Enhanced Facial Emotion Recognition for Autism Spectrum Disorder Using Multi-Model CNN Architectures

Hira Iqbal[1], Muhammad Siraj Rathore[2], Fareeha Ashraf[1]
[1]Software Engineering, Capital University of Science and Technology, Islamabad, Pakistan
[2]Computer Science, Szabist University, Islamabad, Pakistan
*Correspondence: fareeha.ashraf01@gmail.com

Citation | Iqbal. H, Rathore. M. S, Ashraf. F, "Enhanced Facial Emotion Recognition for Autism Spectrum Disorder Using Multi-Model CNN Architectures", IJIST, Vol. 8 Issue. 9 pp 3708-3737, October 2026
Received | Sep 11, 2026 **Revised |** Oct 10, 2026 **Accepted |** Oct 20, 2026 **Published |** Oct 27, 2026.

Facial Emotion Recognition (FER) plays an important role in assistive technologies for individuals with autism spectrum disorder (ASD), particularly in therapeutic and educational settings. However, existing FER systems often rely on generic facial expression datasets that may not adequately represent the distinctive characteristics of emotional expressions associated with ASD. To address this limitation, this study develops an enhanced FER framework by integrating the FER-2013 dataset with the FT-2022 ASD-specific dataset. Following quality filtering, the resulting hybrid dataset contained 30,106 facial images, combining the diversity of general facial expressions with ASD-specific emotional characteristics. Three transfer-learning-based convolutional neural network (CNN) architectures, namely VGG16, ResNet50, and Xception, were systematically evaluated individually and in combination. On the hybrid dataset, the individual models achieved accuracies of 91% (VGG16), 90% (ResNet50), and 68% (Xception), with corresponding F1-scores of 0.90, 0.89, and 0.68, respectively. Three multi-model configurations were subsequently evaluated: ResNet50–VGG16, ResNet50–Xception, and VGG16–Xception. The corresponding accuracies were 99%, 98%, and 65%, with F1-scores of 0.99, 0.98, and 0.65, respectively. The ResNet50–VGG16 configuration achieved the highest accuracy, improving performance by 8 percentage points over the best individual model, VGG16. These results demonstrate that combining complementary CNN architectures can substantially improve FER performance when applied to a diverse hybrid dataset containing both general and ASD-specific facial expressions. The proposed framework provides a promising foundation for ASD-focused assistive technologies supporting emotion recognition and social interaction in therapeutic and educational settings.

Keywords: Facial Emotion Recognition (FER), Autism Spectrum Disorder (ASD), Deep Learning, Transfer Learning, Convolution Neural Network (CNN)

Introduction:

Emotions are an integral part of communication and social interaction. Individuals communicate through facial expression, verbal language, gestures and body language. Of these modes of communication, facial expression is one of the most effective and informative methods of nonverbal communication as it gives the person information about their emotional state at the moment. In fact, the accurate identification of emotional states is crucial for daily life and is applied in healthcare, education, security, behavioral analysis, human–computer interaction and assistive technologies [1].

The Facial Emotion Recognition (FER) is an active research area, as advances in AI, computer vision, and machine learning accelerate in the face of a rapidly changing world. FER's purpose is to automatically recognize human emotions found in an image or a video sequence by analyzing facial features and expression patterns. Due to its potential applications in the fields of smart home, intelligent tutoring systems, healthcare monitoring, surveillance systems, depression assessment, patient monitoring, and human–computer interaction applications, FER systems have received much attention in recent years, both in academia and industry [2]. With the rising demand for intelligent systems that can understand human emotions, research in this area has rapidly gained momentum.

With the recent advances in Deep Learning, the performance of FER systems has significantly been enhanced. Unlike traditional machine-learning methods, deep-learning models are able to learn discriminative facial representations automatically from large-scale image datasets, thus reducing the need of manually engineered features and achieving better performance in facial expression recognition [3]. The various deep learning architectures have been successfully applied to FER tasks, producing better accuracy and robustness in FER. Despite the many FER systems that operate successfully, the limited training data, variations in facial appearance, low quality images, class imbalance, and real-time implementation constraints still limit the performance of many FER systems [4]. To overcome these challenges, researchers have turned to a number of techniques to boost model performance and improve their generalization abilities, such as transfer learning, data augmentation, attention mechanisms, and ensemble learning approaches.

Despite the great combination through ensemble learning development of FER, the application to special populations is still a difficult problem. An important aspect is the ability to recognize feelings in people with autism spectrum disorder. ASD is a neurodevelopmental condition, which impacts social communication, social interaction, and behavior. The World Health Organization (WHO) estimated that one in every 100 children in the world suffer from ASD [5]. The prevalence of ASD has been reported to increase in recent studies and there is a need for development of effective tools that can be used to support diagnosis, intervention and behavioral assessment [6].

People with ASD may have trouble understanding or interpreting social cues and emotions, and trouble expressing them in social situations [7]. These difficulties may impact communication, learning and social interactions. Moreover, studies indicate that the expression of emotions may be different for autistic people compared to non-autistic people, leading to challenges for caregivers, teachers, and health care providers in recognizing autistic emotions [8]. Early identification and intervention have found to enhance social communication skills and quality of life for people with ASD [9]. So, it has become important to work on reliable FER systems for the ASD-related applications.

The applications of artificial intelligence and deep learning algorithms for facial analysis in ASD have been investigated in various studies [10]. Although the current FER models have had good outcomes on public available datasets, most of these datasets rely on a relatively small number of facial expressions from neurotypical individuals. Therefore, the models developed from such data sets might not properly reflect the nuances and atypical emotions that are

commonly seen in people with ASD. Autism-specific datasets can also be smaller, leading to potential overfitting and a lack of model generalization. This suggests that existing FER systems may not achieve the same level of performance for both autistic and non-autistic people [11].

Convolutional neural networks (CNNs) have been widely adopted as a deep learning solution for Facial Emotion Recognition (FER) due to their strong capability to automatically learn discriminative facial features and their consistently high classification performance reported in existing studies [12]. In the context of ASD-related FER, many studies have evaluated individual CNN architectures independently, often using a single facial expression dataset. Specifically, the transfer-learning models including VGG16, ResNet50 and Xception have demonstrated encouraging performance when used individually for facial-expression datasets. However, there are relatively few studies that have systematically evaluated the performance of these architectures under the same experimental conditions or studied their combination through ensemble learning [13]. Similarly, limited research has explored the integration of generic facial expression datasets with ASD-specific datasets to provide more diverse and representative training data. Furthermore, multi-model deep learning approaches that leverage the complementary strengths of multiple CNN architectures remain relatively underexplored in ASD-focused FER. These limitations highlight the need for comprehensive FER frameworks that can improve classification accuracy, robustness, and generalization for ASD-related applications [14].

To address these challenges, the present study proposes an enhanced facial emotion recognition framework based on integrated facial expression datasets and multi-model CNN architectures. The proposed dataset integrates the FER-2013 generic facial expression dataset with the Fatma Talaat ASD-specific facial expression dataset, thereby increasing the diversity and representation of both typical and ASD-related emotional expressions. Three state-of-the-art CNN models, namely VGG16, ResNet50, and Xception, are independently trained and evaluated on the integrated dataset. Subsequently, these CNN models are combined using transfer learning to develop multi-model CNN architectures that leverage their complementary feature representations. The proposed multi-model architectures aim to improve facial emotion recognition accuracy, robustness, and prediction consistency for ASD-focused applications. Experimental results demonstrate that the multi-CNN configurations substantially outperform the individual models, with classification accuracy improving from 91% for the best-performing individual model, VGG16, to 99% for the ResNet50–VGG16 configuration. These findings demonstrate that combining complementary CNN models can significantly enhance facial emotion recognition performance and provide a more robust approach for ASD-focused applications.

This study has made the following major contributions:

Performance Evaluation of individual CNN models:

Three CNN architectures, namely VGG16, ResNet50, and Xception are systematically tested for facial-expression recognition. They are evaluated using Fatma Talaat dataset of facial expressions for the ASD (autism spectrum disorder).

Integrated and balanced dataset:

In order to create a more generalized and robust FER dataset, the FER2013 dataset is combined with the Fatma Talaat ASD-specific dataset. The generic dataset reveals a wider range of expressions and includes facial elements related to the ASD and allows for a more thorough analysis of FER models.

Multi-model CNN architecture:

A multi-model CNN framework is presented, which combines the selected CNN models via transfer learning. The aim of the framework is to take advantage of the complementary feature representation extracted by VGG16, ResNet50 and Xception to achieve better recognition accuracy, robustness and prediction consistency.

Literature Review:

Facial Emotion Recognition for Neurotypical People:

Facial emotion recognition (FER) has become an important research area, with applications in healthcare, education, human--computer interaction, security, and assistive technologies [15]. Computer Vision and Affective Computing focus on automatically recognizing human emotions through facial expressions. Early FER systems relied on handcrafted features such as LBP, HOG, facial landmarks, and geometric measurements combined with classifiers like SVM, Random Forest, and Decision Trees. However, these methods struggled with variations in pose, illumination, occlusion, and individual differences [16].

Facial emotion analysis is a state-of-the-art technique that has advanced significantly with the advent of deep learning. Deep learning has become a cutting-edge solution for facial emotion analysis. This technique has become well-known in the field of facial emotion recognition (FER) in addition to its capacity to automatically learn discriminative features. Deep-learning methods can automatically learn facial features, reducing the need for manually engineered features [17]. Among deep-learning methods, Convolutional neural networks (CNNs) have proven to be especially effective. Hierarchical spatial representations are crucial for performance in FER, and they can be learned via successive convolution, pooling and classification operations [18]. Its hierarchical representation-learning ability allows CNNs to learn subtle and complex patterns in faces related to of different emotional states, which would be suitable for accurate and robust facial emotion recognition.

The CNN-based FER models achieve better performance on benchmark datasets such as CK+, FER2013, RAF-DB, and AffectNet [19]. CNN architectures, like VGG19, ResNet50, MobileNet and InceptionV3, have been shown to perform well on facial-expression recognition, but the reported accuracies vary a lot depending on the dataset and evaluation protocol. For example, on the controlled KDEF dataset, ResNet50 and InceptionV3 reached accuracies of 97.55%, and VGG19 achieved 96.73% with selected test samples [20]. In another comparison on the modified

FER2013 dataset, MobileNetV2 achieved the best testing accuracy of 83.64% among the models compared [21]. Similarly, advanced methods using attention mechanisms and residual networks improve robustness, achieving strong results on datasets like CK+, RAF-DB, FERPlus, and AffectNet [22]. However, most existing FER studies focus on general emotion recognition and do not address the unique facial expression characteristics of individuals with autism spectrum disorder (ASD).

One of the main challenges in facial emotion recognition (FER) is the limited availability of large-scale, well-annotated emotion datasets. Recently, there has been a great deal of interest in learning-based approaches, where a CNN is learned from a large dataset of images of the same class that is available on the Internet. Fine-tuning on a large dataset like ImageNet is done for FER. This strategy can increase recognition accuracy, yet decrease the training time and number of Specific data for a task is needed [23]. Using 10-fold cross-validation on the CK+ dataset, the pretrained models achieved accuracies of 94.65% for VGG19, 91.1% for ResNet50, 97.1% for InceptionV3, and 98.7% for DenseNet-161. DenseNet-161 achieved the highest accuracy among the evaluated models [24].

In another study [25], the use of attention mechanism is proposed in order to increase the accuracy of facial expression recognition. To enhance the accuracy of DenseNet model, the attention mechanism is applied and it is demonstrated that proposed solution improves accuracy on controlled dataset CK+ and real-world dataset RAF-DB, which reach 99.94% and 83.18% respectively. Also, multi-model CNN architectures have been investigated to enhance the performance of facial emotion recognition (FER) systems. For example, a CNN--RNN architecture proposed to glean spatial features from facial image. Model temporal relationships

among frames in temporal sequences of images. The CNN--RNN model obtained an accuracy rate of 82.4% on the KDEF dataset, which means its suitability to sequence or video-based FER. However, this architecture is more complex than the conventional single-frame CNN models because it needs to deal with spatial and temporal information.

