Vision-based Approach for Autism Diagnosis using Transfer Learning and Eye-tracking

Mahmoud Elbattah[1,2], Jean-Luc Guérin[1], Romuald Carette[3], Federica Cilia[4] and Gilles Dequen[1]
[1] *Laboratoire MIS, Université de Picardie Jules Verne, Amiens, France*
[2] *Faculty of Environment and Technology, University of the West of England, Bristol, U.K.*
[3] *Evolucare Technologies, Villers-Bretonneux, France*
[4] *Laboratoire CRP-CPO, Université de Picardie Jules Verne, Amiens, France*

Keywords: Autism Spectrum Disorder, Eye-tracking, Deep Learning, Transfer Learning.

Abstract: The potentials of Transfer Learning (TL) have been well-researched in areas such as Computer Vision and Natural Language Processing. This study aims to explore a novel application of TL to detect Autism Spectrum Disorder. We seek to develop an approach that combines TL and eye-tracking, which is commonly used for analyzing autistic features. The key idea is to transform eye-tracking scanpaths into a visual representation, which could facilitate using pretrained vision models. Our experiments implemented a set of widely used models including VGG-16, ResNet, and DenseNet. Our results showed that the TL approach could realize a promising accuracy of classification (ROC-AUC up to 0.78). The proposed approach is not claimed to provide superior performance compared to earlier work. However, the study is primarily thought to convey an interesting aspect regarding the use of (synthetic) visual representations of eye-tracking output as a means to transfer representations from models pretrained on large-scale datasets such as ImageNet.

1 INTRODUCTION

Autism Spectrum Disorder (ASD) is a neuro-developmental disorder, which is characterized by various impairments, mainly social communication and interaction issues, and repetitive behavior (American Psychiatric Association, 2013). ASD-diagnosed individuals usually suffer from troubles in interaction and communication in multiple forms. The most remarkable symptom is the poor development of non-verbal skills such as the lack or absence of eye contact. With such deficits, a considerable strain can be placed on the well-being of autistic individuals and families as well. From an economic perspective, it was estimated that autism costs the UK, for example, more than heart disease, cancer, and stroke combined (Buescher et al., 2014).

The early intervention for autism is highly favorable to realize common benefits for children and their families (Estes et al., 2015). Multiple studies (e.g., Smith et al., 2000; Dawson et al., 2010) reported improved outcomes of treatment such as intellectual capacity, communication, adaptive behavior, and educational support. However, the diagnosis of autism has been considered as a challenging task. Typically, the diagnosis process includes a variety of cognitive tests that could require hours of intensive clinical examinations. Furthermore, standardized tests require a considerable amount of time and effort, and the diversity of symptoms increase the complexity of identifying an accurate classification.

A major part of the psychological research endeavored to develop assistive instruments based on observational measures or diagnostic interviews. Examples include Childhood Autism Rating Scale (CARS) (Schopler et al., 1980), Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 1989), and Autism Diagnostic Interview (ADI-R) (Lord, Rutter, and Le Couteur, 1994). More recently, computer-aided methods have come into prominence to support the diagnosis process at different levels. In particular, the eye-tracking technology has received an extensive research interest. The literature is rife with studies that analyzed eye-tracking recordings as an effective means to identify the ASD symptoms (e.g., Eraslan et al., 2019; Harrison, and Slane, 2020; Greene et al., 2021). In addition, Artificial Intelligence (AI) is being increasingly combined with eye-tracking. The state-of-the-art techniques, such as Deep Learning, have demonstrated promising

Elbattah, M., Guérin, J., Carette, R., Cilia, F. and Dequen, G.
Vision-based Approach for Autism Diagnosis using Transfer Learning and Eye-tracking.
DOI: 10.5220/0010975500003123
In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Volume 5: HEALTHINF*, pages 256-263
ISBN: 978-989-758-552-4; ISSN: 2184-4305

potentials to advance the development of diagnostic applications (e.g., Heinsfeld et al., 2018; Eslami, and Saeed, 2019).

In this respect, this study explores a Transfer Learning (TL) approach. The TL concept has deemed as an attractive path for a multitude of healthcare and medical applications, which largely suffer from data paucity and imbalance. TL-based methods allow for using sophisticated models pretrained on huge datasets such as ImageNet (Deng et al., 2009). With many successful applications (e.g., Khan et al., 2019; Ahuja et al., 2021), TL is well-poised to be the path forward to deal with such challenges.

We adopt a vision-based methodology using the visual representation of eye-tracking output for learning the patterns associated with ASD. Accordingly, the detection of autism could be approached as a problem of image classification. The approach applicability was experimented with popular pretrained models including VGG-16 (Simonyan, and Zisserman, 2014), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017).

