Early Autism Disorder Detection Through Visualizing Eye-Tracking Patterns Using Compact Convolutional Transformers

1 ABSTRACT

Autism spectrum disorder (ASD) is a group of impairments that includes issues with social communication, challenges with reciprocal social connections, and abnormal sequences of repetitive activity. The identification of autism has been helped by a variety of computer-aided methods in the modern world. 59 school-going children participated in the study where a selection of age-appropriate images and video recordings addressing social behaviour was shown to the participants. 13 autistic children underwent calibration and data-collecting on eye movement. Four of the six who were unable to complete the task did not glance at the screen, and two had calibration issues. The InceptionV3 model gained a better accuracy compared to EfficientNetB7 and MobileNetV2 which is 68.8%. Xception was implemented which is an extension of the Inception architecture excluding the depth-wise separable convolutions. Thereafter, The Visual Transformer Model showed impressive performance acquiring an accuracy of 91%. Compact Convolutional Transformer (CCT) has a parameter of 0.4M which is a lot small-scaled compared to the parameters of all other models. We deployed the Adam optimizer with 32 batches and a 30% dropout rate. 212 samples of ASD were successfully predicted as positive while 7 of them were wrongly predicted as Non-ASD images. Xception is an extension of the Inception architecture that lacks depthwise separable con- volutions. Xception achieves greater accuracy than all previously deployed models. VGG16 outperforms Xception, with an accuracy of 87.22%. Overall accuracy was 96.43%, including 96.29 recall, and 96.0% F1 score.

CCS Concepts: • **Computing methodologies**; • **Artificial intelligence**; • **Computer Vision**;

ACM Reference Format:
. 2023. Early Autism Disorder Detection Through Visualizing Eye-Tracking Patterns Using Compact Convolutional Transformers. . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 INTRODUCTION

The term "autism spectrum disorder" (ASD) refers to a group of impairments known as the "triad of impairments," which includes issues with social communication, challenges with reciprocal social connections, and abnormal sequences of repetitive activity [1]. As a consequence, people with ASD may have difficulty with social communication and interaction deficiencies in a variety of situations. Creating and keeping eye contact while natural engagement is rarely simple or spontaneous when someone has ASD diagnoses. Unfortunately, these worrying inadequacies can put a lot of stress and anxiety on people's lives and those of their families. Nevertheless, intellectual impairment or general developmental delay does not provide a more compelling explanation for these abnormalities [2]. Given this, it's fascinating to note that autistic people occasionally have exceptional talent in the arts, music, problem-solving, or mathematics.

Author's address:

ACM ISBN 978-1-4503-9622-6/22/05…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Early diagnosis of the disorder may result in early therapy, which is often advantageous for the kid and the family. The diagnosis of ASD still needs a series of cognitive testing and maybe hours of clinical examinations, making it a complex and challenging undertaking. The identification of autism has been helped by a variety of computer-aided methods in the modern world. Examples comprise Electroencephalography (EEG), Magnetic Resonance Imaging (MRI), and eye-tracking, which is the subject of this study. While anomalies of eye contact have long been recognized as the distinguishing feature of autism in general, eye-tracking technology has gained quite a bit of interest in the context of ASD.

The usage of AI-enabled technologies is also growing quickly, opening up new opportunities to enhance the diagnostic process of the eye-tracking. The process of "eye-tracking" involves creating, capturing, following, and computing eye movements or the "absolute point of sight". When the eye is fixed on the visual scene, it designates a certain place. According to a number of studies, analyzing eye movements can assist distinguish ASD symptoms from reactions to verbal or visual stimuli. Many diagnostic indices have had their accuracy increased as technology has advanced thanks to eye-tracking. Through recordings of eye movement, they assessed patients' visual attention in order to detect ASD [3]. Normal kids are more aware of certain gaze patterns when they focus on their eyes. However, autistic children exhibit intolerable social behaviors and are unable to produce typical physical reactions [4].

