Foundation University Journal of Engineering and Applied Sciences

FUJEAS
Vol. 5, Issue 2, 2025
DOI:10.33897/fujeas.v5i2.882

Research Article

Article Citation:
Iqbal et al. (2025). "*Autism Spectrum Disorder Detection using Facial Expression*". *Foundation University Journal of Engineering and Applied Sciences*
DOI:10.33897/fujeas.v5i2.882

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright
Copyright © 2025 Iqbal et al.

Published by
Foundation University
Islamabad.
Web: https://fui.edu.pk/

Autism Spectrum Disorder Detection using Facial Expression

Saqib Iqbal [a, *], Ghzanfar Farooq Siddiqui [a], Lal Hussain [b], Musserat Shaheen [c]

[a] Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan.
[b] Department of Computer Science, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan.
[c] Department of Computer Science, Islamic International University, Islamabad, Pakistan.

[*] **Corresponding author**: saqibiqbal@cs.qau.edu.pk

Abstract:

Autism Spectrum Disorder (ASD) is a complex neurological disorder that has an impact on communication, language, and social skills. Early identification of ASD patients, particularly in children, could make it easier to design and implement the best therapy approach at the appropriate time. Analyzing facial characteristics, eye contact, and other aspects of human faces can be used to detect ASD. To better accurately identify children with ASD in the early stages, this paper proposes an improved transfer-learning-based autism face recognition framework. This study improves the accuracy of ASD detection and classification of normal and autistic, using a machine learning and deep learning approach. This study classifies autistic and non-autistic faces using a deep learning-based CNN model. The study also analyzes the pre-trained transfer learning approaches with the proposed model. Results reveal that the proposed model has better detection and classification results, having 99% accuracy. Based on our accuracy, we propose that the diagnosis of autism spectrum disorders can be done effectively using facial images.

Keywords: ASD, Facial, AI, ML, CNN, Transfer Learning, Augmentation.

1. Introduction

Computer vision is one of the very important and widely used research areas that has gained researchers' attention. It is a vast subject to which scientists have made major contributions. The goal is to make computers capable of doing tasks in the same manner as human visual systems can [1]. Extensive research has been done on Artificial Intelligence (AI), and innovative strategies are being developed to leverage existing resources.

Because of improvements in Web technology and the availability of amazing data (both structured and unstructured data, including text, photographs, audio, or video), AI's subfields can now answer intelligently. AI is improving its capacity to derive knowledge and patterns from the available data. Diagnosis of diseases in the medical field is the most rigorous and basic discipline, as it is connected to the life of patients and their treatment. Autism spectrum disorder (ASD) is a cognitive disease, according to the Centers for Disease Control and Prevention. An impairment that can cause substantial social, communicative, and behavioral issues (CDC). In the United States, the prevalence of ASD affects one out of every 59 children aged eight and under, and the number is rising [2]. However, there are significant and ongoing

racialized disparities in the commonly added and access to prevention strategies. Children from ethnic minority groups are less likely to receive a diagnosis with ASD than White children, and they're more likely to receive a diagnosis or have their diagnosis delayed [3]. The total projected ASD prevalence in 2018 was 16.8 per 1000 (1 in 59) youngsters, a significant decrease from the previous year. For children with ASD, these delayed or misdiagnosed diagnoses as a consequence, immediate action being lost. Clinical findings show that early, thorough, and intense intervention can result in significant, long-term improvements in not only cognitive capacity and social etiquette, but also in the intensity of Neuro-developmental disorders [4]. In two cases, children treated with the ESDM did not meet the criteria for an ASD diagnosis [5]. A recent cost comparison analysis of intervention behavior therapy in the Netherlands found that if early diagnosis is started before 30 months of age, lifetime cost reductions might be above EUR 1 million per individual. These findings show how early detection and intensive ASD-specific intervention can enhance long-term outcomes for children with ASD, while also underlining the need to expand this work into underserved community settings to help all children with ASD achieve better outcomes.

ASD traits loose social interactions like eye contact, not recognizing things, and not showing emotions and expression by nine months. By the age of one year, makes little or no gestures (for example, does not wave goodbye). The following are the primary reasons that contribute to the discrepancy in ASD prevalence and delayed diagnoses in the United States [6]. The subjective nature of diagnosis: ASD is currently diagnosed through behavioral observation, which means that only experienced clinicians can accurately diagnose ASD in children as young as two years old, with the average age of diagnosis being four to five years [7]. Many families lack access to experts/specialists, and access is much more limited in underprivileged areas. Lack of knowledge and screening, particularly in rural areas, is also an issue. Furthermore, children from racial and ethnic minorities who match Children that satisfy the requirements for ASD are less likely to receive a diagnosis, so they're more prone to be misdiagnosed.

With advancements in machine learning (ML) and deep learning (DL) technologies, the journey to make computers intelligent has become easier. ML is a subfield of AI. ML is making considerable progress in many areas of daily living, including health, education, and the economy. Deep learning is a branch of machine learning that is making a huge impact on computer vision. Deep learning is important in image classification, object identification, object categorization, edge detection, and other domains. With the help of DL, we can get patterns and features automatically from images and can make accurate predictions and classifications on the basis of our trained models [8]. Here we are working on facial emotion classification and trying to make a thorough analysis and bring improvement in the existing accuracy to achieve better results. Our research studies have worked on pre-trained and custom models to do an evaluation and improve accuracy.

2. Related Work

Human facial images are the most common and promising input source because they include a significant amount of information for expression recognition studies [9]. Facial emotion recognition (FER) is a method of determining an individual's emotional state by analyzing facial expressions in static images and videos. Facial expressions are nonverbal communication to reveal human emotions. Facial expressions can be related to physiological or mental states of mind, and they play an important part in the treatment of psychiatric diseases. The effective use of FER is critical in health care, recommendation systems, personalization of services, employment, public safety, criminal detection, and other critical areas [10], [11]. Many authors have conducted behavioral research on face processing in ASD in recent years, but it is still unclear if those individuals actually have a problem detecting emotional facial expressions [12]. Some research, such as [13], implies that the ability is intact; however, others reported significant impairments in comparison to TD children [14]. There is evidence that people with ASD process faces differently from children with TD. Persons with ASD appear to prefer to apply analytical processing procedures, whereas TD individuals perceive

faces holistically.

Around half a century ago, the pervasive incidence of ASD was questioned. It was considered to only happen in Western developed countries with developed technology. Awareness about ASD and its prevalence has increased in different parts of the world over the last decade; the position in Africa regarding ASD remains ambiguous, with the bulk of content originating from the West. In Africa, no detailed research on the epidemiology of ASD has been conducted. One of two studies focused on the etiology of ASD in Arab countries, although that included two nations in Northern Africa. In African literature, there was a higher percentage of nonverbal instances of ASD compared to verbal cases. Mental retardation, epilepsy, and oculocutaneous albinism were among the founder conditions. Post-encephalitic infection, genetic and auto-immune variables, and vitamin D insufficiency were all proposed as etiological reasons [15]. In the United States, ASDs are thought to affect roughly one percent of children. This number is consistent with estimates from other developed countries, as per statistics from several studies. According to the (ADDM) System of the United States Centers for Disease Control and Prevention (CDC), the usual reports of ASDs have risen significantly in a short span of time [16].

This study will address issues related to the diagnosis of intelligence-based ASD detection in early childhood. Individuals with ASDs should display slower and less efficient emotional expressions, especially for socially complicated emotions, if this is the case. This study put this theory to the test by measuring emotion detection speed and accuracy whilst limiting exposure and reaction window. Multisensory processing is frequently required to comprehend emotions [17]. Speech fluency, as well as facial and bodily gestures, are used to interpret emotions and others' moods. Before birth, the fetus can distinguish speech phonation [18]. For newborns, another major source for understanding emotional states in others is facial expressions. Studies also showed that the capability to understand others' feelings (more precisely) improves with time, peaking around the age of 10–11 when kids reach an adult level of understanding.

