

Autism Spectrum Disorder Equipped with Convolutional-cum-visual Attention Mechanism

Ayesha Shaik*

Center for Cyber Physical Systems, Vellore Institute of Technology, Chennai, 600127, India

E-mail: ayasha.sk@vit.ac.in

ORCID iD: <https://orcid.org/0000-0002-9804-8031>

*Corresponding author

Lavish R. Jain

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, 600127, India

E-mail: lavishjain307@gmail.com

ORCID iD: <https://orcid.org/0009-0000-7584-4062>

Balasundaram A.

Center for Cyber Physical Systems, Vellore Institute of Technology, Chennai, 600127, India

E-mail: balasundaram.a@vit.ac.in

ORCID iD: <https://orcid.org/0000-0001-9746-6357>

Received: 20 May 2025; Revised: 08 July 2025; Accepted: 22 August 2025; Published: 08 October 2025

Abstract: This research work aims to utilize deep learning techniques to identify autism traits in children based on their facial features. By combining traditional convolutional approaches with attention layers, the study seeks to enhance interpretability and accuracy in identifying autism spectrum disorder (ASD) traits. The dataset includes diverse facial images of children diagnosed with ASD and neuro-typical children, ensuring comprehensive representation. Preprocessing techniques standardize and enhance image quality, mitigating biases. Integration of attention layers within the convolutional neural net-work (CNN) architecture focuses on crucial facial features, improving feature extraction and classification accuracy. This approach enhances model interpretability through eXplainable AI (XAI) techniques. Model training involves optimization and validation processes, employing hyper parameter tuning and cross-validation for robustness. The performance of this combined model yielded close to 95% accuracy outperforming existing models in terms of complexity to accuracy ratio.

Index Terms: Explainable AI, Autism Spectrum Disorder, Convolutional Neural Network, Visual Attention, Deep Learning.

1. Introduction

Autism Spectrum Disorder (ASD) affects brain function at many levels and is recognized to have a strong genetic basis. The World Health Organization estimates that there are about 67 million people affected by this condition worldwide. It leads to difficulties with behavior, communication skills, and social interaction because it presents with such a wide range of symptoms and challenges. “Spectrum” refers to the fact that there is great variation in the signs shown or impairments felt, from mild cases of autism where individuals may only need some help in life areas, through severe ones requiring constant support.

Other forms involve abnormal social communication patterns, repetitive actions as well as in-tense sensitivities towards certain things among others like repetitive actions such as rocking back forth obsessively or staring intently at objects for hours on end. Some persons may therefore not be able tell when someone else is joking while another might keep flapping their hands whenever they get excited.

Thus an important point about Autism Spectrum Disorders (ASDs) is that they are actually different types of neurological disorders that can affect people differently depending on each individual’s unique neurology. As a result, screening for ASDs tends to be time-consuming and demanding in terms of resources due mainly but not only limited too medical professionals having to conduct extensive evaluations before making any conclusions.

Technological developments in artificial intelligence have the potential to benefit the detection and recognition of ASD. In children diagnosed with autism, their speech may be delayed, their voice and facial expressions may not follow normal patterns, they may repeat actions or words and they might have difficulties with communication. Usually appearing before a person turns three, these signs continue throughout life. Furthermore, ASD can be inherited from one or both parents as a neurodevelopmental abnormality; apart from that it can also occur sporadically. Males are disproportionately affected by autism spectrum disorder when compared against females statistically.

Early recognition of ASD is vital for managing the condition effectively and reducing its impact on people's lives. To evaluate autistic individuals' behaviors and symptoms health care providers usually employ standard diagnostic tools such as ADOS (The Autism Diagnostic Observation Schedule) or ADI-R (Autism Diagnostic Interview-Revised) [1]. Moreover, diagnosis has been supported by computer-aided additional testing like CT scans (computed tomography), MRI scans (magnetic resonance imaging) and EEGs (electroencephalograms).

Nevertheless, a few diagnostic tools have their own considerations and challenges. For example, ionizing radiation is exposed to patients during CT scans that may pose potential health risks. For this reason, there should be careful evaluation on the use of CT imaging in ASD diagnosis with regards to its hazards and benefits [2]. Additionally, MRI scans can be expensive and time-consuming especially when children have to be still or under anesthesia for an accurate image.

While acknowledging the genetic basis of ASD, researchers are still striving towards finding out which genes or sets of genes together with gene mutations cause autism specifically. This study is aimed at enhancing accuracy as well as precision of ASD diagnosis hence leading to more effective interventions and support systems for individuals living with autism spectrum disorders.

ASD detection poses numerous very noticeable problems which must be dealt with in order for effective and dependable diagnostic tools to be developed. One of the main challenges is the fact that ASD symptoms vary so widely and are so complex that they can look entirely different from one person to another. It is difficult to establish a standardized set of criteria or features for diagnosing autism because of this diversity; therefore, some children may not get diagnosed correctly or at all until much later than they should have been. Another problem lies with traditional diagnostic methods' dependence on subjective observations as well as behavioral assessments, which introduces biases into the process of making diagnoses consistently among clinicians. Moreover, there is a lack of resources and expertise for diagnosis and treatment related to autism spectrum disorders in certain underserved communities or regions where such services are rare if not nonexistent altogether; thus early detection may take longer than necessary leading up to poorer outcomes throughout life quality among individuals affected by ASDs. Furthermore, logistical issues arise with integrating new technologies like deep learning and facial image analysis into clinical practice because questions surrounding data privacy, algorithmic fairness (bias), or AI's ability to explain itself when used as an aide for medical decision-making remain unanswered at this time — especially concerning these areas within healthcare that are already fraught with ethical debates over patient rights versus societal benefits from using them widely. To conquer these tests, we need collaboration between those in the field of medicine, scientists, decision-makers and other concerned parties in order to come up with all-around ASD diagnosis and management methods that are accurate, accessible and just.

In detecting Autism Spectrum Disorder (ASD) deep learning has proved useful especially when applying facial image analysis. Atypical facial expressions or subtle cues can be detected as being related to ASD by deep learning algorithms such as convolutional neural networks (CNNs) which effectively extracts necessary features from facial images while recognizing patterns indicative of the condition.

Through the use of deep learning methods, it is possible for scientists to create automatic systems that can correctly determine whether someone is autistic or not based on their face alone. This method has various benefits one of them being better precision than traditional diagnosis techniques which rely on behavioral assessments and subjective observations hence making it consistent as well as scalable.

