SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Effect of Different Modalities of Facial Images on ASD Diagnosis using Deep Learning-Based Neural Network

Mohammad Shafiul Alam[1,2], Zabina Tasneem[1], Sher Afghan Khan[1], Muhammad Mahbubur Rashid[1,*]

[1] Department of Mechatronics Engineering, Kulliyah of Engineering, International Islamic University Malaysia, 53100 Jln Gombak, Malaysia
[2] Department of Electrical and Electronic Engineering, Faculty of Science and Engineering, Northern University Bangladesh, Dhaka, Bangladesh

ARTICLE INFO

Article history:
Received 27 May 2023
Received in revised form 6 September 2023
Accepted 16 September 2023
Available online 3 October 2023

Keywords:
Deep Learning; depth image; facial image; Autism Spectrum Disorder; explainable AI

ABSTRACT

This paper aims to investigate the effectiveness of different modalities of facial images for diagnosing Autism Spectrum Disorder (ASD) using deep learning-based neural networks. The motivation behind this study is the potential of advanced technologies to aid in accurately diagnosing ASD. The research revolves around the need to explore the performance of deep learning models on different modalities of facial images and to identify the challenges and potential solutions associated with each modality. The methodology involves training and testing the models on the respective datasets and analysing their accuracy and performance. ResNet50V2 achieved a 100% accuracy on the 2D test dataset, while Xception achieved an accuracy of 93.75% on the 3D test set. The detection accuracy suggests that neural networks-based deep learning methods have the potential to diagnose ASD using facial images accurately. However, the models perform better on 2D data, highlighting the need for additional training on larger 3D datasets to improve accuracy on 3D images. The study contributes to the field by providing insights into the performance of different modalities of facial images, emphasizing the need for robust datasets, and suggesting future research directions to enhance the accuracy and efficiency of ASD diagnosis using deep learning techniques.

1. Introduction

Autism Spectrum Disorder (ASD) is a complex brain development disorder affecting social interaction and communication skills, and patients with ASD engage in repetitive behaviour [1]. Early and accurate diagnosis of ASD is crucial for timely intervention and improved outcomes for affected individuals. Children with ASD may benefit from an early diagnosis because of the flexible nature of the brain, which could lead to enhanced social functioning. Only around one-third of children with autism spectrum disorder are identified after age 3, according to a recent study [2]. While behavioural assessments and clinical observations are commonly used for diagnosis, there is growing interest in utilizing computer vision techniques to aid in identifying ASD. The human visage is a

* *Corresponding author.*
E-mail address: mahbub@iium.edu.my

https://doi.org/10.37934/araset.32.3.5974

reflection of the human brain, which can process expression by receiving direct input from the central nervous system. It can therefore serve as a crucial biomarker for detecting brain-related diseases. The capacity to differentiate between various facial expressions is a crucial characteristic that can aid in the detection of brain asymmetry or neurodevelopmental disorders [3]. Over the years, research has shown that facial images carry valuable diagnostic information, making facial image analysis an essential tool in ASD diagnosis. Traditional approaches have primarily focused on 2D facial images, but recent advancements have allowed for the exploration of 3D facial images, providing richer spatial information for analysis.

Deep learning-based neural networks have emerged as powerful tools for image analysis tasks, including facial image recognition. These networks can automatically learn hierarchical representations from raw data, enabling them to capture intricate patterns and extract discriminative features for classification. In the context of ASD diagnosis, deep learning-based neural networks offer great potential to improve accuracy and efficiency by leveraging the information present in facial images [4].

Traditionally, 2D facial images have been the primary modality for ASD diagnosis. Deep learning-based neural networks have been the subject of numerous studies exploring the use of 2D facial images for image diagnosis of Autism Spectrum Disorder (ASD). These images capture facial features and expressions from a frontal perspective, providing valuable insights into emotional cues and social communication. The studies have shown promising accuracy results, indicating that 2D facial images have potential in diagnosing ASD [5,6].

However, 2D images lack depth information, making it challenging to fully represent the three-dimensional nature of the face. This limitation may hinder the accurate characterization and differentiation of facial expressions associated with ASD. In contrast, 3D facial images capture depth information, allowing for a more comprehensive representation of facial geometry and shape [7]. Additionally, 3D face recognition methods can better handle challenges such as disguises, and aging, which can significantly impact the performance of 2D-based systems. Furthermore, 3D face recognition techniques offer improved security by providing a higher level of uniqueness and difficulty in spoofing. Overall, the superiority of 3D face recognition in terms of accuracy, robustness, and security, making it a preferred choice in various applications, is discussed in many studies [8]. This additional dimension provides insights into subtle variations in facial landmarks, contours, and expressions, potentially enhancing the discriminative power of facial image analysis for ASD diagnosis. By incorporating depth information, 3D facial images offer a more holistic view of facial features, enabling the detection of nuanced characteristics related to ASD [9].

Researchers are drawn to this field due to the recent trend of detecting ASD using deep learning and facial characteristics, which has led to significant progress as shown in Table 1. Recent research that used their own CNN to extract ASD features from facial images achieved 91% of prediction accuracy [10]. In a separate study comparing automated and nonautomated machine learning techniques, the automated ML technique achieved an astonishing ASD detection accuracy of 96% [11]. Using Xception, M. S. Alam *et al.,* conducted a comprehensive model-centric ablation investigation and obtained a detection accuracy of 95% [12]. In other studies, the MobileNetV2 algorithm was used to classify ASD and Normal control (NC) children with an accuracy of 92% [13]. The final article demonstrates that ResNet101's prediction accuracy for facial image classification is only 86.7% [14].

Table 1
Recent research on ASD diagnosis using facial images by deep learning methods

Sl No	Ref	Algorithm	Accuracy	Modality	Datatype	Demographics	ASD Assessment
1	[10]	own CNN	91%	2D	Synthetic	Only gender	None
2	[11]	AutoML	96%	2D	Synthetic	Only gender	None
3	[12]	Xception	95%	2D	Synthetic	Only gender	None
4	[13]	MobileNetV2	92%	2D	Synthetic	Only gender	None
5	[14]	ResNet101	86.70%	2D	Synthetic	Only gender	None

All these works are based on a curated dataset from online source named Kaggle ASD children dataset and there were concerns expressed regarding the effectiveness of this current online dataset [4]. The images were noisy, blurred, not aligned properly and so on. The models have been criticized for their inability to perform well on datasets with a lot of noise, and their outcomes are often presented without adequate statistical analysis [15]. In addition, there is no demographic information on the children, only gender-classified images; the same is true for the assessment of ASD status for autistic children. So, there is a need of proper valid dataset to identify autism with high accuracy. The significant contributions of this work are:

i. Acquiring a real-time, non-synthesized dataset containing assessments and demographic data for autistic and normal control children.
ii. It introduces an unprecedented approach by incorporating diverse facial image modalities, 2D and 3D, for training deep learning models.
iii. The study rigorously assesses the modalities' effectiveness through validation against a real-world dataset.
iv. Employs innovative explainable AI techniques to identify crucial facial features emphasized by the model for predicting ASD and normal control children.

The first section of the paper is an Introduction that reviews the relevant literatures, identifies any research gaps in earlier publications, and proposes a likely course of actions with significant contributions. The subsequent section discusses the method of ASD diagnosis, as well as the new real-world dataset acquisition protocol for both 2D and 3D modalities, the transfer learning approach, and the evaluation matrices. Following is a section containing the results of CNN deep learning models after training with collected datasets and evaluating with a distinct test set from a real-world scenario. Section 4 concludes with a discussion of prospective directions and current findings.

