Ingénierie des Systèmes d'Information
Vol. 30, No. 10, October, 2025, pp. 2595-2607
Journal homepage: http://iieta.org/journals/isi

Hybrid Deep Learning Framework for Autism Detection in Children Using Facial Emotion Recognition

Jahanara Shaik*, R. Shekhar, Chetan J. Shelke

Department of Computer Science and Engineering, Alliance University, Bengaluru 562106, India

Corresponding Author Email: jshaikPHD720@ced.alliance.edu.in

https://doi.org/10.18280/isi.301007

Received: 1 July 2025
Revised: 26 September 2025
Accepted: 5 October 2025
Available online: 31 October 2025

Keywords:
autism spectrum disorder, emotion recognition, deep learning, convolutional neural networks, ensemble classification, MobileNetV2, DeepFace, computer vision

ABSTRACT

Health care providers determine diagnosing autism to be a difficult undertaking because it mostly relies on anomalies in brain activities that could not be evident in the initial stages of a young development autism condition. An alternate and effective method to facilitate the earlier identification of autism involves facial emotion. This is because autistic children typically exhibit unique patterns that make it easier to differentiate from typical kids. some of many significant developments in enhancing the state lifestyle for those having autism is technological assistance. This study proposes a hybrid deep learning approach to detect the autism in children using deep facial features and emotional expressions. The proposed model combines the feature extraction from pre-trained CNN models like EfficientNetB0, ResNet50, and MobileNetV3Small and then classify autism and non-autism using a soft-voting ensemble model. The dataset is divided into 80% training and 20% testing. The MobileNetV2 model is used as emotion recognition that integrated on DeepFace model to enhance behavioral interpretation. The proposed model is trained and validated on two datasets: one containing images of autistic and non-autistic children, and another containing six types of emotions. The baseline classifier such as LR obtained the accuracy score of 77.57%, XGBoost of 81.00%, RF of 78.36%, SVM of 79.95%, MLP of 79.68. The proposed model obtained the accuracy score of 84.00% and a ROC-AUC score of 92.29% outperforming as compared to baseline models.

1. INTRODUCTION

The medical conditions collectively referred to as autism spectrum disorders (ASD) are varied in several children's. They are distinguished by a certain level of relationships and communication problems. Atypical patterns of behavior and operations, such as trouble switching between activities, attention to detail, and odd responses to emotions, are additional traits. People with autism have different needs and abilities, and these might change over time [1]. Early diagnosis and intervention are critical for improving developmental outcomes and quality of life for affected children. However, traditional diagnostic methods rely heavily on subjective behavioral assessments by clinicians, which can lead to delays or inconsistencies in diagnosis, particularly in resource-constrained environments [2]. Emotional signs and face emotions are crucial behavioral indicators for diagnosing ASD. Facial study reveals that children having autism frequently display unusual facial expressions, decreased eye interactions, and restricted emotional reactivity [3, 4]. Through the use of computer intelligence and deep learning strategies, these non-verbal indications offer a useful and approachable way to identify individuals for autism [5, 6]. Children with autism may have limited tastes or behaviors, as well as difficulties interacting and interacting with others. They may also employ different attentional, gestural, and psychological strategies. Even while there is now no approved treatment for ASD, early identification is essential for timely medication to lessen symptoms and assist the kid in developing the skills they will need to survive within the future [7]. The researchers' study looked at the potential use of facial features in children to detect autism. Their findings indicate that children with autism display a distinct set of facial characteristics that distinguish them from children without autism. The features include an exceptionally big top face with widely spread eyes and an unusually short central facial area that encompasses the edges of the cheekbones and mouth [8]. contrasting adolescents without autism in the second row with autistic in the first row. The differences in facial features between the two groups are depicted in Figure 1, which was extracted from the Kaggle database. The automation of the detection of ASD using image-based analysis has showed potential through the latest innovations in DL and artificial intelligence (AI). Convolutional neural networks (CNN) among others have proven to be highly effective at extracting intricate visual clues from images of faces [9]. The model's capacity to identify socioemotional deficiencies that are typical of ASD is significantly improved by combining emotion identification with autism screening.

In order to identify autism based on facial phrases, this study suggests a hybrid deep learning model which integrates several pretrained CNN models with an ensemble classification approach. Both MobileNetV2 and DeepFace are used to retrieve and evaluate emotion aspects, which allows the framework to read emotional reactions in addition to classifying autism. The goal of the suggested method is to use facial image research to produce an improved and comprehensible testing tool for autism screening. The main contributions can be summarized as:

- To design a robust model for facial emotion recognition in autistic children.
- To extract deep facial features using EfficientNetB0, ResNet50, and MobileNetV3Small.
- To implement an ensemble learning classifier for autism detection.
- To integrate DeepFace for real-time emotion recognition and fusion with autism prediction.

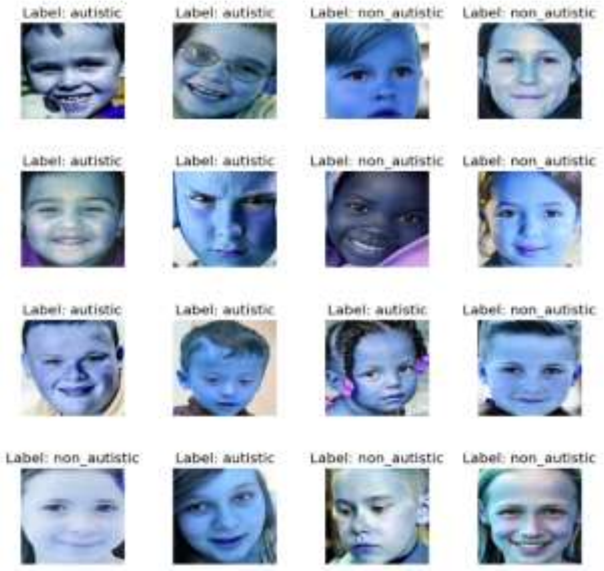

Figure 1. Visualization of dataset categories: Autistic vs. Non-autistic image samples

The remainder of this paper is organized as follows: Section 2 presents related work. Section 3 describes the proposed methodology, including feature extraction, ensemble classification, and emotion recognition. Section 4 discusses experimental results and performance evaluation. Section 5 concludes the paper and outlines future research directions.

2. RELATED WORK

Research on autism detection and emotional behavior analysis has increasingly turned toward deep learning, given its ability to extract complex and expressive facial cues that may be indicative of ASD. Existing studies primarily focus on two directions: (i) autism classification using facial features and (ii) emotion recognition in autistic children. Although these works have advanced the field, most face notable methodological limitations. The following sections outline key findings and methodologies from recent research.

The study explores emotion detection in children with ASD using deep learning techniques, specifically modifying YOLO model YOLOv5s, YOLOv7-tiny, YOLOv8s to achieve high accuracy in recognizing emotional expressions through multimodal data, enhancing therapeutic interventions [10]. The study presents a real-time emotion detection system for children with autism using an Enhanced DL technique. It identifies six emotions anger, fear, joy, natural, sadness, and surprise that achieves 99.99% accuracy through a deep convolutional neural network [11]. The paper presents a hybrid model combining DenseNet121 and MobileNetV2 for emotion detection in autistic children from facial images, utilizing a new dataset with four emotion classes. This approach enhances accuracy compared to traditional DL models [12]. The study focuses on ASD detection through deep learning by analyzing facial features, which indirectly relates to emotion detection. The ResNet34 model achieved an accuracy of 87%, aiding in recognizing ASD traits through facial analysis [13].

