Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 2
Open Access

Lightweight Deep Learning Model for Autism Spectrum Disorder Detection and Expression Recognition in Children Using Facial Images

Nilofer Attar[1], Dr. Shilpa Paygude[2]

[1]Research scholar, School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Kothrud, Pune
Email: kittadnilofer@gmail.com

[2]Professor, School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune
Email: shilpa.paygude@mitwpu.edu.in

Article Info

Article type:

Research

Article History:

Received: 2024-03-22

Revised: 2024-05-13

Accepted: 2024-06-18

Keywords:

Autism, Augmentation, Facial Expression, MobileViT, VGG-16, Accuracy, Loss, Website, Deep Learning

A B S T R A C T

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental illness that affects social skills, pronunciation, and communication abilities. Early diagnosis of ASD relies on detecting brain function abnormalities, which may be modest or absent at the beginning of the condition. Because children with ASD frequently exhibit distinct patterns that set them apart from normally developing children, facial expression analysis has emerged as an alternative and successful tool for the early detection of ASD. Identifying autism using facial expressions is difficult for both parents and physicians. However, Deep Learning (DL) can help address this issue. In this study, we developed a DL system to detect children's conditions, distinguishing between normalcy and autism. Additionally, the system identifies expressions such as happiness, sadness, and anger from facial images of children. For ASD detection, we employed MobileViT, while for expression recognition, we utilized VGG-16. The dataset used for training and testing comprised over 600 facial images sourced from the internet. The DL models' performance was assessed using standard evaluation metrics like accuracy and loss. During training and validation, the MobileViT model achieved peak accuracies of 99% and 98%, respectively, while the VGG-16 model attained peak accuracies of 96% and 99%, respectively. With these promising results, we proceeded to deploy the models on a website. The website interface allows users to simply upload a photo of a child, whereupon the models, MobileViT and VGG-16, discern the child's condition and expression, facilitating easy assessment.

1. INTRODUCTION

ASD is a complex neurological and developmental disorder that affects the ability to grasp and absorb information [1]. It is becoming more prevalent among individuals of all ages. The majority of children with this disorder develop symptoms between the ages of two and five. The frequency has spread from young toddlers to adults and teenagers. Individuals with ASD typically have a decreased capacity to communicate verbally. The emergence of this disorder is linked to both genetic and neurological factors. ASD symptoms include issues in social interaction, reasoning, visualizing, and repetitive behavior. This disorder does not currently have a treatment option. Although the specific methods by which ASD develops are unknown, researchers believe that both environmental triggers and genetic flaws play a significant role in the disorder's progression. Genetic disorders associated with ASD affect multiple brain regions [2].

Based on the Centers for Disease Control and Prevention (CDC), ASD affects around one in every 36 American children [3]. Autism, a disorder that has become more common in recent years, may impact one out of every 70 infants born today. Boys are more likely than girls to be diagnosed with ASD. Between 2009 and 2017, around one-sixth of parents with children aged three to seventeen reported that their child had a developmental disability. 10 to 20% of ASD patients have several gene mutations, and more than one hundred genes have been related to the disorder. An early diagnosis of the neurological disorder can help protect the patient's mental and physical health. In 2022, the CDC reported a 178% increase in the rate of ASD in the US alone.

This disorder is most commonly observed in children over the age of two and is distinguished by a lack of eye contact, stereotypical conduct, and low social interaction [4]. Diagnosis is challenging because this condition currently lacks distinct clinical tests, such as blood testing. Since there are currently no established pathophysiological diagnostic criteria for ASD, several psychological assessments have been created to fill this requirement. A child's social interaction and behavior can be evaluated with the use of these questionnaires, which use questions tailored to their age. To validate the diagnosis, these instruments are used in conjunction with the patient's health record, clinical findings, and IQ tests [5]. While ASD is not yet curable, early diagnosis is critical for symptom management and future skill development. Doctors and nurses may also lack the knowledge and skills necessary to diagnose autistic children, making communication difficult. In terms of disease prediction, DL algorithms surpass human reviewers. In this research, we utilized the MobileViT and VGG-16 models to identify children's mental conditions and expressions. The evaluation of the model was conducted based on accuracy and loss functions. Subsequently, the finalized model was deployed to develop a website for predicting whether a child is affected by ASD or not, along with identifying their expression using their facial images.

2. LITERATURE SURVEY

The application of Artificial Intelligence (AI), Machine Learning (ML), and DL to detect indications of ASD in facial images has recently been the focus of extensive research. Several significant studies have looked into the hopeful future of these technologies for ASD screening and diagnosis. The goal of the research [6] is to investigate the feasibility of applying different ML algorithms to detect and analyze ASD concerns in children, adolescents, and adults. Six ML and DL algorithms were used and tested the proposed approaches on three publicly available datasets that do not contain any clinically significant ASD information. The first dataset for ASD screening in children consists of 292 cases and 21 factors. The second dataset for adults contains 754 occurrences and 21 factors. The third dataset includes 104 cases and 21 factors associated with ASD screening in adolescents. Utilizing a variety of ML methods and addressing missing values, the results demonstrate that Convolutional Neural Network (CNN)-based prediction models outperform all three datasets, achieving the highest levels of accuracy for adolescent, adult, and ASD screening data.

The study [7] demonstrates that data-centric DL algorithms can reliably diagnose ASD using face images. This technique involves training a CNN with a collection of face pictures to distinguish between ASD participants and non-ASD individuals. The approach employs data-centric techniques, including pre-processing and synthesis of training datasets. Performance metrics of various data-centric strategies are evaluated by comparing the trained model to an independent test set. The proposed method enhances previous attempts by combining pre-processing strategies with augmentation approaches on the training dataset. Furthermore, the study enhances the algorithm's clarity and comprehensibility by employing explainable AI techniques, providing physicians with meaningful and interpretable insights into the ASD diagnosis model's decision-making process. The study [8] introduces an excellent prediction model based on ML approaches to develop a mobile app for ASD prediction across all age groups. A mobile application is created using the proposed ASD detection method, which combines Random Forest (RF) and Classification and Regression Trees (CART). The suggested model was evaluated using 250 authentic datasets, including AQ-10 results as well as those from individuals with and without autism. Based on the evaluation findings, the proposed prediction model enhanced both kinds of datasets.