2. Methodology

The objective of this study is to use a transfer learning-based approach for autistic facial to predict Autism Spectrum Disorder (ASD) in children at an early age. In this study, we utilized pre-trained deep learning models with 2D and 3D facial images to automatically extract robust features that were too intricate to be identified through visual inspection. After passing these features through multiple layers, the diagnosis of ASD was obtained through the topmost dense layer.

2.1 Dataset Acquisition

This dataset aims to facilitate research and analysis in the field of facial expression recognition, as well as comparisons between children with ASD and neurotypical children. The sample consists of

3 to 11-year-olds with ASD and 4 to 11-year-olds for NC. During image capturing, the children were shown a visual stimulus and the images were captured by iPhone 14 Pro camera in Portrait mode. Three photographs were taken: front, left side, and right side. Participants for normal control samples were recruited from local schools, while participants for ASD samples came from special care schools and therapy centres. As per the admission policy of these special school a prior clinical assessment was done for each child. The parents or legal custodians of the participants provided informed consent after being informed of the purpose and procedures of the study. All images were preserved on a password-protected and secure device. Any identifying information will be kept private and stored separately from image data. This dataset is referred to as UIFIDV1 and contains 120 samples, with a 2:3 ASD to Normal control ratio and a total of 40 children as shown in Table 2. The male-to-female ratio is 3:1. The accumulated dataset will be made available for research purposes in accordance with all applicable ethical and legal standards. Researchers and organizations interested in utilizing the dataset will be required to submit an application outlining their research objectives and scope.

Table 2
Distribution of the ASD and NC class

Age					Number		
Class	Mean	St. Div.	Min	Max	Male	Female	Total
ASD	7.12	2.12	3	11	15	2	17
NC	6.96	2.16	4	11	14	9	23

2.1.1 Generating 3D depth maps from facial images

Portrait mode on the iPhone 14 employs its dual-camera system to capture images with depth information. One of the cameras captures the principal image, while the other measures the distance between the camera and various scene elements. The dual-camera system uses depth-sensing technology, such as Time-of-Flight (ToF) or structured light [16], to detect the time it takes for light to bounce off objects and return to the camera. The effect of portrait mode, which blurs the background while keeping the subject in focus, is accomplished by identifying the subject and separating it from the background using the depth map, demonstrating the camera system's depth-sensing capabilities [17]. This information is essential for reconstructing a three-dimensional depth map. Dual-camera systems with both a wide-angle and telephoto lens are utilized to capture the depth map on the iPhone camera. These lenses capture images from slightly different vantage points, allowing the device to calculate the difference between the two images. This disparity data is then used to compute the depth values for various scene points, yielding a depth map. A depth map was ultimately created by separating the different channels from an iPhone portrait mode photo [18]. Figure 1 shows the flowchart of generating the depth map from the Portrait mode images captured by the iPhone 14 pro, At the very beginning the images are passed through a face detector MTCNN [19] and a pre-processing queue for cropping and resizing prior to being fed into the 3D depth map generation pipeline for depth map extraction.

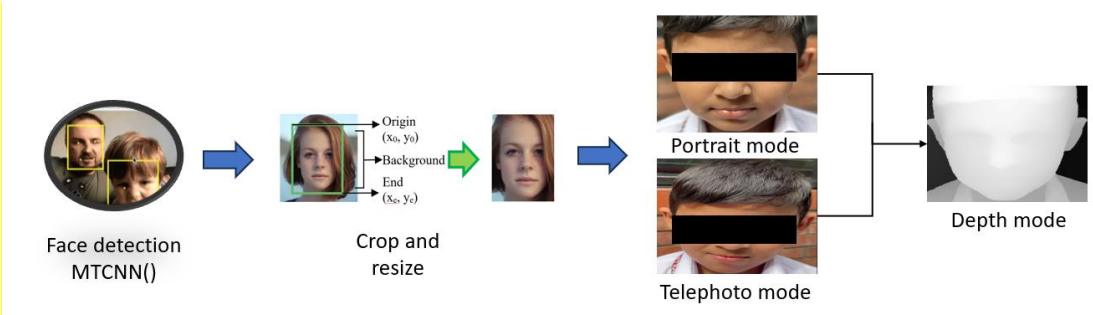


Fig. 1. Proposed flowchart of the transfer learning approaches used in this study

2.2 Data Augmentation Technique

Data augmentation refers to the process of expanding the quantity of data used in data analysis [20]. This is achieved by either creating slightly modified copies of existing data or generating synthetic data based on the existing samples. It has become a widely used technique for increasing the amount of data needed to effectively train machine learning (ML) models. The primary purpose of data augmentation is to act as a regularizer during the training of ML models, helping to prevent overfitting. By introducing variations in the data, the model becomes more robust and generalizes better to new, unseen data [21]. To augment the dataset in this study, the Keras image processing library, specifically the "ImageDataGenerator" function, was utilized. This function offers various options for augmenting images, including rotation, width and height shifting, and flipping. The specific details of these augmentation techniques can be found in Tensorflow (2022)[22]. Table 3 outlines the parameters used to augment the image data in this research. The choice of generator type and facility types was determined through trial and error, based on their impact on the evaluation performance of the model.

Table 3
Image augmentation used for training

Function parameter	Argument value
rotation_range	15
rescale	1./255
shear_range	0.2
zoom_range	0.2
horizontal_flip	TRUE
fill_mode	nearest
width_shift_range	0.1
height_shift_range	0.1

2.3 Transfer Learning in ASD Diagnosis

Transfer learning is a promising approach for diagnosing autism spectrum disorder (ASD). Transfer learning allows for the transfer of learned features and representations to ASD diagnosis tasks, even when there is limited ASD-specific data. This is made possible by leveraging pre-trained models on large-scale datasets. This approach tackles the difficulties posed by limited data availability and helps in creating precise and resilient diagnostic models for ASD. This enables them to differentiate between an ASD and normal control face by learning from an extensive collection of images [23]. In this study, the underlying architecture of the model remains intact, but its utilization is focused on extracting unique features from facial images of individuals with autism and those without autism. The top layers of the model are subsequently modified to facilitate the classification

task. 3D grayscale depth images are structured as matrices, containing depth information per pixel. Initial transformation consists of reformatting these matrices into 2D arrays and aligning them with the (224,224,3,0) input shape. The design of the CNN input layer (224,224,3,0) accommodates three colour channels (typically red, green, and blue). Due to the grayscale nature of depth images, pseudo colour channels are created to match the input configuration. This modification transforms the grayscale depth image into a "grayscale" image containing three identical colour channels. Finally, the images are resized to (224,224) specifications. The input shape for Xception is (299,299,3,0), so the images are reshaped accordingly.

The primary objective of this study is to evaluate the effect of different modality dataset 2D/ 3D on the identification of ASD with facial images. Using the unique characteristics extracted from these images, the research aims to develop a reliable and effective system for autism diagnosis employing non-intrusive and easily accessible visual data. The subsequent sections of this paper will explore the explication of an explainable AI methodology designed to expedite the identification of specific facial features that serve as the centre of attention for CNN models. These facial characteristics are poised to play a crucial role in aiding clinicians' decision-making processes when diagnosing and treating ASD cases. Figure 2 depicts the visual representation of the study's procedural framework, which is a flowchart of the investigation's methodology.