The use of multimodal CNN approaches that integrate facial features with others have also been explored a one such method employed the M-LFW-F and CREMA-D datasets to combine emotional elements of sight and sound. Many problems in the field of multimodal fusion can to boost recognition performance more than a single modality model, and additional speech processing and feature-fusion components added. Requires computational resources and may be restricted to deployment in resource-rich environments real-time applications [26]. In summary, it can be observed that most of existing FER studies achieve high accuracy on benchmark datasets such as CK+, RAF-DB, and FER2013. However, these studies primarily rely on neurotypical facial expression datasets and do not address ASD-specific emotion recognition.

Facial Emotion Recognition for Autistic People:

One major limitation of ASD-related FER is the lack of ASD-specific datasets. Typical FER benchmarks such as FER2013, CK+, RAF-DB and AffectNet are not designed for ASD research and usually lack ASD-related diagnostic labels. To overcome this limitation, Ji et al. proposed the Hugging Rain Man (HRM) dataset that includes around 130,000 video frames of children with ASD and typically developing (TD) children. The dataset provides annotations for 22 action units, 10 action descriptors and perceived atypicality. However, HRM is more focused on facialaction and atypicality analysis than on direct emotion classification [27].

Another dataset related to ASD is the Autistic Children Emotions dataset created by Fatma M. [28]. This dataset includes facial images of autistic children with six classes of emotion. [26]use this custom dataset and propose a deep-learning approach for emotion recognition from facial images of autistic children [29]. They use a convolutional network based on MobileNetV2, data augmentation and meta-learning techniques.

In another work [30], the hybrid CNN model of DenseNet121 and MobileNetV2 has been trained on the FERAC autism-specific dataset. The multi-model solution performed better than the individual CNN backbones, but was tested on a smaller amount of ASD facial images (i.e. 770 images in total, 691 used for training and 79 used for testing). Another study [31] by authors introduced a facial emotion recognition (FER) system based on a CNN-based model and achieved approximately 65% accuracy on ASD facial images, which, however, was restricted by limited size and imbalanced dataset of ASD images.

Subsequent studies investigated more complicated architectures and class-imbalance issues. In a multimodal late-fusion framework, a modified Xception model was tested on children with ASD and presented an emotion accuracy of approximately 68% for the categorical emotion classes, and a significant decrease was observed for minority emotion classes [32]. Likewise, DenseNet121 demonstrated superior performance when compared to other architectures in the overall performance for FER tasks, though models continued to have difficulties with underrepresented emotions [33].

Finally, CNN and Vision Transformer (ViT) models were finally compared in the context of ASD facial images, which showed that accuracy decreased from approximately 74% on the FER2013 dataset to approximately 65% on the ASD dataset, highlighting the need for more representative, ASD specific data [34].

In the multimodal approach, a hybrid model (Xception model modified with facial and vocal modality, dubbed multimodal) reported an accuracy of 79.81% on M-LFW-F and CREMA-D, which demands more computational resources, but achieves better performance with multimodal input [35]. CNN--RNN hybrid models have been demonstrated to outperform

traditional CNN models on general FER datasets like FER2013 and CK+ but they have not been directly applied to the ASD population [36].

Table 1. Representative deep-learning studies on facial emotion recognition in ASDrelated datasets

S.No.	Study (Ref.)	Dataset	Method	Key Result	Limitation
1	[37]	Existing facial expression dataset	CNN	67.50% accuracy	Limited dataset and generalization
2	[28]	HRM: ~130k ASD/TD frames; AU annotations	AU detection and temporal atypicality analysis	AU-detection baselines; no categorical FER accuracy	Not a conventional categorical FER benchmark; partial data access
3	[29]	Autistic Children Emotions dataset	EDL/CNN with GA optimization	99.99% reported accuracy across six emotions	Small dataset; limited external validation
4	[32]	FERAC: 770 images, four emotion classes	DenseNet121 – MobileNetV2 hybrid	75% accuracy	Small and imbalanced dataset
5	[31]	FERAC dataset	CNN and transformer architectures	Swin Transformer: 80.00% accuracy	Small dataset; difficult minority emotions
6	[33]	ASD facial expression and biosignal data	Multimodal late fusion	68% categorical accuracy; 78% likelihood-based accuracy	Strong class imbalance; weak minority-class recognition
7	[32]	FER2013 + ASD facial images	CNN and ViT	CNN: 74% on FER2013 and 65% on ASD images; ViT: 69% on FER2013	Limited ASD data; reduced performance on ASD images
8	Proposed study (this work)	FER2013 + Fatma M. Talaat ASD dataset	Multi-model CNN architecture	99% best accuracy for our multimodal CNN architecture	Further external and real-world validation required

Research Gap and Motivation:

In recent years, many facial expression recognition (FER) studies have been performed on large-scale FER benchmark datasets such as FER2013, CK, RAF-DB, JAFFE and AffectNet with the high accuracy using CNN, transfer learning and hybrid deep learning method, etc. This is because these data sets are mainly composed of facial expressions of typically developing children and may not provide a comprehensive view of the facial expression patterns found in children with autism spectrum disorder (ASD).

Table 1 lists some of the studies that have been conducted on deep learning-based FER with ASD and shows the potential of deep learning models to achieve promising accuracy. For example, Talaat [28] reported 99.99% accuracy using an EDL/CNN model with GA optimization, while Radočaj et al. achieved 80.00% accuracy using a Swin Transformer. [32] reported 75% accuracy using a DenseNet121--MobileNetV2 hybrid model. However, many of

these studies are based on small ASD-related datasets, such as the Fatma M. Talaat ASD dataset and FERAC, as well as other limited collections of ASD images. Small sample sizes, class imbalance, limited validation, and lack of diversity may limit the generalizability and robustness of the developed models. In addition, some works in the literature focus on facial analysis and action unit annotation rather than conventional multi-class facial emotion recognition.

Despite the promising results reported in ASD-oriented FER studies, a clear research gap remains in developing and systematically evaluating a robust FER framework that combines diverse general-purpose and ASD-specific facial-expression data while investigating whether multimodel CNN architectures can provide better performance and generalization than individual CNN models. This gap motivates the proposed approach, which combines FER2013 with the Fatma M. Talaat ASD dataset and systematically evaluates individual and integrated CNN architectures. In this context, three CNN-based architectures namely VGG16, ResNet50 and Xception are first individually evaluated on hybrid dataset. Moreover, transfer-learning based multi-model CNN architectures are designed by integration of various CNN models to explore if model integration improves face emotion classification than individual CNN model.

The study aims at finding the most suitable single model and multi-model CNN architecture and boost the robustness and generalization performance of facial emotion recognition in the context of ASD. As opposed to the method that uses only small dataset for ASD, the proposed hybrid-dataset method offers more types of facial expressions for training models. We believe that the systematic comparison of individual CNNs and their combinations would yield a more robust framework for FER that could reach the maximum accuracy, and tentatively overcome the shortcomings of previous ASD-specific FER studies.

The novelty of this study lies in the development and systematic evaluation of a multimodel deep learning framework for facial emotion recognition in the context of Autism Spectrum Disorder (ASD), addressing limitations identified in both ASD-specific FER and existing multimodel FER studies. Unlike previous ASD-FER studies that primarily rely on relatively small and limited ASD-specific datasets, this study combines the FER2013 dataset with the Fatma M. Talaat ASD dataset to construct a more diverse hybrid dataset containing facial expressions from both general and ASD-specific sources. This dataset strategy is intended to provide broader variation in facial appearance and emotional expression and to improve the diversity of the data available for model training and evaluation.

A second contribution is the systematic comparison of individual and multimodel CNN architectures within the same experimental framework. The study evaluates VGG16, ResNet50, DenseNet121, MobileNetV2, Xception CNN models individually and further investigates combinations of these architectures using transfer learning. Rather than evaluating a single proposed architecture, the study examines whether integrating complementary feature representations from different CNN models can improve facial emotion classification compared with individual models. This provides a direct comparison between single-model and multimodel approaches and enables the identification of the most effective architecture for the considered ASD-oriented FER task.

Therefore, the contribution of this work is not limited to applying existing CNN architectures to an ASD dataset or simply combining multiple CNN models. It lies in the integration of a more diverse hybrid dataset with a systematic individual-versus-multimodel evaluation strategy specifically designed to investigate robust facial emotion recognition in the ASD context. The experimental analysis identifies the best-performing individual and multimodel configurations and provides evidence regarding the effectiveness of CNN model integration for ASD-oriented facial emotion recognition.

Research Objectives:

The primary objective of this study is to develop and systematically evaluate deep learning-based facial emotion recognition models using a hybrid dataset containing both general

facial expression data and ASD-specific facial-expression data. The specific objectives of the study are:

To construct a hybrid facial emotion recognition dataset by combining the FER2013 and Fatma M. Talaat ASD datasets to provide greater diversity in facial expressions and representation for ASD-oriented emotion recognition.

To evaluate individual CNN architectures, including VGG16, ResNet50, DenseNet121, MobileNetV2, Xception, and a custom CNN, for multi-class facial emotion recognition using the hybrid dataset.

To develop multimodel CNN architectures by integrating different CNN models and investigate whether combining their learned feature representations improves facial emotion classification compared with individual CNN models.

To quantitatively compare individual and multimodel architectures using accuracy, precision, recall, F1-score, and confusion-matrix-based analysis, and identify the best performing architecture.

To investigate the effect of increased dataset diversity on ASD-oriented facial emotion recognition, particularly by incorporating both general FER data and ASDspecific facial-expression data rather than relying solely on a limited ASD-specific dataset.

To identify a high-performing and comparatively robust CNN architecture for ASDoriented facial emotion recognition based on the experimental results obtained from the hybrid dataset.

Proposed Multi model CNN Architecture and Integrated Dataset:

In this section, a hybrid facial emotion recognition (FER) framework for autism spectrum disorder (ASD) is proposed. We have two objectives while proposing our solution. Firstly, to develop a system that can provide high accuracy while recognizing/classifying the emotions of autistic people. Secondly, the system should be reliable enough to maintain its accuracy while deploying in a real-world scenario. We know from literature that existing solutions such as [28] provides high accuracy of 91% but CNN models were trained on a small dataset. In this situation, maintaining the same accuracy might be challenging under different real-world scenarios. It is due to the fact the model may perform under or overfitting due to limited training. On the other hand, the generic facial emotion recognition solutions (which are not particularly focused on autistic people but to classify emotions of any person) have lower accuracy when evaluated for autistic people. Although, the CNN models of these solutions are trained on large datasets but there is usually no training on autistic images. In some cases, the facial emotions of autistic people may differ than the other people, since models are not trained on autistic images, they result in lower accuracy when evaluated for autistic people.

To address the aforementioned problem, our methodology is divided into three parts: (i) The performance of three top rated CNN models namely Xception, Resnet and VGG16 is evaluated while using both FT-2022 ASD custom dataset along generic dataset FER-2013. This provides us a baseline case to make a comparison with our proposed solution which is based on hybrid dataset and multi-model architecture. (ii) a Hybrid Dataset is created by combining FT-2022 ASD custom dataset along generic dataset FER-2013, and the three CNN models are trained and tested on the Hybrid Dataset. (iii) the CNN models are then combined to create multi-model to further improve the accuracy. In this way, the proposed framework is based on a multi-model CNN architecture which is trained on a hybrid dataset.