In research [9], the author introduced an ML strategy for detecting children with ASD that combines electroencephalography (EEG) data with behavioral data. Its use can reduce costs while improving detection

efficiency. The study began by collecting features from EEG, facial, and eye expressions using a unique technique. Then, for precise multimodal data fusion, a weighted naive Bayes algorithm-based hybrid fusion method was proposed. According to the findings, the study's ML classification algorithm is effective at detecting ASD in its early stages. Graphs and confusion matrices show that EEG could provide a source of discriminative information; however, eye and facial expressions have varying discriminative abilities for distinguishing children with ASD and normal. Importantly, behavioral and physiological data complement each other. The study's proposed ML strategy, which incorporates supplemental data, can significantly improve classification accuracy. In the article [10], common characteristics such as simulation technique, comparison approach, and input data were investigated and compared for ASD prediction. This study's primary goal is to provide a consolidated framework for researchers working on ASD prediction. The RF technique outperformed more traditional ML methods, thus it was utilized to achieve the greatest results. Readers can refer to the workflow representations of the examined frameworks to gain a better understanding of their underlying architecture and activities.

The study [11] investigates the potential of creating predictive models for using facial images in the diagnosis of ASD in children. To accomplish this, the study examined a dataset of 29,36 face images of both typically developing and autistic children. The program made use of classic ML techniques. They applied two cutting-edge methodologies: DL and automatic machine learning (AutoML). Compared the outcome obtained using the numerous methods that are now available. As a result, it was discovered that AutoML beat the competition while optimizing pipelines with Hyperpot and tree-based approaches. Furthermore, they were able to determine the ideal parameter choices using AutoML techniques, with no human intervention necessary for feature engineering. The research [12] found that facial processing plays an important role in the development of ASD. The author used an event-related potentials (ERP) task to compare a group of 8-month-old babies with an older sibling with ASD (n = 148) to those without (n = 68). Then standard case-control comparisons with ML to predict social features and ASD diagnosis at 36 months, as well as Bayesian hierarchical clustering to divide the newborns into subgroups. Several alterations in the structure of how the brain interprets faces from infancy were linked to ASD later in life, and there was a high overlap in the ERP components that predicted social features and diagnosis. Different patterns of brain reactivity to faces, which varied according to later sensory sensitivity, enabled us to identify two primary subgroups within ASD. When taken as a whole, the findings suggest that the diverse pattern of alterations associated with ASD in the first year of life is driven by individual differences among neonates. To better understand the processes that lead to subsequent neurodevelopmental outcomes, they shift the focus away from group-level comparisons and toward pattern recognition and classification in clinical cohorts.

In the article [13], the DL model that can correctly classify children as either normal or perhaps autistic with a success rate is presented. Autism individuals struggle with verbal and nonverbal communication, repetitive behaviors, and social skills. Although the condition is assumed to be hereditary, the most reliable diagnosis is made after an examination of the child's behavioral and facial traits. Researchers can detect if a kid has the illness based on an image alone since sufferers frequently exhibit a pattern of clearly identifiable facial deformities. The DL model extracts characteristics and classifies images using MobileNet with two Dense Layers (DeL).

From the literature survey, numerous DL models have been developed for ASD detection. However, a primary drawback of the existing systems is their narrow focus solely on diagnosis, without consideration for other crucial aspects of autism management like therapy, intervention, or emotion recognition, all of which are vital for the comprehensive well-being of children with autism. One significant shortcoming lies in the absence of emotion recognition capabilities. Understanding and supporting the emotional development of children with autism are pivotal, yet the current systems overlook this essential aspect. Furthermore, the existing system's scalability could be restricted due to its reliance on manual and expertise-dependent processes, which may hinder its ability to effectively meet the growing demand for autism evaluation and diagnosis. To address these issues, we developed an effective DL model and deployed it on a website capable of predicting not only ASD but also children's emotions.

3. MATERIALS AND METHODS

For ASD detection, children's images are collected and categorized into two labels: ASD and normal. Additionally, expressions such as happy, angry, and sad are also labelled. Pre-processing methods performed on the images consist of resizing, normalization, and augmentation. Following this, the processed images that have been labelled as ASD and normal are employed in the training and validation of MobileNet. Conversely, the images that have been labelled with expressions are utilized in the training of VGG-16. Following the training and validation phases, a website is developed for detecting children's mental conditions and expressions. Test images are uploaded to the site, processed, and then passed to MobileNet for detecting mental conditions and to VGG-16 for recognizing expressions. Finally, the website provides assessments of children's health conditions and expressions. The overall workflow is illustrated in Figure 1.