The FEFA is indeed a computer-based sentiment analysis test that uses black-and-white images of six major emotions (happiness, sorrow, fear, anger, surprise, and disgust) along with impartial images to measure face interpretation and expression identification [19]. To identify ASD by using facial images [20] introduced transfer learning models on the Kaggle dataset to achieve the highest accuracy. The authors used both deep and shallow models for diagnosing Autism with improved MobileNet-V1. To significantly improve the ASD detection \cite{hosseini2021deep} used the MobileNet model. Features from the images were extracted from deep learning pre-trained models, then three fully connected dense layers were used to predict the autism spectrum disorder. To improve the accuracy author removes young children's images from datasets. As a result, they were able to achieve an accuracy of 95%. Authors in [21] used two models, Xception and EfficientNet B, on the image dataset to achieve accuracy, and especially focused on the area under the curve (AUC). When compared to other forms of stimuli, such as schematic drawings, visual search is an ecologically acceptable experimental strategy that resembles everyday scenarios in which one must find a target, and photographic facial expressions are realistic stimuli. It's unclear whether people with ASD have trouble detecting emotional facial expressions quickly. Only a few studies have looked into this, employing the visual search paradigm with images of facial expressions, like in studies of TD people, and none of them have found clear differences in ASD people's performance [22], [23].

Table 1 represents the studies related to ASD and their drawbacks. The proper recognition of the expressions is dependent on a process known as feature extraction. Feature extraction is the process of converting raw data into numerical features that can be processed while retaining the information contained in the original data set [24]. The optimum features can achieve excellent recognition accuracy. Even if we use the best classifier for facial expression identification, the recognition accuracy will suffer if the extracted features are poor [31], [32], [33]. The features are extracted using two methods: texture-based and geometric-based. Geometric-based approaches deal with the placement and form of face components, whereas texture-based methods focus on modifying the

Table 1: ASD detection using facial expression – comparison of recent work

Authors	Problem Identified	Proposed Solution	Limitations in the study
[25]	Accurate facial expression recognition in ASD	Facial expression intervention in handling children with ASD	Fewer participants and fewer factors.
[26]	Fewer participants and fewer factors	Extraction of the accurate emotion of an individual	Other than facial expression, techniques may be used to extract more information.
[27]	Behavioral and physiological subjects contain more information	Physiology-based ASD detection	More subjects are needed for stability and accuracy of the algorithm.
[28]	Child to adult facial changes impact on ASD detection	Adaptation of facial features for better detection of ASD	The model is data-dependent; some features may be useful for classification, but not correlated.
[29]	Compound facial expressions of emotion	Better detection using compound facial expression	The dataset is not sufficient; the method needs to be revised.
[30]	Typically developing (TD) individuals have worse ASD detection	Neural correlation-based ASD detection	Lack of reaction time and different processing in different individuals.

local texture, which is more reliable [34]. Gabor wavelets [35] are quite popular, yet the feature dimension is enormous, resulting in computational complexity. Local binary pattern (LBP) [36], an appearance-based technique used by many researchers owing to its excellent classification capabilities and higher computation efficiency. A significant range of texture-based approaches, such as LBP, LGC, HOG, PCA, and ICA, are employed in the literature for expression recognition. The advancement of computer vision and machine learning techniques has had a significant impact on facial expression recognition [37], [38].

The classification of expressions refers to labeling the new data by building a model set by training data that contains observations whose category is identified. The most common classifiers are SVM, decision trees, KNN, ensemble classifiers, and instance-based learning classifiers [39]. Early attempts at facial expression recognition relied heavily on handcrafted features [39]. However, researchers have paid little attention to handcrafted models because they produce less accurate results than the models of deep learning [40]. However, Deep learning has gained popularity in the computer vision field following the performance of the AlexNet deep neural network on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [41]. Several papers presenting deep learning methods for facial expression identification were presented just at the 2013 Facial Expression Recognition (FER) Competition [42]. Deep features, also called learned features, are features that a machine learning algorithm automatically learns from training examples during training. They are extremely effective while solving specific problems, but we don't have control over the features extracted by the model from the data. In several cases, the features are good for data classification with no real-world elaboration. Some recent papers [31], [32], [33] advised training an ensemble of CNNs for improved performance, but [43] combined deep features and handcrafted features like SIFT [44] or HOG [45].

While the bulk of research concentrated on identifying facial expressions from static photographs, several studies focused on recognizing facial expressions in video [43], [46]. Hasani and Mahoor [46]. Introduced a network design having 3D convolution layers and a Long-Short Term Memory (LSTM) network that extracts the spatial connections within face photographs with the temporal interactions between various frames in a video [43]. Unlike prior efforts,[47] and [48] demonstrated identity-aware FER models. [47] suggested concurrently estimating expression and identification variables using a neural architecture built on identical CNN layers, in order to reduce inter-subject

variability generated by personal attributes and obtain higher performance of FER. Paper [49] presented a trainable system from start to finish. Patch-Gated CNN recognizes the viewable areas of the face while automatically perceiving the regions of a face. The methodology divides an intermediate feature map into numerous patches based on the placements of the associated facial landmarks to discover the viewable portions of the face. Each patch is then reweighted based on its relevance, as decided by the patch itself. Zeng et al [50] proposed a methodology to handle errors in labelling across data sets. Images are tagged with various labels in their framework, which are either given by human annotators or predicted by learning algorithms. Then, from the inconsistent pseudo-labels, an FER model is trained to suit the hidden truth. Hua et al. [33] introduced a deep learning system that is composed of comprises three sub-networks of varying depths. Each sub-network is built on a CNN that has been trained individually. Unlike Hua et al. [33], we use local learning and integrate deep CNN features with handmade features. These deep learning (DL) methods have drastically altered people's perceptions of information processing.

DL is seen to be a superior solution for vision and classification challenges because of its extraordinary capacity to self-learn [51]. DL techniques have been used in FER to solve the challenges raised above, as well as other learning tasks [52]. The technique of feature extraction in deep learning algorithms employs an algorithmic way to identify and extract different characteristics. Deep learning algorithms are composed of a layered data representation architecture. The networks' final layers perform as high-level feature extractors, while the lower levels act as low-level feature extractors [53]. For video processing, recurrent convolution networks (RCNs) [54] were established. Convolutional neural networks (CNNs) are applied to video frames, which are subsequently input into an RNN for information processing with respect to time. These models perform well when the target ideas are complicated, and there is a limited amount of training data, but they have drawbacks when compared to deep networks. To address this issue, a methodology called DeXpression for strong face recognition was developed [55]. It is made up of two feature extraction blocks that work in parallel and have layers like pooling, convolutions, and ReLU. For improved performance, it employs multiple-feature fusion rather than a single feature. Unsupervised learning methods, such as autoencoders, are used to build the model [56]. A hybrid RNN-CNN technique is used in [56] to simulate the spatiotemporal information of human facial expression. [56] fused distinct modalities at the decision and feature levels, producing higher accuracies than single modality classifiers.