Additionally, deep learning models can be educated using huge datasets of pictures of faces containing all sorts of people and conditions to make sure that they are strong enough to work with any-one. This permits the creation of more precise and inclusive ASD detection tools which can help health care workers with their interventions at an early stage as well as provide support for individuals suffering from Autism Spectrum Disorder (ASD) along with their families.

Artificial intelligence has great potential in detecting ASDs as a whole and diagnosing them individually too. The ultimate goal is creating fast, fair-minded, widely available tools for early intervention which would make life better not only for children but also adults living on the autism spectrum.

2. Literature Survey

Autism spectrum disorder (ASD) is a complex condition characterized by repetitive behaviors and difficulties in social communication, often influenced by genetics. Many individuals with ASD find it challenging to live independently or maintain steady employment. While genetics and neuroscience have uncovered interesting risk patterns associated with ASD, practical applications of this knowledge remain limited. Ongoing research aims to identify which children, including those with additional health conditions, could benefit from behavioral and medical treatments, how these interventions can be tailored, and when they should be implemented [3]. Equally important is the development of comprehensive services for adults with ASD to support them through various life stages.

Clinicians play a vital role in this process by anticipating transitions such as changes in family dynamics and school transitions, providing accurate information, and offering timely support to individuals and families affected by ASD. Understanding the unique challenges faced by individuals with ASD at different ages is crucial for clinicians to provide effective assistance and improve their quality of life.

In a recent study focusing on the language aspects of ASD, the condition is characterized by ongoing difficulties in social interaction and communication. One particular linguistic behavior linked to ASD is echolalia, which involves repeating speech patterns. A study explored echolalia types and communication abilities in seven young boys with ASD who used oral communication [4]. Through playful therapy sessions and input from parents and professionals, the study observed compromised communication skills and increased immediate echolalia in children with ASD.

Furthermore, a research analyzed the changes of bad temper, excessive movement and social isolation in children with ASD as well as in their peers who suffer from developmental delays. The study indicated that the kids with more severe symptoms of autism represent higher levels of irritability and activity. Over time there was some reduction in restlessness but it coincided with increasing age when many individuals diagnosed with Autism Spectrum Disorder become even more withdrawn from society than before thus indicating wide-ranging patterns of conduct among persons identified with this condition [5].

Having a child diagnosed with Autism Spectrum Disorder (ASD) can be overwhelming for families and bring about significant disruptions to daily routines. Presently, it is estimated that 1 in every 66 Canadian children aged between five and seventeen years old has ASD which means that general practitioners' offices family physicians clinics medical settings will increasingly encounter such cases [6]. Early intervention is important because interventions at this point are likely to yield better long-term outcomes given these challenges. Therefore, providing evidence-based recommendations alongside tools would help community pediatrician's general practitioners primary care providers recognize early signs make accurate diagnoses plan interventions more effectively when dealing with Autism Spectrum Disorders.

To find out if a child has ASD, the American Academy of Pediatrics (AAP) advises parents to have their child screened at 15, 30, and 9 months old. Not pointing or waving goodbye by the first birthday could be among the earliest signs that a toddler might have autism. Other red flags include poor eye contact, delayed speech and socialization skills as well as limited interest in others or in sharing experiences. The Modified Checklist for Autism in Toddlers (M-CHAT) is one screening tool that can help identify this condition among young children; another is called "CHAT." Children who show any unusual symptoms such as repetitive behaviors like flapping hands should also be evaluated for possible autistic spectrum disorders during early childhood development programs where they may demonstrate restricted pretend play or odd interests along with rigid routines.

Children with autism may lack social skills and are often unable to communicate effectively with their peers. They may not understand how people think or feel which makes it difficult for them to make friends even when they want to. They are usually very good at following rules though and so teachers sometimes don't notice anything different about these kids until they start asking questions like "Why did you do that?" or "How come I can't go?" This concrete thinking combined with an interest in peers but poor social interaction skills should raise suspicions of ASD in older children. Screening tools that might be helpful here include SCQ (social communication questionnaire), SRS (social responsiveness scale) and ASSQ (autism spectrum screening questionnaire) [7].

Ways to identify ASD symptoms can be costly and time-consuming. The diagnostic process has been transformed by the combination of artificial intelligence (AI) with machine learning (ML) which allows for early detection and intervention strategies improvement for ASD. For this reason, Support Vector Machine (SVM), Decision Trees, Linear Discriminant Analysis and Logistic Regression among other machine learning algorithms are increasingly being used for ASD identification which may lead better outcomes in individuals [8].

Recent studies have suggested that EEG data collection should be integrated with eye-tracking tests when diagnosing ASD especially among young children between three to six years old. By using SVM classifiers and feature selection methods that are advanced in nature, it has been possible for researchers to achieve a high level of accuracy when differentiating between kids with autism from those who are typically developing. Therefore this new approach based on machine learning which combines both eye tracking data together with EEG holds promise as an effective means of identifying autism spectrum disorder while directing diagnosis [9].

The research done in [10] is about using resting-state functional magnetic resonance imaging (fMRI) data to detect individuals with autism spectrum disorder (ASD), through deep belief networks (DBNs). This research presents some advanced machine learning techniques to deal with ASD diagnosis complexities.

They chose deep belief network as their model; this kind of artificial neural network uses unsupervised learning in feature extraction and supervised learning for classification purposes. What they mainly wanted was to find a way of tapping into the huge amounts of information that are contained in resting-state fMRI data so as to accurately identify people who have been affected by ASD.

It details the methodology adopted, encompassing data preprocessing protocols, feature extraction methodologies, and the intricate architecture of the deep belief network utilized. To validate their approach, the researchers utilized a comprehensive dataset containing resting-state fMRI scans obtained from individuals both with and without ASD, facilitating the training and rigorous evaluation of the deep belief network model.

The findings presented in the paper serve as a testament to the efficacy of the deep belief network in effectively discerning individuals with ASD from neurotypical individuals based solely on resting-state fMRI data. The model exhibited remarkable accuracy and reliability in ASD classification, underscoring the potential of advanced machine

learning techniques as indispensable aids in the realm of ASD diagnosis.

Another approach employs functional connectivity analysis on fMRI data, which measures the temporal correlation between different brain regions, to capture patterns associated with ASD [11]. They then applied feature selection techniques to identify the most relevant and discriminative features from the fMRI data. These selected features were used as input to deep learning algorithms, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), for classification tasks.