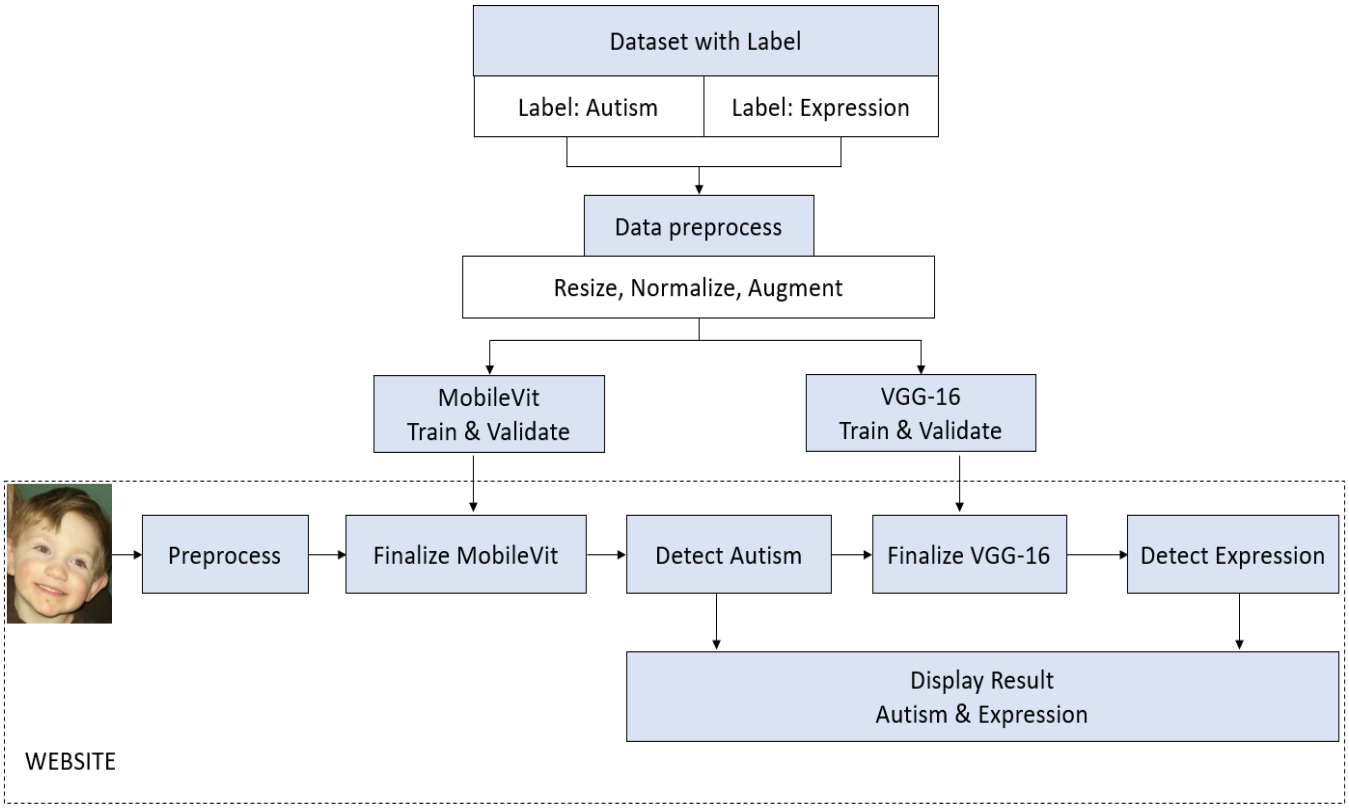


Fig. 1. Overall Framework for Autism and Expression Recognition

A. Data

The study used publicly available data from the internet [14] to identify children with autism and without autism and their facial expressions. We collected 292 images from individuals with autism and 320 images from normal individuals. Among the 292 images, we have 114 happy expressions, 64 sad expressions, and 114 angry expressions. From the 320 images, we have 228 happy, 36 sad, and 56 angry images. Sample images of expressions from both autism and normal individuals are given in Figure 2. Table 1 provides the distribution of the dataset.

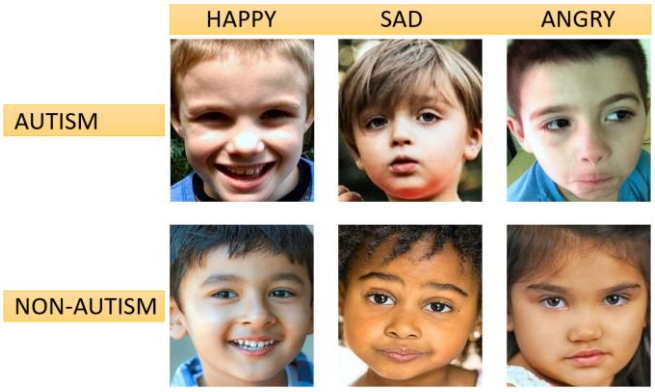


Fig. 2. Sample images from the dataset

Table 1. ASD and normal data distribution

Expression	Autism	Non-Autism
Happy	114	228
Sad	64	36
Angry	114	56

B. Pre-process

The data needed to be pre-processed before being used to train the DL model; it was collected via the Internet. To prepare the dataset for analysis, it is necessary to resize the image to adapt to the MobileViT and VGG-16 architecture [15]. The images are augmented using geometrical transformation, and the dataset is normalized by rescaling the parameters in the Keras pre-processing. All images in the dataset are scaled from [0, 255] to [0, 1]. Normalizing the dataset is essential for training with the DL model. The augmentation techniques like rotation, flip, shift, and zoom are done in the training phase of the research.

C. Autism Prediction – MobileViT

Figure 3 shows the framework of the MobileViT network. With fewer parameters, the MobileViT block attempts to represent both global and local information in an input tensor. MobileViT generates an output for an input tensor $X \in R^{H*W*C}$ using both $n*n$ conventional convolutional layers (CL) and a point-wise (or 1*1) CL. The $n*n$ CL encodes local spatial details by learning linear combinations of input channels, but the point-wise convolution learns to project the tensor to a high-dimensional space. Our goal for MobileViT is to depict non-local dependency over long distances using an effective $H*W$ receptive field. Dilated convolutions are a model for long-range interdependence that has been extensively studied. However, such techniques necessitate careful consideration of dilation rates. If this is not the case, the valid spatial region is ignored, and padding zeros are weighted [16]. Another viable alternative is self-attention [17, 18]. Vision transformers (ViTs) with multi-head self-attention provide an excellent way to focus on visual identification tests. Nonetheless, ViTs are cumbersome and have limited optimizability. This happens because there is no spatial inductive bias in ViTs. They partition X_L into N non-overlapping patches $X_U \in R^{P*N*D}$ so that MobileViT can obtain global models with a spatial inductive bias. Here, $P = wh$, N is the count of patches, and both h and w are integers less than or equal to n. $X_G \in R^{P*N*D}$ as a result of encoding inter-patch interactions using transformers for each p in the set $p \in \{1,\cdots,P\}$.