Table 1. Summary of related work in the field of autism detection

Ref.	Techniques	Findings	Limitations
[10]	Facial emotions and EEG signals for emotion detection Machine learning and deep learning algorithms utilized	High accuracy in emotion recognition using multimodal data. YOLO models effectively identify emotions in children with ASD.	Requires multimodal sensors (EEG + video), increasing system complexity; Focuses only on emotion recognition rather than ASD classification.
[11]	Enhanced deep learning (EDL) technique CNN with optimal hyperparameters selected using GA	Real-time emotion identification system for autistic children with 99.99% accuracy. Enhanced deep learning technique outperforms other algorithms for emotion classification.	Medical diagnosis relies on brain abnormalities not visible early. Limited to facial emotion recognition among children with autism.
[12]	Hybrid model integrating DenseNet121 and MobileNetV2 architectures. Developed and analyzed with four deep-learning models	Proposed hybrid model outperforms individual deep-learning models in accuracy. Introduced new dataset FERAC for autistic children's emotion recognition.	Dataset size is relatively small, reducing model generalization.
[13]	Deep learning with ResNet 34 model for analysis	Achieved 87% accuracy using ResNet 34 for ASD detection. Non-invasive diagnostic aid through facial feature analysis.	Improving model precision is necessary.
[14]	Deep DCNN for facial expression recognition Autoencoder for feature extraction and selection	Developed real-time emotion identification system for autistic children. Xception model achieved 95.23% accuracy in emotion recognition.	Early diagnosis of ASD is challenging due to brain abnormalities. Detection of abnormalities may not be evident in early stages.
[15]	Deep learning model with multi-label categorization Improved I-CNN optimization techniques	DL-ASD model predicts autism spectrum disorder in children aged 1-10. Proposed method achieves classification accuracy up to 98%.	Limited to children aged 1-10 for ASD prediction. Limited to emotion recognition and analysis in facial expressions.

Table 1 shows the summary of some existing studies on autism detection predominantly focuses on multimodal emotion recognition and DL-based facial analysis, yet each approach presents notable constraints. Studies combine facial expressions with EEG signals shows high accuracy through multimodal fusion and YOLO-based emotion detection, but they require complex sensor setups and do not directly address ASD classification. The study developed a real-time emotion detection system for autistic children using a deep CNN and autoencoders that obtained high accuracy in recognizing six emotions, enhancing early diagnosis and quality of life for individuals with ASD [14]. The paper presents a DL-ASD framework utilizing deep learning for emotion detection in children with ASD. It employs an Improved CNN to classify emotions, achieving a classification accuracy of 98% [15]. The SENSES-ASD system utilizes DL, specifically CNN, to enhance emotion detection in individuals with ASD. It classifies seven emotional states, achieving 71% accuracy on training data, aiding social interactions and communication [16]. The research employs a 2-D CNN model to analyze voice-based emotional traits in children with ASD, utilizing Mel Frequency Cepstral Coefficients and Mel Spectrograms to categorize emotions and understand behavioral differences compared to Typically Developing children [17]. Across these studies, three recurring limitations are evident:

Fragmentation between autism detection and emotion recognition: Most methods address either facial-based ASD classification or emotion prediction, but rarely integrate both, despite evidence that emotional affect is behaviorally relevant to ASD.

Restricted feature representation: Many deep learning models rely on a single CNN extractor, leading to suboptimal learning of subtle morphological cues such as eye spacing, upper facial width.

The ease with which various groupings of emotions can be identified is a significant advantage of using this approach. A variety of computer-based technologies have been developed to understand human attitudes and feelings better, thereby improving the user experience [18]. The Author primarily employs cameras to predict significant human facial movements. When someone looks at a camera that someone can infer their emotions having an average level of correctness. Meanwhile, several ML and image-processing experiments have demonstrated that facial traits and eye-glazing behaviors can be used to identify human moods [19]. The Facial Action Coding System (FACS) is a classification system for facial expressions based on facial action. Another deep learning model based on the AffectNet over the RAF-DB dataset to detect the facial emotion [20]. The suggested model obtained the 77.37% detection accuracy. Using a CNN and a modified PSO, the authors of the study [21] propose an efficient dynamic load balancing technique to examine the FC model in the healthcare domain. Wankhede and Selvarani [22] proposed a new effective hyperparameters optimization algorithm for CNN.

Although prior studies have explored facial expression analysis and DL for ASD, there is no unified model that jointly used multi-CNN deep feature extraction, ensemble learning, and emotion-aware interpretation. Existing models either:

- Depend on a single CNN backbone (limiting feature richness),
- Focus solely on emotion recognition without autism classification,
- Ignore emotion cues entirely despite their relevance to ASD behavioral profiling.

3. MATERIAL AND METHODS

This section presents the proposed hybrid DL Model for autism detection and emotion recognition and analysis. The methodology comprises five primary components: (1) feature extraction using pre-trained CNN, (2) ensemble classification for autism prediction, (3) emotion-based clustering and classification, (4) real-time emotion detection using DeepFace, and (5) a fusion strategy to integrate both autism and emotion information into a unified prediction system. A high-level system model is outlining the key stages, including multi-CNN feature extraction, ensemble classifier, MobileNetV2 emotion classification, and DeepFace integration. Additionally, the feature- and decision-level fusion strategy for generating the final autism + emotion prediction.

Dataset Description: In this study two dataset namely ASD and ASD with emotion have been used which is freely available [23, 24].

ASD Dataset: The ASD dataset contains the training and testing directories which also contains autistic and non-autistic subdirectories. For training, and testing forms also have two subdirectories: one for people with autism and another for people without.

Figure 1 illustrates representative facial images from the dataset, showcasing both autistic and non-autistic children. These samples reflect variations in facial geometry, expressions, and visual cues, which are critical for training the deep learning models to distinguish between the two categories. The diversity in lighting, pose, and emotion enhances the robustness of feature extraction and generalization during model training.

Figure 2 presents the class-wise distribution of images in the dataset. It highlights the number of samples available for each class autistic and non-autistic that ensure a balanced or imbalanced of the dataset. This distribution is crucial for evaluating model fairness and guiding the use of cross-validation and class balancing techniques during training.

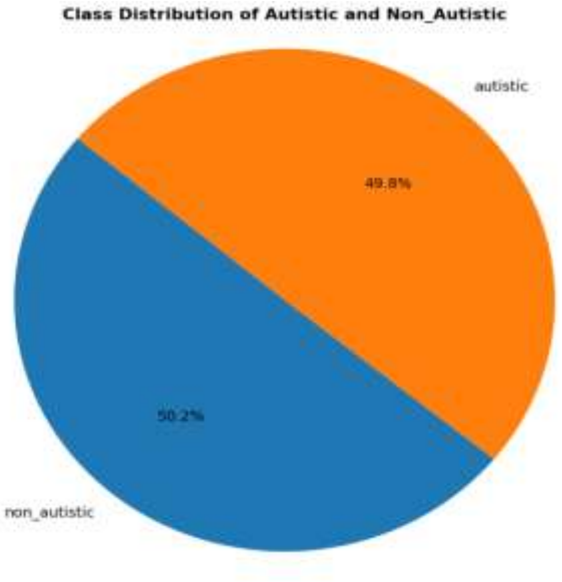

Figure 2. Distribution of dataset

ASD with Emotion Dataset: This dataset is emotion-specific, pre-processed images of autistic kids which contain six type of facial emotional images such as "Natural", "anger", "fear", "joy", "sadness", "surprise". The dataset consists of total 833 images. Each class of Natural of 48, Anger of 67, Fear of 30, Joy of 350, Sadness of 200 and Surprise of 63.

Data Preprocessing: When collecting data for utilization in proposed ML models and other types of investigation, data processing becomes a crucial step. It aids in ensuring that errors and inconsistencies are eliminated along with ensuring the data is formatted consistently. Whenever the data is used for classification or other kinds of investigation, this may produce superior results. In the instance of the ASD dataset, image processing was used to enhance the evaluation outcomes. The ASD dataset was preprocessed using following distinct methods.

The first method involved resizing each image in the collection. Resizing images in DL models enhances the model's effectiveness in a number of ways. By reducing the strain on the machine during training and inference, it first increases its computational effectiveness by accelerating computation and consuming fewer resources. As a result, the entire model may convergence faster when training. The model gains robust features over a range of dimensions when images are enlarged, which enhances its generalisation and situational recognition capabilities. Inputs must be resized to uniform sizes in order for models to be used in output. It will help with integrating and offer uniformity over a wide range of possibilities. Reduced sizes of images reduce memory requirements and improve model adaptability to real-world conditions through the use of efficient data augmentation approaches. This tactic is particularly important in scenarios with minimal memory because smaller image sizes result in lower storage demands. Image size reduction improves computing efficiency, speeds up training, improves generalizability, and increases deployment flexibility, among other advantages. This was accomplished by comparing the outcomes of proposed ML models using 224×224. The images were resized at two distinct configurations using Python programming.

