduced engagement in group attention [4]. So, timely intervention for ASD can improve the child's ability in particular areas, such as interaction and communication. A key factor of ASD is impairment in social communication; facial expressions play a major role in conveying social-emotional information and facilitating effective communication with others [5]. ASD children often show atypical facial expressions such as unclear or limited emotions, a tendency toward negative or neutral affect, and facial expressions that can appear unsuitable for surrounding situations [6]. Such behavioral patterns are associated with psychological theories that highlight challenges in emotional understanding, emotional awareness, empathy, and other related activities supported by neuroimaging studies showing atypical activation and neural connections in the regions associated with facial expressions, including the amygdala and the mirror neuron system [7].

Collectively, these findings highlight that facial expression behavior provides useful information, and it is a non-invasive marker of ASD-based social and emotional differences, emphasizing its potential for supporting ASD recognition and the necessity for objective and systematic assessment of facial expression patterns [8]. Hence, understanding the unique characteristics of facial expression production in this population is significant for aiding their emotional development and social communication [9]. Although facial expressions are an important feature for detecting ASD, they are less frequently utilized because of the limited availability of effective methods. Affective computing has largely focused on automatically identifying ASD from human facial expressions using Facial Expression Recognition (FER) techniques. Both machine learning (ML) and deep learning (DL)-based techniques are widely used to diagnose ASD with significant accuracy. However, the usage of such ML and DL-based techniques in clinical practice is limited due to the lack of model interpretability. Individuals with ASD, therapists, and family caregivers need clear and understandable explanations of these models to build a trustworthy AI system and support proper and informed decision-making. In recent years, Explainable AI (XAI) has emerged as an important approach for supporting the diagnosis of ASD in children and is currently attracting considerable interest. Autism Spectrum Disorder (ASD) diagnosis and detection remain challenging because they involve several developmental and behavioral patterns [13].

To address these challenges, more effective facial emotion recognition (FER) approaches are required for accurately identifying subtle and distinct emotional expressions, particularly sadness, surprise, and fear, in autism-specific scenarios. In addition, conventional methods do not identify these subtle emotional changes. Thus, this paper motivates the design of an efficient architecture that improves precision along with providing a better understanding of the emotional behavior of ASD children. Accordingly, this work introduces an Interpretable Deep Convolutional Framework-based Autism Facial Emotion Recognition (IDCF-AFER). The proposed method's primary objective is to develop an effective framework to detect and classify facial emotions in ASD children, and the major contributions of this research can be summarized as follows.

- Utilize a ConvNeXt-V2-based convolutional neural network (CNN) architecture for extracting deep spatial features and capturing discriminative representations from facial images.
- Design a stacked sparse autoencoder (SSAE) to learn highly compact feature representation, while the Softmax layer categorizes facial expressions into six classes: natural, anger, fear, joy, sadness and surprise.
- Integrate Harris Hawks Optimization (HHO) for optimal hyperparameter tuning and effective convergence of the proposed model.
- Apply a score-CAM for visualizing class-discriminative regions and improving model interpretability.
- Perform simulation analysis of the proposed IDCF-AFER model using the Autism Facial Emotion Recognition dataset, and demonstrate its efficiency and model performance.

2. LITERATURE REVIEW

This section provides an overview of existing studies on autism facial emotion recognition, highlighting key research gaps. Kirsch et al. [14] investigated the impacts of task adaptation and stimulus properties on FER performance in autistic and non-autistic adults utilizing mixed-effects modeling. The results were analyzed and consider stimulus characteristics, trial number, and measures derived from automatic facial exploration. Bhavana et al. [15] designed a new algorithm for identifying emotions through a CNN DL strategy. This model exploits CNN's feature learning abilities to capture the discriminating characteristics from the raw input information, empowering us to find a precise emotion classifier. Convolutional layers monitored by max-pooling functions form a multi-layered CNN framework specifically for this purpose to extract spatial features while maintaining their key characteristics hierarchically. Balasubramani et al. [16] developed an Ensemble of DL algorithm that depends on the ASD in facial images approach. The proposed approach aims to identify the variance among facial images of adults with ASD. For the ASD recognition process, the presented algorithm utilizes an ensemble approach for the classification method. In [17], a dual-branch CNN-based visual transformation model (Db-CNN-VTM) has been presented. Initially, the images are collected from the data are prone to pre-processing for noise reduction. Afterward, the feature representation and the feature fusion residual network strategy are established.

Saunshi et al. [18] presented the implementation of an emotion identification model that will support persons with ASD to meet real-time difficulties. The study employs the ResNet-50 DL algorithm integrated with attention mechanisms and a meta-learner for emotion detection. Melinda et al. [19] examined the use of DL methodology for the recognition of ASD with a facial expression study. The major goal is to determine the results of several DL frameworks, to help the performance of an AI-based ASD detection system in clinical image standards. Farhah [20] addressed an innovative algorithm to identify persons who have autism by evaluating their facial imageries. A variety of DL techniques, comprising InceptionV3, EfficientNetB0, and VGG16, are employed to detect autistic individuals by using facial expression recognition. The study determines the optimal configuration and parameter settings of the deep learning model to enhance prediction accuracy and overall performance. Hosney et al. [21] proposed a deep CNN-based real-world emotion detection mechanisms for autistic children. The presented model is developed to detect six facial emotions namely delight, surprise, fear, sadness, neutral, and joy, and to support clinical experts.

In spite of significant development in FER, conventional system frequently shows limited efficiency in managing subtle and abnormal facial expressions related with ASD. More traditional DL methods require appropriate attribute discrimination ability for fine-grained emotion categories, namely sadness and fear, in autism-specific datasets. Furthermore, several methodologies concentrate only on classifier accuracy without confirming method interpretability, which is essential in healthcare treatments. Moreover, a partial optimization approach is applied for fine-tuning deep frameworks, resulting in suboptimal results and minimized generalization over benchmark autism datasets. Accordingly, a robust and explainable framework is required for improving facial emotion recognition performance while ensuring reliable and interpretable decision-making for ASD identification, which motivates the development of the proposed model.

3. METHODOLOGICAL FRAMEWORK

This research proposes a novel IDCF-AFER framework for the precise identification and classification of facial emotions in children with ASD. Fig. 1 depicts the complete workflow of the proposed IDCF-AFER framework. The proposed IDCF-AFER model follows a multi-stage approach, encompassing image preprocessing, feature extraction, classification, parameter tuning, and model interpretability.