The classification patterns in the brain imaging series that underscored an intriguing anti-correlation of brain function observed between the anterior and posterior brain regions [12]. This discovered anti-correlation holds significant implications as it aligns with existing empirical data highlighting anterior-posterior disruptions in brain connectivity observed in individuals with ASD. By harnessing deep learning methodologies, the extract shows intricate neural signatures and discern unique brain activation patterns characteristic of ASD, contributing to a deeper understanding of the underlying neurobiological mechanisms associated with the disorder.

The identification of autism spectrum disorder (ASD) has been revolutionized with the integration of social media data and biological images, providing valuable insights into the mental health condition. ASD's impact on facial appearance, intricately linked with brain development, has led researchers to explore distinct facial landmarks in teenagers with ASD compared to typically developing children (TD). Leveraging advanced deep learning techniques within the Flask framework and employing convolutional neural networks (CNNs) with transfer learning, this research delves into the effectiveness of models like Xception, Visual Geometry Group Network (VGG19), and NASNETMobile in addressing the complexities of ASD categorization [13].

In clinical settings, physicians rely on a comprehensive assessment that includes behavioral observations and developmental histories to accurately diagnose ASD. Facial features, serving as potential biomarkers reflecting underlying brain conditions, offer a convenient avenue for early diagnosis. This study capitalizes on the power of facial imagery to identify children with autism, employing a range of deep CNN-based transfer learning techniques [14]. Notably, the enhanced Xception model emerged as the top performer, achieving an impressive 95% accuracy on the test dataset post-training and validation with meticulously optimized configurations. Complementing this, models like ResNet50V2, MobileNetV2, and EfficientNetB0 demonstrated accuracies of 94%, 92%, and 85.8%, respectively, showcasing the superiority of transfer learning strategies over traditional methods.

A fast and reliable method for predicting the autism spectrum disorder in the adults using Random Forest Regression is discussed in [15]. The machine learning based Random Forest-CART (Classification and Regression Trees) and Random Forest-Id3(Iterative Dichotomiser 3) autism prediction model is proposed in [16] and also a mobile application also developed for predicting autism traits for different age groups. The performance of single classifiers has been explored and then ensembled to improve the performance for predicting the autism spectrum disorder is presented in [17]. Deep learning [18] and basics of machine learning [19, 20] based models that helps to predict the autism traits are discussed. Facial expression dependent autism spectrum disorder prediction model is presented in [21].

Augmenting the model's capabilities, a Convolutional Neural Network (CNN) leveraging the MobileNet architecture was implemented along-side advanced transfer learning methodologies to bolster recognition tasks related to ASD [22]. The experimental results were notable, showcasing a validation accuracy of 89% and a test accuracy of 87%, underscoring the efficacy of the MobileNet model in conjunction with transfer learning approaches. Moreover, achieving maximum scores for precision value and F1-score at 87% further solidified the reliability and robustness of the proposed recognition methodology. These findings highlight the transformative potential of deep learning techniques fused with transfer learning paradigms in revolutionizing ASD identification and intervention strategies.

3. Proposed Methodology

Simplifying images is a tough task that is evolving every day. Convolutional layers are the face of image processing right now. In our approach we chose VGG16 that won the ILSVR (Imagenet) competition in 2014. VGG16 serves as the back-bone network for the model. Though the current research trend focusses on results, we strongly believe that validation is of utmost importance for any model. The feature maps generated at every stage serves as a purpose to define the characteristics of the image that are taken into account to derive results. To provide a visualized approach for the same, we have introduced attention layers to visualize the intermediate underlying patterns.

To ease our implementation, we have used the prominent technique of transfer learning [23] where VGG16 that has been trained on imagenet dataset serves as the initial weights for further processing. In our base model, we have removed all the dense layers and also removed dropout layers to introduce few attention layers.

VGG16 is divided into 5 pooling layers, with each pooling layer's output is passed on to the next pooling layer. These intermediate feature vectors are passed on to the attention layers to visualize them. Pool-1 and Pool-2 starts to extract the important features, the main focus shifts to-wards pool-3 and pool-4 layers. Pool-3 layer starts to generalize while pool-4 layer starts to concentrate. Pool-5 layer has the most concentrated feature extracted information. So, that serves as a reference layer view point for other layers so as to improve the visualization.

Every pool layer is max-pooled. After max-pooling, pool-5 has a size of 7X7X512 whereas pool-3 and pool-4 has 28X28X256 and 14X14X512. To align the spatial size, we use bi-linear interpolation.

$$O = W * \text{ReLU}(Wf * F + \text{upFactor}(Wg * P5)) \quad (1)$$

W is a convolutional kernel with 256 filter in-put size that outputs a single channel response at the end. The upFactor tries to align the spatial size of the reference and desired pool using bipolar interpolation. $P5$ is the final pool-5 output vector whereas Wf and Wg are the convolutional kernel with 256 filters.

The attention map is then normalized using a sigmoid function.

$$Ai = \text{Sigmoid}(O) \quad (2)$$

The final representation is the dot product of the attention map with the feature vector F . We use a softmax classification at the end to form the final feature.

$$F = Ai.Fi \quad (3)$$

The concatenation at every step is done using the global average pooling between the feature vectors and then the model is trained end-to-end.

3.1. Data Preparation and Preprocessing

The research concentrates on determining autism in children by examining pictures of their faces. A dataset with about 2940 images, conveniently available on Kaggle [24], is a helpful source for scientists and doctors. This data set falls into two main groups: images of kids with autism and with-out it. In order to make sure that the findings are dependable, the researchers randomly divide this dataset into training and testing sets according to an allocation rate of 80%:20%. Such split helps thoroughly validate models as well as assess their performance thus making them strong methods for detecting autism based on facial expressions vali-datable in terms of trustworthiness and effective-ness.

For the purpose of implementation, many pre-processing techniques were performed to enhance the quality and relevance of the images. In the be-ginning, what was done is normalizing pixel intensities so as to standardize input data hence removing any disparities caused by lighting conditions or camera settings alterations. Also, image resizing and cropping methods were used in order to keep image dimensions constant which helped more efficient computation when training models. Still, data augmentation strategies that involve rotation, translation and flipping were applied to artificially increase the dataset thereby making it more di-verse during training thus improving generalization ability of models. Through these pre-processing techniques, dataset was made ready for further analysis.