By eliminating the influence of images of various levels, this standardization guarantees consistent and reliable model training. Since normalization avoids issues like disappearing or inflating gradients which arise when images have extremely wide frequency ranges, it accelerates up resolution during training. Furthermore, maintaining numerical consistency enhances the model's potential to pick up important traits. By lessening the impact of high values and illumination fluctuations, the process increases the model's adaptability to different lighting conditions. Normalization also facilitates the optimal use of pre-trained models, which are usually created on datasets having uniform input. Image normalization enhances the effectiveness, strength, and generalization capabilities of proposed ensemble model, allowing them to operate better over a range of scenarios.

The dataset completed these preprocessing techniques to enhance the investigation's outcomes and guarantee that it was in the optimal shape for ML along with additional analysis techniques.

Feature Extraction: In this study, the feature extraction process forms a critical foundation for representing facial images with high-level abstract patterns that are beneficial for distinguishing between autistic and non-autistic children. We utilized a multi-model feature extraction approach, that used the discriminative capabilities of three existing pretrained CNN models: EfficientNetB0, ResNet50, and MobileNetV3Small. These models are pre-trained on the ImageNet dataset and used here as fixed feature extractors by discarding their classification layers and retaining only the convolutional base with global average pooling $(include\_top = False, pooling = 'avg')$. Each image is first resized to 224×224 and then preprocessed using the respective preprocessing function of each model to match the expected input distribution. For every image, a feature vector is extracted from the output of the final pooling layer of each CNN. These vectors are flattened and concatenated to form a unified representation for further classification. Figure 3 shows architecture, the feature extraction process can be formally represented as follows.

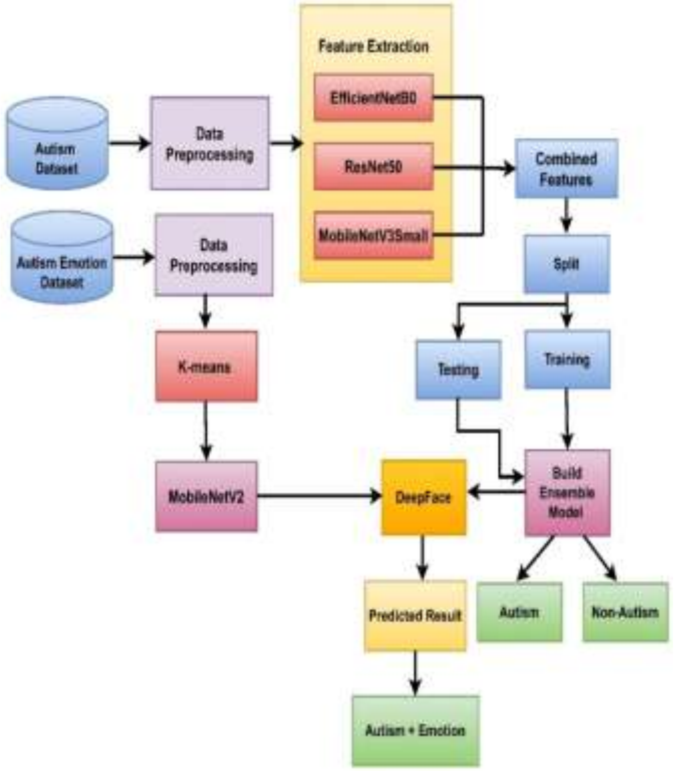

Figure 3. Architecture of autism with emotion detection

Let, $I = \{I_1, I_2, \dots, I_n\}$ be the set of facial images, $M = \{M_1, M_2, M_3\}$ be the set of pre-trained models, where, M_1 is EfficientNetB0, M_2 is ResNet50, M_3 is MobileNetV3Small.

$\phi_j(I_i)$ be the feature extraction function of model M_j applied on image I_i, $Pre_j(I_i)$ be the preprocessing function applied to I_i for model M_j. Then the final feature vector F_i for image I_i is computed as:

$$F_i = [\phi_1(Pre_1(I_i)) \parallel \phi_2(Pre_2(I_i)) \parallel \phi_3(Pre_3(I_i))] \qquad (1)$$

where $\parallel$ denotes the concatenation operator. All extracted feature vectors F_i are combined into a single feature matrix $X \in R^{n \times d}$, where d is the total dimensionality of concatenated features from all models. The corresponding labels $y \in \{0,1\}^n$ are binary, indicating non-autistic (0) and autistic (1) classes. This ensemble feature strategy effectively captures diverse hierarchical representations and enriches the feature space, enhancing the downstream classifier's ability to discern subtle facial patterns associated with autism.

```
Algorithm: Feature Extraction Using Pre-trained CNNs
Input: Facial image dataset of children (Autistic and Non-Autistic)
Output: Autism prediction (Autism / Non-Autism)
1. Initialize Pre-trained CNN Models:
  - Load EfficientNetB0, ResNet50, MobileNetV3Small with ImageNet weights
  - Remove top layers and apply global average pooling to extract features
2. For each image in the dataset:
  a. Resize image to 224 × 224 × 3
3. Initialize:
  - An empty list: $features\_all$
  - An empty list: $labels\_all$
4. Define $models\_info$ as a list of tuples:
  (model, $corresponding\_preprocessing\_function$)
5. For each (model, $preprocess\_function$ ) in models_info:
  a. For each $class\_folder$ in [$autistic\_folder, non\_autistic\_folder$]:
     i. Set label = 1 if folder is 'autistic', else 0
     ii. For each $image\_file$ in $class\_folder$:
        A. Load image and resize to $(224, 224, 3)$
        B. Convert image to array
        C. Expand dimensions to match model input
        D. Preprocess the image using $preprocess\_function$
        E. Extract feature using $model.predict()$
        F. Flatten the feature vector
        G. Append the feature to a temporary list
        H. Append the corresponding label to a temporary label list
  b. Combine features from both classes into one array
  c. Append the combined features to $features\_all$
  d. Append the labels to $labels\_all$
6. Concatenate all features from all models (axis = 1)→X
7. Convert $labels\_all$ into a numpy array → y
8. Return X, y
```

Ensemble Classification: After obtaining the high-dimensional feature vectors through the feature extraction process, we employ an ensemble learning approach to improve the classification performance for detecting autism based on facial features. Ensemble methods combine the predictive capabilities of multiple base classifiers to produce a more robust and generalized model. In this work, we construct a soft voting ensemble classifier composed of three diverse learning algorithms:

Support Vector Machine (SVC) with a Radial Basis Function (RBF) kernel to capture non-linear decision boundaries,

Gradient Boosting Classifier (GBC), a powerful ensemble tree-based learner that builds models sequentially to correct errors from previous models,

Multi-Layer Perceptron (MLP), a feed-forward neural network that captures non-linear patterns through multiple hidden layers.

Figure 4 shows the proposed ensemble model that each classifier is independently trained on the same training dataset and outputs class probability estimates. The final predicted probability for each class is computed as the average of probabilities predicted by all base classifiers, and the class with the highest average probability is selected as the final output (soft voting).

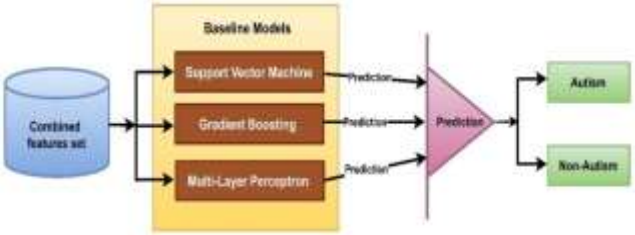

Figure 4. Proposed ensemble model

To ensure robust evaluation, we adopt Stratified K-Fold Cross-Validation with $k = 5$, preserving the class distribution in each fold. This method enables consistent estimation of model performance across different data partitions and reduces variance in performance metrics.

Let, $X \in R^{n \times d}$ be the feature matrix, $y \in \{0,1\}^n$ be the label vector, $C = \{C_1, C_2, C_3\}$ be the set of classifiers, where, C_1 is the SVC, C_2 is Gradient Boosting Classifier, C_3 is Multi-Layer Perceptron.