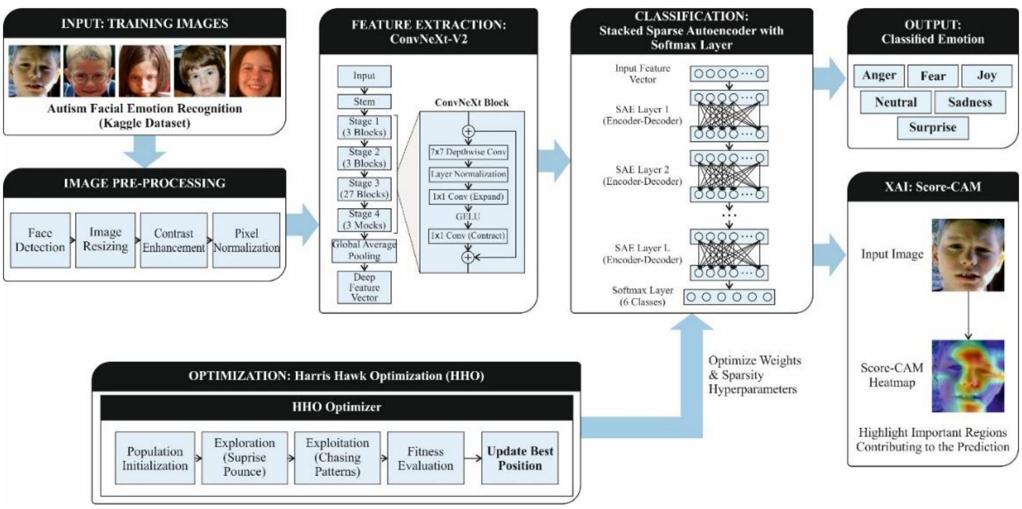

Fig. 1. Complete workflow of the proposed IDCF-AFER framework

3.1. Image Preprocessing Module

Image preprocessing is an essential stage that ensures high-quality input data for subsequent analysis. It involves face detection, image resizing, contrast enhancement using CLAHE, and min-max normalization to improve feature consistency and visual clarity across samples [22].

Face Detection: Face Detection identifies and extracts the facial region from each image while removing irrelevant background information, ensuring that subsequent analysis focuses only on the facial regions related to emotion.

Image Resizing: Image Resizing is used to resize the facial image to a fixed dimension prior to processing. This stage uniformly scales the spatial dimensions of the images to ensure consistent processing in subsequent phases, thereby improving computational efficiency and enabling the DL model to extract discriminative features from the resized facial images.

Contrast Enhancement using CLAHE: Contrast Enhancement using CLAHE enhances the local contrast of facial images without excessively amplifying noise. Unlike traditional global histogram equalization methods, CLAHE operates locally on image sections and applies contrast limitation to prevent the overamplification of pixel values. This approach improves the visibility of facial details and expression-related features under varying illumination conditions, thereby enabling more effective feature extraction and emotion recognition.

Min-Max Normalization: Min-Max Normalization scales the feature values to a range between 0 and 1 using the minimum and maximum values of the data. It uniformly scales the features while preserving their relative relationships.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Here, $X_{min}$and X_{max}signify the minimum and maximum values in the dataset, X_{norm}refers to the normalized value, and Xdenotes the original value.

3.2. ConvNeXt-V2 based Feature Extraction

The processed dataset is passed into the ConvNeXt-V2 model for extracting discriminative features crucial for classification. ConvNeXt-V2 is another CNN model that incorporates some of the design advantages of Transformers while retaining the efficiency of CNNs. This network generates more effective feature representations using improved normalization and a masked auto-encoding strategy, which enables enhanced performance on visual recognition tasks [23]. This method masks certain regions of the image and then trains the system to predict the missing patches, thereby enabling it to better capture feature dependencies, including spatial and contextual information. Another significant step in ConvNeXt-V2 is the introduction of the Global Response Normalization (GRN) technique, which helps reduce the issue of feature collapse, thereby improving the robustness of feature representation learning. The technique can also improve channel-wise competition and normalize global activation levels across the feature mappings. This allows different features to remain informative throughout the deep layers.

$$G(X) = \{||X_1||, ||X_2||, \ldots ||X_c||,\} \quad (2)$$

After, a channel-wise normalization has been operated to determine the qualified significance of every feature channel.

$$N(||X_i||) = \frac{||X_i||}{\sum_{j=1}^{C} ||X_j||} \quad (3)$$

Lastly, the feature map is modified through shifting parameters and learnable scaling, permitting adaptive feature refinement.

$$X_i' = \gamma \cdot X_i \cdot N\big(G(X_i)\big) + \beta + X_i \quad (4)$$

Here β and γ denote learnable parameters. Thus, ConvNeXtV2 is an effective architecture to build a hierarchical feature representation that works for generic image classifier applications, building on progress in convolutional effectiveness and transformer's semantic interpretation.

3.3. Autism Facial Emotion Classification Framework

The extracted deep features are then passed to the SSAE-based classification framework for accurate facial emotion classification. Stacked Sparse Autoencoder (SSAE) is a DL approach formed by stacking multiple SAEs to learn hierarchical and compact feature representations. It enhances discriminative capability by enforcing sparsity constraints, making it effective for classification tasks with high-dimensional data [24]. The SSAE module is a traditional three-layer artificial neural network (ANN) that is utilized for unsupervised learning to obtain the compressed code from the input. The compressed code represents the reduced representation of the original input. Then, the AEs are utilized as feature extractors for the DL framework. The SAE comprises the decoder φ and the encoder ϕ modules:

$$\phi: Q = \sigma(WP + a) \tag{5}$$

$$\varphi: P' = \sigma'(W'Q + a') \tag{6}$$

In the above mathematical form, the parameter W indicates the weight coefficient matrices, the term a refers to the bias vector, and the symbol σ points to the activation function for the encoding process. The factor a' represents the bias vector, the notation σ' denotes the activation function, and the parameter W' specifies the weight coefficient matrices corresponding to the decoding process. The mathematical form $P \in R^N$ considers the input and this term $Q \in R^M$ refers to the compressed code. The term $P' \in R^N$ represents the reconstruction vector. The SAE converts the input P to the latent variable Q and reconstructs P' by employing the decoder module. Thus, the objective of the AE is to learn the W and W', which allows the result of the decoder to closely reconstruct the actual input.

$$L(P, P') = \|P - P'\|^2 = \|P - \sigma'[W'\sigma(WP + a) + a']\|^2 \tag{7}$$

The stacked autoencoder network is a NN encompassing different layers of AEs that employ the result of previous AE layers as the input for the following ones. In the specified training instance $x^{(i)}$, where i refers to the number of instances.

$$J(w, a) = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{2}\left\|\sigma\left(WX^{(i)} + a\right) - Y^{(i)}\right\|^2\right] + \frac{\lambda}{2}\sum_{l=1}^{n-1}\sum_{i=1}^{s_l^c}\sum_{j=1}^{s_l^r}\left[W_{ji}^{(L)}\right] \tag{8}$$

From this mathematical expression, the parameter m is the total number of the instances, the factor Y characterizes the resultant label of P. The second term specifies the weight regularization that is employed to avoid overfitting, and the symbol λ points to the weight attenuation coefficient that can be leveraged to modify the relative weight of the first and second terms. The parameter s_l^c and s_l^r specify the overall neuron counts in layer l, and the parameter W_{ji}^l implies the corresponding weight between neurons in two layers.

The sparsity of features is accomplished through the summation of parameters, and the entire feature extraction method is enhanced. After that, the loss function of the SSAE can be measured by:

$$L_{sparse}(W, a) = J(W, a) + \beta\sum_{j=1}^{s_2} K\,L\left(\rho\hat{\rho}_j\right) \tag{9}$$

$$KL\left(\rho\hat{\rho}_j\right) = \rho\log\frac{\rho}{\rho} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \tag{10}$$

In this condition, the s_2 refers to the number of hidden layer (HLs), the mathematical notations β, ρ signify the divergence constant and sparseness constant, respectively. The term $\hat{\rho}$ represents the mean activation value.