The dataset obtained from the Kaggle source site [24] contains around 2,940 images. To verify the accuracy and the adolescent intentions of the dataset, the authors looked at age and few ethnic and gender proportions. It is important to note that the dataset mostly includes children aged between 3 and 12 years and tilted more towards males due to a higher likelihood of an autism diagnosis in males. However, this could pose a problem as there are some underrepresented racial groups which could bring about biases and narrow generalizability. Given these limitations, the authors describe potential biases that could affect the proposed work and suggest avenues for future research in order to include a more diverse sample and thus increase the cross-cultural applicability of the model.

3.2. Network Training

In our methodology, we utilize the powerful PyTorch framework [25] to implement our approach. To begin, we prime our backbone network with the renowned VGG-16 architecture, leveraging the technique of transfer learning [4]. This involves utilizing the pre-trained weights from the ImageNet dataset, a vast repository of diverse im-ages spanning numerous categories. By leveraging transfer learning, we harness the knowledge and feature representations learned by VGG-16 on ImageNet, which provides a significant head start in training our model for our specific task. Furthermore, we initialize attention modules within the network, ensuring that they are optimally poised to extract salient features from the input data.

The VGG16 model is selected for the proposed work as a backbone due to its balance between interpretability and computational efficiency. The architecture for VGG-16 model is given in Fig. 1. While recent models like ResNet, DenseNet, and Transformer-based architectures offer potential advantages, the authors found that VGG16's simplicity and proven performance in image classification tasks align well with our study's focus. Additionally, VGG16 is more compatible with the attention mechanisms used in the proposed approach, enhancing the model's ability to highlight critical facial features associated with autism. Future studies could explore these alternatives to evaluate whether their structural advantages yield improvements specific to autism classification.

The training process itself is conducted in an end-to-end manner, allowing the entire network to adjust its parameters collectively to minimize the defined loss function. Despite the complexity of our model and the richness of the dataset, we opt for a relatively short training duration of only 3 epochs. This decision is made with careful consideration to balance computational resources and training effectiveness. Throughout the training phase, we employ stochastic gradient descent with momentum as the optimization algorithm, a popular choice known for its effectiveness in navigating complex optimization landscapes. The initial learning rate is set at 0.01, providing a suitable starting point for the optimization process, which may be adjusted dynamically as training progresses.

3.3. Model Evaluation

To test the performance of our model, an evaluation metric called ROC curve is used along with a test dataset. Besides, we select focal loss [26] instead of binary cross entropy [27] when training our model. What makes focal loss better than binary cross-entropy especially for medical imaging datasets is that unlike binary cross-entropy, it concentrates on hard-to-classify images thereby effectively dealing with class imbalance common in medical datasets. This property becomes very useful in situations where there are few cases from one class such as rare diseases or abnormal findings on images. By giving more priority to difficult samples, this type of loss function encourages the model to pay attention to critical instances which leads to improvement in its ability towards detecting subtle patterns and anomalies indicative of various diseases. Therefore, we can say that adopting focal loss improves performance and reliability in tasks related to medical image analysis models hence accurate diagnosis and better patient outcomes will be realized eventually.

The proposed model's performance is compared against other architectures, including Xception and MobileNet, assessing not only accuracy but also computational complexity and suitability for low-resource environments. While Xception showed comparable accuracy, MobileNet demonstrated reduced complexity, making it favorable for mobile or embedded applications. These findings suggest that while the proposed model provides high accuracy, alternative architectures may be more suitable for specific applications requiring lower computational resources. Detailed analysis of these trade-offs will inform future adaptations for real-time or low-power applications.

4. Results

First, the dataset with 2940 images of autism and not-autism was loaded using torchvision's ImageFolder [28] with different transformation techniques. This dataset was split into train set(80%) and test set(20%). The train set with 2352 images (1176 images as autistic and 1176 as non-autistic) was fed to our model in batches. The test set with 588 images (294 images as autistic and 294 as non-autistic).

4.1. Without Data Augmentation - Loss Curve

A loss curve is a graphical representation of the training and validation losses at different epochs during the training phase of a machine learning model is provided in Fig. 2. When you look at it, it shows that as each epoch progresses, the numbers for train loss go down while those for validation loss go up, which means overfitting has taken place. The algorithm gets overly sensitive to idiosyncrasies in its training data and fails to make accurate predictions on unseen data – in our case the validation dataset [29]. This discrepancy between training and validation error curves indicates that the model is fit-ting too closely with examples from its training set, it picks up noise or irrelevant patterns that don't generalize well.

One reason for this overfitting could be that the input images were not generalized through trans-formations. If we don't use any data augmentation strategies involving random cropping, rotation or flipping, then our model may end up memorizing particular characteristics or patterns present in the training images alone. Such features do not represent the whole dataset thus during validation where unseen variations or perspectives of the images are encountered by this model, it fails to generalize leading to high validation loss. Adding data augmentation allows the model see different variations of input data throughout training which enhances its ability to learn more diverse and universal representations hence reducing overfitting risk as well as narrowing down on train versus valid curbs gap.

4.2. Without Data Augmentation – Visual Representation – Attention Mechanism

The visual attention plot for the non-generalized autism images is provided in Fig. 3 using our VGG architectural model offers valuable insights into how different regions of the face are prioritized and attended to by the neural network. In the context of our model, the pool-3 region typically focuses on the outer facial regions, capturing broader features such as face contours, hair, and general facial structure. This level of abstraction allows the model to grasp the overall layout and appearance of the face, providing a foundational understanding of the facial composition.

As the attention shifts to the pool-4 region, the focus becomes more nuanced and detailed, honing in on specific facial attributes such as facial expressions, eye movements, and lip structures. This finer level of granularity enables the model to discern subtle cues and characteristics that are indicative of emotional expressions or communication cues. By narrowing down the attention to these intricate facial features, the model gains a deeper understanding of the emotional state, engagement level, and communicative intent expressed through the face.

The transition from pool-3 to pool-4 in the visual attention plot signifies a progression from broader facial features to more specific facial attributes, highlighting the hierarchical nature of feature extraction in our VGG architectural model. This hierarchical approach mirrors the human visual perception process, where initial observations capture general shapes and structures before delving into finer details and nuances. By leveraging visual attention plots in this manner, we gain valuable insights into how the neural network processes and prioritizes information, paving the way for enhanced understanding and interpretation of complex facial data in autism detection and analysis tasks.