For a given test instance $x_i \in R^d$, let the probability that classifier C_j assigns to class $c \in \{0,1\}$ be: $P_j(c \mid x_i)$ is the probability estimate from classifier C_j. The final ensemble prediction is given by soft voting:

$$P_{ensemble}(c \mid x_i) = \frac{1}{3}\sum_{j=1}^{3} P_j(c \mid x_i) \qquad (2)$$

$$\hat{y}_i = arg \max_{c \in \{0,1\}} P_{ensemble}(c \mid x_i) \qquad (3)$$

This ensemble formulation allows the model to leverage the strengths of individual classifiers while minimizing their weaknesses, leading to improved generalization and more reliable autism detection from facial expressions.

Emotion-Based Processing

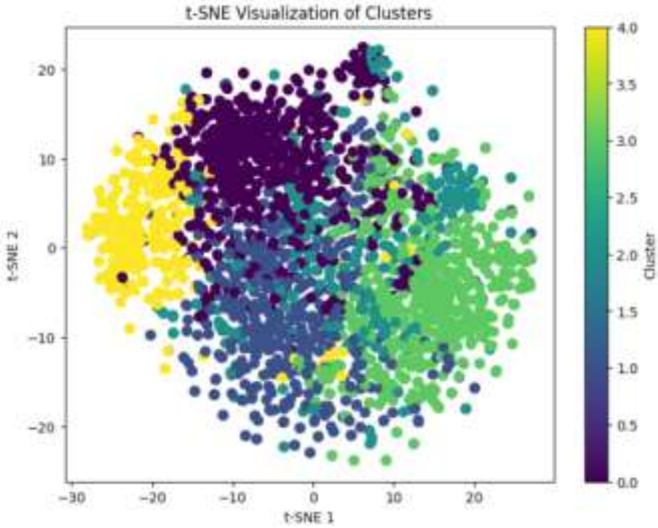

Figure 5. Cluster of emotion

To enhance the autism detection model with emotional cues, we incorporate facial emotion features into our analysis. These features are extracted from an additional autism emotion dataset, which contains images of children's facial expressions labelled with emotion categories. Initially, we apply data augmentation and normalization using a

$ImageDataGenerator$, which rescales pixel values to the range [0,1] and splits the dataset into training and validation subsets. This step helps ensure that the model generalizes well to unseen data and avoids overfitting. Next, we employ K-Means clustering on the training set to group images based on their underlying emotion features. The K-Means algorithm partitions the dataset into k distinct clusters by minimizing the intra-cluster variance. Figure 5 shows cluster that used k=6 to correspond with typical emotions groups (happy, sad, angry, surprised, fear, and neutral), that serve as the semantic framework for classifying related facial emotions. When emotional data is fed into the DeepFace model for categorization and incorporation into the autism detection mechanism, clustering stage helps to organize the data.

Let, $X = \{x_1, x_2, \dots, x_n\} \subset R^d$ denote the set of n preprocessed emotion image vectors, each of dimension d. k is set of clusters. $\mu_j \in R^d$ is the centroid of cluster j, where $j = 1,2, \dots, k$. The K-Means algorithm seeks to minimize the within-cluster sum of squares (WCSS) objective:

$$args \min_{\{\mu_j\}_{j-1}^{k}} \sum_{i=1}^{n} ||x_i - \mu_{c(i)}||^2 \tag{4}$$

$c(i) \in \{1, \dots, k\}$ denotes the cluster assignment for sample x_i, $\mu c(i)$ is the centroid of the cluster assigned to x_i, ‖·‖ is the Euclidean norm. The next DL module is then informed by the cluster labelling from the trained K-Means model, which enables the system to associate particular emotion expressions with possible characteristics associated with autism. MobileNetV2, a lightweight and effective pretrained CNN model created especially for resource-constrained contexts like mobile is then used to process the clustered data. In the behavioral investigation of ASD, the objective of this stage is to extract precise emotional elements from the children's facial images.

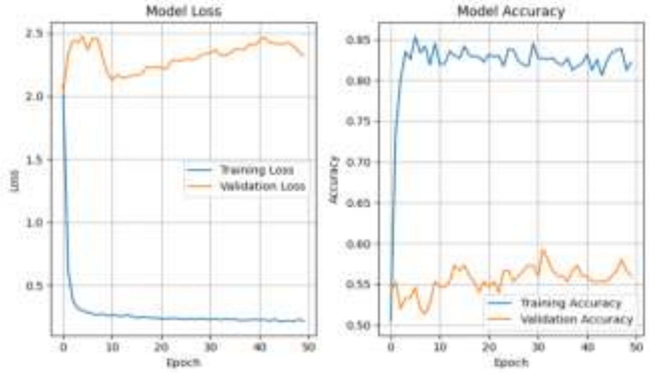

Figure 6. Performance convergence of MobileNetV2 during training: Accuracy and Loss

Figure 6 shows the performance convergence of MobileNetV2 model. The training loss decreases during the initial epochs and stabilizes at a low value that shows effective feature learning and rapid optimization.

DeepFace-Based Emotion Analysis: In this step, we integrate DeepFace, a state-of-the-art facial analysis framework, to perform emotion recognition as part of our autism detection pipeline. After extracting multi-view features using pretrained CNN architectures of ResNet50, EfficientNetB0, and MobileNetV3Small, the image is passed through a trained autism ensemble classifier to detect the presence of autism. Simultaneously, DeepFace analyzes the same image to determine the dominant facial emotion from predefined classes such as anger, fear, joy, sadness, surprise, and neutral. The purpose of incorporating DeepFace is to complement the autism classification with emotional cues. Emotional recognition can provide deeper behavioral insights, especially in children with ASD, where affective expression may differ significantly from neurotypical peers.

Let, $\in R^{224\times224\times3}$: Input facial image. $f_r(x), f_e(x), f_m(x)$: Feature vectors extracted from ResNet50, EfficientNetB0, and MobileNetV3Small respectively.

$F(x) \in R^d$: Concatenated feature vector, $F(x) = [f_r(x), f_e(x), f_m(x)] \in R^{d_r + d_e + d_m}$

C_{autism}: Trained ensemble classifier (Voting Classifier). $C_{deepface}$: Pretrained DeepFace emotion classifier. $\hat{y}_{autism} \in \{0,1\}$: Predicted class for autism (0: Non-Autism, 1: Autism)

$\hat{y}_{emotion} \in E$: Predicted emotion class, where, $E = \{Anger, Fear, Joy, Sadness, Surprise, Neutral\}$

Then:

$$\hat{y}_{autism} = C_{autism}(F(x)), \hat{y}_{emotion} = C_{deepface}(x) \tag{5}$$

The result is:

$$Prediction(x) = (\hat{y}_{autism}, \hat{y}_{emotion}) \tag{6}$$

4. RESULT ANALYSIS

The experimental configuration was carefully structured to ensure reproducibility, fairness, and methodological rigor. For autism detection, the ASD dataset was partitioned into 80% training, 20% testing using a stratified sampling strategy to preserve the original autistic and non-autistic distribution across all subsets. The autism classification component was evaluated using a Stratified 5-Fold Cross-Validation framework, ensuring consistent performance assessment across folds, while MobileNetV2 emotion classification employed a two-phase training process involving frozen-layer feature extraction followed by fine-tuning. All experiments were conducted in Google Colab using an NVIDIA Tesla T4 GPU, TensorFlow 2.12, scikit-learn 1.3, and standard Python scientific packages. Hyperparameters were optimized using the Adam optimizer with a batch size of 32, and random seeds were fixed (seed = 42) across TensorFlow, NumPy, and scikit-learn to guarantee replicability. This comprehensive configuration provides a transparent and reproducible foundation for the proposed hybrid deep learning model.

$$Accuracy = \frac{T\_P + T\_N}{T\_P + T\_N + F\_P + F\_N} \tag{7}$$

$$Precision = \frac{T\_P}{T\_P + F\_P} \tag{8}$$

$$Recall = \frac{T\_P}{T\_P + F\_N} \tag{9}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

Table 2 shows the performance analysis of six baseline classification models LR, XGBoost, RF, SVM, MLP, and the proposed Ensemble model on key evaluation metrics. Among all models, the Ensemble model achieved the highest overall performance with an accuracy of 83.86%, F1-score of 84.16, and an outstanding ROC-AUC of 92.29, indicating excellent discrimination capability. It also led in precision (82.23%), recall (86.17%), Kappa score (67.73), and Matthews Correlation Coefficient (MCC = 67.81), shows balanced classification. While XGBoost and SVM also performed well with ROC-AUC scores of 81.09 and 89.13, respectively their precision and F1-scores were comparatively lower. LR performed the worst on the majority of measures, despite being straightforward and easy to understand. The Ensemble model is unquestionably the most reliable and strong method for identifying autism based on face traits, according to this comparison study.