$$X_G(P) = Transformer\big(X_U(p)\big), 1 \le p \le P \qquad [1]$$

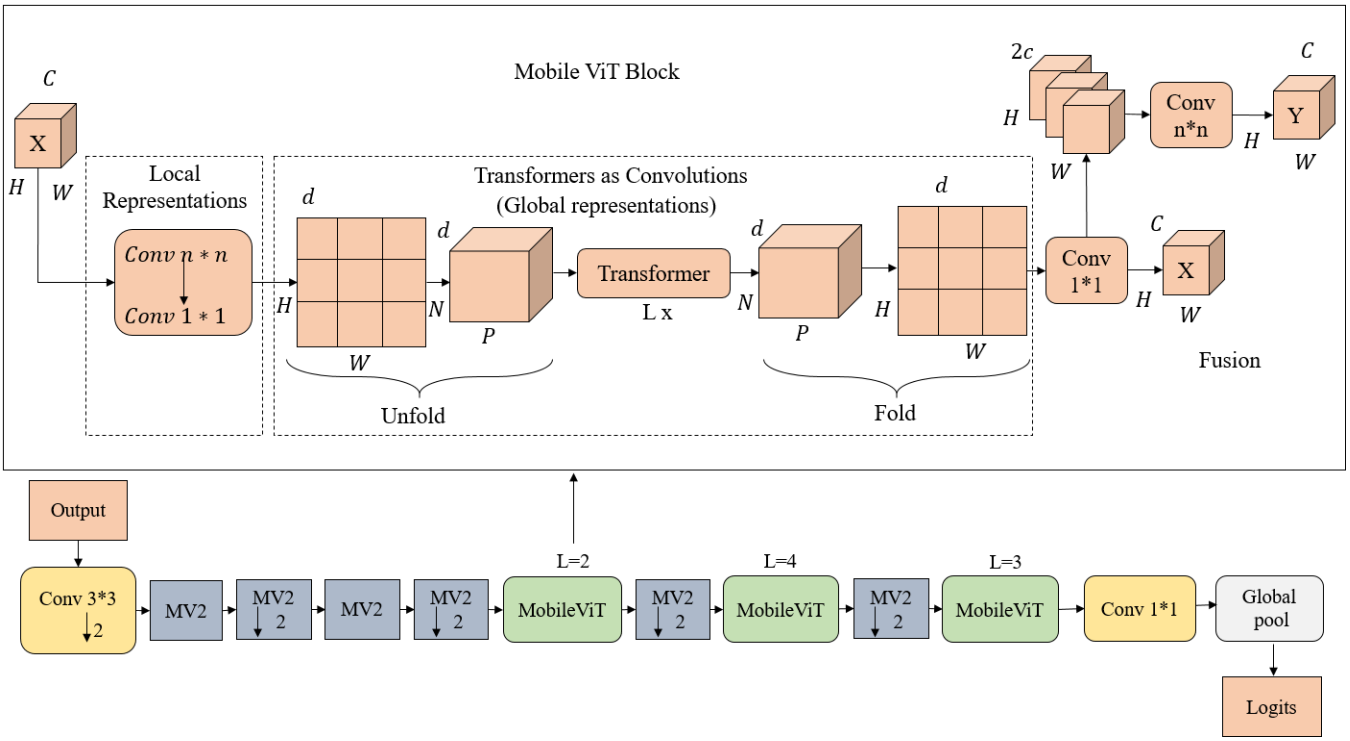


Fig. 3. MobileViT Architecture

When compared to ViTs, which alter the pixel spatial order, MobileViT preserves both the patch order and the pixel spatial order inside each patch. One way to achieve $X_F \in R^{H*W*D}$ is to fold $X_G \in R^{P*N*D}$. A point-wise convolution is used to project X_F to low C-dimensional space, and then it is combined with X. Then, an extra $n * n$ CL is utilized to apply these fused features. Bear in mind that $X_U(p)$ stores local information from an $n * n$ region using convolutions, while $X_G(p)$ utilizes patches to encode global information for the p-th position, so every pixel in X_G can encode from every pixel in X. Thus, $H * W$ is the total effective receptive field of MobileViT.

The three steps comprising a standard convolution are: unfolding, matrix multiplication (for acquiring local representations), and folding. Since they both utilize the same building blocks, MobileViT blocks are similar to convolutions. In convolutions, the MobileViT block employs a stack of transformer layers for deeper global processing instead of matrix multiplication, which is considered local processing. This causes MobileViT to exhibit characteristics (such as spatial bias) often associated with convolutions. Convolutions and transformers are two perspectives on the MobileViT block. Our purposely simplistic design has the advantage of providing effective low-level implementations of convolutions and transformers, making it simple to use MobileViT on a variety of devices. The MobileViT block utilizes transformers to acquire global representations, while regular convolutions are used for acquiring local representations. It is reasonable to question MobileViT's lightweight nature given that previous studies [19, 20] have shown that networks constructed using these layers are somewhat hefty. Learning global representations with transformers is the main issue, in our opinion. To convert spatial data into latent information for a specific patch, earlier studies [21, 22] utilized a linear combination of pixels. Subsequently, the inter-patch information is learned via transformers and used to encode the global information. Consequently, these models do not possess the inherent CNN-specific inductive bias. Therefore, children must have a stronger capacity to learn through visual means. This makes them broad and deep. By using a combination of convolutions and transformers, MobileViT can achieve global processing while still exhibiting features similar to convolutions, distinguishing it from existing models.