The Softmax layer is generally utilized as the final activation function in multiclass classification networks to convert the output scores of the previous dense layer into a probability distribution across the target classes [25]. This transforms the original output values (logits) into standardized probabilities, in which the summation of probabilities across all classes equals 1. The class with the maximum probability is chosen as the predicted class. The mathematical form of the softmax function can be determined by:

$$\sigma(S_i) = \frac{\exp(S_i)}{\sum_{j=0}^{n-1} \exp(S_j)} \tag{11}$$

Here, the term S_j points to the output score of class j, the parameter S_i indicates the output score with respect to class i, and the resultant value $\sigma(S_i)$ denotes the probability that the input corresponds to class i.

3.4. Parameter Optimization using HHO

The proposed model is further optimized using the Harris Hawks Optimization (HHO) algorithm to improve the classification performance of the proposed framework. HHO is a nature-inspired metaheuristic model that simulates the cooperative searching behavior of HHs in nature. It is widely used for solving optimization problems by balancing exploration and exploitation to find optimal or near-optimal solutions [26]. The HHs are a category of raptors known for their higher intelligence and co-operative searching method. The HHO method is a population-based optimizer inspired by the cooperative behavior of HHs. The locations of different hawks in the swarm represent various candidate solutions. However, the optimal solution characterizes the location of the projected prey, like a rabbit. The working method of the HHO is achieved by employing three stages: the exploration, the transition, and the exploitation phases.

Exploration phase: During the exploration stage, every hawk arbitrarily occupies a constant position vector $P(t)$(candidate solution) in an effort to identify the prey (optimal solution). The position vector $(\theta_{on}, \theta_{off})$is allocated for every hawk. The fitness evaluation for every vectorized position is tested, which depends upon the chosen objective function (F_{obj}). The location of the hawk nearest to the projected rabbit is considered the best solution and is represented as $P_{rabbit}(t)$, and it provides the best value.

$$P(t+1) = \begin{cases} P_{rand}(t) - r_1|P_{rand}(t) - 2r_2P(t)| & q \geq 0.5 \\ (P_{rabbit}(t) - P_{av}(t)) - r_3(LB + r_4(UB - LB)) & q < 0.5 \end{cases} \tag{12}$$

Here, the mathematical term $P_{rand}(t)$refers to the arbitrarily designated hawk, $P(t+1)$indicates the next position vector of all the hawks, while r_1, r_2, r_3, r_4, and q represent arbitrary numbers. The parameters LB and UB represent the lower and upper bounds of every variable $(\theta_{on}, \theta_{off})$, and the parameter $P_{av}(t)$represents the average position.

$$P_{\sigma v}(t) = \frac{1}{N}\sum_{i=1}^{N} P_i(t) \tag{13}$$

In this mathematical form, the parameters T and N characterize the number of iterations and the total hawk counts, respectively. The parameter $P_i(t)$signifies the hawk position at the existing iteration (t).

Transition phase: The swarm chooses whether to move to the exploitation stage or remain in the exploration stage. The decision is made based upon the escape energy $(E)$of the target rabbit. The escape energy diminishes with respect to time (iterations), and its mathematical form is given below.

$$E = 2E_0\left(1 - \frac{t}{T}\right) \tag{14}$$

Here, E_0specifies the initial stage. When the parameter $E_0$is produced within the range of (-1, 0).

Exploitation phase: The swarm begins to search for the prey by generating a surprise attack. The prey can be forced to escape from this threat. Various escape behaviors occur under real-world conditions, which are based on the energy of the prey. Thus, diverse chasing techniques are performed. In HHO, there are four chasing methods that comprise the hard besiege and the soft besiege, the hard besiege with rapid dives, and the soft besiege with progressive rapid dives. By applying the chasing approach, the position vectors of hawks are updated, and then the hawks gradually approach the targeted prey (optimum outcome), therefore increasing the probability of co-operatively capturing it. The parameters R and M indicate diverse sub-rules that are chosen based on the specified conditions, and the mathematical terms are represented by:

$$Y = P_{rabbit}(t) - E|JP_{rabbit}(t) - p(t)| \tag{15}$$

$$R = Y + S \times LF(D) \tag{16}$$

$$M = P_{rabbit}(t) - E|JP_{rabbit}(t) - P_{av}(t)| \tag{17}$$

$$LF(p) = 0.01 \times \frac{u \times \sigma}{|v|^{\frac{1}{\beta}}}, \sigma = (\frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2\left(\frac{\beta-1}{2}\right)} 1^{\frac{1}{\beta}} \tag{18}$$

Here, the notation β refers to the default value of 1.5, while the terms u and v point to arbitrary values in the range of (0, 1). The HHO model derives a fitness function (FF) to achieve enhanced classification performance. It determines a positive integer to depict the enhanced performance of the candidate solutions. In this manuscript, the reduction of the classification error rate is measured as the FF.

$$fitness(x_i) = ClassifierErrorRate(x_i)$$

$$= \frac{number\ of\ misclassified\ instances}{Total\ number\ of\ instances} * 100 \tag{19}$$

3.5. Explainability Analysis via Score-CAM

To improve model interpretability, Score-CAM method is integrated into the proposed model. Score-CAM was selected because it provides gradient-free visual explanations and produces more stable localization maps than gradient-based visualization methods. Score-CAM is a gradient-free explainable AI technique leveraged to generate visual explanations for DL methodologies. It highlights important regions in an input image by using class activation scores, improving interpretability of model predictions [27]. Score-CAM is applied to assure that the method's predictions are proper for autism classifiers. Score-CAM generates heatmaps that demonstrate to clinicians and highlighting the regions of the facial image that offered more to the prediction. The method includes:

- **Feature Map Extraction:** Extracted are attribute mappings from the final traditional layer.
- **Class Score Calculation:** Feature maps contribute to the last class score (S_c).

$$w_k = RELU\left(\frac{\partial S_c}{\partial A_k}\right) \tag{20}$$

- **Heatmap Generation:** Class activation map (CAM) computed by the weighted sum of feature maps is projected on the source image:

$$CAM = \sum_k w_k A_k \tag{21}$$

Improving confidence and empowering consistent autism screening and classification, clinicians and researchers can imagine the major sections extract the prediction by covering the CAM on the input image, the transparency of autism identification models for detecting essential regions that influence predictions.

4. EXPERIMENTAL VALIDATION AND DISCUSSION

The performance validation of the IDCF-AFER technique can be explored through the Autism Facial Emotion Recognition dataset [28]. This dataset is a substantially increased and accurately split version of a current dataset, developed for training and evaluating ML approaches to identify ASD from facial images. Through extensive data augmentation and a structured train/test split, this dataset aims to deliver a strong and distinct foundation for developing highly precise and generalizable facial detection methods in the context of ASD research. Table 1 and Fig. 2 show that the dataset comprises seven categories with a total of 7600 sample images. Fig. 3 shows the representative samples of facial emotion from the Autism Facial Emotion Recognition dataset.