Table 2. Performance analysis of several models for autism detection

Models	Accuracy	Precision	Recall	F1-Score	Kappa Score	MCC Score	ROC-AUC
LR	77.57	75.82	77.09	76.45	55.04	55.05	77.55
XGBoost	81.00	78.31	82.68	80.43	61.99	62.08	81.09
RF	78.36	74.37	82.68	78.31	56.84	57.16	87.62
SVM	79.95	76.96	82.12	79.46	59.91	60.03	89.13
MLP	79.68	77.72	79.89	78.79	58.26	58.30	88.80
Ensemble	**83.86**	**82.23**	**86.17**	**84.16**	**67.73**	**67.81**	**92.29**

Table 3. Cross validation of proposed ensemble model

K-Fold	Accuracy	Precision	Recall	F1-Score	Kappa Score	MCC Score	ROC-AUC
1	82.85	81.00	85.71	83.29	65.70	65.82	91.30
2	81.53	79.90	84.13	81.96	63.07	63.15	90.97
3	82.54	82.45	82.45	82.45	65.08	65.08	90.61
4	83.07	84.83	80.32	82.51	66.13	66.22	91.60
5	83.86	82.23	86.17	84.16	67.73	67.81	92.29

Table 3 shows the cross-validation results for the proposed ensemble model that presents its consistent and robust performance across five folds. Accuracy values range from 81.53% to 83.86%, with the highest performance observed in Fold 5. Precision and recall remain well-balanced across folds, with Fold 4 achieving the highest precision (84.83%) and Fold 5 attaining the highest recall (86.17%). The F1-scores, all above 81.9%, reflect strong harmonic mean of precision and recall, while the Kappa scores and MCC across all folds exceed 63%, indicating substantial agreement and balanced classification. The ROC-AUC values remain exceptionally high, ranging from 90.61 to 92.29, confirming the model's excellent discriminative power.

Figure 7 presents a spider (radar) plot that shows the various performance metrics of the proposed ensemble model across five cross-validation folds. The chart highlights the model's consistent performance, with all metrics maintaining values above 80% across folds. The visual representation confirms the stability, robustness, and generalizability of the ensemble classifier in detecting autism using facial expression features.

Figure 8 illustrates the confusion matrices for five baseline classifiers—Logistic Regression (LR), XGBoost, Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP)—used for binary classification of autistic and non-autistic children. Among these, XGBoost demonstrates the most favorable performance with the highest true positive count (TP = 159) and the lowest false negative rate (FN = 31), indicating strong sensitivity in correctly identifying autistic cases. SVM and MLP follow closely with TP values of 156 and 157, respectively, though they suffer slightly from higher false positives or false negatives. Random Forest shows a higher false positive count (FP = 51), suggesting a tendency to over-predict autism. Logistic Regression, while simpler, has the lowest TP (154) and the second-highest FN (41), reflecting limitations in modeling complex nonlinear patterns.

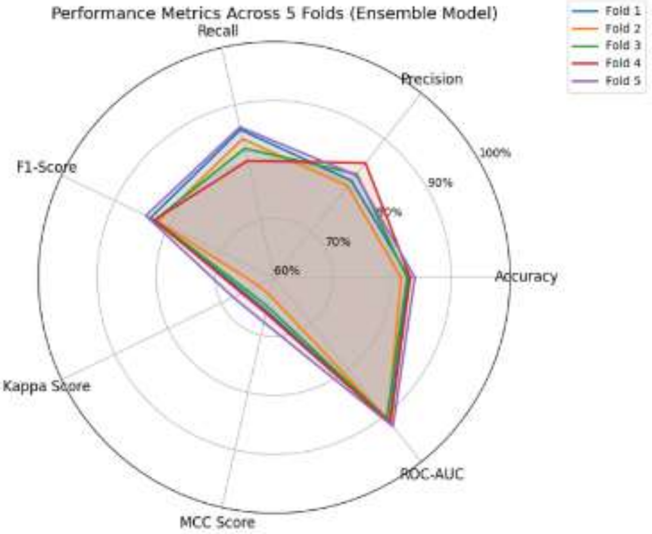

Figure 7. Spider (Radar) plot of performance metrics across 5 folds of ensemble model