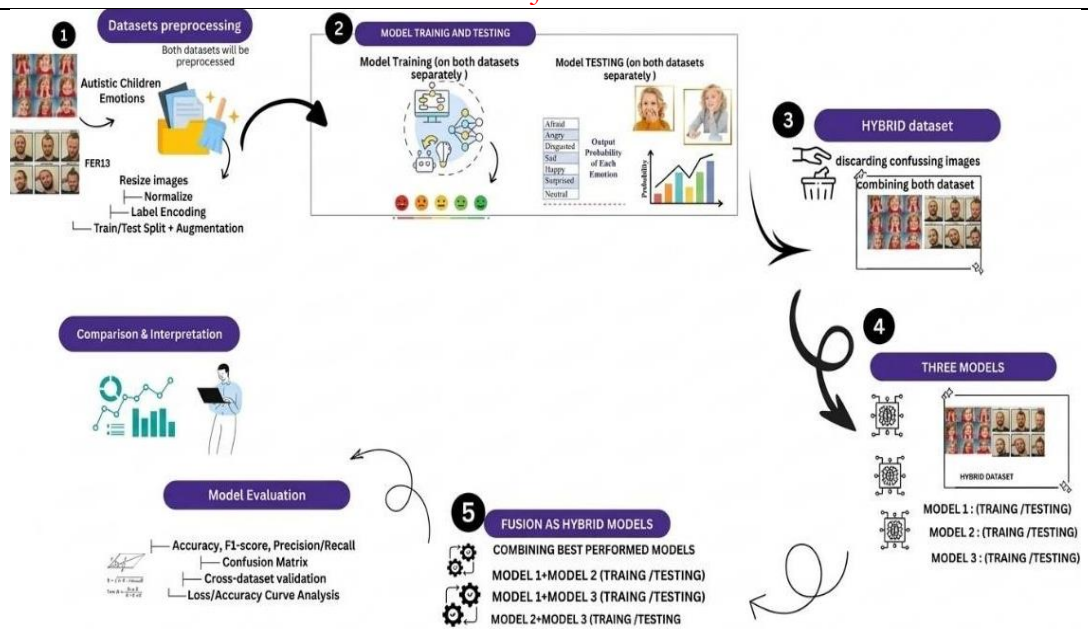

Figure 1. Proposed Hybrid Model Framework for Emotion Recognition Using Autism Children's Emotions and FER13 Datasets.

Preprocessing of FT-2022 and FER 2013 Datasets:

The preprocessing pipeline applied to the FT-2022 and FER-2013 datasets is illustrated in Figure 1. To ensure a consistent input format across both datasets, all facial images were resized to 48 × 48 pixels. This standardized the spatial dimensions of the input images while reducing computational requirements during model training. Subsequently, pixel intensity values were normalized to the range [0, 1] to provide a consistent numerical scale, improve training stability, and facilitate model convergence.

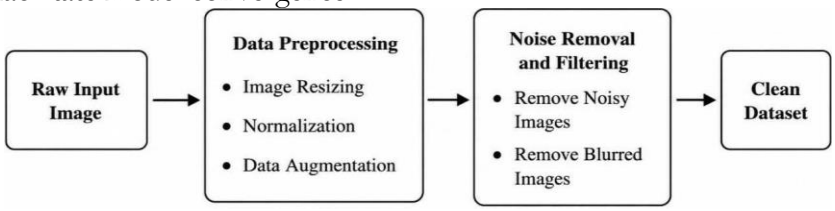

Figure 2. Image preprocessing pipeline for obtaining a clean dataset.

To increase the diversity of the training samples and mitigate the risk of overfitting, data augmentation techniques were applied to the training data. These included random rotation, horizontal flipping, width shifting, and height shifting. These transformations generate additional variations of the original facial images while preserving the essential characteristics of the corresponding emotional expressions. Data augmentation therefore enables the models to learn more robust facial representations and improves their ability to generalize to previously unseen images.

The same preprocessing procedures were consistently applied to both datasets to maintain a standardized data preparation pipeline. This consistency ensures that the CNN models receive inputs with comparable dimensions and numerical characteristics, thereby enabling a fair comparison of model performance. Consequently, differences in the observed results can be more appropriately attributed to the characteristics and performance of the evaluated CNN architectures and the proposed multi-model approach rather than to variations in the preprocessing procedure.

Noise Removal and Filtering from FT-2022 and FER 2013 Datasets:

During the initial inspection of the FT-2022 and FER-2013 datasets, a small number of images were identified as potentially mislabeled, corrupted, excessively noisy, severely blurred,

partially occluded, or containing insufficiently visible facial regions. Such samples can introduce ambiguity into the training process and negatively affect the reliability of the learned facial emotion representations. Therefore, these low-quality and potentially misleading samples were removed to obtain a cleaner and more reliable dataset for subsequent model training and evaluation.

Several automated image denoising and quality-assessment techniques, including Gaussian filtering, median filtering, Non-Local Means filtering, and blur detection, are available for identifying or reducing image noise and low-quality samples. However, these techniques were not applied in the present study. Facial Emotion Recognition relies on subtle facial characteristics, and aggressive filtering or smoothing may inadvertently modify important expression-related details, it is due to the fact that particularly around regions such as the eyes, eyebrows, and mouth. Furthermore, given the manageable size of the datasets, a manual quality assessment was considered more appropriate. Each image was therefore visually inspected, and samples exhibiting excessive noise, severe blur, significant occlusion, poor facial visibility, or incorrect labelling were excluded.

Following this manual quality-control process, 28,709 images from FER-2013 and 1,397 images from the FT-2022 ASD-specific dataset were retained. The resulting dataset contains more than 30,000 facial images, providing a substantially larger and cleaner collection of facial expressions for subsequent preprocessing, model training, and evaluation. This refinement process helped ensure that the retained samples were of sufficient visual quality and representative of the target facial emotion recognition task. We call this integrated dataset a hybrid dataset and evaluate if further in next section.

Handling Class Imbalance in Hybrid Dataset:

Although integrating the FT-2022 and FER-2013 datasets provides a substantially larger collection of facial expression images, the resulting dataset exhibits a significant class imbalance, as illustrated in Figure 2. The Happy class is the most represented, accounting for approximately 25.05% of the total samples, whereas the Disgust class is the least represented, accounting for only 1.52%. Such an uneven distribution can introduce bias during model training, causing the classifier to favor majority classes while producing comparatively weaker predictions for underrepresented emotions. Therefore, appropriate measures are required to mitigate the effects of class imbalance.

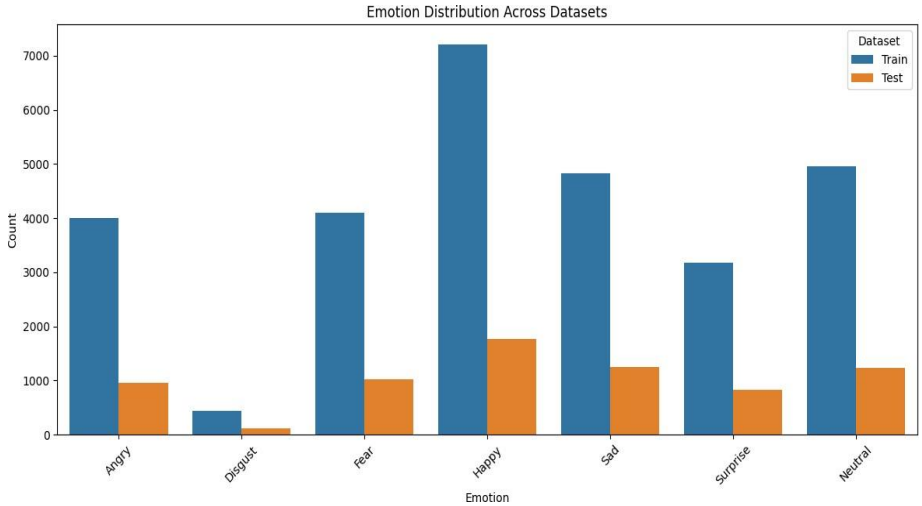

Figure 3. Emotion distribution across the training and testing hybrid dataset.

Several approaches can be used to address class imbalance, including random oversampling, under sampling, and synthetic data generation techniques such as SMOTE and ADASYN. However, each approach has potential limitations. Oversampling may increase the risk of overfitting because minority-class samples may be duplicated, causing the model to learn

specific samples rather than generalizable patterns. In contrast, under sampling reduces the number of majority-class samples and may result in the loss of potentially valuable information. Synthetic sampling techniques such as SMOTE and ADASYN generate additional minority-class samples, but the resulting synthetic examples may not always accurately represent realistic facial expressions and could introduce undesirable patterns into the training data.

To address the class imbalance while preserving the original dataset, this study employs a class weighting strategy during model training. Rather than modifying the number or content of samples in individual classes, class weighting assigns a larger penalty to misclassifications involving minority classes and a relatively smaller penalty to majority classes. Consequently, the model is encouraged to pay greater attention to underrepresented emotions without discarding existing samples or generating artificial images.

The class weight assigned to each class is calculated using the following formulation:

$$\boldsymbol{Wc} = \{\boldsymbol{N}\}/\{\boldsymbol{C} \cdot \boldsymbol{Kc}\}$$

where W_c represents the weight assigned to class c, N denotes the total number of training samples, C represents the total number of classes, and K_c denotes the number of samples belonging to class c. This formulation assigns higher weights to classes with fewer samples and lower weights to classes with larger numbers of sample. As a result, the contribution of minority-class errors to the training objective is increased, helping to reduce the bias toward majority classes while retaining the complete set of available training samples.

The overall process of integrating the two datasets, including dataset preprocessing, quality filtering, and class-imbalance handling, is illustrated in Fig.3.

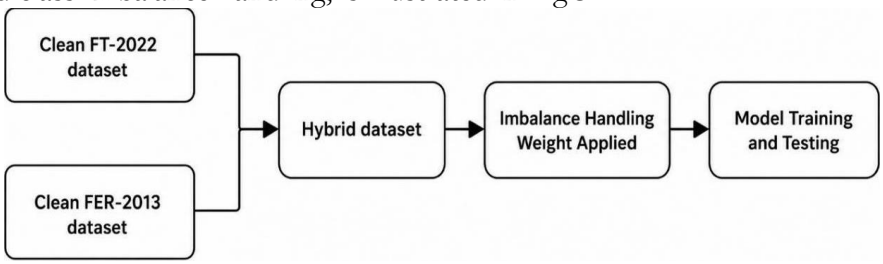

Figure 4. Overview of the proposed hybrid dataset processing and model training workflow.

Before combining the FT-2022 and FER-2013 datasets, the emotion labels from the two datasets were harmonized to establish a common seven-class label space. The FT-2022 dataset originally contains six emotion categories: Natural, Sadness, Fear, Joy, Anger, and Surprise, whereas FER-2013 contains seven categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. To make the two datasets compatible for joint training, semantically equivalent labels were mapped to the corresponding FER2013 terminology. Specifically, Anger was mapped to Angry, Joy to Happy, Sadness to Sad, and Natural to Neutral, while Fear and Surprise retained their original labels. FT-2022 does not contain a Disgust category; therefore, the Disgust class in the resulting hybrid dataset is contributed by FER2013 only. No new emotion category was introduced during the harmonization process. After this label mapping, both datasets were represented using the same seven target classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Table 2 summarizes the original emotion classes of the two datasets and the resulting hybrid dataset.

Table 2. Summary of datasets used in the study including size, emotion classes, and their intended purpose for facial emotion recognition.

Dataset	Number of Images	Emotion Classes	Purpose
FT-2022	~1,400	Anger, Fear, Joy, Sadness, Surprise, Neutral	ASD-specific Facial Emotion Recognition (FER)

FER-2013	32,589	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	General Facial Emotion Recognition (FER)
Hybrid Dataset	30,106	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	General + ASD Facial Emotion Recognition (FER)

Multi-Model CNN Architecture:

A multi-model CNN architecture is proposed in this study by combining three widely used CNN architectures for Facial Emotion Recognition (FER), namely VGG16, ResNet50, and Xception. The motivation behind this design is to exploit the complementary feature-learning capabilities of different CNN architectures and thereby improve the accuracy, robustness, and consistency of facial emotion classification. VGG16 employs a deep sequence of convolutional layers that can effectively learn fine-grained spatial and textural features from facial regions, which are important for distinguishing subtle emotional expressions. ResNet50 incorporates residual connections that facilitate the learning of deeper feature representations while mitigating degradation during training, enabling the extraction of increasingly abstract facial patterns. Xception employs depth wise separable convolutions to efficiently learn spatial feature representations while reducing computational complexity compared with conventional convolutional operations. Integrating the complementary representations learned by these three CNN architectures can provide a richer and more discriminative representation of facial expressions than individual CNN models.

The multi-model design can also improve the robustness of the FER system by allowing different CNN models to respond to facial variations that may arise from differences in pose, illumination, occlusion, or other imaging conditions. Furthermore, the use of pre-trained CNN models through transfer learning allows the proposed framework to leverage previously learned visual representations while adapting them to the target facial emotion recognition task. Consequently, the complementary learning capabilities of the individual models can compensate for their respective limitations and provide a more comprehensive facial representation. The resulting multi-model framework is therefore expected to improve classification accuracy, generalization, and prediction consistency compared with individual CNN models, particularly when applied to diverse facial expression data and ASD-focused FER scenarios.