In this study,

3 RELATED WORK

This section outlines a selection of the relevant work that has already been done on tracking autism spectrum disorder. In [5] and [6], the authors introduced an additional dimension to the representation of eye-tracking scan paths using SensoMotoric instruments to perform eye-tracking functions and examined the dynamics of eye movements captured through images and videos by SMI RED250, and SMI RED mobile respectively. In [5], the authors used static and dynamic stimuli where the addition of nonsocial targets was necessary due to the growing correlation between ASD symptoms and nonsocial attention. The classification accuracy shows successful encapsulation of both the information about gaze motion and its underlying dynamics through scan path visualizations which later on were transformed into visual patterns. Then, Three cross-validation rounds conducted over three epochs were used to train the model with the dataset being split into train and test sets through 3-fold cross-validations using 3 stepwise procedures. Pooling layers reduced the dimensions of feature maps that were extracted by applying a convolutional kernel all over the image. Deep CNN was used in the model's implementation. The CNN model has two fully connected layers, four pooling layers, and four convolutional layers. Dropout layers were also included, helping to lessen the danger of overfitting. A prediction accuracy of around 90%, recall (approximate sensitivity of 83%), and 80% precision could be achieved by the model.

Whereas, in [6] authors used SMI Red mobile to capture the dynamic of eye movements enabling it to record on a frequency of 60Hz while calculating the velocity, and acceleration of the eye where a line has been drawn every time a position changes from [x(t), y(t)] to [x(t+1), y(t+1), with t being a predetermined period in the experiment. The trained model was a conventional logistic regression where AUC of 0.819 was achieved through 10-fold cross-validation. In both [5] and [6] the size of the output data was short. Therefore in [7], the authors suggested a k-means clustering technique that divided the dataset into four clusters to retrieve the optimal outcome from each cluster. Four eye-tracking clusters were implemented with a number of classifiers, including DT, GB, KNN, LR, MLP, NB, RF, SVM, and XGB. While RF exhibits the highest performance(74.22% accuracy, 71.53% AUC, 73.61% f-measure, 71.48% g-mean, 74.22% sensitivity, and 25.78% miss rate) over the whole dataset, MLP(88.89%

accuracy, 79.17% AUC, 87.82% f-measure, 78.57% gmean, 88.89% sensitivity, 69.44% specificity, 30.56% fall out,) showed the best result in cluster 1.

In [8], while computing the velocity of gaze movement using the coordinates/time data, the authors sought to visually convey the dynamics of gaze using color gradients. With five layers starting with a 10K input layer that matches the image dimensions, autoencoders were used to reduce the dimensionality of data. Using the K-Means approach along with various sets of characteristics, four clustering models were created having the quality assessed using the Silhouette method. In [9], the authors' approach was based on converting eye-tracking data into an image-based format. In order to visualize fixations and saccades, scanpath representation builds on this fundamental concept. The empirical machine-learning experiments had two phases where the generative modeling of eye-tracking scanpaths was one of the early stages and the original dataset was augmented using the previously produced VAE-generated images. The latent representation of scanpath pictures was investigated using VAE. The encoder and decoder of the VAE model were both built using two convolutional layers. This design was based on a basic symmetric layout. The decoder model was a "flipped" version of the encoder and CNN model was implemented for the classification where the model had 4 convolutional layers each consisting of 1 maxpool layer setting Relu as the activation function. Without data augmentation, the model had an average accuracy of 67% and after VAE-based augmentation, it reached 70%. However, the study lacks open access to eye-tracking datasets which should be considered in further investigations.

4 METHODOLOGY

4.1 Dataset

4.1.1 Datasaet 1. There are a total of 567 images where 328 images are of the non-ASD category and 219 are ASD-diagnosed images. The default dimensions of those images are 640x480. The raw data generated by the eye-tracking equipment have been used to create the scan path pictures [5].