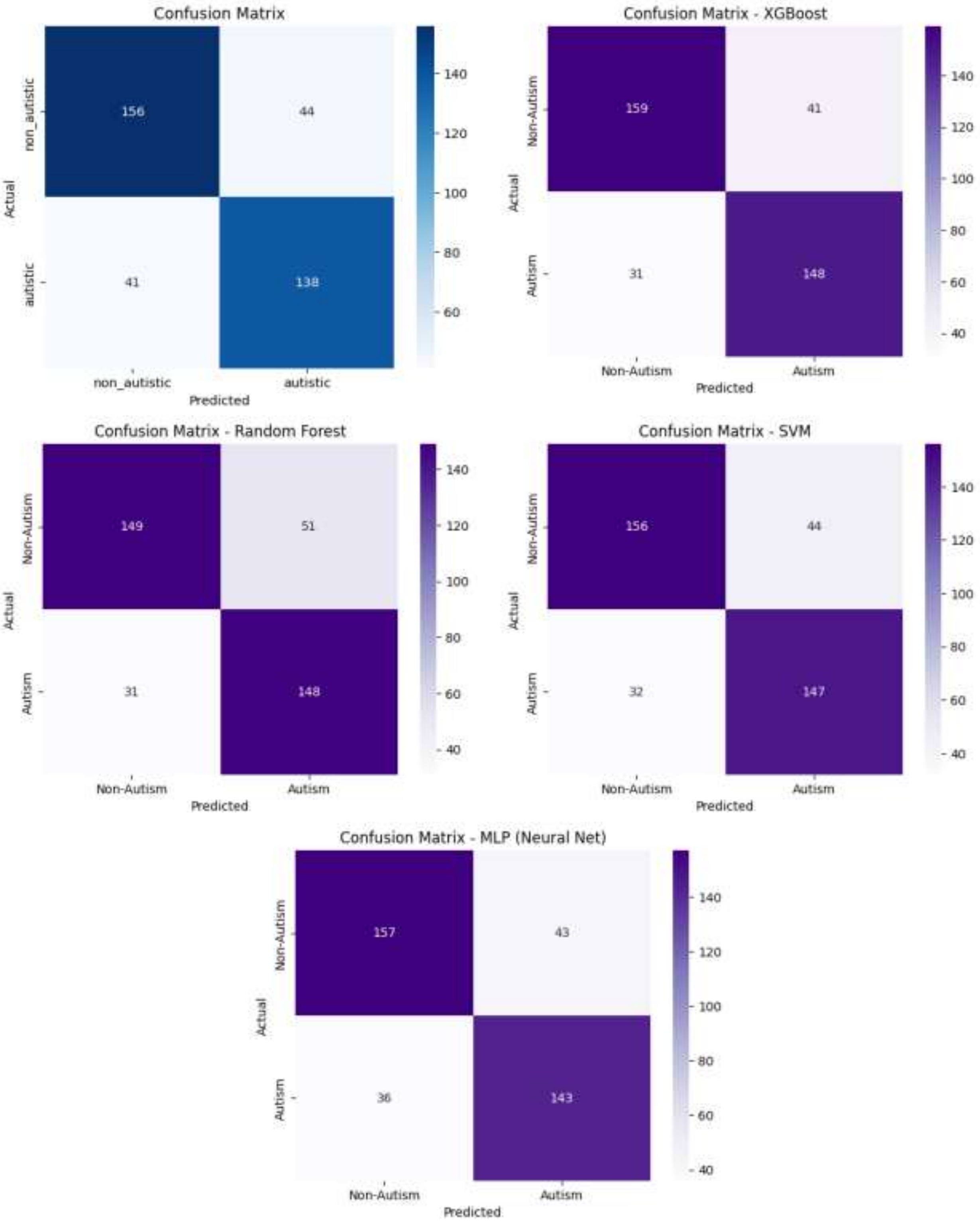

Figure 8. Confusion matrix of baseline classifiers

Figure 9 displays the confusion matrices of the proposed ensemble model evaluated across five stratified folds. The ensemble model consistently demonstrates robust classification performance, with high true positive (TP) and true negative (TN) values in each fold. In Fold 1, the model achieved a TP of 152 and TN of 162, indicating strong sensitivity and specificity. Fold 4 achieved the highest TP (163), showing the model's capacity to correctly identify autistic children, though it also had a slightly elevated FN (37). Fold 5 presented the lowest false negative rate (FN = 26), highlighting its effectiveness in minimizing missed autism cases. Across all folds, false positives (FP) remained moderate, ranging from 27 to 40, suggesting the model maintains a good balance between precision and recall. Overall, the ensemble model's performance is stable and reliable across folds, consistently outperforming individual baseline classifiers in terms of both detection accuracy and generalization.

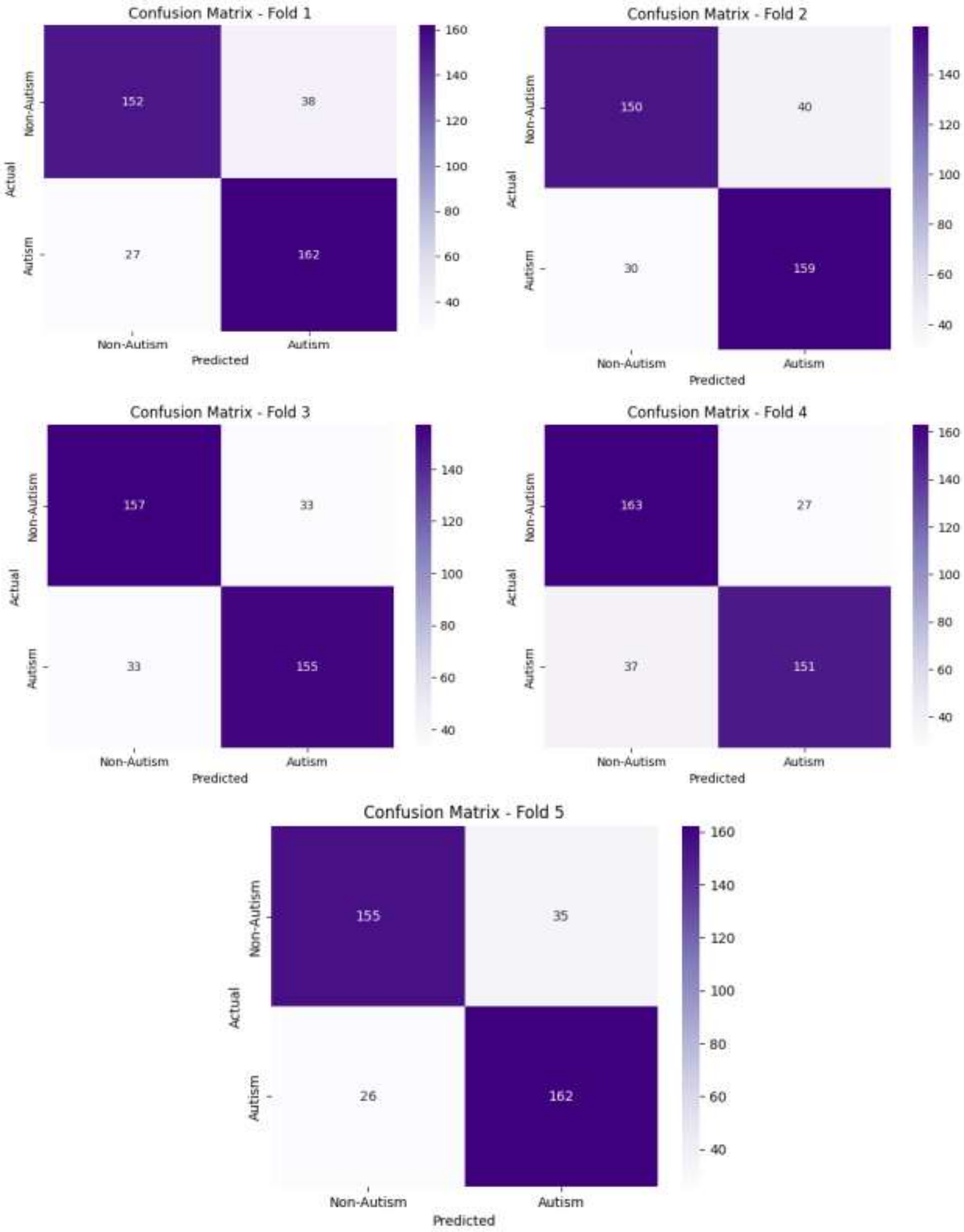

Figure 9. Confusion matrix depicting true vs. predicted labels for the ensemble-based classification

The ROC curved for the suggested ensemble model, averaged over the five cross-validation folds, is shown in Figure 10. Having a mean AUC of 0.91, the curve exhibits strong effectiveness in classification, showing a significant capacity for discrimination across autistic and non-autistic groups. With an AUC of 0.92, Fold 5 in particular demonstrated exceptional sensitivity and specificity. The framework's reliability and dependability for identifying autism based on facial traits are confirmed by the ROC curve, which continuously sits substantially at the random baseline (dashed diagonal). The model's promise for real-world testing screening where reducing false negatives is crucial is further demonstrated by its high true positive rate (TPR) at low false positive rates (FPR).

The projected dominating emotions, including joy, sadness, anger, fear, and surprise, are displayed across various facial emotions using the dataset in Figure 11. These results validate the model’s ability to correctly interpret emotional cues that are behaviorally relevant in autism diagnosis.

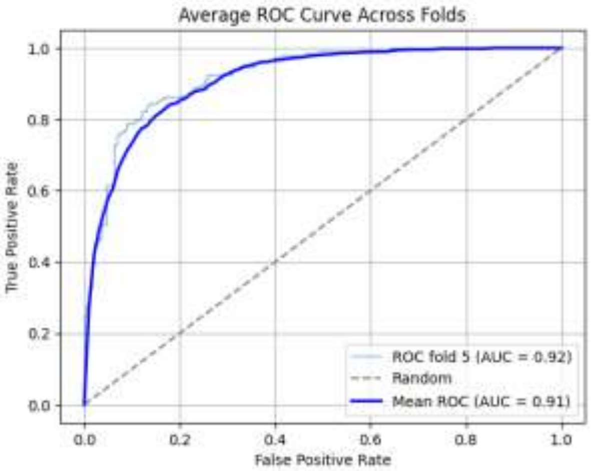

Figure 10. ROC-AUC of proposed ensemble model

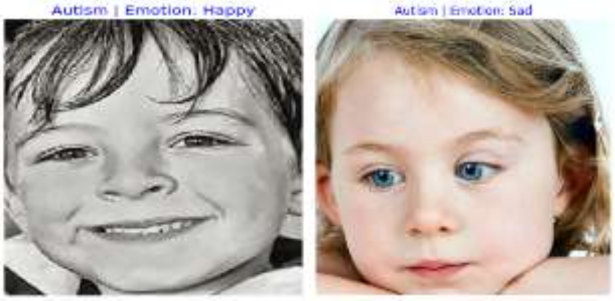

Figure 11. Predicted result of emotion analysis

Statistical Analysis

To assess the robustness and statistical significance of the proposed ensemble model, two statistical tests One-Way ANOVA and the Wilcoxon signed-rank test were conducted on the performance metrics reported in Table 2 (Baseline Models) and Table 3 (Cross-Validation of the Proposed Model).