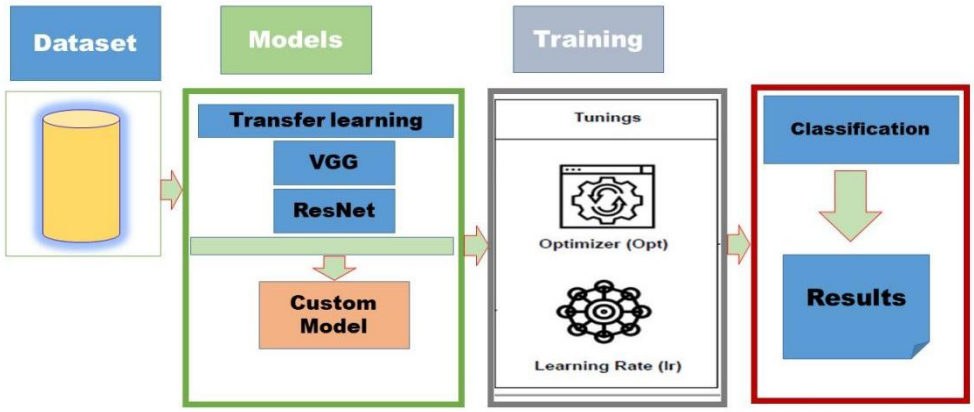
Similarly, a multitask global-local network (MGLN) for FER is proposed in [57], which combines two modules: a global face module (GFM) that extracts spatial features from the frame with peak expression and a part-based module (PBM) that learns temporal features from the nose, mouth, and eyes regions. Extracted features of GFM from a CNN and PBM from an LSTM network are then merged together to capture substantial facial expression variance. In [58], a shallow CNN architecture with dense connections over pooling is suggested, whereas the fully connected layer is dropped to ensure feature sharing. It offers good performance for the effective depiction of facial emotions with little training data, while trained DenseNet40 and DenseNet121 exhibit a performance decrease owing to over-fitting. Although there is a lot of work has been done in image features extraction but there is a need to introduce model models which use for neural network-based facial extraction by using deep features and having generic nature datasets and their performance would preferably high.

3. Materials and Methods

ML and DL are emerging technologies that are widely used in various scientific studies and many areas of today's scientific research [59]. Image processing is broadly benefiting from the emergence of deep learning. Previously, research work has been done to do proper object identification and classification. We are using deep learning algorithms and techniques to acquire better results. We have worked on three pre-trained models and a custom model. We compared the effectiveness of the algorithms on different and varying learning rates (lr) as well as optimizer functions (opt). Our work differentiates the low-performing and high-performing techniques on the two given datasets with

respect to the optimizer and learning rates as well. We have summarized the proposed approach, which is elaborated in Figure 1.

Figure 1: Overview of proposed methodology



3.1. Dataset

A more challenging dataset is more useful in obtaining results that are anticipated in the real world. Several datasets are fabricated under favorable conditions where the variations in light and pose are not significant. In this study, Kaggle datasets will be used, which are publicly available, to validate the proposed model. Kaggle Autism Facial Dataset (Kaggle Dataset) has 2936 face images, evenly distributed between children with ASD and children with TD. The initial dataset had 3014 images [20], which caused apparent issues. All of the photographs in the Kaggle dataset were found through an internet search because the contributor indicated that he was unable to secure any ASD images from organizations or verifiable sources. This study utilized the dataset from [17], which had 2936 photos after removing photographs that were plainly incorrect. There are about 89 percent White children and 11 percent children of color in this sample. This dataset was only used to demonstrate the influence of racial characteristics on deep-learning development based on face images. The Kaggle Autism Facial dataset is limited in demographic diversity. The observed gap between training and validation accuracy (99% vs. 74%) further suggests the presence of dataset-induced bias.

3.2. Preprocessing

We went through a series of steps, such as data collection, data preprocessing, data organization, and many others, while preparing the model. In the case of image processing, preprocessing is of vital importance, which is used to convert raw images into valid images that the model can process for training and make inferences. The images in a dataset can be of various sizes, different contrasts, or in the wrong orientation. Image preprocessing helps us to take preprocessing steps to make sure that our images are formatted correctly for our model. There are a series of different steps in image processing that can be applied, and many more, like noise reduction, image resizing, etc. We have discussed the following steps only.

3.2.1. Augmentation

Image augmentation is another preprocessing step that expands the image dataset significantly and hence increases the diversity of images without collecting new images. There are different types of image augmentation techniques, such as flipping or orientation, padding, etc. Flipping can be applied in two different ways: horizontal and vertical. In our approach, we have used both kinds of flipping. Replication is a process of using an instance twice for training an ML model, as more data to train an ML model results in high accuracy, and hence, we obtain an efficient ML model. Augmentation was

selectively applied to the minority class. In the proposed methodology, we have applied augmentation in order to handle imbalanced data distribution among different classes. We empirically balanced augmentation using rotation, zoom, and brightness adjustments to ensure class balance per epoch. This supported more equitable training without oversampling risks. A key strategy was the use of on-the-fly data augmentation using the ImageDataGenerator API in TensorFlow. The applied transformations included:

- **Random horizontal flipping** – to simulate mirrored faces
- **Random zoom (range: 0--15%** – to emulate changes in camera distance
- **Rotation (up to 15 degrees)** – to introduce pose variation
- **Brightness adjustment (range: 0.8--1.2)–** to mimic lighting variation
- **Width/height shift (range: 0--10\%)** – to simulate subject repositioning

These augmentations effectively increased dataset diversity and helped the model generalize better to unseen conditions.

3.2.2. Feature Selection

The GooleNet-based training network is utilized in YOLO. In this straightforward comparison, the author contrasts GooleNet with VGG16. About computational complexity, GooleNet performs better than VGG16 (8.25 billion operations vs. 30.69 billion operations). In ImageNet, the former is significantly lower than the latter (88 percent vs. 90 %). As the core network of YOLO V2, the author employs a new categorization model, Darknet-19. The complete network structure is shown in Table 6: Only 5.58 billion operations are required for Darknet-19. In YOLO, the network has 19 convolutional layers and five maximum pooling layers. There are 24 convolutional layers and two fully connected layers in the GooleNet employed in V1. As a result, the number of convolutions and convolution operations in Darknet-19 is lower than in GoogleNet. Overall, YOLO is the best option.

3.2.3. Local Binary Patterns Histograms (LBP)

It is a face recognition algorithm that recognizes a person's face. It is well known for its performance and ability to distinguish a person's face from the front and the side. Matrix formats, which are made up of rows and columns, are used to represent all images. The pixel is the most fundamental element of an image. A group of pixels makes up an image. Each of these is made up of little squares. We can build the whole image by putting them side by side. A single pixel in an image is thought to contain the least amount of information possible. The value of pixels in each image ranges from 0 to 255. Local Binary Pattern uses the LBP operator to summarize the local specific structure of a face picture while focusing on local aspects.

3.3. Classification

The classification was done after feature extraction. There are two types of classification: binary and multi-class classification. We worked to classify the emotions and improve the results as well. Our work was performed on two datasets, and the accuracy has improved. Besides, we identify different parameters that could help in classifying the images effectively. Besides, during experimentation, different ANN classifiers have also been used, and performance has been evaluated. At first, pre-trained ANN models were loaded, and then their fully connected sections were replaced by appropriate classification parts. Convolutional neural networks may be used for image analysis in three different ways. The first method involves training the convolutional neural network from scratch. This is the most difficult method since it requires a substantial amount of computer power and hundreds of annotated images. The second method is based on transfer learning, which holds that we can use knowledge of one type of problem to solve a related problem. For instance, we could use a convolutional neural network model that has been trained to recognize animals to start and train a

new model that can differentiate between cars and trucks. Compared to the first approach, this one needs fewer data and processing resources. The third way of extracting features from a pre-trained convolutional neural network uses a machine learning model that has been trained. The ability to train a hidden layer to recognize edges in an image, for instance, may be used to take pictures from a variety of different domains. The least quantity of information and processing power is also needed for this strategy. We ensured clean separation of training and test sets using stratified splits before augmentation. Augmented samples were never shared across sets. Despite aggressive training augmentation, the increase in validation loss confirms that no leakage occurred.