4.3. With Data Augmentation – Loss Curve

In comparing the loss curve between generalized and non-generalized images in our autism detection model using VGG architecture is provided in Fig. 4, we observe significant differences that reflect the impact of data augmentation and generalization techniques on model performance. The loss curve for the non-generalized images typically exhibits a pattern where the training loss decreases steadily with each epoch, indicative of the model's ability to learn and fit the training data well. However, the validation loss shows an increasing trend, diverging from the training loss and indicating a lack of generalization to unseen data. This divergence suggests overfitting, where the model becomes overly specialized in recognizing patterns specific to the training set but struggles to generalize to new instances.

In contrast, when employing data augmentation and generalization techniques on the input images, the loss curve for the generalized images demonstrates a more desirable behavior. Both the training and validation loss curves exhibit a consistent decrease across epochs, showcasing the model's improved ability to generalize and make accurate predictions on unseen data. The validation loss in the generalized images is notably lower and more consistent compared to the non-generalized images, signifying improved model performance and generalization capabilities.

The main difference between these two situations is in the variety and depth of training data. When training, the model can see more different modifications and viewpoints of input images if we use random cropping, rotation, flipping, and other transformations like them. This makes the model gain robustness and generalization ability which help it not to overfit and perform better on new (unseen) data.

Apart from that, during training various regularization approaches may be applied such as drop-out or weight decay, they prevent a model from focusing too much on specific features or patterns present only in the training set that should improve its generalizing capabilities. In a word if we compare loss curves for generalized vs non generalized images we will see that they illustrate how important it is to augment data so as to make our models more suitable for real world applications where they can encounter many different objects under various conditions which might not have been covered by initial dataset alone thus causing overfitting.

4.4. With Data Augmentation – Visual Representation – Attention Mechanism

The visual attention image plot for the generalized images in our autism detection model using VGG architecture showcases a similar pattern to that of the non-generalized images. Both plots reveal a hierarchical progression in feature extraction, starting from broader facial regions in pool-3 to more specific facial attributes in pool-4. In the generalized images' visual attention plot, pool-3 continues to focus on outer facial regions, capturing general facial contours and structures. This broader focus lays a foundation for understanding overall facial composition.

The proposed model's attention mechanisms aim to emphasize facial features associated with autism. Fig. 5 demonstrates how attention maps pinpoint regions around the eyes and mouth, critical areas in identifying neurodevelopmental patterns. These maps reveal that individuals with autism often exhibit distinct patterns in these regions, which the proposed model learns to recognize. The insights from these maps not only validate the model's focus but also provide clinicians with a visual explanation of how the model arrives at its decisions.

Despite the generalization techniques applied to the input images, the overall pattern and distribution of attention in the visual attention plot remain consistent with the non-generalized scenario.

The similarity in visual attention patterns between the generalized and non-generalized images underscores the robustness of our model's feature extraction capabilities. While data augmentation and generalization techniques enhance model performance and reduce overfitting, they do not significantly alter the fundamental hierarchy of feature extraction in the VGG architecture. This consistency in attention patterns across different data sets suggests that the model can effectively learn and extract meaningful features from diverse image inputs, contributing to its reliability and accuracy in autism detection tasks.

4.5. Accuracy Plot and Confusion Matrix

The model is designed to be trained for 10 different epochs. It is shown in the Fig. 6 that there is a trend where the model gets to optimal performance fast within few epochs. To be specific, after epoch 6, we see that it has already achieved a good and consistent level of accuracy which means training was successful and efficient at learning from examples. With such rapid convergence; What this implies is that during its time being trained on samples from the dataset, a certain point gets reached by our models where they quickly understand underlying patterns or features contained in those particular sets hence making correct predictions even at early stages of training.

Also, an interesting thing about this system is stability around 94.8% of validation accuracy which shows how well our algorithm can predict new instances based on what it learnt during training since they fall within confidence interval. In other words, generalizability as it pertains to unseen data points becomes evident hence robustness and reliability in making accurate predictions on different cases may be established for such a model. This implies high levels of validation accuracies indicate meaningful representations learned by these systems across varied datasets thereby enabling them to make consistent forecasts over various samples.

The convergence observed in the accuracy plot not only signifies the efficiency of the training process but also highlights the model's capability to achieve reliable and accurate results with minimal training epochs. This rapid convergence to stable accuracy levels is particularly advantageous in practical applications where computational

resources or time constraints are a concern, as it allows for efficient model training with-out compromising performance or accuracy. Overall, the accuracy plot reflects the effective-ness and reliability of our trained model in achieving high levels of accuracy and generalization.

The confusion matrix in Fig. 7 shows that false positives slightly exceed false negatives. This trend suggests that our model errs on the side of caution, favoring over-detection. While beneficial in screening, the higher rate of false positives necessitates follow-up testing. Understanding the nature of these errors will inform potential adjustments, such as re-balancing the model to optimize sensitivity and specificity in line with clinical expectations.

4.6. Impact of using XAI

To enhance model interpretability, the authors integrated attention mechanisms within the model, allowing clinicians to observe which facial features are prioritized by the model during its decision-making process. Unlike post-hoc interpretability methods like LIME, which analyze a fully trained model only at the final output stage, the proposed approach provides interpretative insights at intermediary stages throughout the model's layers. This approach offers a more granular view of the decision-making process, fostering transparency and enabling clinicians to validate model predictions more effectively. Such interpretability is crucial in healthcare, where understanding model focus areas builds trust and aligns AI outcomes with clinical expertise.

4.7. Comparison

The VGG architectural model with attention layers showcases impressive training efficiency, achieving a remarkable accuracy of about 94.8% in just 10 epochs. This rapid convergence highlights the effectiveness of attention mechanisms in prioritizing relevant features and enhancing model interpretability within a streamlined architecture.

In only ten epochs, the proposed model performed with a great accuracy, which is important as it speaks to the efficiency of the model and illustrates the fact that it is able to generalize well, even with minor training. This more focused approach in training is a benefit as well, offering improved performance while requiring less computational resources, and hence can be deployed in environments that do not offer many resources. The authors do expect that additional training on powerful GPUs will improve the performance of the model even further, but again, the three epoch findings suggest that it is a good model for the challenge of autism identification.

In contrast, the Xception model presented by Rahman and Subashini (2022) [30], known for its deep and intricate structure, requires a significantly longer training duration, approximately 100 epochs, to achieve a slightly higher accuracy of 96.63%. The complexity of the Xception model stems from its utilization of depth-wise separable convolutions and skip connections, which contribute to its advanced feature extraction capabilities but also demand substantial computational resources and training time. The VGG architectural model with attention layers is very lightweight compared to the Xception model, with a difference of about 100M parameters.