One-Way ANOVA Test: The One-Way ANOVA test was applied to compare the performance of different models across all the evaluation parameters. The null hypothesis

$H0$ assumes no significant difference among the means of the compared classifiers, whereas the alternative hypothesis $H1$ assumes at least one significant difference. The ANOVA F-statistic is computed as:

$$F = \frac{Between - group\ variance}{Within - group\ variance} \quad (11)$$

where,

$$Between - group\ variance = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{k-1} \quad (12)$$

$$Within - group\ variance = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_{ij} - \bar{x}_i)^2}{N-k} \quad (13)$$

where, k is the number of groups, ni is the number of observations in group i, N is the total number of observations, $\bar{x}_i$ is the mean of group i, and $\bar{x}$ is the overall mean.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test, a non-parametric paired difference test, was used to compare the proposed ensemble model with each baseline classifier. This test is appropriate for non-normally distributed paired data. The Wilcoxon test statistic W is given by:

The ANOVA results revealed p-values < 0.05 for all metrics, indicating statistically significant differences among the models. This validates that the proposed ensemble model provides statistically higher performance than baseline models.

$$W = \sum_{i=1}^{n} R_i . sgn(d_i) \quad (14)$$

where, di is the difference between paired observations, R_i is the rank of $| d_i |$, and $sgn(d_i)$ is the sign function.

The results showed p-values < 0.05 for all pairwise comparisons, confirming that the proposed ensemble model significantly outperforms the individual baseline models in terms of all key performance metrics.

Descriptive Stability Analysis

To further analyze stability, the mean and standard deviation (SD) of each performance metric were computed from the five-fold cross-validation results in Table 4. The descriptive stability table is shown below:

Table 4. Descriptive stability of ensemble model across 5 folds

Metric	Mean	Std. Dev.	Mean ± SD
Accuracy	82.77	0.88	82.77±0.8882.77 \pm 0.88
Precision	82.08	1.73	82.08±1.7382.08 \pm 1.73
Recall	83.76	2.09	83.76±2.0983.76 \pm 2.09
F1-Score	82.87	0.88	82.87±0.8882.87 \pm 0.88
Kappa Score	65.94	1.71	65.94±1.7165.94 \pm 1.71
MCC Score	65.99	1.71	65.99±1.7165.99 \pm 1.71
ROC-AUC	91.75	0.61	91.75±0.6191.75 \pm 0.61

Table 5. Performance comparison of the proposed model with state-of-the-art autism detection

Author	Methods	Dataset Used	Findings
Smitha and Vinod [25]	PCA	JAFFE	82.3%
Awatramani and Hasteer [26]	CNN	FER2013	67.7%
Haque and Valles [27]	Texture features + SVM	Autism Images	77.96%
Reddy [28]	VGG19	Autism Images	51.44%
Reddy [28]	VGG16	Autism Images	54.15%
Farooq et al. [29]	SVM	Adult ASD	81%
Farooq et al. [29]	LR	Adult ASD	78%
Kadhum and Tawfeeq [30]	SVM	Autism Images	80.4%
Proposed	**Facial Features + Ensemble**	**Autism Images**	**84.00%**

The descriptive stability analysis of the proposed ensemble model across five-fold cross-validation, as shown in Table 4, demonstrates the model's robustness and consistency in performance. The results indicate that the ensemble model maintains high accuracy (82.77±0.88), precision (82.08±1.73), recall (83.76±2.09), and F1-score (82.87±0.88) with minimal variability across folds. Similarly, the Kappa score (65.94±1.71) and MCC score (65.99±1.71) show stable agreement and classification reliability, while the ROC-AUC (91.75±0.61) reflects consistently strong discriminatory power. The low standard deviation values across all metrics confirm the ensemble model's ability to deliver stable results regardless of data partitioning.

Table 5 shows the comparative analysis of the proposed ensemble model with existing models for autism and emotion recognition. Traditional ML and DL models such as PCA on JAFFE (82.3%), DCNN on FER2013 (67.7%), and texture feature–SVM models on autism images (77.96%) show moderate to good performance, while deep CNN model like VGG19 and VGG16 on autism datasets yield relatively lower accuracies of 51.44% and 54.15%, respectively. For adult ASD data, SVM and LR obtained the 81% and 78% accuracy. The proposed facial feature–driven ensemble model obtained the 84.00% accuracy on autism images that shows outperforming prior autism-focused methods and indicate superior discriminative capability for ASD-related facial patterns.

5. DISCUSSION

The effectiveness of the proposed method is further validated through a comparative analysis with existing autism detection approaches, as summarized in Table 4. Traditional methods like Principal Component Analysis (PCA) and handcrafted texture features combined with classical classifiers such as SVM have shown moderate success but are limited by their reliance on manually engineered features and domain-specific constraints. For instance, Smitha K.G. achieved 82.3% accuracy on the JAFFE dataset using PCA, while Haque M.I.U. reported 77.96% accuracy using texture features on ASD data. Deep learning models like VGG16 and VGG19, as implemented by Reddy P.J., performed significantly lower with 54.15% and 51.44% accuracy respectively, indicating limited transferability to autism-related facial cues. On the other hand, Farooq M.S. demonstrated improved performance with SVM and Logistic Regression on adult datasets, achieving 81% and 78% accuracy, respectively. In contrast, the proposed hybrid DL model used CNN-based feature extraction and an ensemble classifier achieved of 84.00% accuracy, outperforming all compared models. The combination of several CNN frameworks to capture various face expressions and the addition of emotion identification via DeepFace and MobileNetV2, which improves interpretability, are responsible for this better performance. These outcomes attest to the suggested framework's resilience and usefulness in actual autism assessments. The outcomes of the experiment validate the efficacy of the suggested hybrid DL technique for facial expression-based autism detection. The ensemble classifier consistently outperformed individual models such as LR, RF, SVM, and MLP based on evaluation parameters. A more thorough facial expression was achieved by the employment of multiple models for integrated deep feature extraction, and the ensemble learning approach improved generalization and decreased bias. Furthermore, a substantial interpretability layer was added by integrating emotion recognition using MobileNetV2 and DeepFace, which allowed the model to evaluate emotional expressions in addition to detecting autism. This multimodal analysis emphasizes emotional variations commonly seen in children with ASD and facilitates a more comprehensive interpretation of behavioral clues. Even though DeepFace did a good job at recognizing prevailing emotions, more training on datasets unique to ASD might improve its domain-specific capabilities. The proposed approach demonstrates a practical and effective solution for supporting early autism screening using readily available image data, combining the benefits of transfer learning, ensemble modeling, and emotion-aware classification.

6. CONCLUSION AND FUTURE SCOPE

This study presented a hybrid deep learning framework integrating multi-CNN feature extraction, a soft-voting ensemble classifier, and dual-stage emotion recognition to support early, non-invasive autism screening using facial images. The model achieved strong predictive performance, with 83.86% accuracy and 92.29% ROC-AUC, demonstrating the effectiveness of combining deep facial representations with emotion-aware cues to enhance the interpretability and reliability of ASD detection. While the findings reaffirm the potential of computer vision–based screening tools, several limitations must be acknowledged to contextualize the scope and applicability of the proposed approach.

First, the datasets used although publicly available and widely adopted are relatively limited in size and demographic diversity. Variations in ethnicity, age distribution, camera quality, and environmental conditions were not fully represented, which may restrict the model's generalizability to broader real-world populations. Second, the model relied exclusively on static facial images and did not incorporate multimodal behavioral signals such as voice patterns, body movements, or eye-tracking trajectories, which carry essential diagnostic value in early ASD assessment. Lastly, the framework was evaluated under controlled experimental conditions and has not yet been validated in clinical or in-the-wild settings where naturalistic behavior may differ substantially.

Future work should focus on expanding the dataset with larger, more diverse, and clinically validated samples to enhance robustness across populations and imaging conditions. Addressing emotion class imbalance through curated data collection or advanced augmentation strategies will further improve emotional inference reliability. Integrating multimodal behavioral cues such as gaze patterns, EEG, speech prosody, or micro-motion analysis may strengthen the model's diagnostic completeness. Deploying the framework as a mobile or cloud-based screening tool and evaluating it in clinical and educational environments will provide critical insights into its real-world feasibility. Exploring explainable AI techniques may also improve transparency, enabling clinicians to interpret model decisions with greater confidence. Overall, the proposed approach represents a promising step toward accessible, AI-driven autism screening, while highlighting the need for continued research to ensure fairness, scalability, and clinical effectiveness.