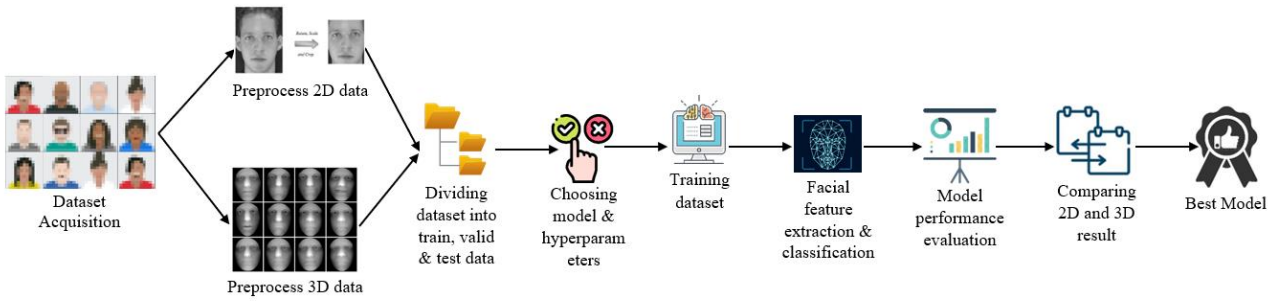


Fig. 2. Proposed flowchart of the transfer learning approaches used in this study

Facial feature extraction from 2D or 3D images has been revolutionized by the widespread use of transfer learning models. Several previously trained models have shown promising results in this area. Researchers can overcome the challenge of limited annotated 3D facial data and achieve state-of-the-art performance in various applications, such as facial expression recognition, age estimation, and gender classification, by utilizing transfer learning models. This study is based on peer-reviewed literature as there is currently no standardized procedure or general criteria for determining the most effective pre-trained algorithms. We have chosen five pretrained models which are ResNet50V2, VGG19, EfficientNetB0, MobileNetV2, and Xception because several referenced literatures have introduced promising performance of these models [4,12,13].

2.4 Evaluation Matrices

Commonly used parameters, including accuracy, recall, precision and Area Under Curve (AUC) were used to evaluate and prepare a comparative study with recent literatures. The formula for calculating these evaluation parameters where True Positive, T_p = number of autistic children predicted as autistic, True Negative, T_n = NC children detected as NC, False Positive, F_p = NC children predicted as autistic and False Negative, F_n = ASD children taken as NC.

$$Accuracy = \frac{T_p+T_n}{T_p+T_n+F_p+F_n} \quad (1)$$

$$Precision = \frac{T_p}{T_p+F_p} \quad (2)$$

$$Recall = \frac{T_p}{T_n+F_n} \quad (3)$$

2.5 Explainable AI

The goal of the field of explainable AI is to create AI systems that can not only make reliable forecasts, but also justify their actions in a way that humans can comprehend. The purpose of Explainable AI is to make machine learning more open, accountable, and understandable [24].

A visualization method called Grad-CAM (Gradient-weighted Class Activation Mapping)[25] can be used to explain the predictions of deep neural networks. The most relevant parts of the input image for the network's prediction are highlighted in a heatmap that Grad-CAM creates. The gradients of the output class score relative to the feature maps are computed in the final convolutional layer of the network, where the heatmap is produced. The feature maps are assigned weights based on these gradients, and the final heatmap is computed as an average of these weighted feature maps. Using Grad-CAM, one may see how a deep neural network functions and get an explanation for the network's predictions. This can be especially helpful in the field of medical imaging, where not only do precise forecasts matter, but so does knowing the reasoning behind the network's conclusion.

3. Result

In this section, the different phases of the experiment that we carried out to enhance the effectiveness of ASD detection using static RGB and RGB-D facial image data were discussed. The first part is a compendium of the results from the various evaluation matrices set for the 2D and 3D facial image dataset parameters. The discussion part examines the diagnosis of ASD using deep learning models and provides a comparative analysis of the contemporary research works.

3.1 Evaluation of 2D and 3D Facial Image Dataset

To determine the best parameters, optimizer, and hyperparameters for evaluating the outcome, we combed through a variety of scholarly articles. The chosen models were trained using Keras API library and other libraries such as pandas, sklearn and matpolib have been utilized to examine and visualize the performance of the models.

The experiment was conducted using a set of fixed hyperparameters, including 25 epochs, a learning rate of 0.001, and a batch size of 16 and these were chosen because these hyperparameters had shown great accuracy in previous studies.

After training and validation, the accuracy graph for different models is shown in Figure 3.

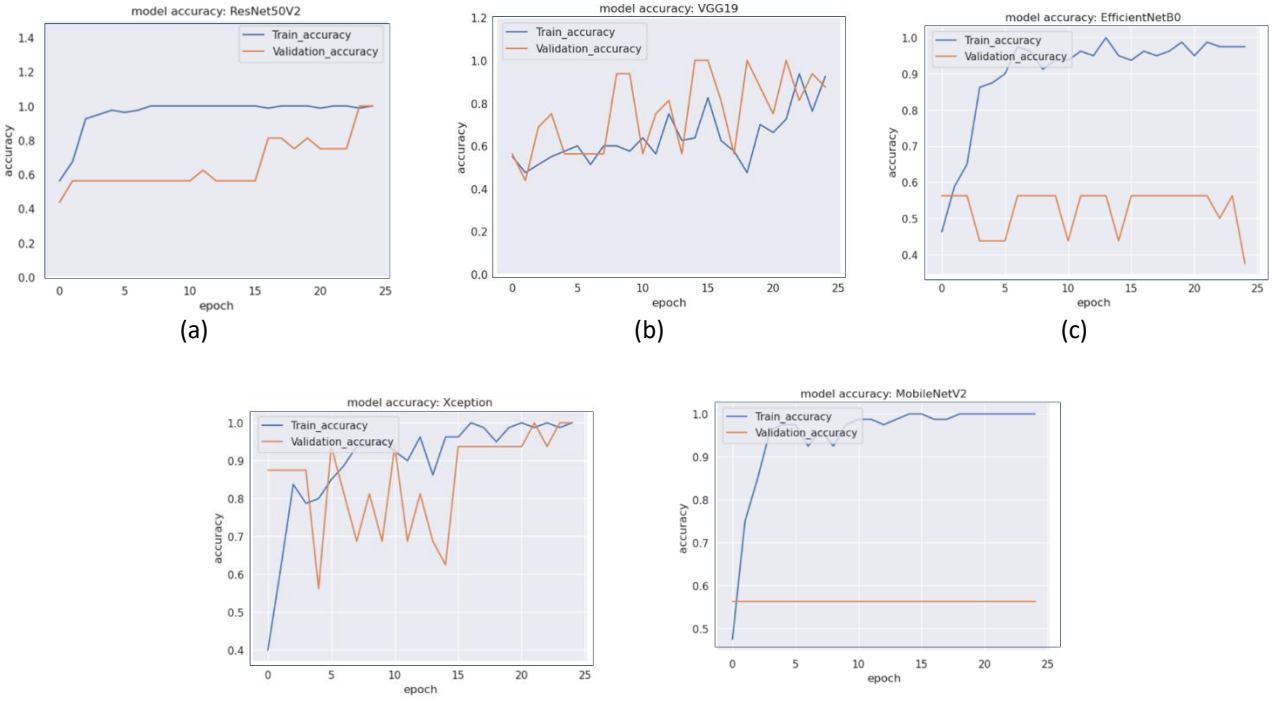


Fig. 3. Graphical representation of accuracy of (a) Resnet50V2, (b) VGG19, (c) EfficientNetB0, (d) Xception, and (e) MobileNetV2 for 2D

There are different optimizers which had great result but for our study we have used Adagrade with an initial accumulator value of 0.01 [12].Utilizing a batch size of 16 with a limited dataset can provide advantages such as faster learning because of more frequent weight updates, decreased memory usage, and a decreased risk of overfitting. The speed at which a model can acquire features from a dataset is determined by its learning rate, which is linked to other hyperparameters like epoch and batch size.