4.1.2 Datasaet 2. There are a total of 481 images with resolutions of 1,280 × 960, 1,024 × 1,024, or 768 × 1,024 (width × height) where the images include faces of different sizes, poses, emotions, ages, genders, and so on. Among 481 images 300 has been selected. In this particular dataset Tobii T120 Eye Tracker has been used as apparatus while recording the eye movement data [10].

4.2 Data Processing

We have used zoom, shear, width shift, and other image enhancement techniques to enhance our photographs. Also, we have resized the photos to 48 × 48 pixels RGB. We raised the size of our datasets to 519 for Dataset 1 and 3480 for Dataset 2.

4.3 Proposed Model Architecture

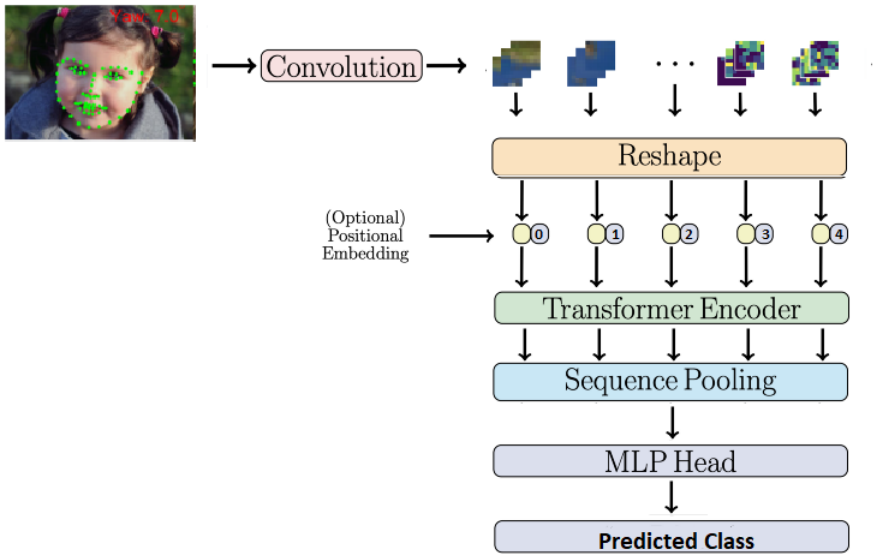

Fig. 1. Prposed Compact Convolutional Transformer (CCT) Model

For the classification job in this work, a Compact Convolutional Transformer (CCT) [11] was utilized. The Vision Transformer (Vit) is a small model for image processing difficulties, and the CCT is the most current iteration. For image processing applications, the traditional transformer is considered data-hungry. Many writers offered various solutions to this problem, including DeiT , ConViT, CvT , and T2T-ViT. The CCT model, on the other hand, outperformed all previous state-of-the-art approaches. The model was trained and tested using three datasets: small-scale and low-resolution pictures (FashionMNIST, MNIST, and CIFAR-10/100), medium-sized (Image Net), and small-scale high-resolution images (Flowers-102). It used minimal parameters, which reduced the model's temporal complexity.

First, the enhanced RGB pictures are delivered into an input layer having 48 48 3 dimensions. To retain the boundary-level information, a dual tokenizer with a kernel size of 3x3 convolutional layers and a stride size of 1 is used. The convolutional layer's resulting feature map aids in reducing the complexity of self-attention computation. As an activation function, the rectified linear activation (Relu) is employed in combination with

the ”he normal” kernel initializer with padding size 1.

Because convolutional blocks build a feature map and maintain local partial information, the model is not dependent on picture resolution. Then, on a high-dimensional feature map region, an attention-based encoder-decoder mechanism is used to extract the spatial and intensity-based relevant features in the multi-class ASD datasets. The sequence embedding layer receives all of the features, which are flattened into a 1D array. It finds information about the relative placements of picture patches within the sequence. This class embedding predicts each class of an input picture after going through the self-attention module. Each MLP block in our model includes a 30% dropout layer. We also employ MaxPooling in the convolutional blocks to cut computing costs and BatchNormalization to allow the model to learn more independently.