The proposed multi-model CNN-based architecture is illustrated in Figure 4. The architecture employs a transfer-learning-based, multi-scale feature extraction strategy that combines sequential convolutional processing with parallel feature extraction and skip connections. The overall design is intended to capture both fine-grained facial details and broader contextual patterns while progressively transforming the input image into a compact and discriminative representation for seven-class facial emotion classification.

The architecture receives an input facial image of size 48 times 3, which is first processed by a Standardized Convolutional Block. This block consists of a Conv2D layer, Batch Normalization (BN), and MaxPooling2D. The Conv2D layer learns spatially meaningful facial features such as edges, textures, contours, and local patterns. Unlike traditional handcrafted feature extraction methods, these representations are learned automatically during model training. Batch Normalization normalizes the intermediate feature representations, helping to stabilize training and improve convergence, while the MaxPooling2D operation reduces the spatial dimensions of the feature maps and retains the most salient information.

The architecture incorporates transfer learning by initializing the convolutional backbone with pretrained weights, such as weights learned from ImageNet. The initial layers of the pre-trained network are kept frozen during the initial training stage so that the generic visual representations learned during pre-training can be reused. Freezing these layers also reduces the number of trainable parameters, lowers computational requirements, and helps mitigate

overfitting, particularly when training on a relatively limited target dataset. The later layers can subsequently learn task-specific representations relevant to facial emotion recognition,

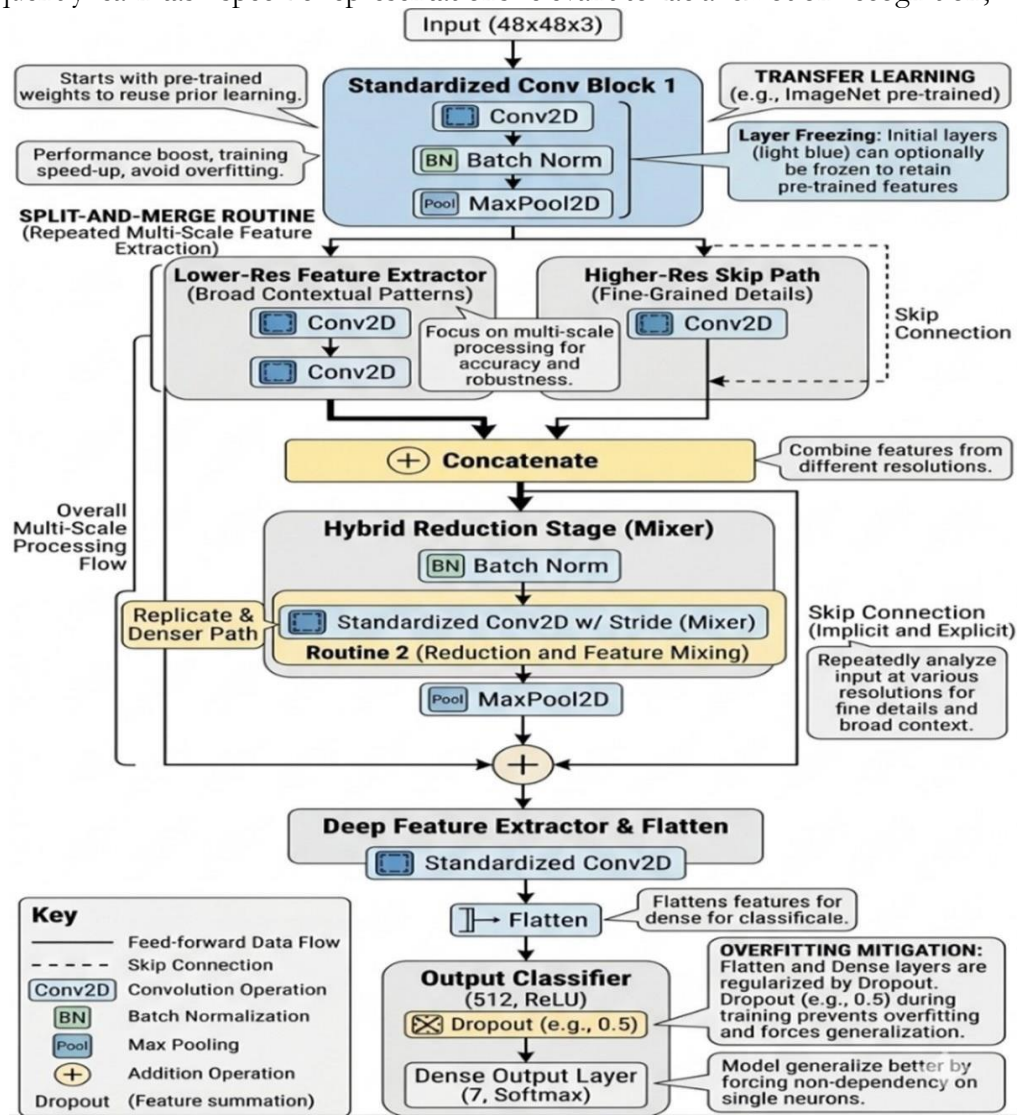

Figure 5. Proposed Convolutional Neural Network architecture using transfer learning.

Following the initial convolutional block, the architecture employs a Split-and-Merge Routine for multi-scale feature extraction. The feature representation is divided into two processing paths: a Lower-Resolution Feature Extractor and a Higher-Resolution Skip Path. The lower-resolution path applies successive convolutional operations to learn broad contextual and high-level facial patterns. Because the feature maps in this path have reduced spatial resolution, the network can capture larger-scale structures and contextual information efficiently.

In parallel, the higher-resolution skip path applies convolutional processing while preserving relatively fine-grained spatial information. This path is particularly important for facial emotion recognition because subtle changes in facial regions, such as the eyes, eyebrows, and mouth, may provide important cues for distinguishing between emotional states. By maintaining higher resolution information, the skip path helps reduce the loss of fine facial details that may occur through repeated pooling and down sampling.

The feature representations generated by the two paths are subsequently combined using a Concatenation operation. This operation merges information from different resolutions into a unified feature representation, allowing the network to simultaneously exploit detailed facial

information and broader contextual patterns. The resulting representation is therefore richer and potentially more discriminative than the representation obtained from either path independently.

The concatenated features are then processed by the Hybrid Reduction Stage (Mixer). This stage consists of Batch Normalization, a stride-based standardized Conv2D operation, and MaxPooling2D. The stride-based convolution, referred to as the Mixer in the architecture, performs additional feature extraction while progressively reducing the spatial dimensions of the feature maps. This operation enables information from different feature scales to be efficiently integrated into a more compact representation. The subsequent pooling operation further reduces the spatial resolution while retaining the most informative features.

A skip connection is also incorporated around the hybrid reduction stage. Rather than requiring all information to pass exclusively through the deeper processing layers, the skip connection provides a direct pathway for previously extracted features to later stages. The resulting feature representations are combined through an addition operation. This design helps preserve useful information, facilitates gradient propagation during training, and reduces the potential loss of important facial features during successive transformations. The replicate-and-dense-path mechanism shown in the figure further supports the processing of feature representations at different resolutions, allowing the architecture to retain both fine-grained details and broader contextual information.

Following the multi-scale processing and feature aggregation stages, the resulting representation is passed to the Deep Feature Extractor and Flatten stage. A standardized Conv2D operation performs further high-level feature extraction before the feature maps are flattened into a one-dimensional vector. The Flatten layer does not perform feature learning; rather, it reshapes the extracted feature maps into a format suitable for the subsequent fully connected classification layers.

The flattened representation is then fed into an Output Classifier consisting of a fully connected Dense layer with 512 neurons and ReLU activation. This layer learns higher-level relationships among the extracted facial features and provides a nonlinear representation for classification. To reduce overfitting, a Dropout layer with a dropout rate of 0.5 is applied after the Dense layer. Dropout randomly deactivates a proportion of neurons during training, reducing the model's dependence on individual neurons and improving its ability to generalize to unseen facial images.

Finally, the architecture uses a Dense output layer containing seven neurons with Softmax activation, corresponding to the seven target emotion classes. The Softmax function converts the output scores into class probabilities whose values sum to one. The emotion class associated with the highest predicted probability is selected as the final classification result.

Overall, the proposed architecture combines transfer learning, multi-scale feature extraction, parallel processing, feature concatenation, convolutional reduction, and skip connections within a unified multi-model CNN framework. This design enables the network to exploit complementary facial representations at different spatial resolutions while preserving important fine-grained facial details. Such characteristics are particularly relevant to ASD-focused facial emotion recognition, where subtle facial expression patterns may play an important role in distinguishing among emotional states.

For reproducibility, the evaluated fusion architectures use VGG16, ResNet50, and Xception as CNN feature extractors, with the models combined in three pairwise configurations: ResNet50– VGG16, ResNet50–Xception, and VGG16–Xception. The original classification heads of the individual networks are removed, and the feature representations obtained from the selected final feature-extraction layers are combined through feature-level fusion. The resulting feature representations are concatenated and subsequently passed to the classification component for prediction of the seven emotion classes: Angry, Disgust, Fear, Happy, Sad,

Surprise, and Neutral. The same feature-fusion procedure is applied to each of the three evaluated pairwise configurations.

All three CNN architectures, VGG16, ResNet50, and Xception, were initialized using weights pre-trained on ImageNet. The pre-trained networks were adapted to the facial emotion recognition task using transfer learning. During the initial training stage, the early layers of the pre-trained backbones were kept frozen to retain generic visual representations, while later layers were allowed to learn task-specific facial-expression representations. A relatively low initial learning rate was used to support stable optimization, and the models were trained using the Adam optimizer with a batch size of 32 for 40–100 epochs. Early stopping was employed when validation performance ceased to improve.

Performance Evaluation of Proposed Multi-Model CNN Architecture:

This section evaluates the Facial Emotion Recognition (FER) performance of the proposed multi-model CNN architecture using the integrated hybrid dataset developed in the preceding section. To assess the effectiveness of the proposed approach, appropriate baseline configurations are required for comparative evaluation. Therefore, a systematic experimental framework consisting of three configurations is adopted. These configurations enable the performance of the individual CNN models to be examined under different dataset settings and provide a direct comparison with the proposed multi-model approach.

Configuration 1:

Individual CNN Models on the FT-2022 Dataset:

In the first configuration, the three selected CNN models, namely VGG16, ResNet50, and Xception, are independently trained and evaluated using the FT-2022 ASD-specific dataset. This configuration establishes a baseline for assessing the performance of individual CNN architectures when trained exclusively on ASD-specific facial expression data.

Configuration 2:

Individual CNN Models on the hybrid Dataset:

In the second configuration, VGG16, ResNet50, and Xception are independently trained and evaluated using the hybrid dataset comprising FT-2022 and FER-2013. This configuration investigates the effect of increasing the size and diversity of the training data on FER performance of individual CNN models compared with training solely on the ASD-specific dataset.

Configuration 3:

Multi-Model CNN Architectures on the hybrid Dataset:

In the third and final configuration, different combinations of the three CNN models are developed to form the proposed multi-model CNN architectures. These architectures are trained and evaluated using the integrated FT-2022 and FER-2013 hybrid dataset. Their performance is subsequently compared with the results obtained from Configurations 1 and 2 to determine whether combining complementary CNN models provides additional performance benefits beyond those achieved by individual models and increased dataset size.

This three-stage experimental design enables the contribution of the proposed approach to be evaluated systematically by distinguishing the effects of ASD-specific training data, increased dataset diversity, and multi-model CNN integration on overall FER performance. The rest of this section is organized as follows.

Implementation Environment and Experimental Settings:

All experiments are conducted in a controlled and consistent computational environment to ensure fair and reproducible evaluation of the CNN models. Google Colab Pro+ is used as the development and execution platform, with a Tesla T4 GPU with 16 GB dedicated memory and 16 GB system RAM. The deep learning models are implemented and trained using TensorFlow and Keras, while Scikit-learn is used for performance evaluation and generation of classification reports. OpenCV is employed during dataset preprocessing,

particularly for image resizing and face detection, whereas Matplotlib and Seaborn are used for performance visualization.