Our networks were influenced by the concept of lightweight CNNs. Three network sizes often used for mobile vision tasks—small (S), extra small (XS), and extreme extra tiny (XXS)—are used to train MobileViT models. The first layer of MobileViT is a strided $3 * 3$ standard convolution, while subsequent layers are derived from MobileNetV2 (or MV2) and MobileViT. Swish is used as the activation function. To keep up with the CNN models, we utilize $n = 3$

in the MobileViT block. Usually, the spatial dimensions of feature maps are multiples of 2, and h and w are less than or equal to n. Consequently, we achieve $h = w = 2$ on all spatial levels. In a MobileViT network, the MV2 blocks are responsible for downsampling the data. This is why MobileViT employs thin and shallow blocks. The geographical level-wise parameter distribution of MobileViT shows that MV2 blocks contribute almost nothing to the overall network parameters across all network topologies.

D. Expression Prediction – VGG16

In the ILSVRC-2014 competition, Simonyan and Zisserman's VGG-16 network structure achieved 92.7% accuracy on the ImageNet data sample [23]. While the VGG-16 network shares many similarities with other models that attained great accuracy, including AlexNet and LeNet, it has undergone extensive improvements [24]. When assigned assignments involving image recognition and larger datasets, this network shines. This approach was used to extract additional features from the facial expression dataset. This allows it to learn to reliably characterize facial expressions in a variety of profiles, tilts, and lighting conditions. As a result, it is ideal for identifying the emotions of online class participants sitting in front of cameras. This problem is addressed in the training dataset, which includes faces in a variety of situations and expressions.

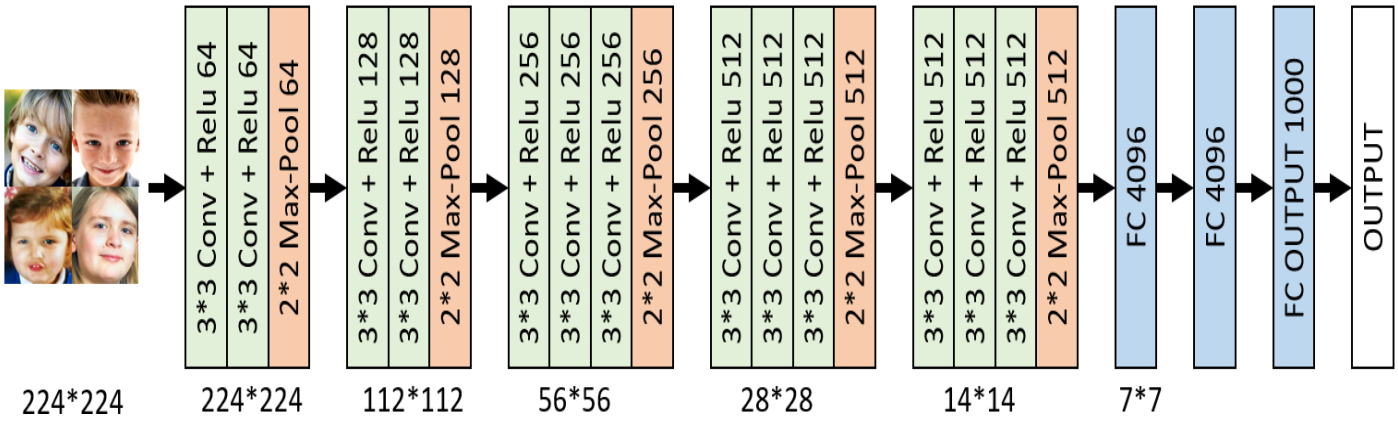


Fig. 4. VGG-16 Architecture

Figure 4 depicts the network design, which consists of thirteen CLs, three Fully Connected Layers (FCL), and five Pooling Layers (PL) [25]. The CLs employ a 3 × 3 kernel with a stride of one. The PL utilizes a 2 × 2 kernel with a 2 × 2 step size. CLs use the rectified linear unit (ReLU) as their activation function. The VGG-16 accepts 224 × 224-pixel images with three channels. The initial part consists of two CLs and one PL. These CLs have 64 kernels.

The PL reshapes the input image to 112 × 112. The next section consists of two CLs, each with 128 convolution kernels. The image is reduced to 56 × 56 pixels using a PL. Following that, there will be three CLs and a PL. The PL, which comes after the CLs, decreases the image dimension to 28 × 28 while increasing the number of kernels to 256. Then the three CLs increase the depth to 512 kernels and reshape the given image to 14 × 14 dimensions using a PL. In addition, the fifth set includes three 512-depth CLs. The last PL reshapes the dimension to 7 × 7. Two FCLs, each with 4096 nodes, complete the structure. To complete their design, the output layer employs a 1000-unit SoftMax, which represents the 1000 classes present in the ImageNet dataset. In our research, we convert this to 3 for identifying expressions of autistic children.

Due to hardware limitations, TL is used to train a CNN. TL uses the pre-trained VGG-16 network's ImageNet weighted layers to train on a different dataset. Nonetheless, the required bottleneck features are not created since the layer weights are not changed. To obtain bottleneck features, the VGG-16 architecture's fully connected and classification layers were removed, leaving only the first five sets of CLs and PLs. A NumPy array holds the bottleneck's properties.