5 RESULTS AND ANALYSIS

Table 1. Model Hyperparameters:

Optimizer	**ADAM**
Loss Function	Binary Cross Entropy
Batch Size	32
Dropout	30
Total Parameters	407,746

The proposed model has been compiled using the Adam optimizer for both datasets while setting Binary-Cross Entropy as the loss function as the classification has been done among two classes(ASD and Non-ASD). Moreover, a dropout rate of 30% was set to prevent overfitting and the batch size was kept as 32.

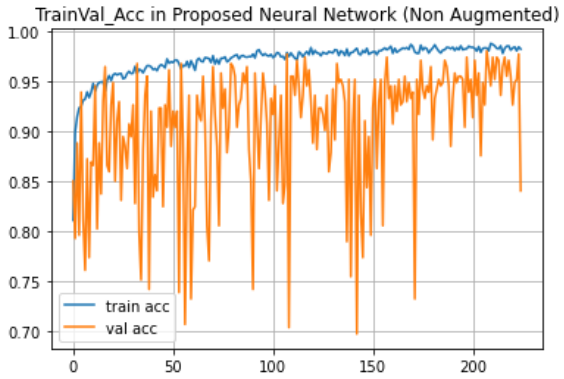

Fig. 2. Training vs validation accuracy graph of the proposed CCT model

The training was done

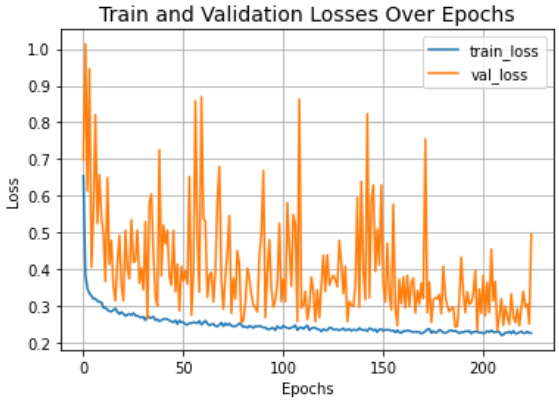

Fig. 3. Training vs validation loss graph of the proposed CCT model of dataset-1

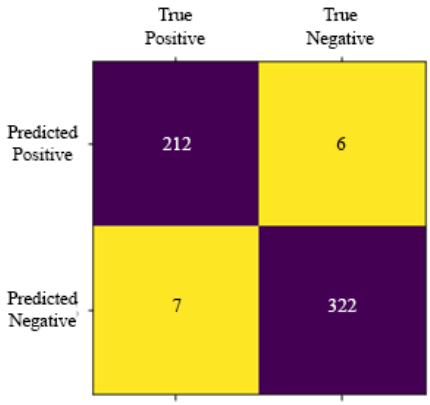

Fig. 4. Confusion matrix of our proposed model in dataset-1

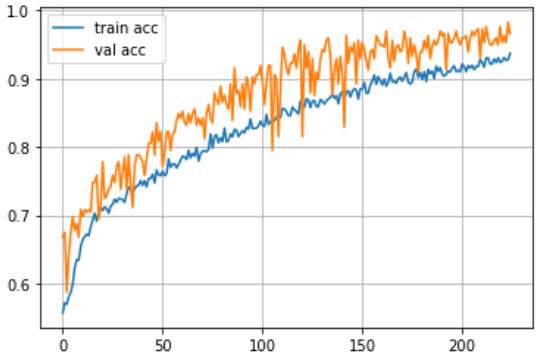

Fig. 5. Training vs validation accuracy graph of the proposed CCT model of dataset-2