The train image set was divided into a validation set with a ratio of 80-20 percent, which is a typical practice in machine learning fields. The obtained dataset contains a total of 120 images, where 80 images have been used for training, 20 images for validation and 20 images for testing. The classification of the images is shown in Table 4. For the following experiments, the optimal set of training parameters are a learning rate of 0.001, Adagrad as the optimizer, categorical cross-entropy is considered as a loss function, and we have trained our models for 25 epochs. To ensure accuracy, we divided the training set into 80:20 ratio for validation purposes.

Table 4
Characteristics of Dataset

Dataset	Number	Class	Label
Train	80	Autistic	Autistic-0
Valid	20	NC	NC-1
Test	20		

The training and validation loss graph is shown in Figure 4. From the figures, we can see that the graph for EfficientNetB0 and MobileNetV2 have underfitting which means that the model is not trained properly and thus cannot provide good detection accuracy while validating.

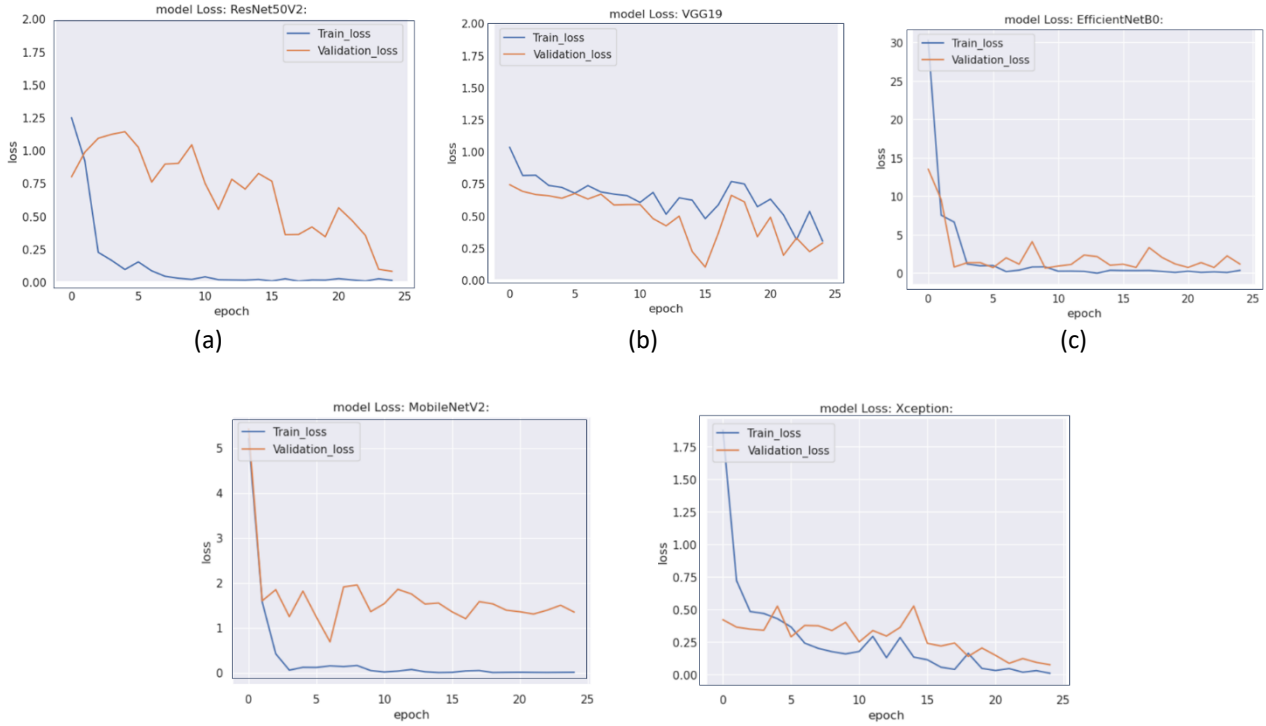


Fig. 4. Graphical representation of loss of (a) Resnet50V2, (b) VGG19, (c) EfficientNetB0, (d) MobileNetV2, and (e) Xception for 2D

After training and validation, the accuracy graph for different models is shown in Figure 5.

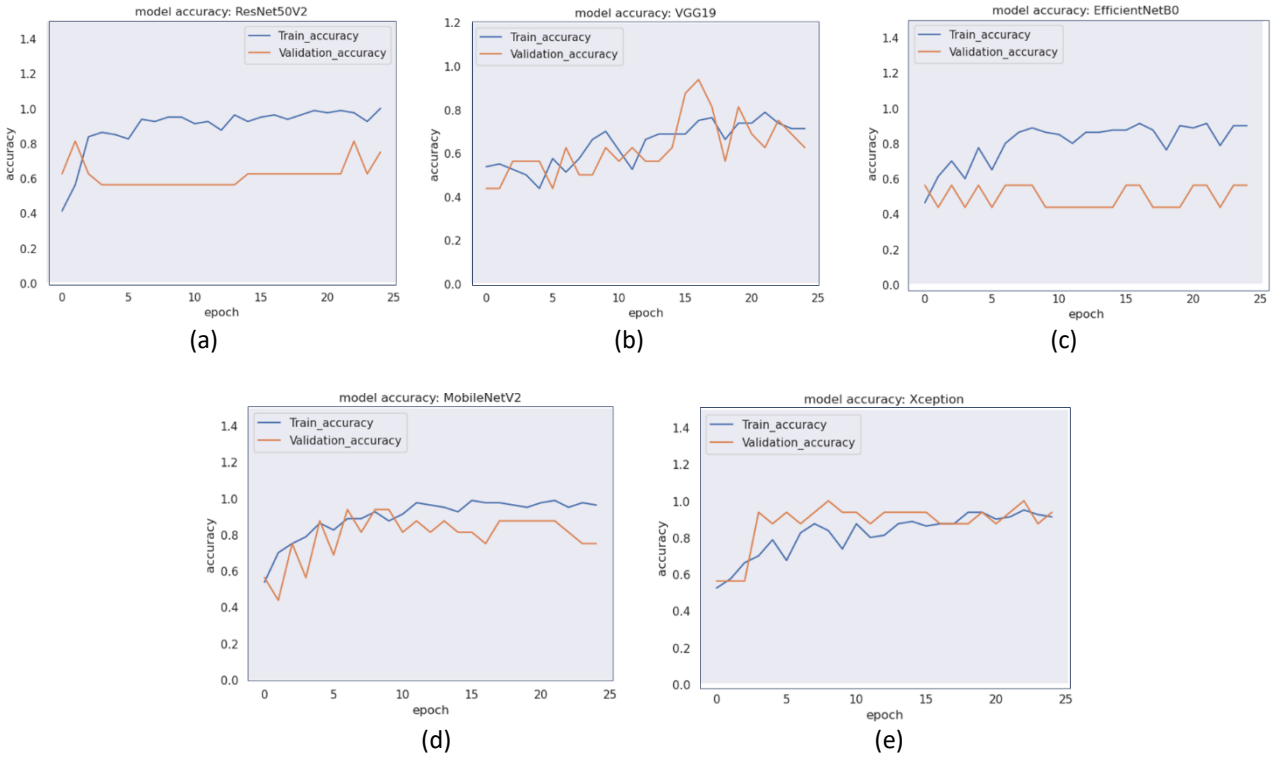


Fig. 5. Graphical representation of accuracy of (a) Resnet50V2, (b) VGG19, (c) EfficientNetB0, (d) MobileNetV2, and (e) Xception for 3D

The training and validation loss graph is shown in Figure 6 for the depth facial images of the same children as of 3D facial images.