Table 1 Details of dataset

Classes	No. of Samples
Natural	480
Anger	690
Fear	300
Joy	3500
Sadness	2000
Surprise	630
Total Samples	**7600**

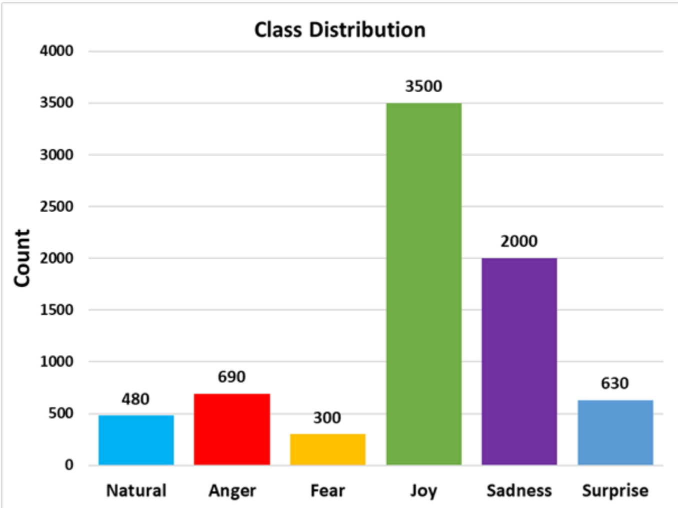

Fig. 2. Class distribution

Fig. 3. Representative samples of facial emotion from the autism facial emotion recognition dataset

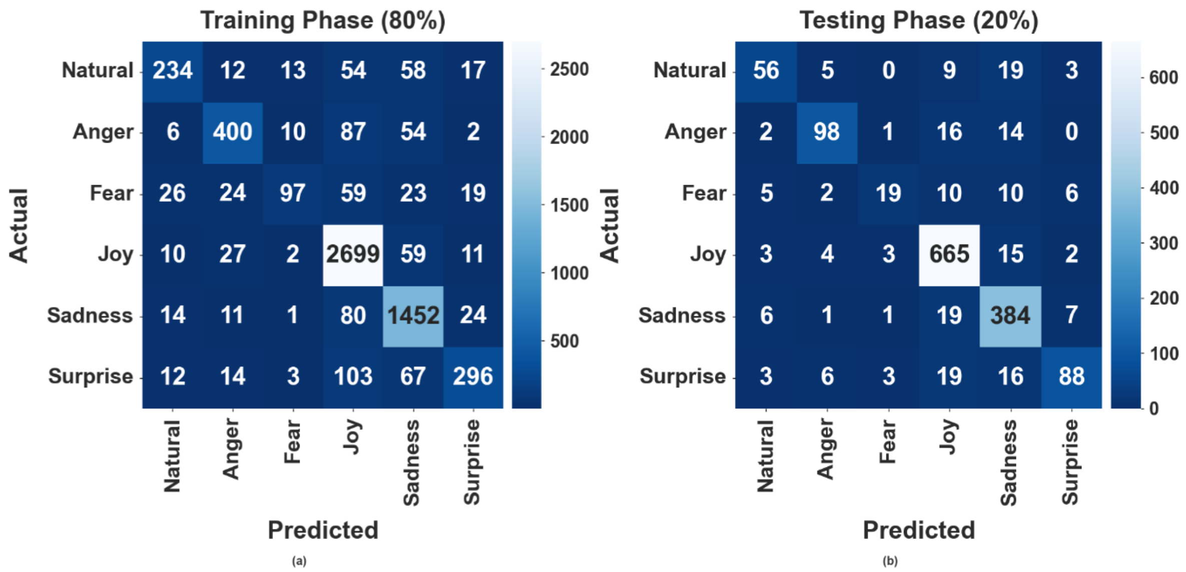

Fig. 4. Confusion matrix of the IDCF-AFER method with 80:20 split

Fig. 4 presents the confusion matrix of the IDCF-AFER approach with an 80:20 split across emotion classes. The higher diagonal values demonstrate accurate prediction, whereas relatively few off-diagonal errors indicate limited confusion between emotion classes.

Table 2 presents the facial emotion recognition results of the IDCF-AFER approach with an 80:20 train-test split. The outcome implies that the IDCF-AFER method accurately identifies the samples. The proposed IDCF-AFER model achieved an average accuracy of 95.05% in TRAIP and 95.39% in TESTP, while achieving an MCC of 71.59% for TRAIP and 72.33% for TESTP, demonstrating high efficiency and reliable performance in recognizing facial emotions.

Table 2 Facial emotion recognition result of the IDCF-AFER approach with 80:20

Classes	Accuracy	Precision	Recall	F1-Score	AUC Score	MCC
TRAIP (80%)						
Natural	96.35	77.48	60.31	67.83	79.56	66.50
Anger	95.94	81.97	71.56	76.41	84.98	74.40
Fear	97.04	76.98	39.11	51.87	69.31	53.62
Joy	91.91	87.57	96.12	91.65	92.21	84.17
Sadness	93.57	84.76	91.78	88.13	92.99	83.86
Surprise	95.53	80.22	59.80	68.52	79.25	66.99
Average	**95.05**	**81.50**	**69.78**	**74.07**	**83.05**	**71.59**
TESTP (20%)						
Natural	96.38	74.67	60.87	67.07	79.77	65.55
Anger	96.64	84.48	74.81	79.35	86.76	77.70
Fear	97.30	70.37	36.54	48.10	68.00	49.53
Joy	93.42	90.11	96.10	93.01	93.64	86.97
Sadness	92.89	83.84	91.87	87.67	92.58	82.87
Surprise	95.72	83.02	65.19	73.03	81.94	71.35
Average	**95.39**	**81.08**	**70.89**	**74.70**	**83.78**	**72.33**

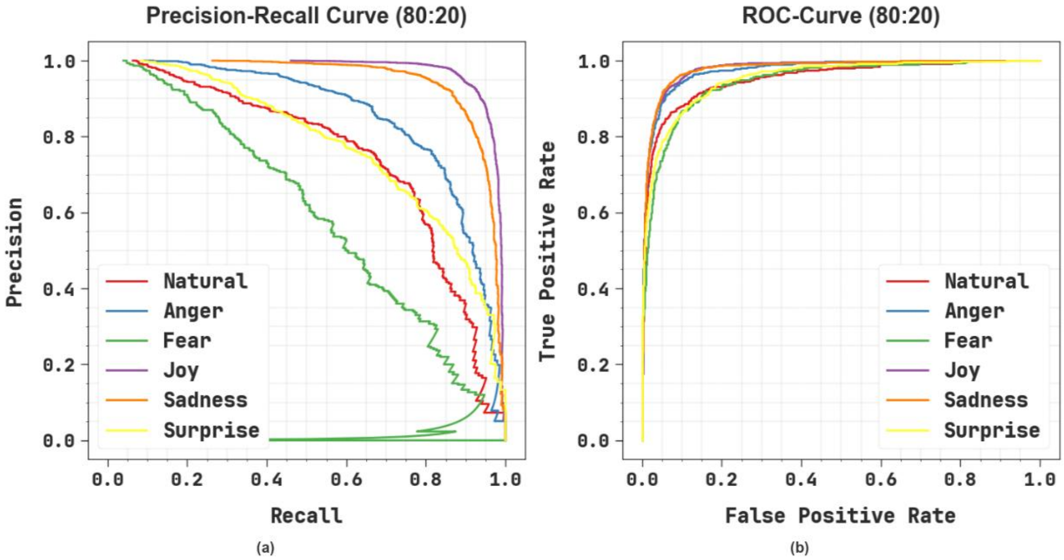

Fig. 5. Classifier result of the IDCF-AFER approach with (a-b) PR/ROC curve of the 80:20