3.3.1. Transfer Learning

Transfer learning is used to apply information from a previous activity to improve performance on a new task. Transfer can be evaluated in three ways. The first is the first performance in the target task attained with solely transferred knowledge before any further learning, as contrasted to the initial performance of an ignorant agent. The time it takes to fully comprehend the target task utilizing transferred information versus learning it from scratch is the second aspect to consider. The difference between the final level of performance attained in the target task and the final level without transfer is the third aspect to consider.

3.3.2. ResNet-101

The superiority of deep networks has been documented in various papers in recent years. The Residual Network (ResNet101) was devised by the authors in [60], and it is based on ImageNet, which is ImageNet's deepest architecture. For the same output features, ResNet 101 utilizes the same number of layers and filters. After applying the integration chain rule, ResNet-101 employs residual connections.

3.3.3. VGG-19

The CNN known as VGG-19 was first developed in 2014 by Andrew Zisserman and Karen Simonyan of the Visual Geometry Group Lab at Oxford University. Compared to the top performance model, this model employed a relatively narrow receptive field over the whole network with a stride of 1 pixel. VGG was first introduced for the image classification of various disorders like MRI or X-Ray, but it may also be used to recognize traffic signals. It won the 2014 ImageNet detection competition [61]. 19 layers make up the VGG-19 model. The size of this network's input image is (224, 224, 3). With a kernel size of (3 x 3), 64 channels, and the same padding for the first two layers, the kernel is constructed. Following two layers with convolution filters of filter sizes (3, 3, 256), two layers with a stride (2, 2), and then the max-pooling layer. Multiple filters are utilized while performing a 3 x 3 convolution. This is the most popular method and is quickly establishing itself as the standard for extracting information from images. With 138 million various variables, it's difficult to keep track of everything. Additionally, a model that has been previously trained via transfer learning can have its parameters improved. The maximum-pooling layer. The ability to decrease volume size is provided by the max-pooling layer. There are 4096 nodes in the completely linked layer. Anyone interested in learning more about feature classification is encouraged to use our training technique, which combines the CNN architecture with the VGG-19 architecture [62].

3.3.4. Convolutional Neural Networks

Convolutional neural networks are one of the most prevalent types of deep neural networks. The term "deep" often refers to the neural network's number of hidden layers; in contrast, a regular neural network only contains 2 to 3 hidden layers. A chain model for deep learning is a convolutional neural network, or CNN. A convolutional neural network is comprised of multiple layers that analyze and reconstruct or build input to generate an output, allowing it to immediately learn from photos. Convolutional neural networks may be trained to perform image analysis tasks such as segmentation,

object detection, and classification. Local receptive fields, Haar weights, biases, and activation and pooling are the final three (3) key principles. However, in regions known as local receptive fields, only a very small proportion of the input layer neurons in a convolutional neural network are linked to the hidden layer neurons. The convolution layer field is altered across the image to produce a feature map from the input layer to the hidden layer of neurons. CNN has neurons with biases and weights, whereas shared weights and biases are similar to a traditional neural network. In the case of convolutional neural networks, this technique, which learns values throughout the training process, will update continually with each new training model. All hidden neurons in a layer have bias, weight values that are the same, indicating that they are all recognizing the same feature edge in various parts of the image. It will enable networks to translate items in images, such as a network trained to recognize cats, if the animals are present in the image. Each neuron with an activation function output is transformed by activation and pooling the activation measure. A common activation function is the rectified linear function (ReLU), which needs a neuron's output and maps it to the highest positive value or, if the output is negative, to zero. A pooling step can be used to further modify the activation step's output. By combining the output of several tiny groups of neurons into one output, pooling reduces the dimensionality of the feature map, which makes the succeeding layers simpler and lowers the number of parameters that the model must learn. In a convolutional neural network, these three concepts can be used to configure the layers, which may have tens or hundreds of hidden layers to learn how to recognize distinctive features in a picture. In the feature map shown below, each hidden layer increases the sophistication of the learned image features; for example, the first hidden layer learns how to recognize edges, and the final hidden layer learns. How to find more complex forms similar to a standard neural network? The last layer binds each neuron from the last hidden layer to the output neurons, creating the last output.

In this study, we have customized a model based on CNN, depicted in Table 2. Details are also depicted in Figure 2.

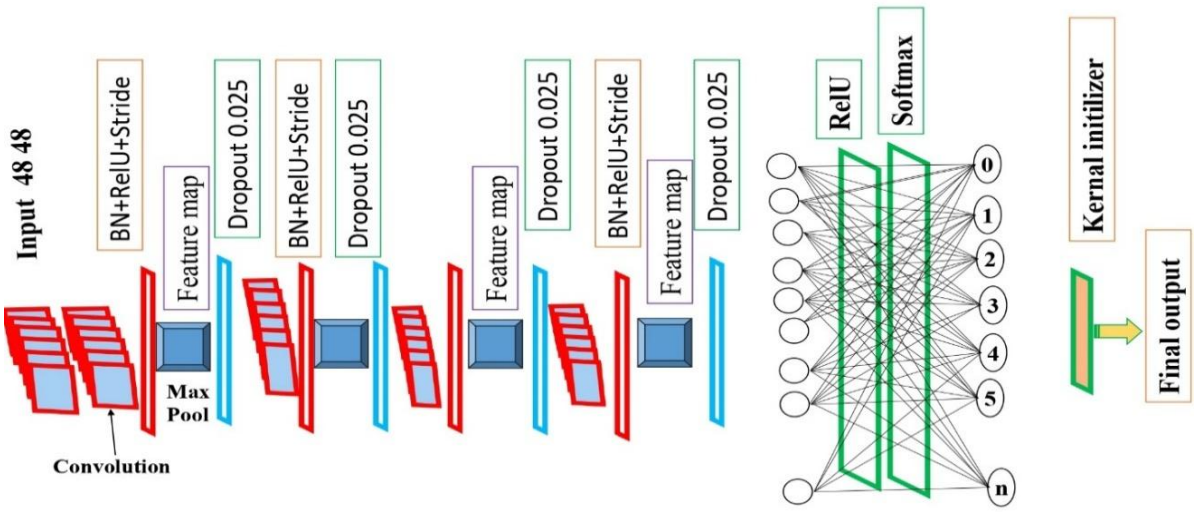


Figure 2: Architecture of the improved CNN model

Freezing Layers: Most of the characteristics are often left alone while learning a new task; therefore, in this method, the first layer contains the common generic features, and the subsequent layers gradually become more specialized to the target dataset.

Convolutional Layer: The foundation of the Convolution Neural Network is this layer. A generalized linear model for the fundamental local picture patch serves as the convolution filter in the fundamental convolutional neural network. It functions at the abstraction level when lower-dimensional instances of latent are present. The filter settings for this layer can be learned.