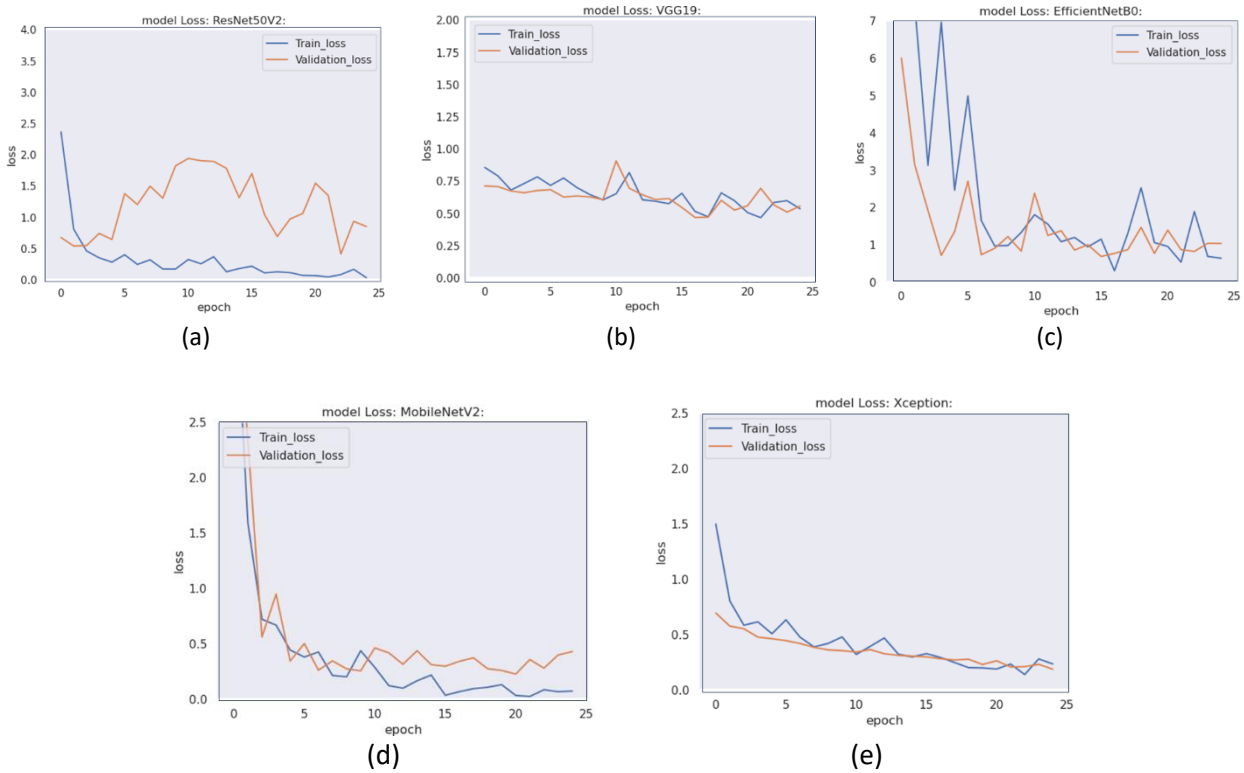


Fig. 6. Graphical representation of loss of (a) Resnet50V2, (b) VGG19, (c) EfficientNetB0, (d) MobileNetV2, and (e) Xception for 3D

The Accuracy, AUC, precision, recall, and loss evaluated on RGB (2D) facial image dataset for each pretrained model on test and train data is shown in Table 5.

Table 5
Result RGB (2D) facial image dataset

	Accuracy		AUC		Precision		Recall		Loss	
Model	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
ResNet50V2	**1.000**	**1.000**	1.000	1.000	1.000	1.000	1.000	1.000	0.004	0.025
VGG19	0.812	0.950	0.871	0.984	0.812	0.850	0.812	0.950	0.425	0.239
EfficientNetB0	0.375	0.425	0.246	0.336	0.375	0.600	0.375	0.425	1.178	1.190
MobileNetV2	0.562	0.525	0.808	0.712	0.562	0.400	0.562	0.525	1.347	2.005
Xception	0.950	0.950	0.997	0.993	0.950	0.950	0.950	0.950	0.102	0.104

For the Depth image dataset, the accuracy, AUC, precision, recall, and loss values on test and train data for each model are displayed in Table 6.

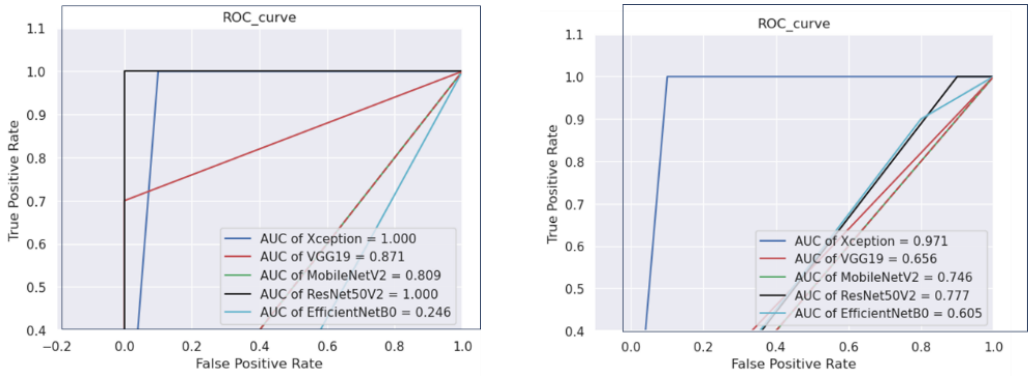
Xception is showing the highest accuracy with 93.7 % test accuracy while evaluating on the test set. As the models are pre-trained on RGB images earlier the dataset seems to be very small for training which cause the lower values in prediction accuracy.

The accuracy graph of Xception shows the optimum training thus shows the highest accuracy value while evaluating with depth image test set. ROC curve is a graphical representation of a binary classifier's performance, especially in image classification tasks. The ROC curve illustrates the compromise between true positive rate (sensitivity) and false positive rate (1 - specificity). Better classification performance is indicated by a steeper ROC curve, while an AUC value closer to 1

suggests greater accuracy. Figure 7's ROC plot indicates that the area under the curve is greater, indicating that the prediction rate for different test samples is higher in the real-world scenario. For 2D samples the Resnet50V2 is the best performer while for depth images Xception algorithm shows the better accuracy.

Table 6
Result on 3D data

Model	Accuracy		AUC		Precision		Recall		Loss	
	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
ResNet50V2	0.625	0.512	0.777	0.673	0.625	0.512	0.625	0.512	1.390	1.432
VGG19	0.500	0.612	0.656	0.732	0.500	0.612	0.500	0.612	1.260	1.135
EfficientNetB0	0.625	0.450	0.605	0.500	0.625	0.450	0.625	0.450	0.690	0.704
MobileNetV2	0.562	0.487	0.746	0.627	0.562	0.487	0.562	0.487	1.609	1.965
Xception	**0.937**	**0.837**	0.970	0.928	0.937	0.837	0.937	0.837	0.256	0.342



(a) (b)

Fig. 7. ROC curve of the models on (a) 2D data and (b) 3D data

Figure 8 displays confusion matrices that aid in comprehending the overall prediction performance.

According to the data presented in the figure, our ResNet50V2 model performed the best, with no misclassified images for 2D facial image test dataset, while Xception comes next with only 1 misclassified image. In contrast, Figure 9 demonstrates that Xception has the greatest performance for 3D images, misclassifying only one image out of a total of twenty samples.