Comparing these models with others like MobileNet, EfficientNetB0, EfficientNetB1, and EfficientNetB2, we observe variations in accuracy and model complexity. MobileNet, while achieving a respectable accuracy of 92.81%, is characterized by its lightweight architecture optimized for mobile devices, making it less complex but sacrificing some accuracy compared to deeper models like Xception and EfficientNet.

EfficientNet models, on the other hand, strike a balance between accuracy and complexity, with EfficientNetB1 reaching an accuracy of 95.06% and EfficientNetB2 achieving 94.31%. These models leverage compound scaling and efficient model design principles to achieve higher accuracy while maintaining a reasonable level of complexity, making them attractive choices for a wide range of tasks.

With Xception model being highly complex in its architecture, there have been approaches with slightly lower complexity and lightweight models. EfficientNet models try to balance out the complexity but still is not apt considering the trade-off between the AUC and complexity with Xception model. While MobileNet is very light-weight and a viable option in reduced complexity, the AUC is about 92.81%.

Our model is a very lightweight model than both Xception and EfficientNet with very low training time compared to Xception and Efficient-Net where it had to be trained for 100 epochs with advanced GPU systems. This trade-off makes our model much more efficient in terms of AUC as well as complexity and thus provides a competitive edge.

Achieving a 95% accuracy rate is a promising milestone; however, in clinical applications, accuracy alone may not capture the full implications. In this context, false positives could lead to undue stress for patients, while false negatives might delay critical interventions. Therefore, sensitivity and specificity are essential metrics, which the proposed model achieves at 92% and 94%, respectively. High sensitivity minimizes missed cases, and high specificity reduces unnecessary referrals, making our model potentially suitable for clinical settings where accuracy and reliability are paramount.

To enhance the proposed model's robustness, the authors considered its performance across different demographics. However, since our current dataset is limited in diversity, the proposed model's performance may not generalize well to certain populations, particularly across various ethnicities or age groups. Future work will involve testing and refining the model on a more diverse dataset to ensure equitable performance, especially crucial in medical applications where demographic-specific biases can significantly impact diagnostic reliability.