Fig. 5 illustrates the classifier outcome of the IDCF-AFER algorithm with an 80:20 split. Fig. 5a shows the PR analysis of the IDCF-AFER system. The results indicate that the IDCF-AFER methodology achieves higher PR values. Lastly, Fig. 5b demonstrates the ROC analysis of the IDCF-AFER method. The figure shows that the IDCF-AFER approach achieves higher ROC values.

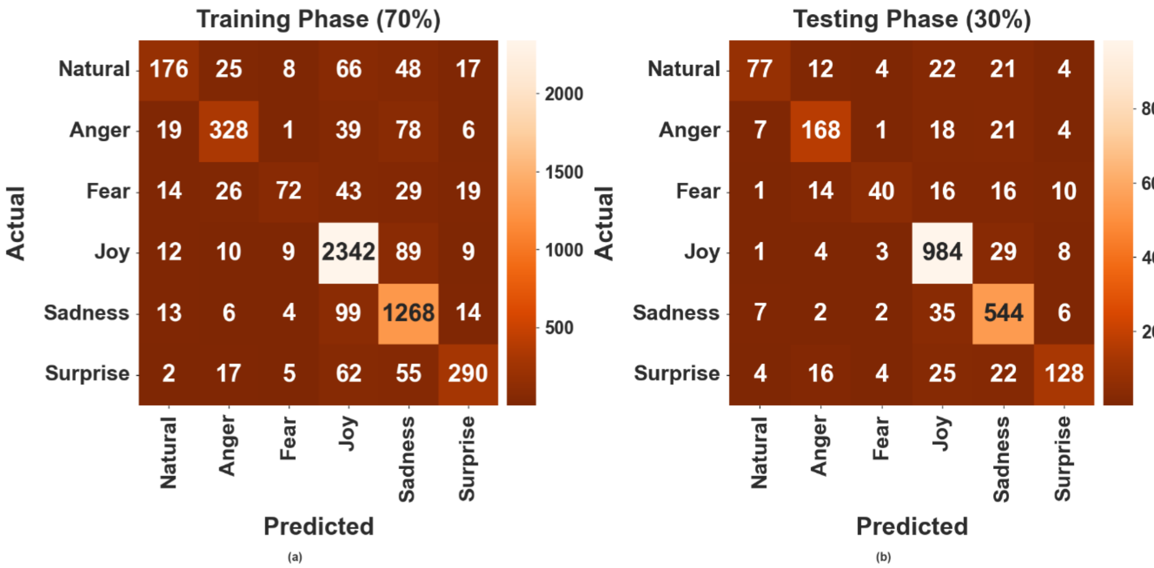

Fig. 6. Confusion matrix of the IDCF-AFER method with 70:30 split

Fig. 6 represents the confusion matrix of the IDCF-AFER technique with a 70:30 split over emotion classes. The highest diagonal values determine accurate prediction, while relatively few off-diagonal errors show limited confusion among emotion classes.

Table 3 depicts the facial emotion recognition performance of the IDCF-AFER system with a 70:30 train-test split. The results indicate that the IDCF-AFER model accurately identifies the samples. The proposed IDCF-AFER algorithm attained an average accuracy of 94.71% in TRAIP and 95.04% in TESTP, while achieving an MCC of 69.59% for TRAIP and 71.80% for TESTP, demonstrating higher efficacy and consistent performance in detecting facial emotions.

Natural Resources for Human Health
eISSN: 2583-1194

Table 3 Facial emotion recognition result of the IDCF-AFER approach with 70:30

Classes	$Accur_y$	$Preci_n$	$Recal_l$	$F1_{score}$	AUC_{score}	MCC
TRAIP (70%)						
Natural	95.79	74.58	51.76	61.11	75.28	60.06
Anger	95.73	79.61	69.64	74.29	83.95	72.17
Fear	97.03	72.73	35.47	47.68	67.47	49.53
Joy	91.77	88.34	94.78	91.45	91.97	83.72
Sadness	91.82	80.92	90.31	85.36	91.34	79.94
Surprise	96.13	81.69	67.29	73.79	82.98	72.12
Average	**94.71**	**79.64**	**68.21**	**72.28**	**82.16**	**69.59**
TESTP (30%)						
Natural	96.36	79.38	55.00	64.98	77.03	64.31
Anger	95.66	77.78	76.71	77.24	87.19	74.84
Fear	96.89	74.07	41.24	52.98	70.30	53.88
Joy	92.94	89.45	95.63	92.44	93.18	86.00
Sadness	92.94	83.31	91.28	87.11	92.40	82.42
Surprise	95.48	80.00	64.32	71.31	81.39	69.37
Average	**95.04**	**80.67**	**70.70**	**74.34**	**83.58**	**71.80**

Fig. 7 illustrates the classifier results of the IDCF-AFER approach with a 70:30 split. Fig. 7a shows the PR analysis of the IDCF-AFER technique. The performance indicates that the IDCF-AFER methodology achieves higher PR values. Finally, Fig. 7b shows the ROC analysis of the IDCF-AFER method. The figure shows that the IDCF-AFER algorithm achieves higher ROC values.