From the graphs above in Figure 10, it can be seen that among all the five models, VGG16, EfficientNetB0, MobileNetV2, Xception, and ResNet50V2, Xception model has outperformed having 95% accuracy on 2D data and 93.25% accuracy on 3D data. Next for ResNet50V2, though it has achieved 100% accuracy on 2D data, it only achieved 62.5% accuracy on 3D data. EfficientNetB0 has the lowest performance having only 37.5% accuracy on 2D data but 62.5% accuracy on 3D data. For VGG16, it has 82.25% accuracy in 2D but only 50% accuracy in 3D. Only MobileNetV2 has same accuracy percentage which is 56.25% on both 2D and 3D data.

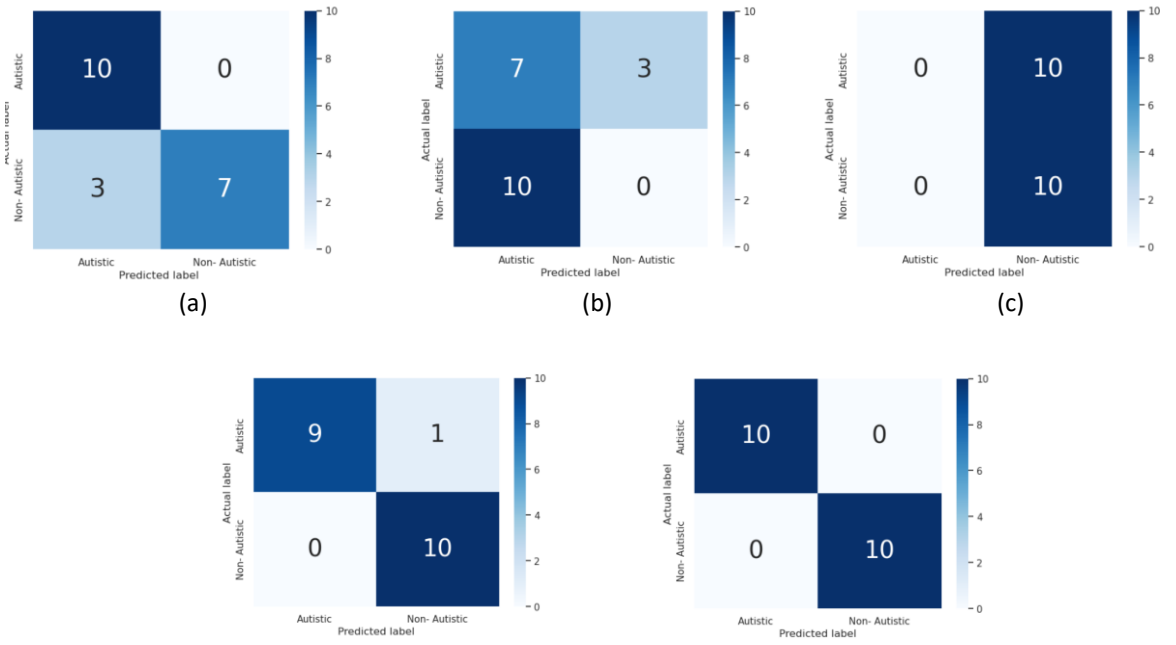


Fig. 8. Confusion matrix of 2D test samples (a) VGG16, (b) EfficientNetB0, (c) MobileNetV2, (d) Xception, and (e) Resnet50V2

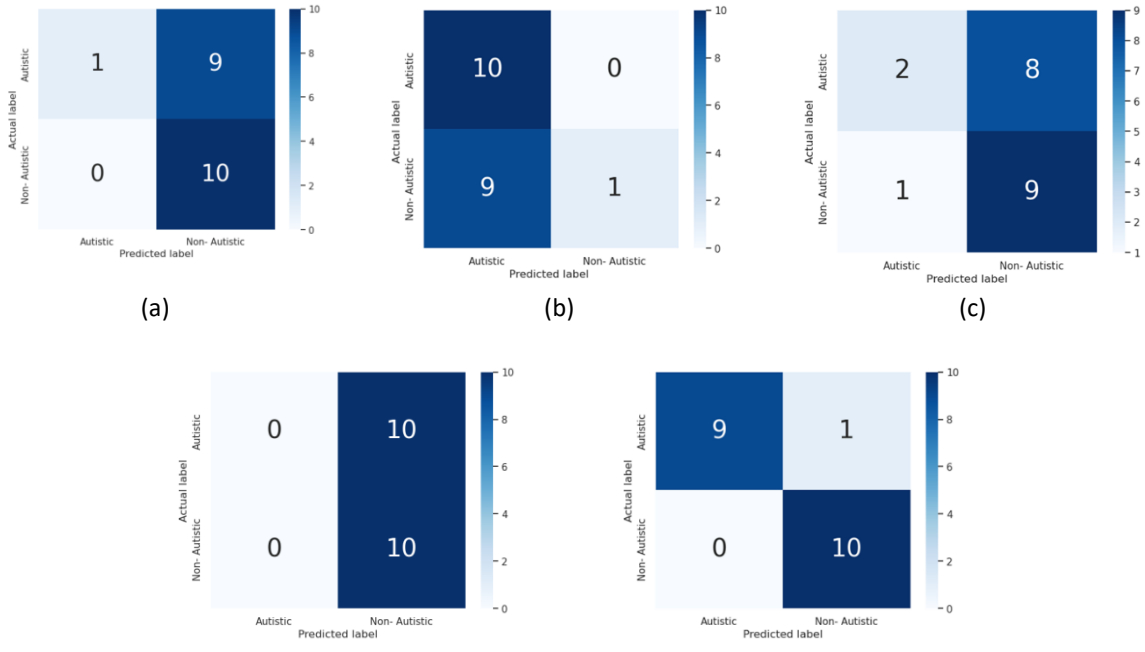
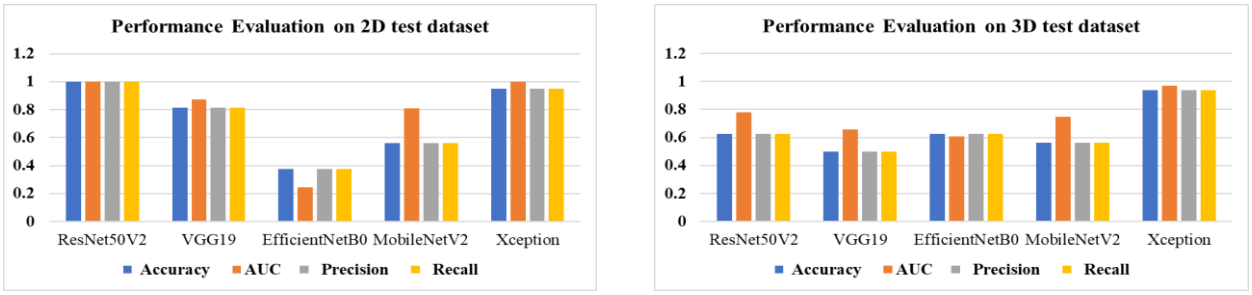


Fig. 9. Confusion matrix of 3D test samples (a) Resnet50V2, (b) VGG16, (c) EfficientNetB0, (d) MobileNetV2, and (e) Xception



(a) (b)

Fig. 10. Graph showing accuracy percentage for five models on test data of (a) 2D and (b) 3D

3.2 Discussion

This research is significant not only for its innovative approach to ASD diagnosis, but also for its use of real-world data collected from both autistic individuals and typically developing children as controls. This aspect lends a high level of legitimacy and relevance to the proposed strategy, thereby enhancing the real-world applicability and dependability of deep learning models in this domain. By utilizing authentic samples, the study has addressed a crucial factor that has the potential to substantially affect the accuracy and generalizability of facial image-based ASD diagnosis.