The training sequential model consists of five DeLs, an overfitting adjustment dropout layer, and a batch normalization layer [26]. The first four DeLs include 512, 256, 128, and 64 ReLUs. Equation (1) represents the ReLU function:

$$a = ReLUb = b; if\ b \geq 0\ 0;\ if\ b < 0 \quad [2]$$

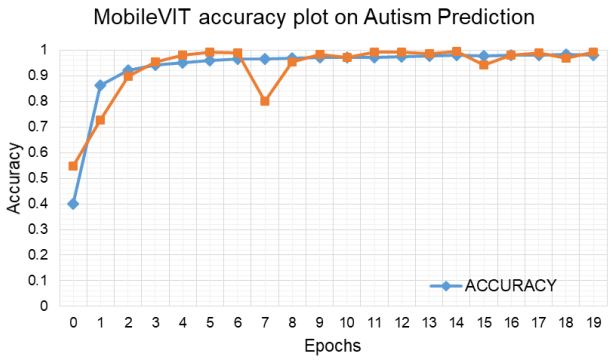
Here, the output and input signals are represented by "a" and "b". For negative input signals, the resultant signal is zero. Thus, for values less than or equal to zero, the slope of "b" (the input signal) remains constant. Because of this, ReLU can improve its training performance while avoiding the vanishing gradient problem. Instead of ReLU, the output layer, the fifth DeL, has a SoftMax function with six categories. This layer will provide a unique probability value for each of the six different categories.

4. EXPERIMENTAL OUTCOME

The processed image is given to MobileViT and VGG-16 for ASD and expression prediction. First, the MobileViT model is trained using images labelled with ASD and normal classes. The training phase exhibits a notable decrease in loss value from 1.22 to 0.054, coupled with an increase in accuracy from 0.39 to 0.98. This significant improvement underscores the model's efficiency. Across epochs 0 to 4, the model's performance shows gradual enhancement, characterized by rising accuracy and declining loss. Subsequently, the model's performance stabilizes around a value of 0.96 and above. Furthermore, the model's accuracy in checking new data improves significantly from 0.84 at the 0th epoch to 0.99 at the 20th epoch. Concurrently, over the final epoch, the validation loss steadily decreases, reaching a minimal value of 0.02. For a visual representation of the MobileViT model's performance during training and validation, Figure 5 provides a snapshot, while Figures 6 and 7 illustrate the accuracy and loss plots, respectively.

```
Found 612 images belonging to 2 classes.
Found 612 images belonging to 2 classes.
C:\Users\Admin\AppData\Local\Temp\ipykernel_2400\2159821025.py:35: UserWarning: `Model.fit_generator` is deprecated and will be removed in a future version. Please use
`MobileVITModel.fit`, which supports generators. r = model.fit_generator(
Epoch 1/20 20/20 [==============================] - 528s 27s/step - loss: 1.2298 - accuracy: 0.6007 - val_loss: 0.3958 - val_accuracy: 0.8473
Epoch 2/20 20/20 [==============================] - 488s 25s/step - loss: 0.3958 - accuracy: 0.8626 - val_loss: 0.2114 - val_accuracy: 0.9282
Epoch 3/20 20/20 [==============================] - 488s 25s/step - loss: 0.2339 - accuracy: 0.9203 - val_loss: 0.2549 - val_accuracy: 0.8988
Epoch 4/20 20/20 [==============================] - 480s 25s/step - loss: 0.1719 - accuracy: 0.9409 - val_loss: 0.1217 - val_accuracy: 0.9532
Epoch 5/20 20/20 [==============================] - 516s 26s/step - loss: 0.1405 - accuracy: 0.9522 - val_loss: 0.0669 - val_accuracy: 0.9805
```

Fig. 5. Screenshot of MobileViT training and validation phase



1.

Fig. 6. Accuracy plot of MobileViT

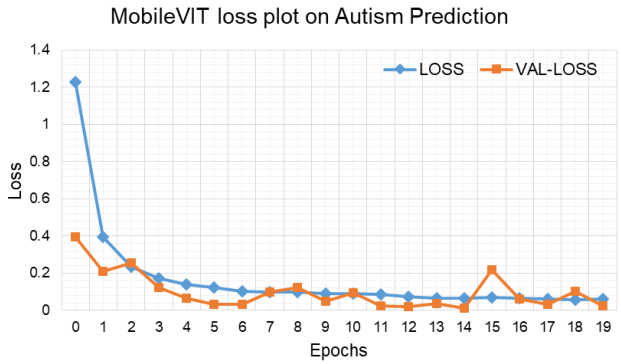


Fig. 7. Loss plot of MobileViT

The processed image is given to VGG-16 for training and validation. We use the categorical entropy loss and the Adam optimizer, running for 20 epochs. The training phase's loss value decreases from 7.4915 to 0.4027 in the first four epochs, enhancing accuracy from 0.45 to 0.85. This demonstrates the model's capability to learn and adapt based on input data. From epoch 5 to epoch 9, the model's performance raised gradually, with increasing accuracy and decreasing loss. Achieving a high accuracy of 0.89 is a noteworthy accomplishment. In training epochs 10-14, the model's accuracy improves to over 0.90. Training accuracy values peak at 0.96 in later epochs 15-19, indicating that the model successfully identifies the dataset's underlying patterns. The model's validation accuracy on unknown data improves from 0.61 at the initial epoch to an impressive 0.99 at the end of the 20th epoch. Over the last epoch, the validation loss steadily decreases until it reaches a low of 0.0311. Figure 8 depicts a snapshot of the VGG-16 model performance during training and validation. Figures 9 and 10 show VGG-16 accuracy and loss plots.

```
Found 612 images belonging to 3 classes.
Found 612 images belonging to 3 classes.
C:\Users\Admin\AppData\Local\Temp\ipykernel_2400\2159821025.py:35: UserWarning: `Model.fit_generator` is deprecated and will be removed in a future version. Please use
`VGGModel.fit`, which supports generators. r = model.fit_generator(
Epoch 1/20 20/20 [==============================] - 528s 27s/step - loss: 7.4915 - accuracy: 0.4526 - val_loss: 1.2146 - val_accuracy: 0.6176
Epoch 2/20 20/20 [==============================] - 488s 25s/step - loss: 1.2705 - accuracy: 0.6748 - val_loss: 0.5648 - val_accuracy: 0.8105
Epoch 3/20 20/20 [==============================] - 488s 25s/step - loss: 0.7003 - accuracy: 0.7614 - val_loss: 0.3836 - val_accuracy: 0.8513
Epoch 4/20 20/20 [==============================] - 480s 25s/step - loss: 0.6130 - accuracy: 0.7712 - val_loss: 0.2932 - val_accuracy: 0.8922
Epoch 5/20 20/20 [==============================] - 516s 26s/step - loss: 0.4027 - accuracy: 0.8562 - val_loss: 0.2983 - val_accuracy: 0.8889
```