All three CNN architectures VGG16, ResNet50, and Xception were initialized using weights pretrained on ImageNet and adapted to facial emotion recognition through transfer learning. The dataset was divided into an 80% training subset and a 20% testing subset. Model training was performed using the Adam optimizer with a batch size of 32 and an initial learning rate of 0.0001. During the initial training stage, the early layers of the pretrained CNN backbones were kept frozen, while the later layers were allowed to learn task-specific facial-expression representations. The models were trained for a maximum of 100 epochs, with early stopping applied when validation accuracy plateaued or exhibited repetitive values over a patience period of 10 epochs. Consequently, 100 epochs served as an upper threshold rather than a fixed duration for all models, allowing training to terminate earlier once accuracy gains stabilized. A learning-rate scheduler was not used. These settings were applied consistently across all experimental configurations. To minimize the possibility of data leakage, the dataset was manually inspected for duplicate images during the data preparation stage, and identified duplicate images were removed before model training and evaluation. Following this duplicate-removal process, the remaining images were divided at the image level into training and testing subsets using an 80/20 split. Therefore, the reported evaluation results are based on the deduplicated dataset, with no intentionally duplicated images shared between the training and testing subsets.

In all experiments, model performance is evaluated using four widely used classification metrics: accuracy, precision, recall, and F1-score. In addition to the overall performance measures, class-wise performance analysis is conducted to evaluate the model's ability to recognize individual emotion classes. This analysis helps determine whether the overall performance is consistently maintained across different classes and ensures that the results are not disproportionately influenced by majority classes, particularly given the class imbalance present in the integrated hybrid dataset.

To improve the reliability of the comparative evaluation, each individual CNN model and multiCNN configuration was independently trained and evaluated five times under the same experimental setup. The reported performance values represent the mean results obtained across the five runs. The repeated experiments showed only minor variation across runs, indicating that the observed performance patterns were consistent and were not attributable to a single training run or random initialization. This repeated-run procedure was applied consistently to both the individual CNN models and the multi-CNN configurations.

In addition, training and validation learning curves are plotted to monitor the learning behaviour of proposed multi-model architecture throughout the training process. These curves illustrate changes in metrics such as accuracy and loss across successive epochs for both the training and validation sets, enabling the identification of convergence behaviour and potential overfitting. Consistent training and evaluation settings are maintained across experimental configurations so that observed performance differences could primarily be attributed to the CNN architectures, dataset characteristics, and proposed multi-model approach rather than variations in the experimental environment.

Results and Discussion:

Configuration 1: Evaluating CNN Models using FT-2022 ASD Dataset:

In this setup, we use FT-2022 (preprocessed and clean dataset as described in previous section) to train and test three individual CNN models as shown in Figure 6. Each CNN model is trained and tested individually on this dataset to generate accuracy, precision, recall and F1 score.

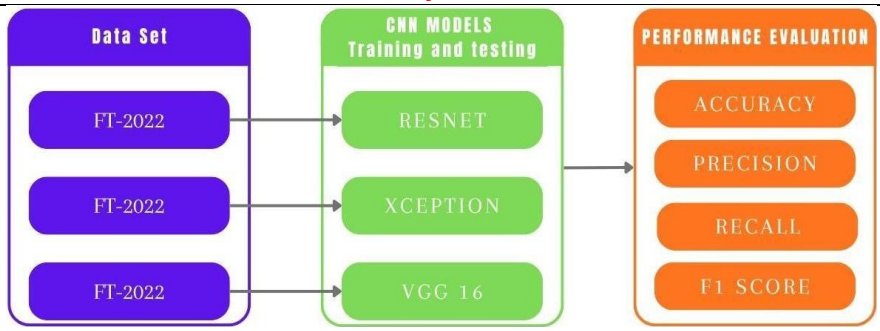

Figure 6. Performance Evaluation of Three CNN Models using FT-2022 Dataset.

The per-class classification report in Table 3 demonstrates exceptional overall and class-wise performance for the Xception model on the ASD FT-2022 dataset, achieving an overall accuracy of 95%. Class-wise evaluation reveals strong predictive power across all emotion categories, with F1scores consistently exceeding 0.90. The model attained perfect precision for the "anger" class (1.00) alongside a high F1-score of 0.97, while the "Natural" class yielded the highest recall at 0.97. Categories such as "fear", "sadness", and "joy" exhibited high stability with F1-scores of 0.96, 0.96, and 0.94, respectively. Even the lowest-performing class, "surprise", maintained strong metrics with a precision of 0.90 and an F1-score of 0.92. Furthermore, the averages for precision, recall, and F1-score all converged at 0.95, confirming that the model delivers balanced classification accuracy without skewing toward any specific category across the dataset's class distribution.

Table 3. Per-Class Classification Report for Xception using the FT-2022 Dataset

Class	Precision	Recall	F1-score
Anger	1.00	0.95	0.97
Fear	0.97	0.94	0.95
Joy	0.94	0.94	0.94
Neutral	0.93	0.97	0.95
Sadness	0.97	0.95	0.96
Surprise	0.90	0.95	0.92
Overall Average	**0.95**	**0.95**	**0.95**
Accuracy: 0.95			

Similar results were observed for ResNet50 and VGG16, as presented in Tables 4 and 5, respectively.

Although their overall accuracy and F1-scores were slightly lower than those of Xception, their perclass performance was generally similar. As shown in Table 3, Table 4, and Table 5, the models demonstrated relatively consistent performance across the emotion classes. This suggests that their performance was not strongly biased toward any particular class.

Table 4. Per-Class Classification Report for ResNet50 Using the FT-2022 Dataset

Class	Precision	Recall	F1-score
Sadness	0.90	0.88	0.89
Joy	0.90	0.88	0.89
Anger	0.90	0.88	0.89
Neutral	0.90	0.88	0.89
Fear	0.87	0.88	0.87
Surprise	0.90	0.88	0.89
Overall Average	**0.90**	**0.88**	**0.89**
Accuracy: 0.90			

Table 5. Per-Class Classification Report for VGG16 Using the FT-2022 Dataset

Class	Precision	Recall	F1-score

Anger	0.97	0.95	0.96
Fear	0.97	0.92	0.94
Joy	0.91	0.91	0.91
Neutral	0.93	0.95	0.94
Sadness	0.95	0.92	0.93
Surprise	0.88	0.95	0.91
Overall Average	0.94	**0.93**	**0.93**
Accuracy: 0.93			

The overall performance of the three individual CNN models on the FT-2022 ASD-specific dataset is summarized in Table 6. The results show that all three models achieved strong facial emotion recognition performance. Xception achieved the best results, with an overall accuracy and F1-score of 0.95. VGG16 achieved an accuracy of 0.93 and an F1-score of 0.93, whereas ResNet50 achieved an accuracy of 0.90 and an F1-score of 0.89. These results demonstrate that the individual CNN models were effective in learning discriminative facial emotion features from the FT2022 dataset.

Overall, the results indicate that all three CNN architectures performed effectively on the ASD specific facial expression data. The findings are also consistent with the results reported by [27] where CNN-based models achieved accuracies above 90% on the same dataset. The strong performance of the individual CNN models therefore provides a suitable baseline for comparison with the proposed multi-model CNN architectures.

However, despite these promising results, the FT-2022 dataset contains only 1,397 retained images after the data-quality filtering process. Its relatively limited size may restrict the diversity of facial expressions and real-world variations represented in the data. Consequently, models trained exclusively on this dataset may achieve strong performance under the experimental conditions while having limited exposure to the broader range of variations that may occur in real-world facial emotion recognition applications. This limitation motivates the use of the larger integrated (hybrid) dataset in the subsequent experiments, allowing the individual and multi-model CNN architectures to be evaluated using a more diverse collection of facial expression data.

Table 6. Performance Comparison of CNN model using FT-2022 ASD Dataset

Model	Accuracy	F1-score	Precision	Recall
ResNet50	0.90	0.89	0.90	0.88
VGG16	0.93	0.93	0.94	0.93
Xception	0.95	0.95	0.95	0.95

Configuration 2: Evaluation of CNN Models using Hybrid Dataset:

In this setup, we use Hybrid Dataset (preprocessed and clean as described in previous section) to train and test three individual CNN models as shown in Figure 7. Each CNN model is trained and tested individually on this dataset to generate accuracy, precision, recall and F1 score.

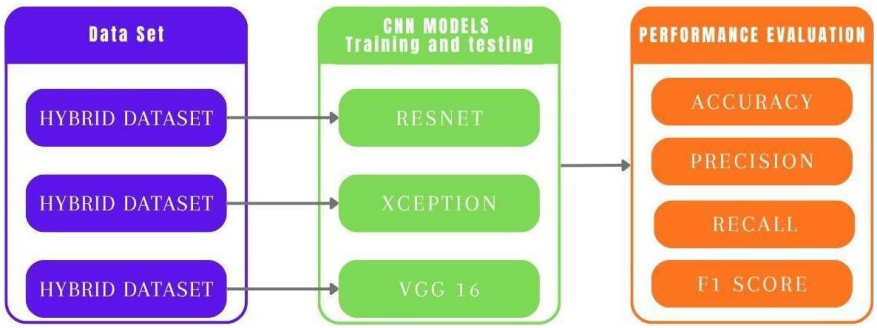

Figure 7. Performance Evaluation of Three CNN Models using Hybrid Dataset

The per-class performance results of VGG16, ResNet50, and Xception on the integrated hybrid dataset are presented. Overall, the results indicate relatively consistent performance across the individual emotion classes, suggesting that the evaluated models do not exhibit a substantial bias toward the majority classes. This observation is particularly important given the class imbalance present in the integrated dataset.

Both VGG16 and ResNet50 achieved strong performance, with accuracy and F1-score values exceeding 90%. These results are comparable to their performance on the smaller FT-2022 dataset in Experiment 1. Achieving similar performance on a substantially larger dataset containing more than 30,000 images is an encouraging result, as it demonstrates that these models can maintain strong classification performance when trained and evaluated on a more diverse collection of facial expressions. The results therefore support the effectiveness of VGG16 and ResNet50 for the ASD-focused facial emotion recognition task considered in this study.

In contrast, Xception achieved a substantially lower accuracy of approximately 68% on the integrated dataset. This indicates that, although Xception performed strongly on the smaller FT2022 dataset, its performance decreased considerably when evaluated using the larger and more diverse integrated dataset. The underlying reasons for this performance reduction require further investigation, including an analysis of the model's class-wise errors and its interaction with the characteristics of the integrated dataset. Table 7 presents the per-class classification report.

Table 7. Per-class classification report for Xception on Hybrid dataset

Class	Precision	Recall	F1-score
Angry	0.68	0.68	0.68
Disgust	0.70	0.68	0.69
Fear	0.68	0.68	0.68
Happy	0.72	0.68	0.70
Neutral	0.68	0.68	0.68
Sad	0.68	0.68	0.68
Surprise	0.66	0.68	0.67
Overall avg	**0.69**	0.68	**0.68**
Accuracy: 0.68			

Table 8 presents the per-class classification report for the ResNet model on the Hybrid dataset, which achieved an overall accuracy of 0.90.

Table 8. Per-class classification report for ResNet on Hybrid dataset

Emotion	Precision	Recall	F1-score
Sad	0.90	0.88	0.89
Happy	0.90	0.88	0.89
Disgust	0.90	0.88	0.89
Angry	0.90	0.88	0.89
Neutral	0.90	0.88	0.89
Fear	0.90	0.88	0.89
Surprise	0.90	0.88	0.89
Overall avg	**0.90**	**0.88**	**0.89**
Accuracy: 0.90			

Table 9 reports the classification performance of the VGG16 model on the Hybrid dataset, which achieved the highest overall accuracy of 0.91 among the three models.

Table 9. Per-class Classification Report for VGG16 on Hybrid Dataset

Class	Precision	Recall	F1-score
Angry	0.94	0.92	0.93

Disgust	0.84	0.80	0.82
Fear	0.85	0.82	0.83
Happy	0.96	0.95	0.95
Neutral	0.91	0.92	0.92
Sad	0.92	0.92	0.92
Surprise	0.95	0.94	0.94
Overall avg	**0.91**	0.89	**0.90**
Accuracy: 0.91			

Despite the comparatively weaker performance of Xception as an individual model, the strong results achieved by VGG16 and ResNet50 provide a solid basis for the next stage of the study. The subsequent experiments therefore investigate whether combining the complementary feature-learning capabilities of the selected CNN models within the proposed multi-model CNN architecture can further improve the accuracy, robustness, and generalization of the facial emotion recognition system. The summary of results for this experiment is presented in Table 10.