Table 7 shows the prediction accuracy of the recent researches with the deep learning CNN and the modality of facial data. This accuracy evaluation is based on a curated dataset from a distinct online platform referred to as the Kaggle ASD facial image dataset with a single-modality dataset containing only 2D facial images, but no validation on real-time dataset was possible. Derbali *et al.,* trained the data with a VGG model and used a platform based on video games to detect autistic children from their facial images with a prediction accuracy of 92.3%. Using the same dataset, Yadav K *et al.,* achieve a 95.3% accuracy with EfficientNet. El Mouatasim *et al.,* achieve a higher prediction accuracy of 96.88% when evaluating the same testset. Consequently, Alkahtani *et al.,* and Venkata Sai *et al.,* only obtained 92% and 88% accuracy using MobileNet and EfficientNet, respectively. It is clearly seen that our proposed model achieved the highest accuracy while validating with clinical dataset collected from real world scenario. The experimental results clearly demonstrate that the meticulous curation of the training and testing datasets from the pool of clinical subjects significantly improves the prediction accuracy for ASD diagnosis.

Table 7
Comparative analysis of the contemporary research

Ref	Authors	Dataset	Deep CNN	Accuracy	Modality
[26]	Derbali *et al.,*	Kaggle	VGG	92.3%	2D
[27]	Yadav K *et al.,*	Kaggle	EfficientNet	95.3%	2D
[28]	El Mouatasim *et al.,*	Kaggle	DensNet-121	96.8%	2D
[13]	Alkahtani *et al.,*	Kaggle	MobileNet	92.0%	2D
[29]	Venkata Sai *et al.,*	Kaggle	Efficient Net	88.0%	2D
	Proposed 2D	UIFIDV1	ResNet50V2	100%	2D
	Proposed 3D	UIFIDV1	Xception	93.7%	3D

Figure 11 depicts the Grad-CAM heatmap, which reveals the conspicuous facial landmarks that attract the primary attention of the best-performing models in general. Our analysis incorporates the entire dataset and includes representative samples from both the ASD and typical control child subsets. The facial regions scrutinised by the Resnet50V2 and Xception models, which are

distinguished by their high prediction accuracy for 2D and 3D images, respectively, are meticulously examined.

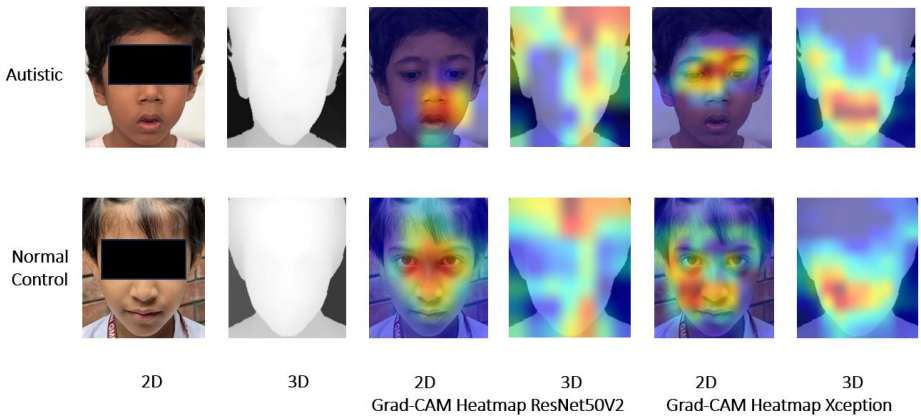


Fig. 11. Grad-CAM result of the autistic and normal control samples for best performing models

Resnet50V2 focuses its attention primarily on the lower nasal area, the cheekbones, and the lips in 2D images. In contrast, the model focuses on the eyes and cheeks of autistic individuals when trained with 3D images. In contrast, Xception favours the eyes and the region between the eyes and the upper nasal regions in 2D images. For 3D images, the model focuses primarily on the eyes and lips of autistic individuals. In the context of normal control children, the attentional pattern for 2D facial images remains consistent, with a focus on the eyes, the region between the eyes and the entire nasal structure, and a portion of the cheekbones. However, for 3D facial images, both ASD and normal control subjects fixate similarly on the lips and eyes, according to the Xception model. Notably, for 3D normal control children, the attentional distribution is inconsistent with the Resnet50V2 model, which contributes to a reduced detection accuracy for 3D images. By carefully selecting and organizing the data used to train the deep learning-based neural network, the model becomes more adept at identifying patterns and features indicative of ASD in facial images. This focused dataset curation ensures that the model is exposed to relevant and representative examples, enabling it to make more accurate and reliable predictions during the diagnosis process.

However, the research work has some limitations. Firstly, the prediction accuracy is higher for 2D data than 3D images when training models on the widely-used ImageNet dataset, which consists predominantly of 2D photographs. This observation raises a significant concern regarding the capacity of these models to completely capture the distinguishing characteristics of 3D facial images. In future, the use of a larger 3D facial image training dataset and advanced models like vision transformers, ConvNeXt during the model training phase might enhance the performance of 3D modality.

Secondly, some models exhibit perfect accuracy despite a small dataset size, as the test set may lack a variety of challenging cases, might make it difficult for the model to generalise to new, unseen images. As sample collection in the field of medical imaging is hindered by obstacles such as patient cooperation, ethical regulation, diversity and imbalance of data pattern, the dataset for this study is of small size. With additional data acquisition efforts in the future, some data augmentation and generalisation technique can be implemented.

4. Conclusions

This study examines deep learning-based neural networks for face image-based Autism Spectrum Disorder (ASD) diagnosis. ResNet50V2 and Xception models performed well on 2D and 3D test datasets, respectively. ResNet50V2 scored with 100% accuracy on the 2D dataset, whereas Xception scored with 93.75% accuracy on the 3D dataset indicates that DL algorithms can effectively diagnose ASD from facial images. The models were pretrained using the ImageNet dataset, which is mostly 2D photos, demonstrating the superiority of 2D data over 3D images. As the models may not properly capture 3D facial photos' distinctive qualities thus the study advises training models with a larger 3D facial picture dataset to increase 3D data performance. This discovery paves the door to more advanced 3D modality handling studies. This research collects real-life data from autistic and normal control youngsters, which lends legitimacy and relevance to the proposed strategy. The study improves real-world applicability and reliability of deep learning models by using authentic samples. Such datasets can improve facial image-based ASD diagnosis.

Acknowledgement

Author would like to express our deepest gratitude to the International Islamic University Malaysia for their support through the Tuition Fee Waiver Scheme 2021. The dataset was collected from the institutions named – “Society for the Welfare of the Intellectually Disabled, Bangladesh”, “Society for The Welfare of Autistic Children” for ASD children and “SOS Hermann Gmeiner College” for normal control children.