Table 2: CNN-based custom model details

Input Layer input size 300, 300,1					
Convolution 1	64,3,3	Padding	Same	Normalization	BN
Activation F	reLU	Pooling	Max Pool	Dropout	0.2
Convolution 2					
Convolution	128,5,5	Padding	Same	Normalization	BN
Activation F	reLU	Pooling	Max Pool	Dropout	0.01
Convolution 3					
Convolution	128,5,5	Padding	Same	Normalization	BN
Activation F	reLU	Pooling	Max Pool	Dropout	0.25
Convolution 4					
Convolution	512,3,3	Padding	Same	Normalization	BN
Activation F	reLU	Pooling	Max Pool	Dropout	0.025
FLATTEN					
FC 1 Dense	256	Padding	Same	Normalization	BN
Activation F	reLU	Pooling	Max Pool	Dropout	0.25
FLATTEN					
FC 1 Dense	512	Padding	Same	Normalization	BN
Activation F	reLU	Pooling	Max Pool	Dropout	0.025
Output Layer	Cross Entropy	Activation F	Soft-Max1	Learning	0.001

Pooling Layer: The Pooling layer comes after the Convolution layer. This layer is in charge of shrinking the input image's spatial size and working independently on each depth slice. This layer, which is nonparametric, creates the output layer by sliding filters with an initial fixed value for the stride. The following filtering operations were applied. Max Pooling: By obtaining the highest input cost within the filter's size, Average Pooling: By providing the average input amount inside the filter size.

Fully Connected Layer: The combined features in the class score that are present in the convolution neural network are converted before the output of the network by the fully connected (FC) layer. By taking the mesh topology technique into account, every neuron in this layer is connected to every other neuron in the layers below it. This layer's primary job is to learn the weights and biases needed to map the input layer to its appropriate output layer. The output O for FC layer i can be computed using the following equation.

$$O^i = AF\left(O^{i-1} \times w^i + \beta^i\right) \quad (1)$$

Here, O is the output of layer i, AF is the activation function, w represents the weight, and β represents bias.

Activation Function: The activation function is used to determine the non-linearity in the network to learn more complex functions. In the deep learning framework, the nonlinear transformation from input to output is performed using the activation functions from the nonlinear layers and their combination with other layers [63]. Therefore, an appropriate activation function is required for a better feature-extracting strategy [64], [65], [66].

Rectified Linear Unit (ReLU): There are different activation functions that exist in the literature; however, Rectified Linear Unit (ReLU) is the most extensively used activation function. This function is denoted by using the following equation.

$$\mathrm{Rl} = \max\{(0, \mathrm{a}) \quad (2)$$

Sigmoid Function: Where a denotes the input from the front layer. The values of the sigmoid function are transformed with values ranging from 0 to 1 and are commonly used to produce a Bernoulli distribution.

$$sig(i) = \frac{1}{1 + e^{-i}} \tag{3}$$

Sig represents sigmoid, and *e* represents Euler's number

Hyperbolic Tangent:

$$\mathrm{Y} = \tanh(\mathrm{x}) \tag{4}$$

The elements of X's hyperbolic tangent are returned. Arrays are handled element-by-element using the tanh function. The function takes inputs that are both real and complicated. Every angle is measured in radians.

Softmax:

$$AF(x) = \frac{e^{x_j}}{\sum_{n=1}^{N} e^{x_j}} \tag{5}$$

This layer is commonly used as the final output layer that can be considered as a probability distribution over the categories. Table 3 contains all details of our Custom model illustrated in Figure 2, which depicts input details, convolution layers, Normalization type, nature of pooling, Activation functions used in each layer, nature of Fully connected FC layers, dropout, and Learning rate Lr parameters. Custom model input images, followed by four convolution layers, using padding, batch normalization, ReLU activation function, and max pooling, where dropout of different ranges is used to avoid over-fitting and to reduce its effects. Three connected layers using 0.25 dropout, where all connected layers are dense, and finally, the output layer contains a Softmax activation function along with a 0.001 learning rate. The learning rate was set to 0.0001 to ensure gradual and stable convergence, especially important when fine-tuning pre-trained models like MobileNetV2. Higher learning rates (e.g., 0.001) caused instability and sharp fluctuations in training loss. A low learning rate allowed the network to fine-tune deep layers without catastrophic forgetting, preserving valuable features from ImageNet initialization while adapting to our autism facial classification task. A dropout rate of 0.3 was applied after the global average pooling layer to mitigate overfitting. This value was selected after testing dropout rates of 0.2, 0.4, and 0.5. A value of 0.3 effectively regularized the fully connected layers, reducing reliance on specific neurons while maintaining learning capacity. It also correlated with a slower increase in validation loss compared to more aggressive dropout.

3.4. Evaluation Measures

Any framework or network's performance is evaluated using evaluation metrics [67]. Specific evaluation measures are used for particular tasks. Standard measurements such as the Dice coefficient matrices, the Jacquard index, Hausdorff, and PSNR SSIM can be used to evaluate segmentation, denoising, and image fusion.

3.4.1. F1 measures

The F1 Measure and the Jacquard index [68] are the most widely used performance measures in

medical image segmentation. Instead of pixel-wise accuracy, these measures are utilised since they are a better indicator of a segmentation's perceptual quality. According to Zijdenbos [69] . For object

segmentation, the Dice score better shows size and location consistency. They furthermore show that when the number of background voxels is taken into account, the Dice score is indeed a subset of the Kappa statistic, considerably outnumbers the number of foreground voxels. Let segmentation in image I be $\{S_1 S_1 S_1 \dots S_n\}$

$$S_i \cap S_j = \emptyset \; for \; i \neq j \; U_i^n S_i = I \tag{6}$$

S=U S_i

$$\rho(S_i A) = \max\left[\frac{|S_i \cap A|}{|S_i|}, \frac{|S_i \cap A|}{|A|}\right] \tag{7}$$

3.4.2. Confusion Matrix

True positive rate / Sensitivity (TPR)

The True Positive Rate, also known as sensitivity, is the number of subjects who have been positively tested. TPR = sum of true positives/sum of positive conditions can be expressed mathematically as:

TPR = sum of true positives/sum of positive conditions

$$\text{Sensitivity} = \frac{\text{TrPos}}{\text{Tr Pos} \; + \; \text{Fa Neg}} \tag{8}$$

True negative rate / Specificity (TNR)

TNR, also known as specificity, indicates that negative tests were correctly recognized. It can be mathematically formulated as:

TNR = Total number of true negative values / Total number of negative conditions)

$$\text{specificity} = \frac{\text{TN}}{\text{FP} \; + \; \text{TN}} \tag{9}$$

Positive predictive value (PPV)

PPV =Total true positive / Total number of predicted positive

$$\text{PosPerVal} = \frac{\text{TrPos}}{\text{Tr Pos} \; + \; \text{Fal Pos}} \tag{10}$$

Negative predictive value (NPV)

NPV=Total true Negative / predicted Negative Condition

$$\text{NegPerVal} = \frac{\text{TrNeg}}{\text{Tr Neg} \; + \; \text{FaNe}} \tag{11}$$

3.4.3. Accuracy (AC)

$$\text{AC} = \frac{\text{TrPos} + \text{TrNe}}{\text{Tr Pos} \; + \; \text{Fal Pos} + \text{FaNe} + \text{TrNeg}} \tag{12}$$

4. Results and Discussion

A machine learning algorithm starts the learning process by using training data. In supervised

learning, data consists of features along with their corresponding labels. Training in machine learning means adjusting weights in such a way that they can map features to corresponding outputs. Now to tell the machine learning algorithm what to do with the input data so that the mapping is performed in the best possible way. We used loss functions and two different optimizers. The loss function tells us to what extent the algorithm's prediction is right or wrong. The Adamax optimizer is used in the proposed algorithm along with sparse categorical cross-entropy. Another important parameter while training a deep neural network is the learning rate, as the core purpose of every ML model is to minimize the loss for each iteration weights are adjusted so that we move towards a smaller loss. The size of the steps toward the optimum is decided by the learning rate. We trained our models with 3 learning rates, which are (0.1, 0.01, 0.001), and then we evaluated the models' results on the basis of different learning rates. Batch size and Epochs are two other hyperparameters. Batch size determines the number of examples used in each iteration, while epoch decides how many times an individual example is used in the training process. Finally, it is also important to note that among them, EfficientNet outperforms in the case of both datasets. Besides, we also concluded that Adamax performed exceptionally well compared to Adam.