Table 10. Performance Comparison of Different Models on Hybrid Dataset

Model	Accuracy	F1-score	Precision	Recall
Xception	0.68	0.68	0.69	0.68
ResNet50	0.90	0.89	0.90	0.88
VGG16	0.91	0.90	0.91	0.89

Configuration 3: Multi-Model CNN Architecture Evaluation using Hybrid Dataset:

In this setup, we use the Hybrid dataset to train and test three different combinations of CNN models, as shown in Figure 8. We have used the same CNN models which are used in the previous experiment; however, alternate combinations are tested in this experiment. As we see in Figure 8, we combine VGG16 with ResNet, ResNet with Xception, and VGG16 with Xception, which are the three multi-model CNN alternatives investigated in this experiment.

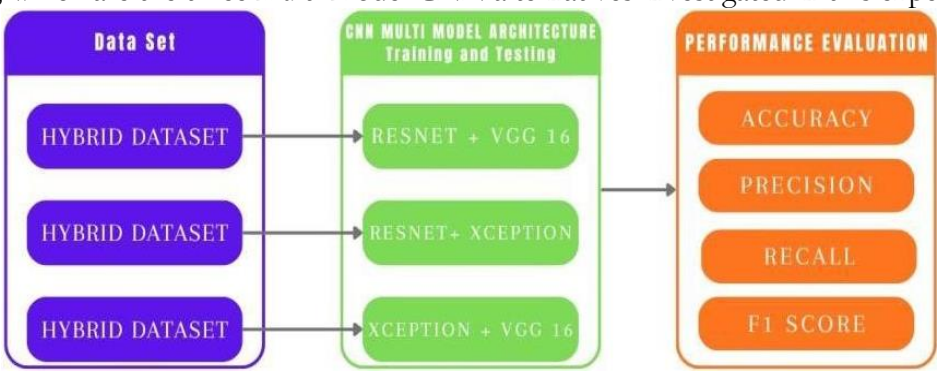

Figure 8. Performance Evaluation of three multi model CNN architecture on hybrid dataset

ResNet + VGG16

The training and validation curves shown in Figure 9 demonstrate stable and effective learning behavior of the proposed multi-model CNN architecture. The training and validation loss decrease rapidly during the initial epochs, followed by a gradual and stable convergence, indicating successful model learning and optimization. Although minor fluctuations may be observed in the validation loss during the early stages of training, the training and validation loss curves become closely aligned in the subsequent epochs, suggesting that the model does not exhibit significant overfitting. Similarly, both training and validation accuracy show a progressive increase and eventually reach high and comparable values. The close correspondence between the training and validation performance indicates a small generalization gap and suggests that the model is able to learn discriminative facial emotion features that generalize effectively to

unseen data. Overall, the learning curves demonstrate stable convergence and provide evidence of good generalization performance of the proposed multi-model CNN architecture.

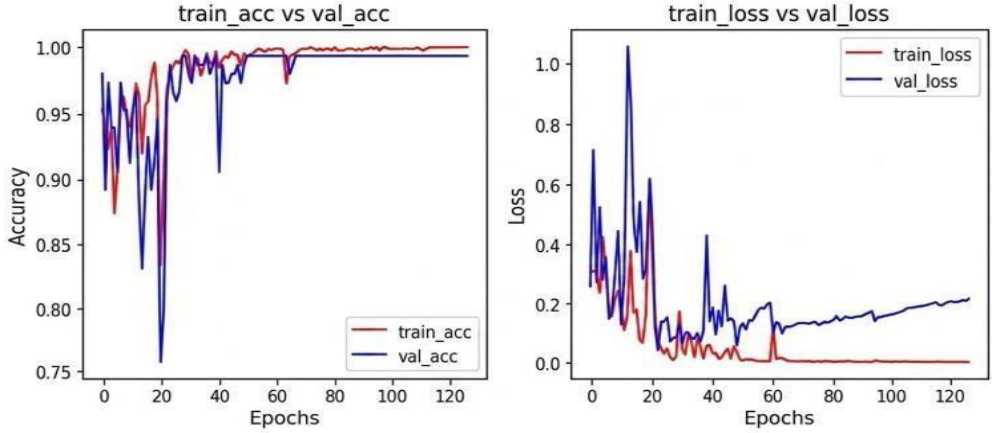

Figure 9. ResNet + VGG16: Model Training and Validation Loss and Accuracy

The classification results presented in Table 11 demonstrate the strong performance of the proposed multi-model CNN architecture.

Table 11. Per Class Classification Report for ResNet + VGG16 Multimodel using hybrid dataset

Emotion	Precision	Recall	F1-score
Sad	0.99	0.99	0.99
Happy	0.99	0.99	0.99
Disgust	0.98	0.92	0.93
Angry	0.99	0.98	0.98
Neutral	0.99	0.98	0.99
Fear	0.99	0.99	0.99
Surprise	0.99	0.97	0.97
Overall avg	**0.98**	**0.97**	**0.98**
Accuracy: 0.99			

The model achieved an overall classification accuracy of 99%, with precision, recall, and F1-score values of approximately 0.99 across the evaluated emotion classes. These consistently highperformance measures indicate that the model can effectively distinguish among different facial emotions while maintaining balanced classification performance across the individual classes.

ResNet + Xception:

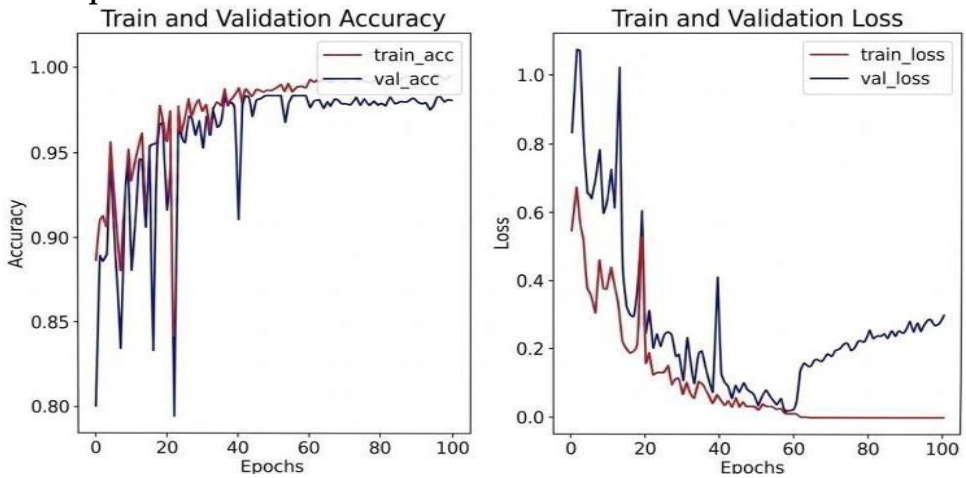

Figure 10. ResNet + Xception: Model Training and Validation Loss and Accuracy

The training and validation curves of the ResNet50–Xception multi-model CNN architecture, presented in Figure 10, exhibit a learning pattern similar to that observed for the previously discussed ResNet50–VGG16 configuration. The loss curves decrease rapidly during the initial training epochs and subsequently converge to stable values, while the training and validation losses remain closely aligned. Similarly, both training and validation accuracy increase consistently and reach high, comparable values. These results indicate stable model convergence and good generalization, with no evidence of significant overfitting.

The classification results presented in Table 12, further demonstrate the strong performance of the ResNet50–Xception configuration, achieving an overall accuracy of 98%. Most emotion classes, including Sad, Happy, Angry, Neutral, Fear, and Surprise, achieve precision, recall, and F1-score values of approximately 0.98, indicating consistent classification performance across the majority of emotion categories. The Disgust class shows comparatively lower recall, which may be associated with its smaller number of samples and the relatively subtle characteristics of this emotion. Nevertheless, the high average performance measures indicate that the model maintains strong overall performance without substantial bias toward the majority classes.

Table 12 Per Class Classification Report for ResNet + Xception Multi-Model using Hybrid Dataset

Emotion	Precision	Recall	F1-score
Sad	0.98	0.98	0.98
Happy	0.98	0.99	0.98
Disgust	0.98	0.92	0.93
Angry	0.98	0.98	0.98
Neutral	0.98	0.98	0.98
Fear	0.98	0.98	0.98
Surprise	0.98	0.97	0.97
Overall avg	**0.98**	**0.97**	**0.98**
Accuracy: 0.98			

The balanced precision, recall, and F1-score values suggest that the model does not exhibit a strong bias toward particular emotion classes. This is particularly important in ASD-focused Facial Emotion Recognition (FER), where facial expressions may exhibit subtle or atypical characteristics that can make accurate emotion classification more challenging. The results indicate that the proposed multi-model CNN architecture is capable of learning discriminative facial features from the integrated dataset, which combines the diversity of the general FER data with ASD-specific facial expression information.

Overall, the findings demonstrate that the proposed approach provides robust and generalized facial emotion recognition performance. The combination of complementary CNN featurelearning capabilities and the integrated dataset contributes to improved classification performance, enabling the model to effectively capture both general and ASD-relevant facial expression characteristics.

VGG16 + Xception:

The training and validation curves of the VGG16–Xception multi-model CNN architecture demonstrate stable and consistent learning behavior as shown in Figure 11. The training and validation accuracy curves remain relatively close throughout the training process, indicating a small generalization gap and no substantial evidence of overfitting. Similarly, both training and validation loss decrease rapidly during the initial epochs and subsequently stabilize, suggesting effective convergence and a stable optimization process. However, despite this stable learning behavior, the model achieves a comparatively moderate validation accuracy of approximately 65.22%, indicating that the VGG16–Xception configuration is less effective at

discriminating among the different facial emotion classes than the other evaluated multi-model configurations.

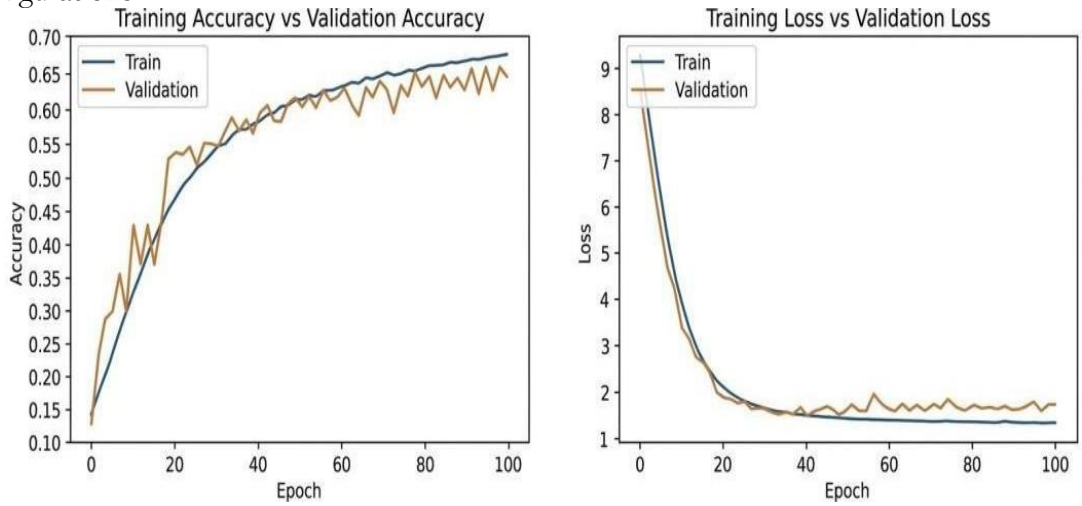

Figure 11. Xception + VGG16: Model Training and Validation Accuracy and Loss

The classification results presented in Table 13 further confirm this observation. The model achieves precision, recall, and F1-score values of approximately 0.65, reflecting moderate and relatively balanced performance across the evaluated emotion classes. The similarity between the average performance measures suggests that the model does not exhibit substantial bias toward particular classes. However, the overall accuracy of approximately 65% indicates that a considerable proportion of facial emotion samples are still misclassified.