References

[1] Ghosh, Tapotosh, Md Hasan Al Banna, Md Sazzadur Rahman, M. Shamim Kaiser, Mufti Mahmud, ASM Sanwar Hosen, and Gi Hwan Cho. "Artificial intelligence and internet of things in screening and management of autism spectrum disorder." *Sustainable Cities and Society* 74 (2021): 103189. https://doi.org/10.1016/j.scs.2021.103189

[2] Zuckerman, Katharine E., Sarabeth Broder-Fingert, and R. Christopher Sheldrick. "To reduce the average age of autism diagnosis, screen preschoolers in primary care." *Autism* 25, no. 2 (2021): 593-596. https://doi.org/10.1177/1362361320968974

[3] LaMantia, Anthony-Samuel. "Why does the face predict the brain? Neural crest induction, craniofacial morphogenesis, and neural circuit development." *Frontiers in Physiology* 11 (2020): 610970. https://doi.org/10.3389/fphys.2020.610970

[4] Akter, Tania, Mohammad Hanif Ali, Md Imran Khan, Md Shahriare Satu, Md Jamal Uddin, Salem A. Alyami, Sarwar Ali, A. K. M. Azad, and Mohammad Ali Moni. "Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage." *Brain Sciences* 11, no. 6 (2021): 734. https://doi.org/10.3390/brainsci11060734

[5] Mohanty, Ashima Sindhu, Priyadarsan Parida, and K. C. Patra. "Identification of autism spectrum disorder using deep neural network." In *Journal of Physics: Conference Series*, vol. 1921, no. 1, p. 012006. IOP Publishing, 2021. https://doi.org/10.1088/1742-6596/1921/1/012006

[6] de Belen, Ryan Anthony J., Tomasz Bednarz, Arcot Sowmya, and Dennis Del Favero. "Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019." *Translational psychiatry* 10, no. 1 (2020): 333. https://doi.org/10.1038/s41398-020-01015-w

[7] Faizabadi, Ahmed Rimaz, Hasan Firdaus Mohd Zaki, Zulkifli Zainal Abidin, Muhammad Afif Husman, and Nik Nur Wahidah Nik Hashim. "Learning a Multimodal 3D Face Embedding for Robust RGBD Face Recognition." *Journal of Integrated and Advanced Engineering (JIAE)* 3, no. 1 (2023): 37-46. https://doi.org/10.51662/jiae.v3i1.84

[8] Faizabadi, Ahmed Rimaz, Hasan Firdaus Bin Mohd Zaki, Zulkifli Bin Zainal Abidin, Nik Nur Wahidah Nik Hashim, and Muhammad Afif Bin Husman. "Efficient Region of Interest Based Metric Learning for Effective Open World Deep Face Recognition Applications." *IEEE Access* 10 (2022): 76168-76184. https://doi.org/10.1109/ACCESS.2022.3192520

[9] Tan, Diana Weiting, Murray T. Maybery, Syed Zulqarnain Gilani, Gail A. Alvares, Ajmal Mian, David Suter, and Andrew JO Whitehouse. "A broad autism phenotype expressed in facial morphology." *Translational psychiatry* 10, no. 1 (2020): 7. https://doi.org/10.1038/s41398-020-0695-z

[10] Arumugam, Sajeev Ram, Sankar Ganesh Karuppasamy, Sheela Gowr, Oswalt Manoj, and K. Kalaivani. "A deep convolutional neural network based detection system for autism spectrum disorder in facial images." In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 1255-1259. IEEE, 2021.
[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. https://doi.org/10.4218/etrij.2021-0097
[12] Alam, Md Shafiul, Muhammad Mahbubur Rashid, Rupal Roy, Ahmed Rimaz Faizabadi, Kishor Datta Gupta, and Md Manjurul Ahsan. "Empirical study of autism spectrum disorder diagnosis using facial images by improved transfer learning approach." *Bioengineering* 9, no. 11 (2022): 710. https://doi.org/10.3390/bioengineering9110710
[13] Alkahtani, Hasan, Theyazn HH Aldhyani, and Mohammed Y. Alzahrani. "Deep Learning Algorithms to Identify Autism Spectrum Disorder in Children-Based Facial Landmarks." *Applied Sciences* 13, no. 8 (2023): 4855. https://doi.org/10.3390/app13084855
[14] Zheng, Yuyu, and Leyuan Liu. "Rapid Screening of Children With Autism Spectrum Disorders Through Face Image Classification." In *2022 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pp. 266-271. IEEE, 2022. https://doi.org/10.1109/IEIR56323.2022.10050070
[15] Mujeeb Rahman, K. K., and M. Monica Subashini. "Identification of autism in children using static facial features and deep neural networks." *Brain Sciences* 12, no. 1 (2022): 94. https://doi.org/10.3390/brainsci12010094
[16] Zhu, Qingpeng, Wenxiu Sun, Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Qianhui Sun *et al.,* "MIPI 2023 Challenge on RGB+ ToF Depth Completion: Methods and Results." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2863-2869. 2023. https://doi.org/10.1109/CVPRW59228.2023.00287
[17] Wang, Lijun, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. "DeepLens: shallow depth of field from a single image." *arXiv preprint arXiv:1810.08100* (2018). https://doi.org/10.1145/3272127.3275013
[18] Apple. "Capturing Photos with Depth.," developer.apple.com, 2023.
[19] Zhang, Ning, Junmin Luo, and Wuqi Gao. "Research on face detection technology based on MTCNN." In *2020 international conference on computer network, electronic and automation (ICCNEA)*, pp. 154-158. IEEE, 2020. https://doi.org/10.1109/ICCNEA50255.2020.00040
[20] Li, Bohan, Yutai Hou, and Wanxiang Che. "Data augmentation approaches in natural language processing: A survey." *Ai Open* 3 (2022): 71-90. https://doi.org/10.1016/j.aiopen.2022.03.001
[21] Ahsan, Md Manjurul, Muhammad Ramiz Uddin, Md Shahin Ali, Md Khairul Islam, Mithila Farjana, Ahmed Nazmus Sakib, Khondhaker Al Momin, and Shahana Akter Luna. "Deep transfer learning approaches for Monkeypox disease diagnosis." *Expert Systems with Applications* 216 (2023): 119483. https://doi.org/10.1016/j.eswa.2022.119483
[22] Yang, Suorong, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. "Image data augmentation for deep learning: A survey." *arXiv preprint arXiv:2204.08610* (2022).
[23] Ghosh, Tapotosh, Md Istakiak Adnan Palash, Mohammad Abu Yousuf, Md Abdul Hamid, Muhammad Mostafa Monowar, and Madini O. Alassafi. "A Robust Distributed Deep Learning Approach to Detect Alzheimer's Disease from MRI Images." *Mathematics* 11, no. 12 (2023): 2633. https://doi.org/10.3390/math11122633
[24] Shin, Donghee. "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI." *International Journal of Human-Computer Studies* 146 (2021): 102551. https://doi.org/10.1016/j.ijhcs.2020.102551
[25] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In *Proceedings of the IEEE international conference on computer vision*, pp. 618-626. 2017. https://doi.org/10.1109/ICCV.2017.74
[26] Derbali, Morched, Mutasem Jarrah, and Princy Randhawa. "Autism spectrum disorder detection: Video games based facial expression diagnosis using deep learning." *International Journal of Advanced Computer Science and Applications* 14, no. 1 (2023). https://doi.org/10.14569/IJACSA.2023.0140112
[27] Yadav, Bhavana, Shreya Vishwas, Nikitha Anand, Rishab Kashyap, and Raghu Bangalore. "Automated Identification and Classification of Autism Spectrum Disorder using Behavioural and Visual Patterns in Children." In *2023 4th International Conference for Emerging Technology (INCET)*, pp. 1-5. IEEE, 2023. https://doi.org/10.1109/INCET57972.2023.10170707
[28] El Mouatasim, Abdelkrim, and Mohamed Ikermane. "Control learning rate for autism facial detection via deep transfer learning." *Signal, Image and Video Processing* (2023): 1-8. https://doi.org/10.1007/s11760-023-02598-9
[29] Narala, Mohana Sree Venkata Sai Krishna, Sandeep Vemuri, and Chandrika Kattula. "Prediction of Autism Spectrum Disorder Using Efficient Net." In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 1139-1143. IEEE, 2023. https://doi.org/10.1109/ICACCS57279.2023.10112807