The results in Table 3 represent a comparison between the proposed framework and other techniques. Proposed techniques achieve optimal values of Accuracy, confusion matrix, and F1 Score.

Table 3: Summary of the confusion matrix

Classifier	Precision	Recall	Accuracy	AUC	F1
ResNet	0.9	0.977	0.98	0.9523	0.9603
VGG	0.89	0.96	0.96	0.966	0.9650
Proposed	1.0	0.98	0.99	0.9869	0.989

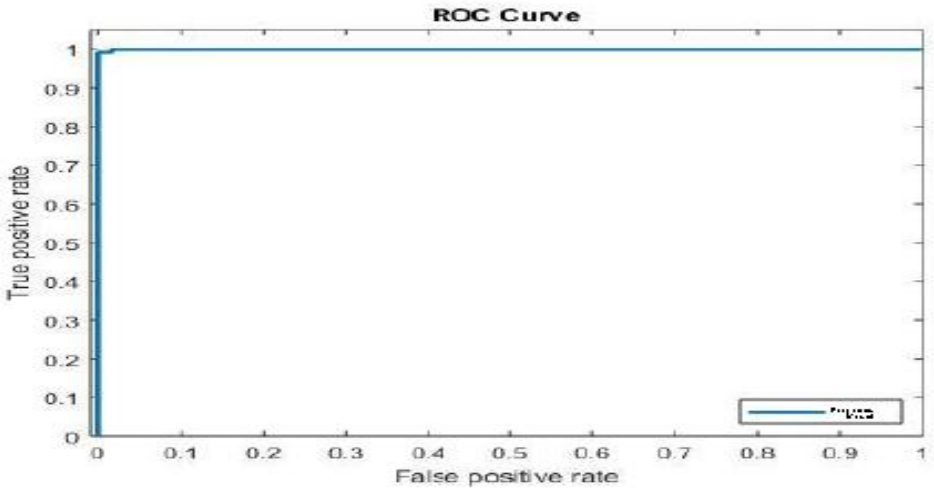


Figure 3: ROC value using the proposed CNN-based model

Despite the training accuracy reaching 99% by epoch 10, validation accuracy plateaued around 74-75% after epoch 4. The divergence between training and validation losses indicated that overfitting began to emerge. This was detected early via continuous monitoring, and training was stopped at epoch 10 to avoid further overfitting. Rather than stacking additional dense layers, the architecture was kept intentionally lightweight. We used a single fully connected output layer with sigmoid activation for binary classification. This helped avoid unnecessary over-parameterization.

Table 3 depicts the overall results of the confusion matrix, accuracy, and F1 score. Results show that the pre-trained transfer learning approach (ResNet-101 and VGG-19) has low accuracy, AUC, and F1 score as compared to our custom CNN-based model. We have improved our learning and detection procedure by using CNN, Batch normalization, and ReLU as activation functions for hidden layers and strides, four convolution layers, two fully connected layers, one input layer, and one output layer. We have also improved results by extracting different facial features by LBP and pre-processing our data with a custom CNN model. Figure 3 depicts that we have achieved a 0.99 AUC value, using a custom model that optimally classifies autistic and non-autistic cases among children's faces. Similarly, Table 4 shows the dataset, model utilized, and accuracy; our customized model outperforms all research.

Results presented in Figures 4, 5, and 6 represent the confusion matrix of each model. It is obvious from the results that the CNN custom model generates more robust results in terms of ASD classification and detection.

Table 4: Comparison between previous techniques and the proposed model

Author	Dataset	Methods	ACC
[20]	Kaggle	MobileNet	90%
[70]	Kaggle	CNN	95%
[71]	Kaggle	MobileNet	78%
[21]	Kaggle	Xception	90%
[72]	Kaggle	Xception	91%
Proposed Model	Kaggle	CNN	99%

Figure 6 indicates the maximum, positive predictive value, and minimum false negative values using a CNN-based custom model, which is a solid contribution to this study. This study is based on DL and ML-based feature extraction, followed by pre-processing steps to avoid under- and over-fitting. We have also performed augmentation steps, and pre-trained models have been employed to tackle the vanishing gradient problems. Results also depict that the custom model outperformed as compared to conventional approaches.

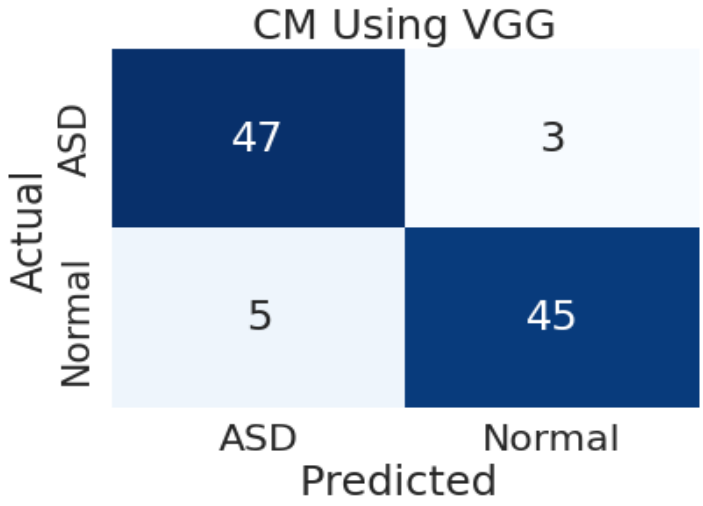


Figure 4: Confusion matrix using VGG

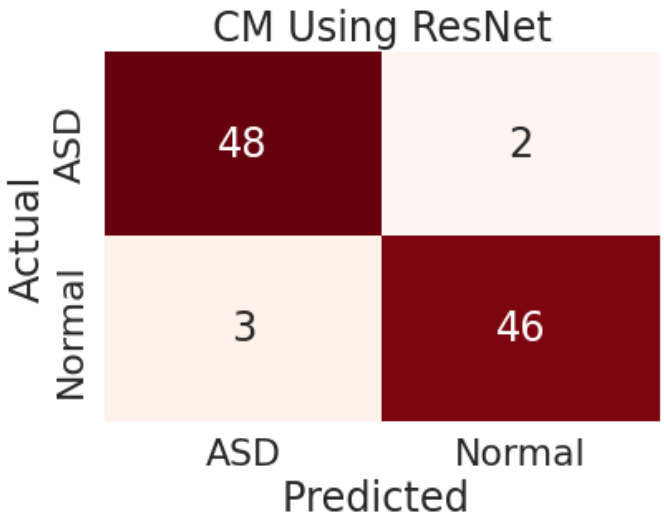


Figure 5: Confusion matrix using ResNet

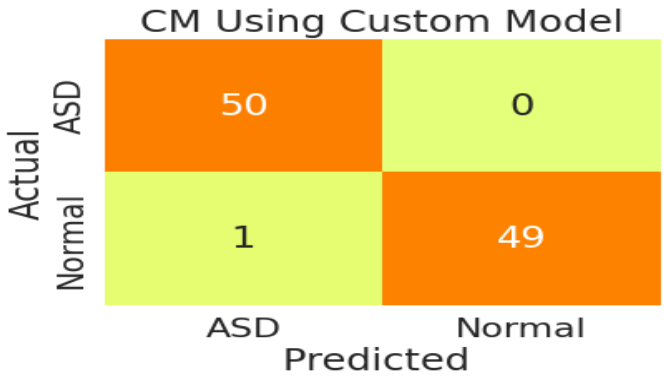


Figure 6: Confusion matrix using the proposed CNN-based model

Accuracy is considered a Major concern in image analysis, in terms of facial expression and emotion detection. Figure 7 shows that our CNN-based custom model has achieved an optimal accuracy of 99% by training for fifty epochs. Results also reveal that after a gradual increment in training, training as well as validation both gradually improved and outperformed compared to pre-trained models.