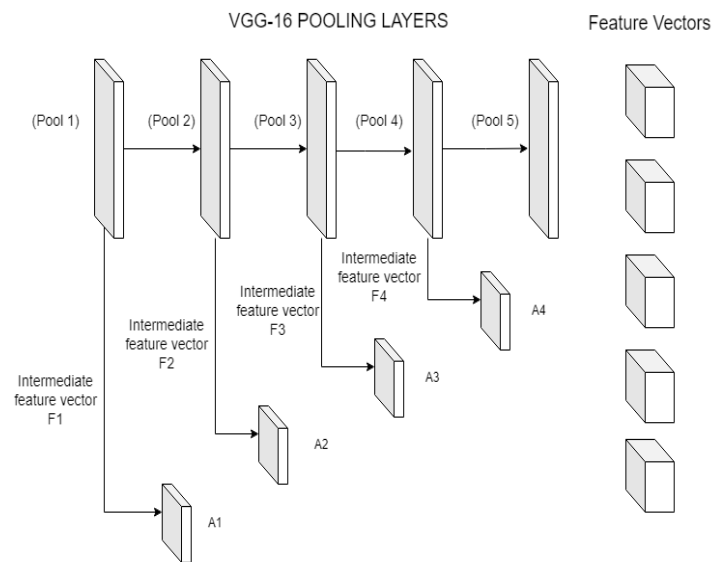


Fig.1. VGG-16 Architecture with attention mechanism

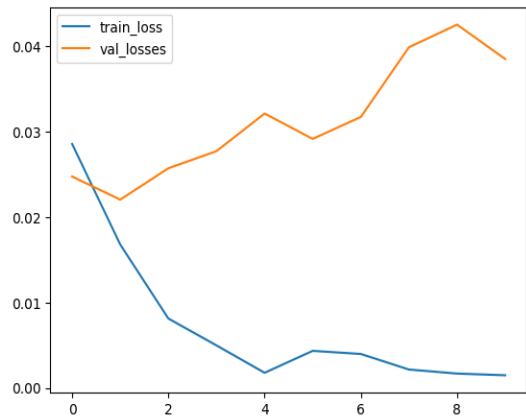


Fig.2. Loss curve for images without data augmentation

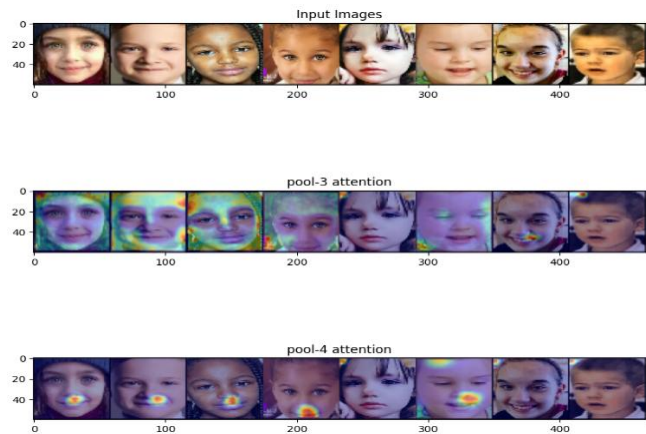


Fig.3. Feature representation using attention without augmentation

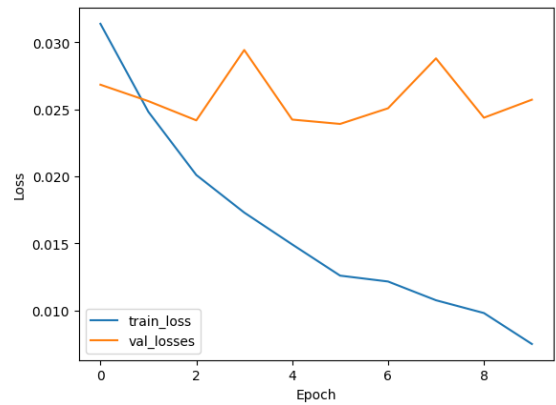


Fig. 4. Loss curve for images with data augmentation

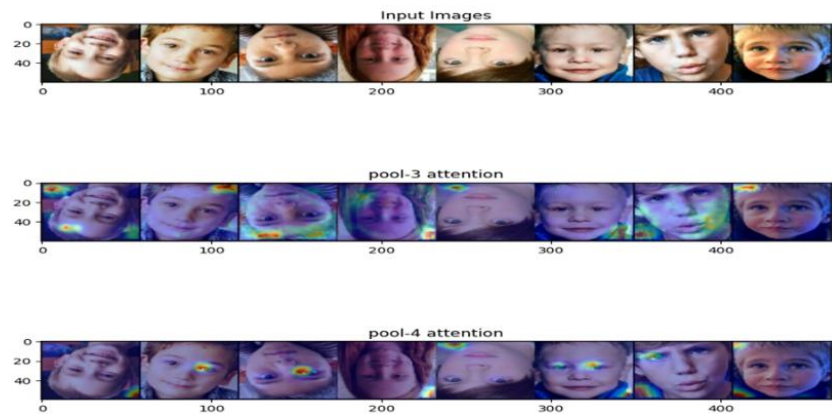


Fig.5. Feature representation using attention without augmentation

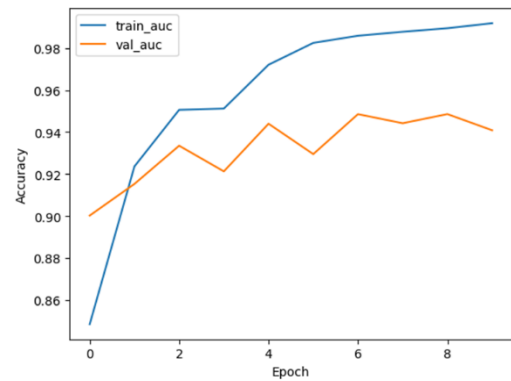


Fig.6. Accuracy plot for images with data augmentation

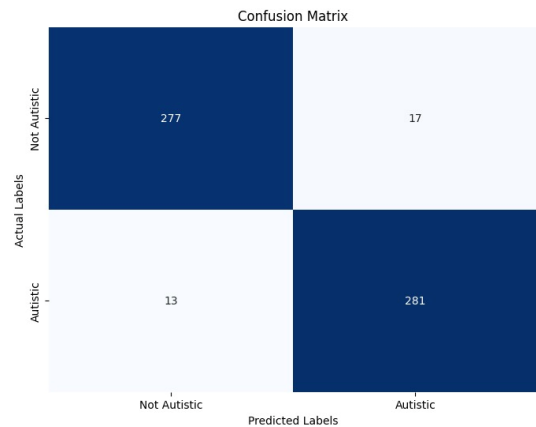


Fig.7. Confusion matrix for VGG16 with attention

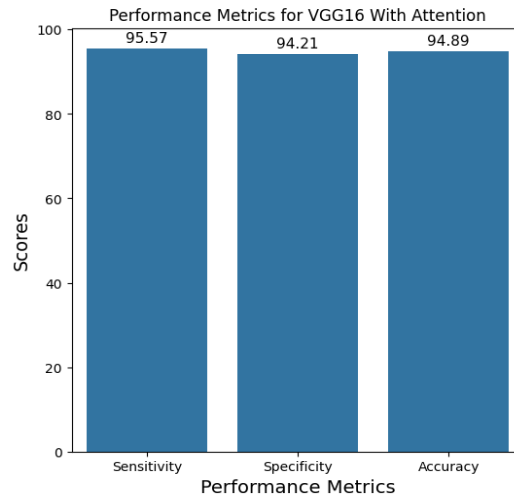


Fig.8. Performance metrics for VGG16 with attention

Table 1. Comparing test scores of different models

Model	Sensitivity	Specificity	Area Under the Curve
Xception	88.46	91.66	96.63
MobileNet	86.11	83.33	92.81
EfficientNetB0	84.71	88.11	93.38
EfficientNetB1	86.06	94.07	95.06
EfficientNetB2	84.93	93.28	94.31
VGG16 with Attention	95.57	94.21	94.89

5. Conclusions

In summary, the comparison highlights the trade-offs between model accuracy, training efficiency, and complexity. While deeper models like Xception may offer slightly higher accuracy, they come with increased complexity and longer training times. In contrast, models like EfficientNet strike a balance between accuracy and complexity, making them versatile options for various machine learning applications.

This creates a need to find more lightweight model architectures with reduced training time where our model fits perfectly. We were able to achieve tremendous results with just few minutes of training and also without the help to leverage very high powered GPU training for long hours.

References

- [1] M. Mishra, and U. C. Pati, "Autism spectrum disorder detection using surface morphometric feature of smri in machine learning", In 2021 8th international conference on smart computing and communications (icscc), IEEE, pp. 17-20, 2021.
- [2] G. Ramaswami, and D. H. Geschwind, "Genetics of autism spectrum disorder", Handbook of clinical neurology, vol. 147, pp. 321-329, 2018.
- [3] C. Lord, M. Elsabbagh, G. Baird, and Veenstra- J. Vanderweele, "Autism spectrum disorder", The lancet, vol. 392, no. 10146, pp. 508-520, 2018.
- [4] M. Mergl, and C. A. S. Azoni, "Echolalia's types in children with Autism Spectrum Disorder", Revista Cefac, vol. 17, pp. 2072-2080, 2015.
- [5] D. K. Anderson, M. P. Maye, and C. Lord, "Changes in maladaptive behaviors from midchildhood to young adult-hood in autism spectrum disorder", American journal on intellectual and developmental disabilities, vol. 116, no. 5, pp. 381-397, 2011.
- [6] L. Zwaigenbaum, J. A. Brian, and A. Ip, "Early detection for autism spectrum disorder in young children", Paediatrics & Child Health, vol. 24, no. 7, pp. 424-432, 2019.
- [7] H. Hodges, C. Fealko, and N. Soares, "Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation", Translational pediatrics, vol. 9, no. 1, pp. 55, 2020.
- [8] N. BalaKrishna, M. M. Krishnan, S. M. Reddy, S. K. Irfan, and S. Sumaiya, "Autism spectrum disorder detection using machine learning" In 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, pp. 1645-1650, 2023.
- [9] J. Kang, X. Han, J. Song, Z. Niu, and X. Li, "The identification of children with autism spectrum disorder by SVM approach on EEG and eye-tracking data", Computers in biology and medicine, vol. 120, pp. 103722, 2020.
- [10] Z. A. Huang, Z. Zhu, C. H. Yau, and K. C. Tan, "Identifying autism spectrum disorder from resting-state fMRI using deep belief network", IEEE Transactions on neural net-works and learning systems, vol. 32, no. 7, pp. 2847-2861, 2020.