Fig. 8. Screenshot of VGG-16 training and validation phase

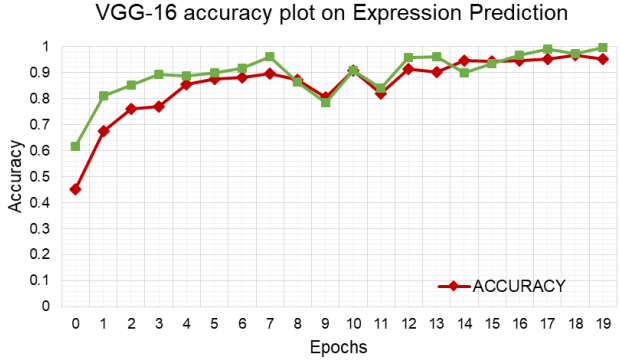


Fig. 9. Accuracy plot of VGG-16

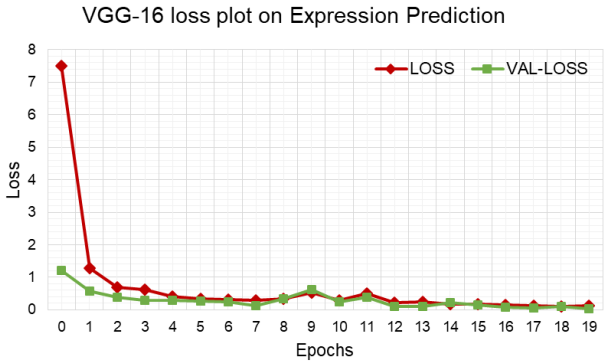
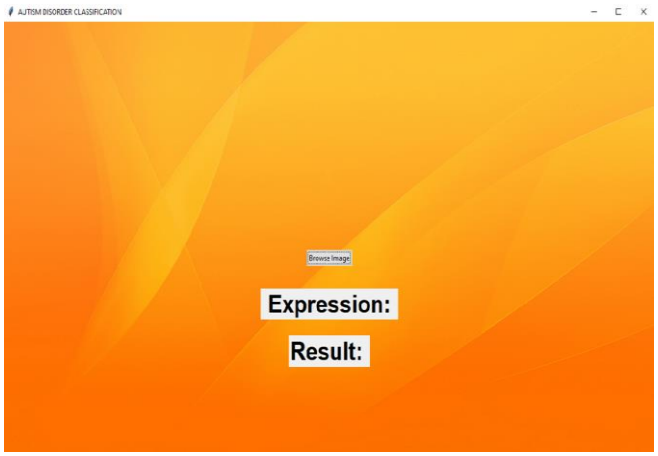


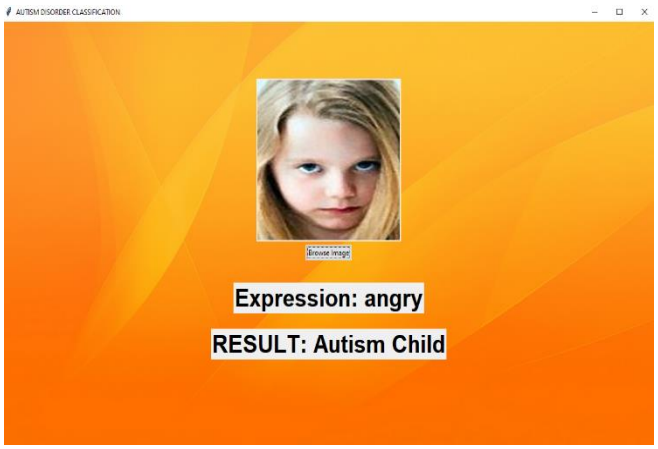
Fig. 10. Loss plot of VGG-16

The training and validation results play a crucial role in finalizing the deployment model. In this research, we have developed a website that is hosted locally. The website is designed in a user-friendly manner, featuring a simple interface with only an upload option. Users can easily upload an image by clicking the browse button. Once the upload is complete, the website provides information on whether the child is affected by ASD or not. Additionally, the facial expression associated with the prediction is displayed.

The home page of the developed website, dedicated to ASD prediction, is presented in Figure 11.a, showcasing the user interface. Subsequently, Figure 11.b illustrates the predicted result, providing users with a clear indication of the model's assessment.



a)



b)

Fig. 11. Working of developed website

5. CONCLUSION

Early diagnosis is crucial for providing effective treatment, especially considering the relatively low prevalence of autism in children. The current diagnostic methods may contribute to this low detection rate. Many DL models were developed for ASD detection, often focusing solely on diagnosis. Emotion recognition, especially, is overlooked despite its pivotal role in supporting the emotional development of children with autism. Additionally, existing systems may struggle to scale effectively due to manual and expertise-dependent processes, hindering their ability to meet the increasing demand for autism assessment and diagnosis. To address these challenges, we've designed an effective DL model deployed on a website capable of predicting both ASD and children's emotions. Our classifiers, MobileViT and VGG-16, achieved remarkable accuracies of 99.8% and 99.51% on ASD and expression prediction. This demonstrates their ability to accurately identify children's mental conditions and facial expressions. Preprocessing methods such as resizing, rescaling, and augmentation were used to improve model performance, and this could potentially increase accuracy even further. We created a user-friendly website that utilizes the DL model to diagnose children's expressions and autism conditions. This research has important implications for real-time ASD screening, with the potential to change the diagnostic process. Future research could explore integrating image and video approaches to create a comprehensive solution for detecting both behavioral and facial phenotype distinctions in ASD.