As compared to state-of-the-art research, we have to improve the classification and detection accuracy of ASD data. In terms of loss computation, over a custom model exponentially reduces the loss, and after some epochs, it remains consistently as low as possible. This also shows that the proposed framework produced more realistic and better results by achieving the gold standards of accuracy and loss.

5. Conclusions

In this study, we have employed transfer learning and a CNN-based custom model to detect and classify Autistic and non-autistic classes from the facial dataset. We have come up with the following conclusions. Firstly, our custom model, which consists of convolution layers, BN, Max pooling, Relu activation function, and drop-out of 0.025, has generated optimal results, maximum accuracy, maximum PPV, and Minimum NPV as compared to the transfer learning approach. Secondly, Res-Net has better accuracy than VGG, where VGG has a lower NPV, which indicates the unique behavior of the dataset. Lastly, our custom model generates optimal results at the 15th epoch, whereas transfer learning generates better results at the 50th epoch.

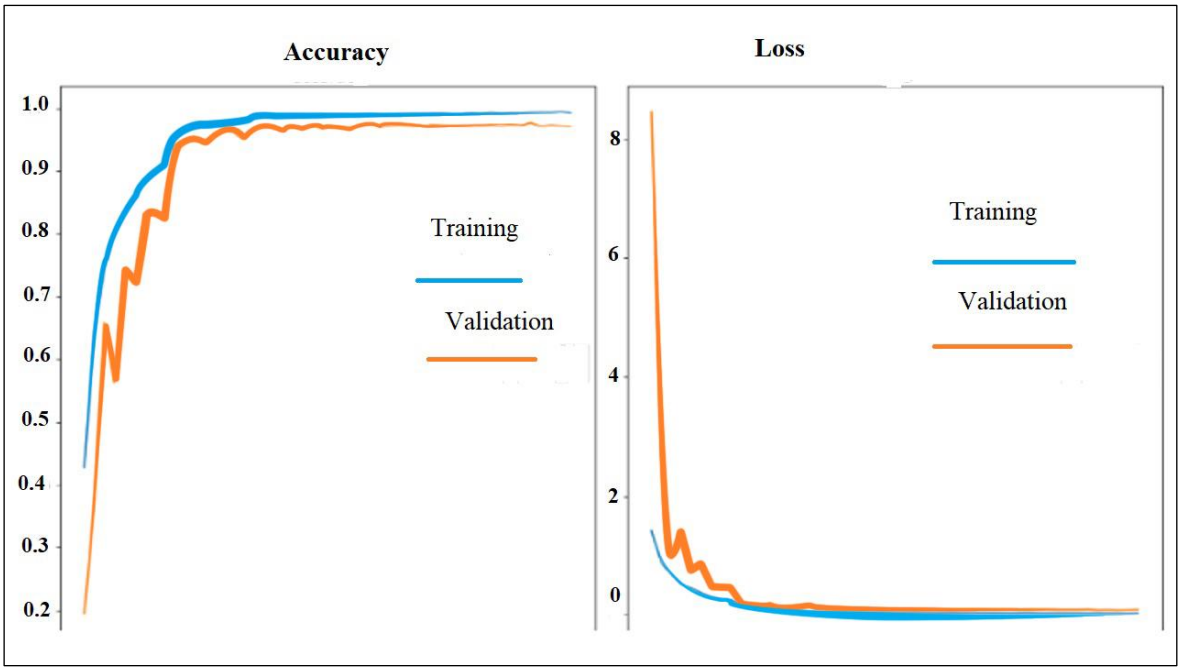


Figure 7: Accuracy using the proposed CNN-based model

6. References

[1] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2002.

[2] J. Baio *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014," *MMWR Surveill. Summ.*, vol. 67, no. 6, p. 1, 2018.

[3] D. S. Mandell *et al.*, "Racial/ethnic disparities in the identification of children with autism spectrum disorders," *Am. J. Public Health*, vol. 99, no. 3, pp. 493–498, 2009.

[4] A. Estes, J. Munson, S. J. Rogers, J. Greenson, J. Winter, and G. Dawson, "Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 54, no. 7, pp. 580–587, 2015.

[5] A. Estes, J. Munson, S. J. Rogers, J. Greenson, J. Winter, and G. Dawson, "Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 54, no. 7, pp. 580–587, 2015.

[6] A. M. Angell, A. Empey, and K. E. Zuckerman, "A review of diagnosis and service disparities among children with autism from racial and ethnic minority groups in the United States," *Int. Rev. Res. Dev. Disabil.*, vol. 55, pp. 145–180, 2018.

[7] L. Zwaigenbaum and M. Penner, "Autism spectrum disorder: advances in diagnosis and evaluation," *Bmj*, vol. 361, 2018.

[8] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, 2018.

[9] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, 2019.

[10] C. Moret-Tatay, A. G. Wester, and D. Gamermann, "To Google or not: Differences on how online searches predict names and faces," *Mathematics*, vol. 8, no. 11, p. 1964, 2020.

[11] S. Jabeen, Z. Mehmood, T. Mahmood, T. Saba, A. Rehman, and M. T. Mahmood, "An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model," *PloS One*, vol. 13, no. 4, p. e0194526, 2018.

[12] K. M. Rump, J. L. Giovannelli, N. J. Minshew, and M. S. Strauss, "The development of emotion recognition in individuals with autism," *Child Dev.*, vol. 80, no. 5, pp. 1434–1447, 2009.

[13] F. Castelli, "Understanding emotions from standardized facial expressions in autism and normal development," *Autism*, vol. 9, no. 4, pp. 428–449, 2005.

[14] H. Akechi, A. Senju, Y. Kikuchi, Y. Tojo, H. Osanai, and T. Hasegawa, "The effect of gaze direction on the processing of facial expressions in children with autism spectrum disorder: an ERP study," *Neuropsychologia*, vol. 48, no. 10, pp. 2841–2851, 2010.

[15] M. O. Bakare and K. M. Munir, "Autism spectrum disorders (ASD) in Africa: a perspective," *Afr. J. Psychiatry*, vol. 14, no. 3, pp. 208–210, 2011.

[16] C. E. Rice *et al.*, "Evaluating changes in the prevalence of the autism spectrum disorders (ASDs)," *Public Health Rev.*, vol. 34, no. 2, pp. 1–22, 2012.

[17] V. Klucharev and M. Sams, "Interaction of gaze direction and facial expressions processing: ERP study," *Neuroreport*, vol. 15, no. 4, pp. 621–625, 2004.

[18] L. S. Smith, P. A. Dmochowski, D. W. Muir, and B. S. Kisilevsky, "Estimated cardiac vagal tone predicts fetal responses to mother's and stranger's voices," *Dev. Psychobiol. J. Int. Soc. Dev. Psychobiol.*, vol. 49, no. 5, pp. 543–547, 2007.

[19] S. Bölte and F. Poustka, "The recognition of facial affect in autistic and schizophrenic subjects and their first-degree relatives," *Psychol. Med.*, vol. 33, no. 5, pp. 907–915, 2003.

[20] T. Akter *et al.*, "Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage," *Brain Sci.*, vol. 11, no. 6, p. 734, 2021.