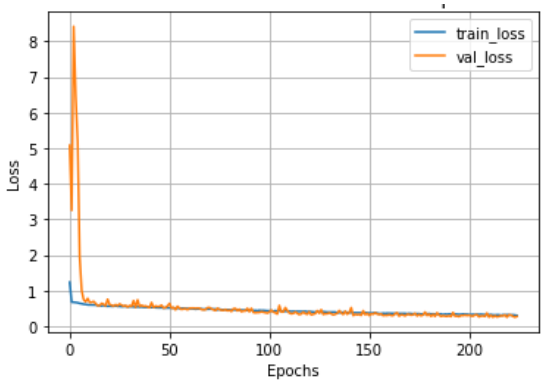

Fig. 6. Training vs validation loss graph of the proposed CCT model of dataset-2

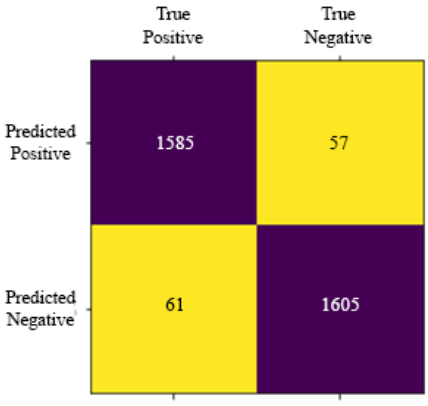

Fig. 7. Confusion matrix of our proposed model in dataset2

5.1 Result Analysis of models on Dataset-1

After examining our data on several models, some have worked splendidly achieving high accuracy with impressive precision, recall, and F1-score rate. As shown in table 2, among all of the models EfficientNetB7 and MobileNetV2 have the lowest accuracy which demonstrates that these two models' aptitude to distinguish among the spectrums of autism disorder data from normal data is quite average. These two transfer-learning models which are previously trained on Imagenet weights have 64 Million and 2.298 Million parameters respectively. The average precision, recall, and F1 score of both these model denotes that the number of False positives and False negatives have been high. Afterwards, the InceptionV3 model gained a better accuracy compared to EfficientNetB7 and MobileNetV2 which is 68.8%. The precision, recall, and F1 score show better performance as well in terms of classification. Therefore, Xception was implemented which is an extension of the Inception architecture excluding the depth-wise separable convolutions. Subsequently, in Xception the result shows higher accuracy than all the previously implemented models which is 82.18% representing better results in the other three parameters. VGG16 does slightly better than Xception achieving an accuracy of 83%. Thereafter, The Visual Transformer Model showed impressive performance on our dataset acquiring an accuracy of 91% besides gaining precision, recall, and F1 score of 90.88%, 92.22%, and 91.54% correlatively. Lastly, DenceNet201 performs slightly better with an accuracy of 91.95% where the precision rate has been 91.96%, and obtained a recall of 92.21, and an F1-score of 92.08. Densenet201 works splendidly better than all other transfer learning models used. The high accuracy denotes that this model performs to classify autism spectrums from usual spectrums and the high rate of precision, recall, and F1 score in this model denotes its capability to keep False positive and False negative rates low.

Table 2. Acquired results of different models on Dataset 1

Model	Accuracy	Precision	Recall	F1-Score	Parameters
ViT	91	90.88	92.22	91.54	21.66M
InceptionV3	68.88	72.56	64.45	68.26	21.86M
VGG16	83	80.89	84.04	82.43	14.7M
ResNet50	94.25	94.12	94.12	94.12	23.64M
EfficientNetB7	59.77	59.74	59.74	59.74	64M
MobileNetV2	59.77	59.9	59.9	59.9	2.298M
Xception	82.18	82.50	82.18	82.33	20.8M
DenseNet201	91.95	91.96	92.21	92.08	18.35M
Proposed CCT	97.62	97.25	96.80	97.03	0.4 M

5.2 Result Analysis of models on Dataset-2

As can be seen from the confusion matrix in figure 7, 212 samples of ASD were successfully predicted as positive while 7 of them were wrongly predicted as Non-ASD images. Meanwhile, 322 samples of Non-ASD were predicted as Non-ASD images successfully where only 6 images were wrongly predicted as ASD-positive images.