Table 13. Per Class Classification Report for VGG16 + Xception Multi-Model using Hybrid Dataset

Emotion	Precision	Recall	F1-score
Angry	0.65	0.65	0.65
Disgust	0.65	0.65	0.65
Fear	0.67	0.63	0.65
Happy	0.69	0.66	0.67
Neutral	0.65	0.65	0.65
Sad	0.65	0.65	0.65
Surprise	0.65	0.65	0.65
Overall avg	**0.66**	**0.65**	**0.65**
Accuracy: 0.65			

Overall, although the VGG16–Xception configuration demonstrates stable convergence and reasonable generalization, its classification performance is substantially lower than that of the ResNet50–VGG16 and ResNet50–Xception configurations. This finding suggests that stable training and closely aligned training and validation curves do not necessarily guarantee superior classification performance. Further investigation is therefore required to determine the factors contributing to the comparatively weaker performance of the VGG16–Xception combination and to examine whether the feature representations learned by these two architectures are less complementary for the facial emotion recognition task considered in this study.

The comparative results of the three multi-model CNN configurations are summarized in Table 14. The results reveal substantial differences in performance among the evaluated combinations, demonstrating that combining multiple CNN models does not necessarily result in improved performance. The effectiveness of a multi-model architecture depends on the compatibility and complementarity of the feature representations learned by the individual CNN models.

Table 14. Performance Comparison of Multi-Model CNN Alternatives on Hybrid Dataset

Model	Accuracy	F1-score	Precision	Recall
ResNet + VGG16	0.99	0.99	0.97	0.98
ResNet + Xception	0.98	0.98	0.97	0.97
Xception + VGG16	0.65	0.65	0.66	0.65

Among the evaluated configurations, the two combinations incorporating ResNet50 achieved the strongest performance. The ResNet50–Xception configuration obtained the highest overall performance, achieving 99% accuracy and a F1-score of 0.99. This is followed by the ResNet50– VGG16 configuration, which achieved 98% accuracy and F1-score of 0.98. These results indicate that combining ResNet50 with either VGG16 or Xception can provide complementary feature representations that significantly improve facial emotion classification performance compared with the individual CNN models. In contrast, the VGG16–Xception configuration achieved substantially lower performance, with 65% accuracy and an F1-score of 0.65.

The results suggest that ResNet50 plays an important role in the most successful multi-model configurations evaluated in this study. Its combination with either VGG16 or Xception produced substantially better results than the VGG16–Xception configuration. This may indicate that the feature representations learned by ResNet50 complement those extracted by the other two CNN models more effectively. However, determining the precise reason for the weaker performance of the VGG16–Xception combination would require further architectural and feature-level analysis.

Overall, the results demonstrate that the proposed multi-model CNN approach is most effective when the combined architectures provide complementary feature-learning capabilities. Among the evaluated configurations, ResNet50–Xception achieved the best overall FER performance, with an accuracy of 99%. A comparison with the individual CNN models (Table 8) further highlights the effectiveness of the proposed approach. The best-performing individual model, VGG16, achieved an accuracy of 91%, whereas the ResNet50–Xception multi-model configuration improved the accuracy to 99%. This substantial improvement demonstrates the potential performance benefit of combining complementary CNN architectures within a multi-model framework for enhancing facial emotion recognition.

The findings of this study have potential practical implications for the development of FERbased assistive technologies for individuals with Autism Spectrum Disorder (ASD). The integration of general facial-expression data with ASD-specific facial-expression data and the improved performance achieved by the multi-CNN configurations suggest that models trained on more diverse and ASD-relevant facial expressions may provide a useful foundation for future emotion-aware applications. In particular, the ResNet50–VGG16 configuration achieved 99% classification accuracy on the hybrid dataset, demonstrating the potential of multi-model architectures for improving emotion-recognition performance in ASD-focused applications.

From a therapeutic perspective, FER systems based on such models could potentially support interactive systems that provide emotion-related feedback during therapeutic activities designed to improve emotional recognition and social communication. In educational settings, similar technologies could potentially be incorporated into assistive learning systems to help support the interpretation of facial expressions during classroom or peer interactions. These applications, however, represent potential future uses rather than outcomes directly demonstrated by the present study, as the proposed framework was evaluated experimentally using facial-expression datasets rather than in clinical or educational environments.

More broadly, the demonstrated improvement in recognition performance provides a basis for further development of FER-enabled assistive technologies intended to reduce communication barriers and support social interaction among individuals with ASD. However, practical deployment would require additional validation under real-world conditions, including

testing with diverse users, assessment of robustness to variations in lighting, pose, and facial appearance, evaluation of computational and latency requirements, and collaboration with relevant clinical and educational stakeholders. Therefore, the current results should be interpreted as evidence of the technical feasibility and potential applicability of the proposed FER framework rather than as evidence of clinical or educational effectiveness. In the present study, the integrated dataset was divided into 80% training and 20% testing data using an image-level random split. Before partitioning, the dataset was manually inspected and duplicate images were removed to minimize the possibility of identical samples occurring across the training and testing subsets. However, the image-level splitting strategy did not consider the identity or subject associated with individual facial images. Consequently, different images of the same individual could potentially be distributed across both the training and testing subsets. Such subject overlap may reduce the independence between the training and testing data and potentially contribute to higher measured classification performance. Therefore, the reported accuracy should be interpreted as performance under the specific dataset composition, preprocessing procedure, duplicate-removal process, and image-level splitting strategy used in this study, rather than as evidence of equivalent performance under all real-world conditions. To further investigate the robustness and generalizability of the proposed approach, future work should consider subject-level data partitioning, in which all images belonging to a given individual are restricted to either the training or testing subset. Additional evaluation under variations in pose, illumination, occlusion, image quality, and spontaneous facial expressions would also help determine whether the high performance is maintained under conditions that differ from those represented in the current dataset.

The substantial performance differences among the multi-CNN configurations suggest that the effectiveness of model combination depends not only on the individual performance of the constituent networks but also on the complementarity of the feature representations they provide. The ResNet50–VGG16 configuration achieved the highest classification accuracy (99%), while ResNet50–Xception also achieved strong performance (98%). In contrast, the VGG16–Xception configuration achieved only 65% accuracy, which was substantially lower than the performance of VGG16 (91%) alone and Xception (68%) alone. This indicates that combining two CNN architectures does not necessarily produce a complementary representation or improve classification performance.

A possible explanation is that the feature representations generated by VGG16 and Xception may contain a greater degree of redundancy or may not be effectively complementary for the facial emotion characteristics represented in the hybrid dataset. In contrast, the stronger performance of the ResNet50-based combinations may indicate that the residual feature representations learned by ResNet50 provide complementary information when combined with the representations learned by VGG16 or Xception. Differences in network architecture, depth, feature extraction characteristics, and the manner in which learned representations are combined may therefore influence the effectiveness of the resulting multi-CNN configuration. However, the present study did not perform a detailed feature-level analysis to experimentally establish the underlying cause of these differences. Therefore, these observations should be interpreted as plausible explanations based on the observed performance patterns, and further investigation using feature-visualization, representation-similarity, and explainability techniques is required to determine why certain architectural combinations are more effective than others.

Furthermore, the relatively stable convergence of the VGG16–Xception configuration does not necessarily imply effective discriminative feature learning. The training and validation curves indicate that the model can learn consistently from the available data, but its lower classification performance suggests that the learned representations may not adequately separate the emotion classes. Thus, convergence behavior and generalization trends alone are insufficient

to explain the effectiveness of a multi-CNN architecture; the quality and complementarity of the learned representations are also important factors.

The substantial performance differences among the multi-CNN configurations suggest that the effectiveness of model combination depends not only on the individual performance of the constituent networks but also on the complementarity of the feature representations they provide. The ResNet50–VGG16 configuration achieved the highest classification accuracy (99%), while ResNet50–Xception also achieved strong performance (98%). In contrast, the VGG16–Xception configuration achieved only 65% accuracy, which was substantially lower than the performance of VGG16 (91%) alone and Xception (68%) alone. This indicates that combining two CNN architectures does not necessarily produce a complementary representation or improve classification performance.

A possible explanation is that the feature representations generated by VGG16 and Xception may contain a greater degree of redundancy or may not be effectively complementary for the facial emotion characteristics represented in the hybrid dataset. In contrast, the stronger performance of the ResNet50-based combinations may indicate that the residual feature representations learned by ResNet50 provide complementary information when combined with the representations learned by VGG16 or Xception. Differences in network architecture, depth, feature extraction characteristics, and the manner in which learned representations are combined may therefore influence the effectiveness of the resulting multi-CNN configuration. However, the present study did not perform a detailed feature-level analysis to experimentally establish the underlying cause of these differences. Therefore, these observations should be interpreted as plausible explanations based on the observed performance patterns, and further investigation using feature-visualization, representation-similarity, and explainability techniques is required to determine why certain architectural combinations are more effective than others.

Furthermore, the relatively stable convergence of the VGG16–Xception configuration does not necessarily imply effective discriminative feature learning. The training and validation curves indicate that the model can learn consistently from the available data, but its lower classification performance suggests that the learned representations may not adequately separate the emotion classes. Thus, convergence behavior and generalization trends alone are insufficient to explain the effectiveness of a multi-CNN architecture; the quality and complementarity of the learned representations are also important factors.

Conclusion and Future Work:

This study investigates how multi-model CNN architectures can be used to enhance Facial Emotion Recognition (FER) of autistic people. Although CNN-based FER systems have demonstrated strong performance in previous studies, existing approaches often rely on individual CNN architectures and are evaluated using either general facial expression datasets or relatively small ASD-specific datasets. The limited availability of ASD-specific facial expression data and the diversity of emotional expressions in autistic people make it challenging to develop an accurate and robust FER system.

To address these challenges, this study integrated the FER-2013 benchmark dataset with the FT2022 ASD-specific facial expression dataset. The resulting integrated dataset contained more than 30,000 quality-filtered facial images and provided a larger and more diverse collection of facial expressions while incorporating ASD-specific emotional characteristics. Appropriate preprocessing, data augmentation, quality filtering, and class-weighting techniques were applied to prepare the data and class imbalance was addressed in the integrated hybrid dataset.

Three state-of-the-art CNN architectures, namely VGG16, ResNet50, and Xception, were initially evaluated individually. The first set of experiments using the FT-2022 ASD-specific dataset demonstrated that all three CNN models achieved strong performance, with Xception achieving the best result of approximately 95% accuracy and F1-score. These results established a baseline for evaluating the proposed approach. However, the relatively small size of the FT-

2022 dataset highlighted the need to investigate model performance using a larger and more diverse dataset.

The individual CNN models were subsequently evaluated using the integrated dataset. VGG16 and ResNet50 maintained strong performance, with VGG16 achieving the best individual-model accuracy of approximately 91%. In contrast, the performance of Xception decreased substantially to approximately 68% when evaluated on the larger and more diverse dataset. This result demonstrates that strong performance on a relatively small ASD-specific dataset does not necessarily guarantee similar performance when the model is evaluated on a larger and diverse collection of facial expressions.

The main contribution of this study was the development and evaluation of three multi-model CNN configurations: ResNet50–VGG16, ResNet50–Xception, and VGG16–Xception. The results demonstrated substantial differences among these configurations. The ResNet50–Xception architecture achieved the best overall performance, with an accuracy of 99% and a F1score of 0.99. The ResNet50–VGG16 configuration also achieved very strong performance, reaching 98% accuracy. In contrast, the VGG16–Xception configuration produced considerably lower performance, with an accuracy and F1-score of 65%.

These findings demonstrate that combining multiple CNN models does not automatically guarantee improved performance. Instead, the effectiveness of a multi-model approach depends on the complementary feature-learning capabilities of different CNN models. In this study, we discovered that a model combined with ResNet50 produced better results, suggesting that the feature representations learned by ResNet50 complemented those extracted by VGG16 and Xception more effectively. The comparison between the best-performing individual model, VGG16 with 91% accuracy, and the best multi-model configuration, ResNet50–Xception with 99% accuracy, demonstrates a substantial performance improvement through the proposed multi-model approach.