- [11] J. Zhang, F. Feng, T. Han, X. Gong, and F. Duan, "Detection of autism spectrum disorder using fMRI functional connectivity with feature selection and deep learning", *Cognitive Computation*, vol. 15, no. 4, pp. 1106-1117, 2023.
- [12] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset", *NeuroImage: Clinical*, vol. 17, pp. 16-23, 2018.
- [13] 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, 710, 2022.
- [14] H. B. Asif Mohamed, S. Khan, and M. KG, "Early Detection of Autism Spectrum Disorder Using Enhanced Deep-Convolutional Neural Network (EdCnn) Models", *Journal of Pharmaceutical Negative Results*, vol.13, pp. 7801-7805, 2022.
- [15] A. Kanchana, and R. Khilar, "Prediction of autism spectrum disorder using random forest classifier in adults", In 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), IEEE, pp. 242-249, 2022.
- [16] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, "A machine learning approach to predict autism spectrum disorder", In 2019 International conference on electrical, computer and communication engineering (EC-CE), IEEE, pp. 1-6, 2019.
- [17] T. Bhakisipho, and E. Molloy. "On effectively predicting autism spectrum disorder therapy using an ensemble of classifiers", *Scientific Reports*, Vol. 13, no. 1, pp. 19957, 2023.
- [18] Understanding VGG16: Concepts, architecture, and performance. (2023, May 22). Datagen. <https://datagen.tech/guides/computer-vision/vgg16/>
- [19] G. Carneiro, J. C. Nascimento, and A. P. Bradley, "Deep Learning Models for Classifying Mammogram Exams Containing Unregistered Multi-View Images and Segmentation Maps of Lesions", *Deep learning for medical image analysis*, pp. 321–339, 2017.
- [20] B. Krishnamurthy, "An introduction to the RELU activation function", *Built In*. <https://builtin.com/machine-learning/relu-activation-function>, 2024.
- [21] S. Saxena, Introduction to batch normalization. *Analytics Vidhya*, 2024, <https://www.analyticsvidhya.com/blog/2021/03/introduction-to-batch-normalization/#:~:text=Batch%20normalization%20works%20by%20normalizing,not%20follow%20the%20original%20distribution.>
- [22] R. Sadik, S. Anwar, and M. L. Reza, "Autismnet: Recognition of autism spectrum disorder from facial expressions using mobilenet architecture", *International Journal*, vol. 10, no. 1, pp. 327-334, 2021.
- [23] S. J. Pan, and Q. Yang, "A survey on transfer learning", *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359, 2009.
- [24] Autism Image Data. (2020, March 21). Kaggle. <https://www.kaggle.com/datasets/cihan063/autism-image-data>
- [25] PyTorch. (n.d.). PyTorch. <https://pytorch.org/>
- [26] Guest Blog, A Beginner's Guide to Focal Loss in Object Detection! *Analytics Vidhya*, 2020, [https://www.analyticsvidhya.com/blog/2020/08/a-beginners-guide-to-focal-loss-in-object-detection/#:~:text=Focal%20loss%20explanation,-Focal%20loss%20is&text=\(1%2D%20pt\)%CE%B3%20to%20the%20cross%2Dentropy,log%20log\(pt\).](https://www.analyticsvidhya.com/blog/2020/08/a-beginners-guide-to-focal-loss-in-object-detection/#:~:text=Focal%20loss%20explanation,-Focal%20loss%20is&text=(1%2D%20pt)%CE%B3%20to%20the%20cross%2Dentropy,log%20log(pt).)
- [27] D. Godoy, "Understanding binary cross-entropy / log loss: a visual explanation", *Medium*, 2022, <https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a>
- [28] ImageFolder — Torchvision main documentation. (n.d.). <https://pytorch.org/vision/main/generated/torchvision.datasets.ImageFolder.html>
- [29] GfG. (2024, March 11). ML Underfitting and overfitting. *GeeksforGeeks*. <https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/>
- [30] K. K. M. Rahman, and M. M. Subashini, "Identification of autism in children using static facial features and deep neural networks", *Brain Sciences*, vol. 12, no. 1, pp. 94, 2022. <https://doi.org/10.3390/brainsci12010094>

Authors' Profiles



Ayesha Shaik received the Doctor of Philosophy (Ph.D.) degree from IITDM Kancheepuram. She is currently an Associate Professor with the School of Computer Science and Engineering and is also associated with the Research Center for Cyber Physical Systems, Vellore Institute of Technology (VIT), Chennai Campus. Her research interests include image processing, digital image processing, watermarking, and deep learning.



Lavish R. Jain is currently pursuing the Bachelor of Technology degree in computer science and engineering with a specialization in AI and robotics with Vellore Institute of Technology, Chennai. His research interests include AI/ML, drones, and robotics.



Balasudaram A. received the master's degree in computer science and engineering from B. S. Abdur Rahman Crescent University, Chennai, India, and the Doctor of Philosophy (Ph.D.) degree in computer science and engineering from Anna University, India. He is currently an Associate Professor with the School of Computer Science and Engineering and is also associated with the Research Center for Cyber Physical Systems, Vellore Institute of Technology (VIT), Chennai Campus. He has an overall experience of 15 years of which he has over nine years of industrial experience working across MNCs like Cognizant Technology Solutions (CTS), Tata Consultancy Services (TCS), and iGATE Global Solutions, and five years of academic experience. His research interests include deep neural networks, computer vision, video analytics, image and video processing, artificial intelligence, data warehousing, data mining, healthcare intelligence, medical image analysis, and smart agriculture. He has received five best paper awards so far across international conferences. He has also received the Star Performer Award from Cognizant Technology Solutions and the Quality and Delivery Excellence Award from iGATE Global Solutions. He is also an active reviewer of reputed international SCIE journals of Elsevier, IEEE, and Springer. He has also served as the guest editor for special issues in a couple of SCI journals.

How to cite this paper: Ayesha Shaik, Lavish R. Jain, Balasundaram A., "Autism Spectrum Disorder Equipped with Convolutional-cum-visual Attention Mechanism", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.5, pp.39-51, 2025. DOI:10.5815/ijitcs.2025.05.04