After reviewing our data on numerous models, we observed that some performed brilliantly, reaching high accuracy with great precision, recall, and F1-score rate. According to table 3, EfficientNetB7 and MobileNetV2 have the lowest accuracy of all models, showing that their ability to distinguish between autism spectrum disorder data and normal data is quite average. These two previously trained transfer-learning models with Imagenet weights have 64 Million and 2,298 Million parameters, respectively. Both of these models have average precision,

Table 3. Acquired results of different models on Dataset 2

Model	Accuracy	Precision	Recall	F1-Score	Parameters
ViT	87	89.88	90.22	90.04	21.66M
InceptionV3	68.68	72.56	64.45	68.26	21.86M
VGG16	87.22	81.55	85.55	83.55	14.7M
ResNet50	93.79	94.22	94.79	94.5	23.64M
EfficientNetB7	55.36	58.18	58.18	58.18	64M
MobileNetV2	56.77	55.7	55.9	55.8	2.298M
Xception	80.18	80.79	79.38	80.07	20.8M
DenseNet201	90.95	92.76	91.41	92.08	18.35M
Proposed CCT	96.43	95.53	96.29	96.43	0.4M

recall, and F1 scores that imply a high number of False positives and False negatives. The InceptionV3 model then achieved 68.8% accuracy in comparison to EfficientNetB7 and MobileNetV2. In terms of categorization, the precision, recall, and F1 score all demonstrate enhanced performance. Consequently, Xception, an extension of the Inception architecture that lacks depthwise separable convolutions, was developed. Consequently, Xception achieves greater accuracy than all previously deployed models, with 79.38% of the other three criteria yielding superior outcomes. VGG16 outperforms Xception, with an accuracy of 87.22%. The Visual Transformer Model then performed brilliantly on our dataset, attaining an accuracy of 87% along with precision, recall, and F1 scores of 89.88, 90.22, and 90.04%, respectively. Dencenet201 has an accuracy of 90.95%, a precision rate of 92.76 a recall of 91.41%, and an F1-score of 92.08%. Densenet201 outperforms all other transfer learning models examined. This model's excellent accuracy indicates its capacity to differentiate autism spectrums from other spectrums, while its high precision, recall, and F1 score suggest its capacity to maintain low False positive and False negative rates.

The usage of a Compact Convolutional Transformer (CCT) led to more accuracy than any other model in this dataset version. The overall accuracy was 96.43%, including 95.% precision, 96.29 recall, and 96.% F1 score.
The parameter of the Compact Convolutional Transformer (CCT) is 0.4M, which is significantly less than the parameters of all previously developed models. We deployed the Adam optimizer with 32-person batches and a 30% dropout rate. Because we sought to separate ASD scan routes from ordinary scan paths, we employed binary cross entropy as the loss function in our binary classification. The training procedure was carried out for 280 epochs, as depicted in the dataset-2 Training versus Validation accuracy and loss graph, and in our proposed non-augmented neural network model, the validation and training accuracy and losse overlap substantially which shown in figure 5, 6 indicating that overfitting is minimal.

6 CONCLUSION AND FUTURE WORKS

Autism spectrum disorder (ASD) refers to a group of impairments known as the "triad of impairments," which consists of difficulties with social communication, difficulties with reciprocal social connections, and abnormal patterns of repetitive behaviour. As a result, individuals with ASD may experience difficulties with social communication and interaction in a variety of settings. When a person has been diagnosed with ASD, making and maintaining eye contact during natural activities is rarely effortless or spontaneous. Unfortunately, these issues can cause a great deal of stress and anxiety in the lives of individuals and their families. However, impairments in intelligence or general developmental delay do not provide a more convincing explanation for these anomalies. Given this, it is fascinating to note that autistic individuals occasionally possess exceptional talent in the arts, music, problem-solving, and mathematics. Early diagnosis of the disorder may result in early treatment,