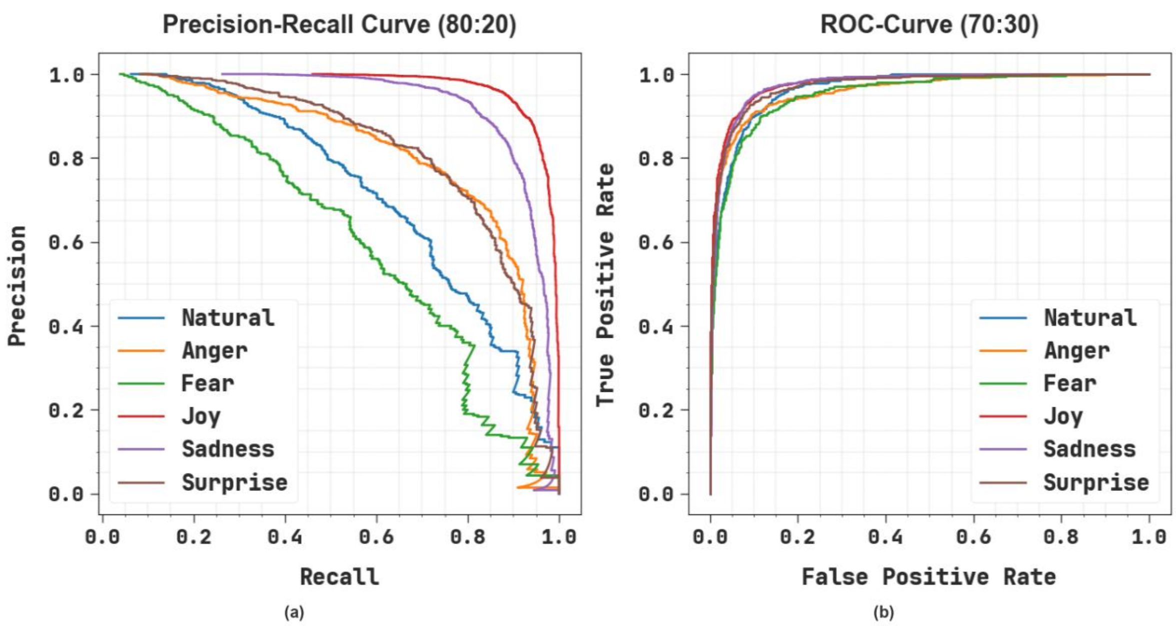

Fig. 7. Classifier result of the IDCF-AFER approach with (a-b) PR/ROC curve of the 70:30

Table 4 and Fig. 8 reveal the comparison study of the IDCF-AFER approach with existing methodologies using evaluation metrics and computational time (CT) [29-31]. The results indicates that VGG19, Inception-ResNet-V2, PCA, and SVM attain worse performance, whereas the ResNet34, MobileNet, CvT + Attention, and Xception techniques attain slightly higher and moderate performance compared to the proposed system. Similarly, the IDCF-AFER method achieves superior performance with an average $accur_y$ of 95.39%, $preci_n$ of 81.08%, and an AUC_{score} of 83.78%, indicating the effectiveness and robustness of the proposed IDCF-AFER method. These comparative results demonstrate the superiority of the proposed IDCF-AFER framework over existing state-of-the-art approaches.

Table 4 Comparison result of the IDCF-AFER method with other approaches

Methods	$Accur_y$	$Preci_n$	AUC_{score}	Computational Time (sec)
PCA	82.30	79.85	76.16	6.54
VGG19	51.44	66.79	60.24	7.37
SVM	81.00	76.59	77.08	5.94
CvT + Attention	90.73	79.76	79.87	7.07
Inception-ResNet-V2	76.50	78.84	80.24	4.52
ResNet34	84.00	76.40	79.24	6.81
MobileNet	85.00	78.53	80.67	8.94
Xception	91.00	76.30	75.55	5.75
IDCF-AFER	95.39	81.08	83.78	3.90

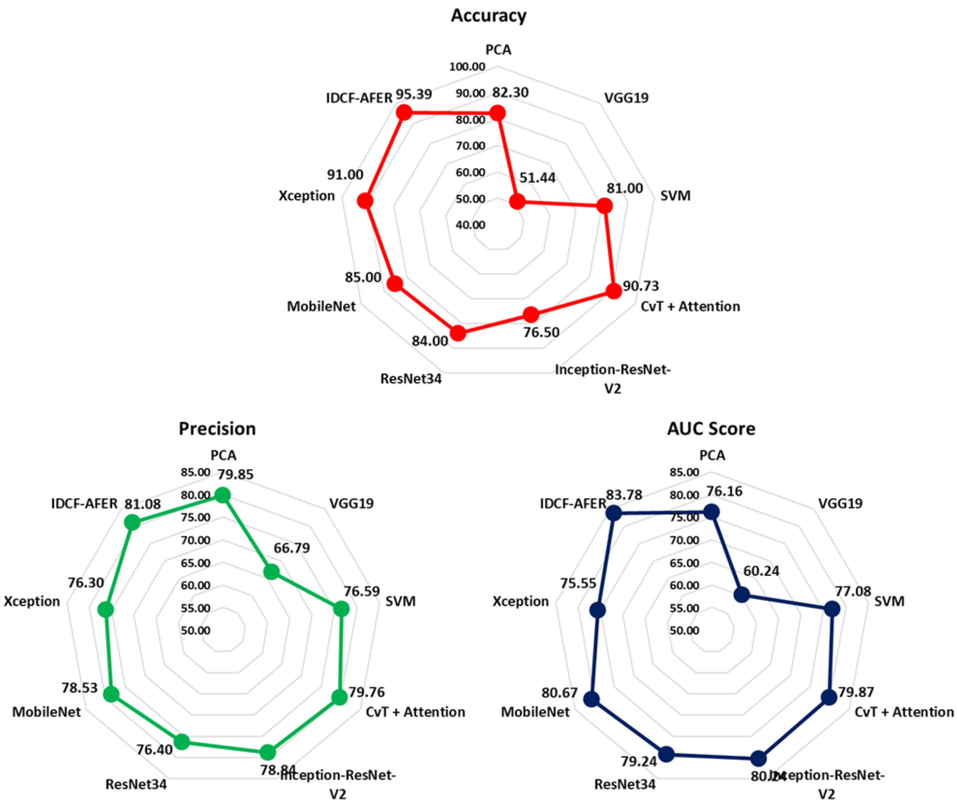

Fig. 8. Comparative result of the IDCF-AFER method with metrics

An ablation study estimates how distinct components of a system impact its outcome by eliminating or changing them. It helps detect which parts are more significant and which contribute less. By comparing solutions, researchers can better understand the method's behavior. Table 5 shows the ablation study of the IDCF-AFER method with various metrics. The findings show that the ConvNeXt-V2 + Softmax (without SSAE), SSAE only, and ConvNeXt-V2 + SSAE (without HHO) approaches have obtained minimal performance across several measures, whereas the IDCF-AFER (ConvNeXt-V2 + SSAE + HHO) method has achieved maximal performance with an average $accur_y$of 95.39%, $preci_n$ of 81.08% and AUC_{score} of 83.78%. The ablation analysis confirms that each component contributes positively to the overall performance of the proposed framework.

Table 5 Ablation study of the IDCF-AFER approach

Methods	$\boldsymbol{Accur_y}$	$\boldsymbol{Preci_n}$	$\boldsymbol{AUC_{score}}$
ConvNeXt-V2 + Softmax (without SSAE)	93.69	78.86	81.94
SSAE only	94.32	79.52	82.58
ConvNeXt-V2 + SSAE (without HHO)	94.88	80.31	83.14
IDCF-AFER (ConvNeXt-V2 + SSAE + HHO)	95.39	81.08	83.78

Table 6 presents the computational efficiency of the IDCF-AFER approach with other methodologies using params, FLOPs, and inference time [32]. The existing algorithms have the highest computational cost and the slowest inference time, while the proposed IDCF-AFER method has fewer parameters of 2.07M, FLOPs of 0.20, and a faster inference time of 0.82 ms, making it suitable for real-time Autism Facial Emotion Recognition.