The findings of this study have potential practical implications for the development of FER-based assistive technologies for individuals with Autism Spectrum Disorder (ASD). The integration of general facial-expression data with ASD-specific facial-expression data and the improved performance achieved by the multi-CNN configurations suggest that models trained on more diverse and ASD-relevant facial expressions may provide a useful foundation for future emotion-aware applications. In particular, the ResNet50–VGG16 configuration achieved 99% classification accuracy on the hybrid dataset, demonstrating the potential of multi-model architectures for improving emotion-recognition performance in ASD-focused applications.

From a therapeutic perspective, FER systems based on such models could potentially support interactive systems that provide emotion-related feedback during therapeutic activities designed to improve emotional recognition and social communication. In educational settings, similar technologies could potentially be incorporated into assistive learning systems to help support the interpretation of facial expressions during classroom or peer interactions. These applications, however, represent potential future uses rather than outcomes directly demonstrated by the present study, as the proposed framework was evaluated experimentally using facial-expression datasets rather than in clinical or educational environments.

More broadly, the demonstrated improvement in recognition performance provides a basis for further development of FER-enabled assistive technologies intended to reduce communication barriers and support social interaction among individuals with ASD. However, practical deployment would require additional validation under real-world conditions, including testing with diverse users, assessment of robustness to variations in lighting, pose, and facial appearance, evaluation of computational and latency requirements, and collaboration with relevant clinical and educational stakeholders. Therefore, the current results should be interpreted as evidence of the technical feasibility and potential applicability of the proposed FER framework rather than as evidence of clinical or educational effectiveness.

References:

[1] A. Rehman, M. Mujahid, A. Elyassih, B. AlGhofaily, and S. A. O. Bahaj, "Comprehensive Review and Analysis on Facial Emotion Recognition: Performance Insights into Deep and Traditional Learning with Current Updates and Challenges," *Comput. Mater. Contin.*, vol. 82, no. 1, pp. 41–72, Jan. 2025, doi: 10.32604/CMC.2024.058036.

[2] A. L. Cîrneanu, D. Popescu, and D. Iordache, "New Trends in Emotion Recognition Using Image Analysis by Neural Networks, A Systematic Review," *Sensors 2023, Vol. 23, Page 7092*, vol. 23, no. 16, p. 7092, Aug. 2023, doi: 10.3390/S23167092.

[3] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[4] T. Khan, M. Yasir, and C. Choi, "Attention-enhanced optimized deep ensemble network for effective facial emotion recognition," *Alexandria Eng. J.*, vol. 119, pp. 111–123, Apr. 2025, doi: 10.1016/J.AEJ.2025.01.078.

[5] N. P. Madali and S. Hawamdeh, "Autism Information Progression and the Impact of Misinformation on Autism Knowledge, Awareness and Stigmatization," *Behav. Sci. (Basel, Switzerland)*, vol. 16, no. 5, May 2026, doi: 10.3390/BS16050660.

[6] J. Zeidan *et al.*, "Global prevalence of autism: A systematic review update," *Autism Res.*, vol. 15, no. 5, pp. 778–790, May 2022, doi: 10.1002/AUR.2696.

[7] K. Briot, A. Pizano, M. Bouvard, and A. Amestoy, "New Technologies as Promising Tools for Assessing Facial Emotion Expressions Impairments in ASD: A Systematic Review," *Front. psychiatry*, vol. 12, May 2021, doi: 10.3389/FPSYT.2021.634756.

[8] American Psychiatric Association, "Diagnostic and Statistical Manual of Mental Disorders," *Diagnostic Stat. Man. Ment. Disord.*, May 2013, doi: 10.1176/APPI.BOOKS.9780890425596.

[9] A. H. Al-Nefaie, T. H. H. Aldhyani, S. Ahmad, and E. M. Alzahrani, "Application of artificial intelligence in modern healthcare for diagnosis of autism spectrum disorder," *Front. Med.*, vol. 12, p. 1569464, May 2025, doi: 10.3389/FMED.2025.1569464/TEXT.

[10] I. Ahmad, J. Rashid, M. Faheem, A. Akram, N. A. Khan, and R. ul Amin, "Autism spectrum disorder detection using facial images: A performance comparison of pretrained convolutional neural networks," *Healthc. Technol. Lett.*, vol. 11, no. 4, pp. 227–239, Aug. 2024, doi: 10.1049/HTL2.12073.

[11] C. Qian, J. A. Lobo Marques, A. R. de Alexandria, and S. J. Fong, "Application of Multiple Deep Learning Architectures for Emotion Classification Based on Facial Expressions," *Sensors 2025, Vol. 25, Page 1478*, vol. 25, no. 5, p. 1478, Feb. 2025, doi: 10.3390/S25051478.

[12] J. M. Mayor Torres, T. Clarkson, K. M. Hauschild, C. C. Luhmann, M. D. Lerner, and G. Riccardi, "Facial Emotions Are Accurately Encoded in the Neural Signal of Those With Autism Spectrum Disorder: A Deep Learning Approach," *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, vol. 7, no. 7, pp. 688–695, Jul. 2022, doi: 10.1016/j.bpsc.2021.03.015.

[13] S. mohammed, hesham ali, W. El-Shafai, A. D. Algarni, naglaa soliman, and fathi abd-elsamie, "Predicting of Uncertainty Facial Expressions in Children with Autism Spectrum Disorder Based on an Improved Fuzzy Convolutional Neural Network Model," 2024, doi: 10.2139/SSRN.4846036.

[14] T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets," *Inf. 2024, Vol. 15, Page 135*, vol. 15, no. 3, p. 135, Feb. 2024, doi: 10.3390/INFO15030135.

[15] S. Ullah, J. Ou, Y. Xie, and W. Tian, "Facial expression recognition (FER) survey: a vision, architectural elements, and future directions," *PeerJ. Comput. Sci.*, vol. 10, 2024,

doi: 10.7717/PEERJ-CS.2024.

[16] A. P. Munanday, N. Sazali, A. S. Veerendra, and K. Kadirgama, "Facial Emotion Recognition: Methods, Applications and Future Challenges," *Int. J. Integr. Eng.*, vol. 18, no. 4, pp. 22–39, Jun. 2026, doi: 10.30880/IJIE.2026.18.04.002.

[17] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Comput. Methods Programs Biomed.*, vol. 215, p. 106621, Mar. 2022, doi: 10.1016/J.CMPB.2022.106621.

[18] and S. J. K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, "Facial emotion recognition using convolutional neural networks," *Mater. Today Proc.*, vol. 8, no. 3, pp. 3560–3564, 2023, doi: https://doi.org/10.1016/j.matpr.2021.07.297.

[19] C. Gupta, "Techniques for Transfer Learning to Enhance Image Classification in Facial Emotion Recognition," *Lect. Notes Electr. Eng.*, vol. 1528 LNEE, pp. 487–503, 2026, doi: 10.1007/978-981-95-5831-5_50/SAVE-RESEARCH.

[20] B. Ashwini, A. Sarkar, P. R. Behera, and J. Shukla, "Multi-source transfer learning for facial emotion recognition using multivariate correlation analysis," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/S41598-023-48250-X.

[21] R. Angeline and A. A. Nithya, "Multimodal Human Facial Emotion Recognition Using DenseNet-161 and Image Feature Stabilization Algorithm," *Trait. du Signal*, vol. 39, no. 6, pp. 2165–2172, Dec. 2022, doi: 10.18280/TS.390630.

[22] I. Bah and Y. Xue, "Facial expression recognition using adapted residual based deep neural network," *Intell. Robot.*, vol. 2, no. 1, p. 72, Mar. 2022, doi: 10.20517/IR.2021.16.

[23] H. M. Shahzad, S. M. Bhatti, A. Jaffar, M. Rashid, and S. Akram, "Multi-Modal CNN Features Fusion for Emotion Recognition: A Modified Xception Model," *IEEE Access*, vol. 11, pp. 94281–94289, 2023, doi: 10.1109/ACCESS.2023.3310428.

[24] Y. Ji *et al.*, "Hugging Rain Man: A Novel Facial Action Units Dataset for Analyzing Atypical Facial Expressions in Children with Autism Spectrum Disorder," *IEEE Trans. Affect. Comput.*, vol. 16, no. 3, pp. 2287–2302, Nov. 2024, doi: 10.1109/TAFFC.2025.3558914.

[25] "Autistic Children Emotions - Dr. Fatma M. Talaat." Accessed: Oct. 23, 2026. [Online]. Available: https://www.kaggle.com/datasets/fatmamtalaat/autistic-children-emotions-dr-fatma-m-talaat

[26] D. Madhusoodanan, N. Sagar, R. Vismaya, H. Santhosh, and S. Palaniswamy, "Facial Emotion Detection in Autistic Children: A Deep Learning Approach," *2nd Int. Conf. Electron. Comput. Commun. Control Technol. ICECCC 2025*, 2025, doi: 10.1109/ICECCC65144.2025.11063965.

[27] P. Radočaj and G. Martinović, "Emotion Recognition in Autistic Children Through Facial Expressions Using Advanced Deep Learning Architectures," *Appl. Sci. 2025, Vol. 15, Page 9555*, vol. 15, no. 17, p. 9555, Aug. 2025, doi: 10.3390/APP15179555.

[28] F. E. El Rhatassi, B. El Ghali, and N. Daoudi, "Deep learning approaches for recognizing facial emotions on autistic patients," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 4, pp. 4034–4045, Aug. 2024, doi: 10.11591/IJECE.V14I4.PP4034-4045.

[29] D. Kiejdo, M. Depka Prądzinska, and T. Zawadzka, "Late Fusion Model for Emotion Recognition from Facial Expressions and Biosignals in a Dataset of Children with Autism Spectrum Disorder," *Sensors 2025, Vol. 25, Page 7485*, vol. 25, no. 24, p. 7485, Dec. 2025, doi: 10.3390/S25247485.

[30] J. Awatramani and N. Hasteer, "Facial Expression Recognition using Deep Learning for Children with Autism Spectrum Disorder," *2020 IEEE 5th Int. Conf. Comput. Commun. Autom. ICCCA 2020*, pp. 35–39, Oct. 2020, doi: 10.1109/ICCCA49541.2020.9250768.

[31] B. Awaji *et al.*, "Hybrid Techniques of Facial Feature Image Analysis for Early Detection of Autism Spectrum Disorder Based on Combined CNN Features," *Diagnostics 2023, Vol. 13, Page 2948*, vol. 13, no. 18, p. 2948, Sep. 2023, doi: 10.3390/DIAGNOSTICS13182948.

[32] M. Afrin, K. E. Hoque, and R. Das Chaiti, "Emotion Recognition of Autistic Children from Facial Images Using Hybrid Model," *2024 15th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2024*, 2024, doi: 10.1109/ICCCNT61001.2024.10724499.

[33] "sklearn.utils.class_weight.compute_class_weight — scikit-learn 1.2.2 documentation." Accessed: Oct. 22, 2026. [Online]. Available: https://scikit-learn.org/1.2/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

[34] H. Alkahtani, T. H. H. Aldhyani, and M. Y. Alzahrani, "Deep Learning Algorithms to Identify Autism Spectrum Disorder in Children-Based Facial Landmarks," *Appl. Sci. 2023, Vol. 13, Page 4855*, vol. 13, no. 8, p. 4855, Apr. 2023, doi: 10.3390/APP13084855.

[35] "Deep learning models and techniques for facial emotion recognition | Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering." Accessed: Oct. 23, 2026. [Online]. Available: https://dl.acm.org/doi/10.5555/3507788.3507792

[36] Muhammad Kamran Abid, Rabia Sajjad, Muhammad Fuzail, Ahmad Naeem, Naeem Aslam, and Kiran Shahzadi, "Integrating temporal dynamics in facial emotion recognition using hybrid CNN-RNN models for enhanced human-computer interaction," *Kashf J. Multidiscip. Res.*, vol. 2, no. 06, pp. 1–15, Jun. 2025, doi: 10.71146/KJMR463.

[37] J. Shaik, R. Shekhar, and C. J. Shelke, "Hybrid Deep Learning Framework for Autism Detection in Children Using Facial Emotion Recognition," *Ing. des Syst. d'Information*, vol. 30, no. 10, pp. 2595–2607, Oct. 2025, doi: 10.18280/ISI.301007.

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.