[21] K. K. Mujeeb Rahman and M. M. Subashini, "Identification of autism in children using static facial features and deep neural networks," *Brain Sci.*, vol. 12, no. 1, p. 94, 2022.

[22] K. O'connor, J. P. Hamm, and I. J. Kirk, "The neurophysiological correlates of face processing in adults and children with Asperger's syndrome," *Brain Cogn.*, vol. 59, no. 1, pp. 82–95, 2005.

[23] T. May, K. Cornish, and N. J. Rinehart, "Exploring factors related to the anger superiority effect in children with Autism Spectrum Disorder," *Brain Cogn.*, vol. 106, pp. 65–71, 2016.

[24] A. O. Salau and S. Jain, "Feature extraction: a survey of the types, techniques, applications," in *2019 International Conference on Signal Processing and Communication (ICSC)*, 2019, pp. 158–164.

[25] K. Zhang, Y. Yuan, J. Chen, G. Wang, Q. Chen, and M. Luo, "Eye Tracking Research on the Influence of Spatial Frequency and Inversion Effect on Facial Expression Processing in Children with Autism Spectrum Disorder," *Brain Sci.*, vol. 12, no. 2, p. 283, 2022.

[26] A. Marotta, B. Aranda-Martín, M. De Cono, M. Á. Ballesteros-Duperón, M. Casagrande, and J. Lupiáñez, "Integration of facial expression and gaze direction in individuals with a high level of autistic traits," *Int. J. Environ. Res. Public. Health*, vol. 19, no. 5, p. 2798, 2022.

[27] M. A. Witherow, M. D. Samad, N. Diawara, H. Y. Bar, and K. M. Iftekharuddin, "Deep Adaptation of Adult-Child Facial Expressions by Fusing Landmark Features," *ArXiv Prepr. ArXiv220908614*, 2022.

[28] S. Du and A. M. Martinez, "Compound facial expressions of emotion: from basic research to clinical applications," *Dialogues Clin. Neurosci.*, 2022.

[29] M. Liao, H. Duan, and G. Wang, "Application of Machine Learning Techniques to Detect the Children with Autism Spectrum Disorder," *J. Healthc. Eng.*, vol. 2022, 2022.

[30] S. Uono *et al.*, "The structural neural correlates of atypical facial expression recognition in autism spectrum disorder," *Brain Imaging Behav.*, vol. 16, no. 3, pp. 1428–1440, 2022.

[31] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 435–442.

[32] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang, and E. Xun, "Ensemble of deep neural networks with probability-based fusion for facial expression recognition," *Cogn. Comput.*, vol. 9, no. 5, pp. 597–610, 2017.

[33] W. Hua, F. Dai, L. Huang, J. Xiong, and G. Gui, "HERO: Human emotions recognition for realizing intelligent Internet of Things," *IEEE Access*, vol. 7, pp. 24321–24332, 2019.

[34] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 454–459.

[35] J. Whitehill, M. S. Bartlett, and J. R. Movellan, "Automatic facial expression recognition," *Soc. Emot. Nat. Artifact*, vol. 88, 2013.

[36] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[37] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2450–2460, 2020.

[38] Y. Xue, X. Mao, and F. Zhang, "Beihang university facial expression database and multiple facial expression recognition," in *2006 International Conference on Machine Learning and Cybernetics*, 2006, pp. 3282–3287.

[39] I. M. Revina and W. R. S. Emmanuel, "A survey on human face expression recognition techniques," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 6, pp. 619–628, 2021.

[40] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[41] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[42] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*, 2013, pp. 117–124.

[43] H. Kaya, F. Gürp\inar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image Vis. Comput.*, vol. 65, pp. 66–75, 2017.

[44] E. Ferrara, M. JafariAsbagh, O. Varol, V. Qazvinian, F. Menczer, and A. Flammini, "Clustering memes in social media," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013, pp. 548–555.

[45] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 886–893.

[46] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 30–40.

[47] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face \& Gesture Recognition (FG 2017)*, 2017, pp. 558–565.

[48] X. Liu, B. V. K. Vijaya Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 20–29.

[49] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated CNN for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2209–2214.

[50] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 222–237.

[51] H. Yu, Z. Luo, and Y. Tang, "Transfer learning for face identification with deep face model," in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, 2016, pp. 13–18.

[52] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim, Y. Li, and X. Liu, "An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips," *IEEE Trans. Nanotechnol.*, vol. 18, pp. 819–829, 2019.

[53] F. Zhi-Peng, Z. Yan-Ning, and H. Hai-Yan, "Survey of deep learning in face recognition," in *2014 International Conference on Orange Technologies*, 2014, pp. 5–8.

[54] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[55] J. Li, T. Qiu, C. Wen, K. Xie, and F.-Q. Wen, "Robust face recognition using the deep C2D-CNN model based on decision-level fusion," *Sensors*, vol. 18, no. 7, p. 2080, 2018.

[56] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 467–474.

[57] M. Yu, H. Zheng, Z. Peng, J. Dong, and H. Du, "Facial expression recognition based on a multi-task global-local network," *Pattern Recognit. Lett.*, vol. 131, pp. 166–171, 2020.

[58] J. Dong, H. Zheng, and L. Lian, "Dynamic facial expression recognition based on convolutional neural networks with dense connections," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3433–3438.

[59] S. Iqbal *et al.*, "Prostate cancer detection using deep learning and traditional techniques," *IEEE Access*, vol. 9, pp. 27085–27100, 2021.

[60] R. U. Khan, X. Zhang, and R. Kumar, "Analysis of ResNet and GoogleNet models for malware detection," *J. Comput. Virol. Hacking Tech.*, vol. 15, no. 1, pp. 29–37, 2019.

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv Prepr. ArXiv14091556*, 2014.

[62] R. N. S. Husna, A. R. Syafeeza, N. A. Hamid, Y. C. Wong, and R. A. Raihan, "Functional magnetic resonance imaging for autism spectrum disorder detection using deep learning," *J. Teknol.*, vol. 83, no. 3, pp. 45–52, 2021.

[63] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, no. 187, pp. 1–30, 2018.

[64] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[65] L. C. Adams *et al.*, "Dataset of prostate MRI annotated for anatomical zones and cancer," *Data Brief*, vol. 45, p. 108739, 2022.

[66] Y. Tayal, P. K. Pandey, and D. B. V Singh, "Face recognition using eigenface," *Int. J. Emerg. Technol. Comput. Appl. Sci. IJETCAS*, vol. 3, no. 1, pp. 50–55, 2013.

[67] K. Suzuki, "Overview of deep learning in medical imaging," *Radiol. Phys. Technol.*, vol. 10, no. 3, pp. 257–273, 2017.

[68] T. Eelbode *et al.*, "Optimization for medical image segmentation: theory and practice when evaluating with Dice score or Jaccard index," *IEEE Trans. Med. Imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[69] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: method and validation," *IEEE Trans. Med. Imaging*, vol. 13, no. 4, pp. 716–724, 1994.

[70] M. S. Alam, M. M. Rashid, R. Roy, A. R. Faizabadi, K. D. Gupta, and M. M. Ahsan, "Empirical study of autism spectrum disorder diagnosis using facial images by improved transfer learning approach," *Bioengineering*, vol. 9, no. 11, p. 710, 2022.

[71] Y. Khosla, P. Ramachandra, and N. Chaitra, "Detection of autistic individuals using facial images and deep learning," in *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 2021, pp. 1–5.

[72] F. W. Alsaade and M. S. Alzahrani, "Classification and Detection of Autism Spectrum Disorder Based on Deep Learning Algorithms," *Comput. Intell. Neurosci.*, vol. 2022, 2022.