Table 6 Computational efficiency of the IDCF-AFER method with other techniques

Models	**Params (M)**	**FLOPs (G)**	**Inference Time (ms)**
CNN Alone	12.40	8.20	3.70
PCA-MLP	5.10	0.90	1.20
PCA-CNN	6.30	2.10	2.40
PCA+PosEnc-CNN	6.50	2.30	2.60
YOLOv8	7.22	4.87	1.04
YOLOv8+SA	13.90	6.09	2.66
IDCF-AFER	2.07	0.20	0.82

Fig. 9 visualizes the activation maps generated by the Score-CAM method on a selection of images taken from the Autism Facial Emotion Recognition dataset. These images reveal which areas of the face are considered most important by the proposed model when classifying facial emotions, confirming the system's focus on features such as the eyes, nose, and mouth regions to produce interpretable and accurate predictions. The intensity of activation within these mapped regions provides an indication of the contribution of each facial area to the decision-making process of the model, enabling a clearer understanding of how different facial expressions are recognized in children with ASD.

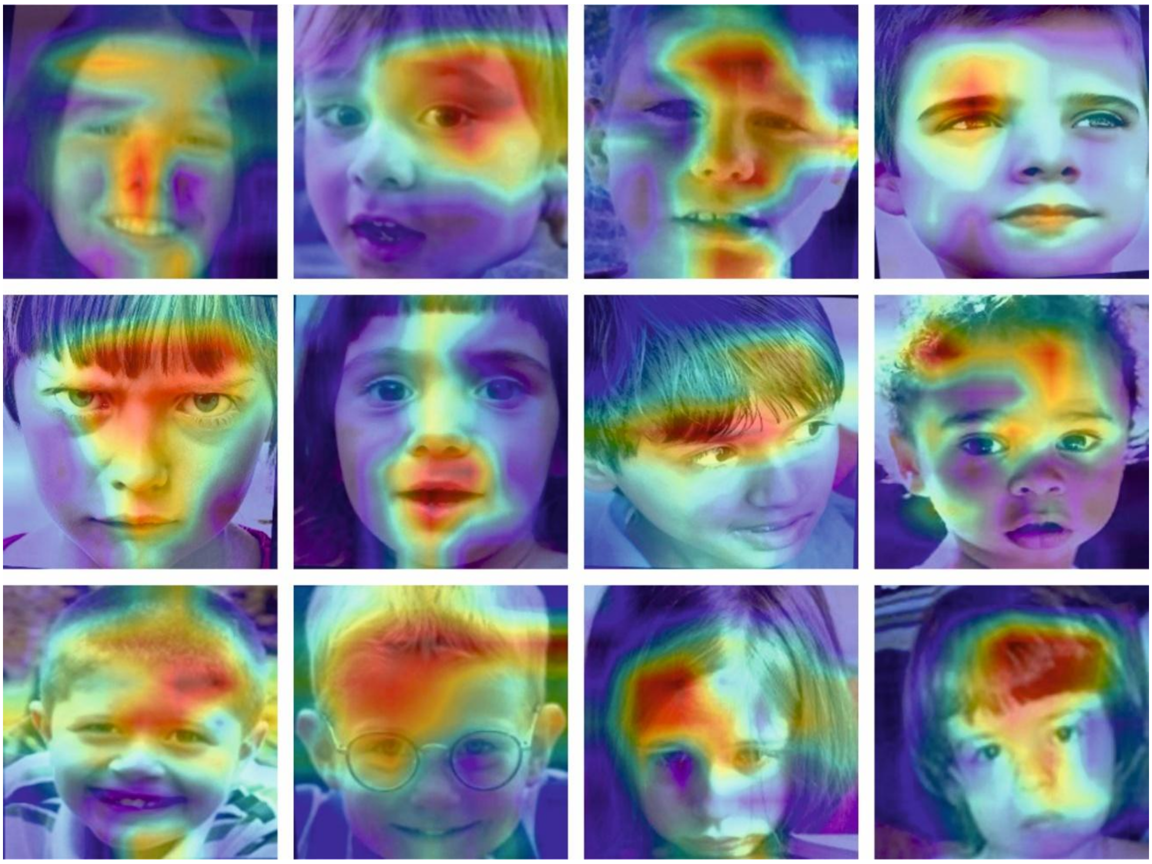

Fig. 9. Score-CAM-Based explainability visualization for autism facial emotion recognition

The proposed IDCF-AFER architecture shows a robust ability to identify and classify facial emotions in children with ASD with enhanced accuracy. ConvNeXt-V2 ensures effective spatial feature extraction, and the sequential sparse autoencoder effectively learns discriminative features needed for classifying specific emotions, such as joy and sadness. Furthermore, the application of HHO optimizes the model parameters and contributes to better classification performance. The overall results on the benchmark dataset show that the proposed system performs more successfully than many current methods in terms of accuracy and robustness and can be applied to real autism facial emotion recognition systems.

5. CONCLUSION AND FUTURE WORKS

This paper proposed a system called IDCF-AFER to recognize six types of facial expressions, including neutral, angry, fear, joy, sadness, and surprise expressions. The architecture was built using ConvNeXt-V2 to capture deep spatial features from facial images. In addition, the SSAE is used as the classifier and is combined with a Softmax activation function to enable the approach to successfully learn discriminative features. Moreover, the HHO algorithm is applied to improve the fine-tuning process of the model parameters, thereby improving classification accuracy. Lastly, the Score-CAM method is employed to highlight the important regions in the facial image. Experiments were carried out on the well-known Autism Facial Emotion Recognition benchmark dataset to confirm the effectiveness of the proposed model, which achieved an accuracy of 95.39%, outperforming other prior methods.

Some potential limitations of the current method include limited data diversity and a lack of generalizability to ASD individuals in real-world situations, which could hinder its effectiveness in clinical settings. The current architecture may also be computationally demanding for real-time deployment on low-resource devices. Future work includes exploring larger and more diverse ASD-specific facial expression datasets to enhance model generalization and developing a lighter and more efficient architecture for real-time operation. Further investigations could involve the development of more sophisticated transformer-based models or enhanced Explainable Artificial Intelligence (XAI) techniques to improve both accuracy and transparency for ASD facial emotion recognition applications.

REFERENCES

[1] Soni, P. and Kumar, R., 2025, August. Facial Expression and Image-Based Early Detection of Autism Spectrum Disorder Using Deep Learning Techniques. In 2025 Seventeenth International Conference on Contemporary Computing (IC3) (pp. 1-5). IEEE.

[2] Palomino, D.R., Ovalle, C. and Cordova-Berona, H., 2025. Deep learning for the identification of autism traits in children through facial expressions: a systematic review. International Journal of Electrical & Computer Engineering (2088-8708), 15(3).

[3] Muthukkumar, R., 2025. Enhancing the identification of autism spectrum disorder in facial expressions using DenseResNet-Based transfer learning approach. Biomedical Signal Processing and Control, 103, p.107433.

[4] Newlin, D.R., G, R., S, A. and T, J., 2025. Trio-DL: a novel deep learning framework for enhanced facial emotion recognition in autistic children. Journal of the Chinese Institute of Engineers, 48(6), pp.732-746.

[5] Sucharita, S. and Malathi, G., 2025. Unravelling traditional machine learning to advanced deep learning: autism detection using vision based gaze estimation and facial analysis. In Integrated Technologies in Electrical, Electronics and Biotechnology Engineering (pp. 123-128). CRC Press.