which is often advantageous for the child and the family. The diagnosis of autism spectrum disorder (ASD) still requires a battery of cognitive tests and possibly hours of clinical examinations, making it a complex and challenging endeavour. Modern identification of autism has been aided by a variety of computer-assisted techniques. Electroencephalography (EEG), Magnetic Resonance Imaging (MRI), and eye-tracking, the subject of this study, are a few examples of neuroimaging techniques. While abnormalities in eye contact have long been recognized as a defining characteristic of autism in general, eye-tracking technology has recently received a great deal of attention in relation to ASD. Utilization of AI-enabled technologies is also on the rise, creating new opportunities to improve eye-tracking diagnostics. Eye-tracking is the process of generating, capturing, observing, and computing eye movements or the "absolute point of sight." When the eye is fixed on a visual scene, a specific location is designated. Multiple studies have found that analyzing eye movements can help differentiate ASD symptoms from verbal or visual stimulus reactions. Due to technological advancements such as eye-tracking, the precision of numerous diagnostic indices has increased. In order to detect ASD, they assessed the patient's visual attention using eye movement recordings. When typical children focus on their eyes, they become more cognizant of particular gaze patterns. In contrast, autistic children display intolerable social behaviours and are incapable of producing typical physical responses.

REFERENCES

[1] Lorna Wing and Judith Gould. Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders*, 9:11–29, 1979.

[2] A American Psychiatric Association, American Psychiatric Association, et al. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 10. Washington, DC: American psychiatric association, 2013.

[3] Gail J Walker-Smith, Alastair G Gale, and John M Findlay. Eye movement strategies involved in face perception. *Perception*, 6(3):313–326, 1977.

[4] Jiannan Kang, Xiaoya Han, Jiajia Song, Zikang Niu, and Xiaoli Li. The identification of children with autism spectrum disorder by svm approach on eeg and eye-tracking data. *Computers in biology and medicine*, 120:103722, 2020.

[5] Federica Cilia, Romuald Carette, Mahmoud Elbattah, Gilles Dequen, Jean-Luc Guérin, Jérôme Bosche, Luc Vandromme, Barbara Le Driant, et al. Computer-aided screening of autism spectrum disorder: eye-tracking study using data visualization and deep learning. *JMIR Human Factors*, 8(4):e27706, 2021.

[6] Romuald Carette, Mahmoud Elbattah, Gilles Dequen, Jean-Luc Guérin, and Federica Cilia. Visualization of eye-tracking patterns in autism spectrum disorder: method and dataset. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, pages 248–253. IEEE, 2018.

[7] Tania Akter, Mohammad Hanif Ali, Md Imran Khan, Md Shahriare Satu, and Mohammad Ali Moni. Machine learning model to predict autism investigating eye-tracking dataset. In *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, pages 383–387. IEEE, 2021.

[8] Mahmoud Elbattah, Romuald Carette, Gilles Dequen, Jean-Luc Guérin, and Federica Cilia. Learning clusters in autism spectrum disorder: Image-based clustering of eye-tracking scanpaths with deep autoencoder. In *2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 1417–1420. IEEE, 2019.

[9] Mahmoud Elbattah, Colm Loughnane, Jean-Luc Guérin, Romuald Carette, Federica Cilia, and Gilles Dequen. Variational autoencoder for image-based augmentation of eye-tracking data. *Journal of Imaging*, 7(5):83, 2021.

[10] Huiyu Duan, Xiongkuo Min, Yi Fang, Lei Fan, Xiaokang Yang, and Guangtao Zhai. Visual attention analysis and prediction on human faces for children with autism spectrum disorder. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–23, 2019.

[11] Aqeel Iftikhar Jajja, Assad Abbas, Hasan Ali Khattak, Gniewko Niedbała, Abbas Khalid, Hafiz Tayyab Rauf, and Sebastian Kujawa. Compact Convolutional Transformer (CCT)-Based Approach for Whitefly Attack Detection in Cotton Crops. https://www.mdpi.com/2077-0472/12/10/1529, sep 23 2022.