REFERENCES

[1] Mukherjee, Sharmila Banerjee. "Autism spectrum disorders—diagnosis and management." The Indian Journal of Pediatrics 84 (2017): 307-314.

[2] Hashem, Sheema, Sabah Nisar, Ajaz A. Bhat, Santosh Kumar Yadav, Muhammad Waqar Azeem, Puneet Bagga, Khalid Fakhro, Ravinder Reddy, Michael P. Frenneaux, and Mohammad Haris. "Genetics of structural and functional brain changes in autism spectrum disorder." Translational psychiatry 10, no. 1 (2020): 229.

[3] Maenner, Matthew J., Zachary Warren, Ashley Robinson Williams, Esther Amoakohene, Amanda V. Bakian, Deborah A. Bilder, Maureen S. Durkin et al. "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2020." MMWR Surveillance Summaries 72, no. 2 (2023): 1.

[4] Zwaigenbaum, Lonnie, Susan Bryson, Tracey Rogers, Wendy Roberts, Jessica Brian, and Peter Szatmari. "Behavioral manifestations of autism in the first year of life." International journal of developmental neuroscience 23, no. 2-3 (2005): 143-152.

[5] Charman, Tony, and Katherine Gotham. "Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders–lessons from research and practise." Child and adolescent mental health 18, no. 1 (2013): 52-63.

[6] Raj, Suman, and Sarfaraz Masood. "Analysis and detection of autism spectrum disorder using machine learning techniques." Procedia Computer Science 167 (2020): 994-1004.

[7] Alam, Mohammad Shafiul, Muhammad Mahbubur Rashid, Ahmed Rimaz Faizabadi, Hasan Firdaus Mohd Zaki, Tasfiq E. Alam, Md Shahin Ali, Kishor Datta Gupta, and Md Manjurul Ahsan. "Efficient Deep Learning-Based Data-Centric Approach for Autism Spectrum Disorder Diagnosis from Facial Images Using Explainable AI." Technologies 11, no. 5 (2023): 115.

[8] Omar, Kazi Shahrukh, Prodipta Mondal, Nabila Shahnaz Khan, Md Rezaul Karim Rizvi, and Md Nazrul Islam. "A machine learning approach to predict autism spectrum disorder." In 2019 International conference on electrical, computer and communication engineering (ECCE), pp. 1-6. IEEE, 2019.

[9] Liao, Mengyi, Hengyao Duan, and Guangshuai Wang. "Application of machine learning techniques to detect the children with autism spectrum disorder." Journal of Healthcare Engineering 2022 (2022).

[10] Qureshi, Muhammad Shuaib, Muhammad Bilal Qureshi, Junaid Asghar, Fatima Alam, and Ayman Aljarbouh. "Prediction and analysis of autism spectrum disorder using machine learning techniques." Journal of healthcare engineering 2023 (2023).

[11] Elshoky, Basma Ramdan Gamal, Eman MG Younis, Abdelmgeid Amin Ali, and Osman Ali Sadek Ibrahim. "Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images." ETRI Journal 44, no. 4 (2022): 613-623.

[12] Tye, Charlotte, Giorgia Bussu, Teodora Gliga, Mayada Elsabbagh, Greg Pasco, Kristinn Johnsen, Tony Charman, Emily JH Jones, Jan Buitelaar, and Mark H. Johnson. "Understanding the nature of face processing in early autism: a prospective study." Journal of Psychopathology and Clinical Science 131, no. 6 (2022): 542.

[13] Hosseini, Mohammad-Parsa, Madison Beary, Alex Hadsell, Ryan Messersmith, and Hamid Soltanian-Zadeh. "Deep learning for autism diagnosis and facial analysis in children." Frontiers in Computational Neuroscience 15 (2022): 789998.

[14] https://www.sendspace.com/file/6v2vnx

[15] Wang, Qi, and Yuan Yuan. "Learning to resize image." Neurocomputing 131 (2014): 357-367.

[16] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).

[17] Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He. "Non-local neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803. 2018.

[18] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[19] Howard, Andrew, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang et al. "Searching for mobilenetv3." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 1314-1324. 2019.

[20] Mehta, Sachin, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. "Delight: Deep and light-weight transformer." arXiv preprint arXiv:2008.00623 (2020).

[21] Touvron, Hugo, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. "Training data-efficient image transformers & distillation through attention." In International conference on machine learning, pp. 10347-10357. PMLR, 2021.

[22] Graham, Benjamin, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. "Levit: a vision transformer in convnet's clothing for faster inference." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 12259-12269. 2021.

[23] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[24] Naseer, Iftikhar, Sheeraz Akram, Tehreem Masood, Arfan Jaffar, Muhammad Adnan Khan, and Amir Mosavi. "Performance analysis of state-of-the-art cnn architectures for luna16." Sensors 22, no. 12 (2022): 4426.

[25] Tammina, Srikanth. "Transfer learning using vgg-16 with deep convolutional neural network for classifying images." International Journal of Scientific and Research Publications (IJSRP) 9, no. 10 (2019): 143-150.

[26] Santos, Claudio Filipi Gonçalves Dos, and João Paulo Papa. "Avoiding overfitting: A survey on regularization methods for convolutional neural networks." ACM Computing Surveys (CSUR) 54, no. 10s (2022): 1-25.