[6] Meera, S., 2025. An Enhanced Detection System of Autism Spectrum Disorder Using Thermal Imaging and Deep Learning. SN Computer Science, 6(2), p.169.

[7] Polavarapu, B.L., Reddy, V.D., Morampudi, M.K. and Abdul, A., 2026. Explainable AI for autism spectrum disorder detection in children using facial images: a multi-scale feature extraction and transfer learning approach. Evolutionary Intelligence, 19(2), p.53.

[8] Sayed, G.I., Dahy, G. and Hassanien, A.E., 2026. Optimized deep learning in a metaverse environment for autistic child support. International Journal of Information Technology, 18(1), pp.561-568.

[9] Kushwaha, A. and Thulal, A.N., 2026, May. CNN-Based Facial Emotion Recognition for Productivity Analysis in Autism Spectrum. In Proceedings of International Conference on Recent Trends in Computing: ICRTC 2025, Volume 1 (Vol. 1, p. 391). Springer Nature.

[10] Spoorthi, M. and Rai, B.K., 2026. A comprehensive review of autism detection using deep learning. Health Services and Outcomes Research Methodology, pp.1-40.

[11] Rao, P. and Uma, S., 2026. Optimized gated spatiotemporal graph attention network-based diagnosis of facial expression through video games. Pattern Recognition, 175, p.112991.

[12] Agha, S.A., Bazai, S.U. and Naushad, A., 2026. Facial expression and key landmarks detection using deep learning techniques. International Journal of Pattern Recognition and Artificial Intelligence, p.2657004.

[13] Marzouk, M.A. and Ibrahim, S.S., 2026. AI-Driven Eye-Tracking and Emotion Recognition to Improve Attention Across Storytelling Modalities in Children with Autism Spectrum Disorder. International Journal of Human–Computer Interaction, pp.1-23

[14] Kirsch, S., Drimalla, H., Saakyan, W., Sajonz, B.E.A., Gritzmann, J., Maier, S., Fangmeier, T., Lin, M., Guendelman, S., Kaufmann, C. and Dziobek, I., 2026. Reduced task adaptation and contextual awareness in autistic adults during facial emotion recognition: evidence from mixed-effects modeling and automated facial analysis. Molecular autism, 17(1), p.16.

[15] Bhavana, N., Guthur, A.S., Reddy, K.S., Ahmed, S.T. and Ahmed, A., 2025. Cognizance through convolution: a deep learning approach for emotion recognition via convolutional neural networks. Procedia Computer Science, 259, pp.1336-1345.

[16] Balasubramani, J., Rajendran, S., Zakariah, M. and Alnuaim, A., 2025. Ensemble of Deep Learning with Crested Porcupine Optimizer Based Autism Spectrum Disorder Detection Using Facial Images. Computers, Materials & Continua, 83(2).

[17] Alamgir, F.M., Zaman, T., Hossain, M.S., Hassan, M.M. and Alam, M.S., 2026. ASDnet: Classification model for individuals with autism spectrum disorder using facial grid-wise expressions features and dual-branch visual transformation. Biomedical Signal Processing and Control, 120, p.109999.

[18] Saunshi, G., Yakkundimath, R., Elemmi, M., Chini, S. and Sajjan, S., 2025. Emotion recognition system for individuals with autism spectrum disorder using deep learning techniques. Proceedings of the Indian National Science Academy, pp.1-13.

[19] Melinda, M., Andriyani, N.A.C., Nurdin, Y., Khariyunnisa, V., Yulita, Y. and Enriko, I.K.A., 2024. Deep learning performance analysis for facial expression-based autism spectrum disorder identification. Radioelectronic and Computer Systems, 2024(2), pp.30-40.
[20] Farhah, N., 2024. Utilizing Deep Learning Models in an Intelligent Facial Expression Classification System for Autism Disorder Diagnosis. International Journal of Advances in Soft Computing & Its Applications, 16(2).
[21] Hosney, R., Talaat, F.M., El-Gendy, E.M. and Saafan, M.M., 2024. AutYOLO-ATT: An attention-based YOLOv8 algorithm for early autism diagnosis through facial expression recognition. Neural Computing and Applications, 36(27), pp.17199-17219.
[22] Mohapatra, S., Swain, B.K. and Mishra, M., 2025. Early Parkinson's disease identification via hybrid feature selection from multi-feature subsets and optimized CatBoost with SMOTE. *Systems Science & Control Engineering*, *13*(1), p.2498909.
[23] Tsai, Y.Y., Lee, K.H., Mukundan, A., Karmakar, R., Nagisetti, Y., Seyoum, D.G., Naing, S.L., Huang, C.W. and Wang, H.C., 2026. Enhancing Esophageal Cancer Detection via Virtual NBI: A Novel Spectrum-Aided Vision Enhancer (SAVE) and Deep Learning Framework.
[24] Zhou, C.A., Zhang, K., Xu, J., Guo, K., Liu, X., Hu, B. and Wang, G., 2025. Tool wear monitoring for robotic milling based on multi-dimensional stacked sparse autoencoders and bidirectional LSTM networks with singularity features. The International Journal of Advanced Manufacturing Technology, 136(7), pp.3543-3557.
[25] Nawaz, M., Javed, A. and Saudagar, A.K.J., 2026. Residual-SwishNet: a deep learning-based approach for reliable lung cancer classification. *Frontiers in Oncology*, *15*, p.1729021.
[26] Saleh, A.L., Al-Amyal, F. and Számel, L., 2024. An enhanced current chopping control strategy for SRM drives using Harris Hawks optimization algorithm. *ISA transactions*, *150*, pp.338-358.
[27] Jadon, R., Gollapalli, V.S.T., Srinivasan, K., Chauhan, G.S. and Budda, R., 2025. Interpretable AI for skin lesion detection: Enhancing diagnostic accuracy with CNN and Score-CAM in IoMT systems. J. Ubiquitous Comput. Commun. Technol, 7(1), pp.1-18.
[28] https://www.kaggle.com/datasets/hasibur013/autism-facial-emotion-recognition/data
[29] Shaik, J., Shekhar, R. and Shelke, C.J., 2025. Hybrid Deep Learning Framework for Autism Detection in Children Using Facial Emotion Recognition. *Ingénierie des Systèmes d'Information*, *30*(10).
[30] Wang, Y., Pan, K., Shao, Y., Ma, J. and Li, X., 2025. Applying a convolutional vision transformer for emotion recognition in children with autism: fusion of facial expressions and speech features. *Applied Sciences*, *15*(6), p.3083.
[31] Ahmad, I., Rashid, J., Faheem, M., Akram, A., Khan, N.A. and Amin, R.U., 2024. Autism spectrum disorder detection using facial images: A performance comparison of pretrained convolutional neural networks. *Healthcare technology letters*, *11*(4), pp.227-239.
[32] Rashidy, N.E., Allogmani, E., Hassan, E., Alnowaiser, K., Elmannai, H. and Ali, Z.H., 2025. Toward real-time emotion recognition in fog computing-based systems: leveraging interpretable PCA_CNN, YOLO with self-attention mechanism. *Frontiers in Computer Science*, *7*, p.1714394.