REFERENCES

[1] Hodges, H., Fealko, C., Soares, N. (2020). Autism spectrum disorder: Definition, epidemiology, causes, and

clinical evaluation. Translational Pediatrics, 9(Suppl 1): S55. https://doi.org/10.21037/tp.2019.09.09

[2] Solek, P., Nurfitri, E., Sahril, I., Prasetya, T., Rizqiamuti, A.F., Rachmawati, I., Gunawan, K. (2025). The role of artificial intelligence for early diagnostic tools of autism spectrum disorder: A systematic review. Turkish Archives of Pediatrics, 60(2): 126. https://doi.org/10.5152/TurkArchPediatr.2025.24183

[3] Ruan, M., Zhang, N., Yu, X., Li, W., Hu, C., Webster, P.J., Li, X. (2024). Can micro-expressions be used as a biomarker for autism spectrum disorder? Frontiers in Neuroinformatics, 18: 1435091. https://doi.org/10.3389/fninf.2024.1435091

[4] Chen, J., Chen, C., Xu, R., Liu, L. (2024). Autism identification based on the intelligent analysis of facial behaviors: An approach combining coarse-and fine-grained analysis. Children, 11(11): 1306. https://doi.org/10.3390/children11111306

[5] Haque, N., Islam, T., Erfan, M. (2025). An exploration of machine learning approaches for early Autism Spectrum Disorder detection. Healthcare Analytics, 7: 100379. https://doi.org/10.1016/j.health.2024.100379

[6] Alsaidi, M., Obeid, N., Al-Madi, N., Hiary, H., Aljarah, I. (2024). A convolutional deep neural network approach to predict autism spectrum disorder based on eye-tracking scan paths. Information, 15(3): 133. https://doi.org/10.3390/info15030133

[7] Grzadzinski, R., Amso, D., Landa, R., Watson, L., Guralnick, M., Zwaigenbaum, L., Piven, J. (2021). Pre-symptomatic intervention for autism spectrum disorder (ASD): Defining a research agenda. Journal of Neurodevelopmental Disorders, 13(1): 49. https://doi.org/10.1186/s11689-021-09393-y

[8] Kang, J., Han, X., Song, J., Niu, Z., Li, X. (2020). The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data. Computers in Biology and Medicine, 120: 103722. https://doi.org/10.1016/j.compbiomed.2020.103722

[9] Mahmood, M.A., Jamel, L., Alturki, N., Tawfeek, M.A. (2025). Leveraging artificial intelligence for diagnosis of children autism through facial expressions. Scientific Reports, 15(1): 11945. https://doi.org/10.1038/s41598-025-96014-6

[10] Talaat, F.M. (2023). Real-time facial emotion recognition system among children with autism based on deep learning and IoT. Neural Computing and Applications, 35(17): 12717-12728. https://doi.org/10.1007/s00521-023-08372-9

[11] Afrin, M., Hoque, K.E., Chaiti, R.D. (2024). Emotion recognition of autistic children from facial images using hybrid model. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, pp. 1-6. https://doi.org/10.1109/icccnt61001.2024.10724499

[12] Walujo, I.Y., Syarif, I., Fariza, A. (2024). Autism detection based on deep learning. Indonesian Journal of Computer Science, 13(6): 9370-9383. https://doi.org/10.33022/ijcs.v13i6.4552

[13] Ahmed, F., Srinivasa, G., John, D., Prince, S. (2024). Emotion detection in children with autism spectrum disorder: leveraging EEG signals and companion bots for enhanced interaction. In 2nd International Conference on Computer Vision and Internet of Things (ICCVIoT 2024), Coimbatore, India, 2024: 307-312. https://doi.org/10.1049/icp.2024.4440

[14] Talaat, F.M., Ali, Z.H., Mostafa, R.R., El-Rashidy, N. (2024). Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children. Soft Computing, 28(9-10): 6695-6708. https://doi.org/10.1007/s00500-023-09477-y

[15] Mittal, R., Malik, V., Rana, A. (2022). DL-ASD: A deep learning approach for autism spectrum disorder. In 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, pp. 1767-1770. https://doi.org/10.1109/IC3I56241.2022.10072429

[16] Abu-Nowar, H., Sait, A., Al-Hadhrami, T., Al-Sarem, M., Qasem, S.N. (2024). SENSES-ASD: A social-emotional nurturing and skill enhancement system for autism spectrum disorder. PeerJ Computer Science, 10: e1792. https://doi.org/10.7717/peerj-cs.1792

[17] Poornima, S., Kousalya, G. (2022). Deep learning based behavioral analysis and exploration of emotions in ASD children. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, pp. 504-509. https://doi.org/10.1109/ICAIS53314.2022.9742842

[18] Shvimmer, S., Simhon, R., Gilead, M., Yitzhaky, Y. (2022). Classification of emotional states via transdermal cardiovascular spatiotemporal facial patterns using multispectral face videos. Scientific Reports, 12(1): 11188. https://doi.org/10.1038/s41598-022-14808-4

[19] Na, I.S., Aldrees, A., Hakeem, A., Mohaisen, L., Umer, M., Al Hammadi, D.A., Alsubai, S., Innab, N., Ashraf, I. (2024). FacialNet: Facial emotion recognition for mental health analysis using UNet segmentation with transfer learning model. Frontiers in Computational Neuroscience, 18: 1485121. https://doi.org/10.3389/fncom.2024.1485121

[20] Huang, Z.Y., Chiang, C.C., Chen, J.H., Chen, Y.C., Chung, H.L., Cai, Y.P., Hsu, H.C. (2023). A study on computer vision for facial emotion recognition. Scientific Reports, 13(1): 8425. https://doi.org/10.1038/s41598-023-35446-4

[21] Bartlett, M., Viola, P., Sejnowski, T.J., Golomb, B., Larsen, J., Hager, J., Ekman, P. (1995). Classifying facial action. Advances in Neural Information Processing Systems, 8: 823-829.

[22] Wankhede, D.S., Selvarani, R. (2022). Dynamic architecture based deep learning approach for glioblastoma brain tumor survival prediction. Neuroscience Informatics, 2(4): 100062. https://doi.org/10.1016/j.neuri.2022.100062

[23] Maria Arockia Dass, J., Sirisha, K., Hema, B., Girisha, A., Hareesh, K., Jyothish Kumar, P. (2024). Autism spectrum disorder detection using face features based on deep neural network. In 2024 International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET), Gautam Buddha Nagar, India, pp. 1008-1012. https://doi.org/10.1109/I3CEET61722.2024.10994079

[24] Alshathri, S., Talaat, F.M., Nasr, A.A. (2022). A new reliable system for managing virtual cloud network. Computers, Materials & Continua, 73(3): 5863-5885. https://doi.org/10.32604/cmc.2022.026547

[25] Smitha, K.G., Vinod, A.P. (2015). Facial emotion recognition system for autistic children: A feasible study based on FPGA implementation. Medical & Biological

Engineering & Computing, 53(11): 1221-1229. https://doi.org/10.1007/s11517-015-1346-z

[26] Awatramani, J., Hasteer, N. (2020). Facial expression recognition using deep learning for children with autism spectrum disorder. In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 35-39. https://doi.org/10.1109/ICCCA49541.2020.9250768

[27] Haque, M.I.U., Valles, D. (2018). A facial expression recognition approach using DCNN for autistic children to identify emotions. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, pp. 546-551. https://doi.org/10.1109/IEMCON.2018.8614802

[28] Reddy, P. (2024). Diagnosis of autism in children using deep learning techniques by analyzing facial features. Engineering Proceedings, 59(1): 198. https://doi.org/10.3390/engproc2023059198

[29] Farooq, M.S., Tehseen, R., Sabir, M., Atal, Z. (2023). Detection of autism spectrum disorder (ASD) in children and adults using machine learning. Scientific Reports, 13(1): 9605. https://doi.org/10.1038/s41598-023-35910-1

[30] Kadhum, S.W., Tawfeeq, M.A. (2025). Early detection of autism spectrum disorder in children using different machine learning algorithms. medRxiv. https://doi.org/10.1101/2025.04.13.25323013