The contributions of this study can be considered as follows. We explore a novel application of TL using a visual representation of eye-tracking, which has not been proposed yet, to the best of our knowledge. The interesting aspect of our methodology is using a synthetic visual representation as a means to allow for transferring representations from Deep Learning models pretrained on large-scale datasets. Further, the study could be generally regarded as an addition to the ongoing contributions of applying Machine Learning (ML) in the context of ASD diagnosis.

2 RELATED WORK

Increasing evidence suggests that the application of ML could have a strong impact on the autism research. This section aims to review part of the recent advances in terms of data representation and model architectures.

A variety of feature sets has been experimented for developing ML models to help detect the symptoms of ASD. This review focuses on two modalities of data including facial expressions and eye-gaze movements. The review is unavoidably selective rather than exhaustive whereas it only highlights potential ML approaches in this context. However, comprehensive literature reviews are provided by excellent survey studies such as (Ahmed, and Jadhav, 2020) and (de Belen et al., 2020)..

2.1 Facial Expressions

Seminal psychology studies (e.g., Hobson, 1986; Weeks, and Hobson, 1987) intensively discussed the impairments of social interaction among ASD-diagnosed individuals. Moreover, clinical studies (e.g., Loth et al., 2018) largely confirmed that facial expressions and emotions could effectively serve as markers for autism.

Building upon the hypothesis of discriminative facial attributes, a variety of ML applications were developed to detect ASD. For instance, (Li et al., 2019) applied Deep Learning for ASD classification using facial expressions. They developed an end-to-end system using Convolutional Neural Networks (CNN). Their results showed that the representation of different facial attributes was statistically significant. Specifically, the ASD classification could be improved by about 7% with F1-score of 76%.

Likewise, (Shukla et al., 2017) utilized facial images for recognizing developmental disorders including ASD. A multi-channel CNN architecture was developed where facial images were divided into four parts and fed separately into CNNs to perform the feature extraction. The extracted representations were combined altogether into a global feature map. Eventually, a linear Support Vector Machine was trained to perform the classification task. In general, the literature includes similar contributions (e.g., Han et al., 2018; Tang et al., 2018) that experimented different ML approaches using facial images.

2.2 Eye-gaze Movement

The behavior of eye gaze has been widely recognized as one of the hallmarks of ASD. The psychology literature has consistently included autism studies that analyzed the eye gaze in response to verbal or visual cues (e.g., Habayeb et al., 2021; Fukui et al., 2021). Therefore, the eye-tracking technology has been adopted in multiple aspects in the ASD research.

In this respect, plentiful contributions aimed to combine eye-tracking methods with a variety of ML applications in various ways and for different purposes. A number of studies (e.g., Liu et al., 2021; Liu, Li, and Yi, 2016) developed ML approaches to detect the autistic patterns of eye movement. Their results presented promising evidence for applying ML methods to identify ASD-diagnosed children, with a classification accuracy up to around 88.5%.

More sophisticated approaches were employed as well. For example, a Deep Learning model was implemented to detect autism using eye-tracking tasks of free-image viewing (Jiang, and Zhao, 2017).

Deep Learning was utilized to extract features automatically from a collection of discriminative images. Likewise, (Chong et al., 2017) applied a CNN-based architecture for the detection of eye contact during social interactions. Their results reported a precision and recall of 76% and 80%, respectively. Another Deep Learning-based framework was developed for ASD screening using photo-taking tasks (Chen, and Zhao, 2019). LSTM models were implemented for encoding the temporal information of eye movements.

More recently, (Elbattah et al., 2020) adopted sequence learning in eye-tracking data. A set of Natural Language Processing (NLP) methods were applied to transform the saccadic eye movements into text-like sequences. The NLP-based approach could allow for transforming high-dimensional eye-tracking data into an amenable representation for ML. The applicability of their approach was experimented using a set of classification models including CNN and LSTM. The best performing CNN model could achieve a ROC-AUC of about 0.84.

The authors undoubtedly recognize all those efforts that sought to advance ASD diagnosis with the help of ML. We hope for this study to contribute to the mounting efforts by exploring the TL approach and its applicability.

3 DATA DESCRIPTION

The dataset under consideration was collected as part of earlier work (Carette et al., 2018) related to the study of autism using eye-tracking. The following sections describe the characteristics of participants and experimental settings of the original study.

3.1 Participants

A group of 59 school-aged children were recruited from French schools in the region of Hauts-de-France. The age of participants was generally in the range of 3 to 12 years old. It was highly aimed for the participants to be at an early stage of development.

A parental permission was acquired for every participant to take part in the eye-tracking experiments. Further, the parents were acquainted with the research goals through orientation sessions.

Initially, the participants were organized based on a basic binary grouping as: i) Typically Developing (TD), and ii) ASD. In addition, the CARS score (Schopler et al., 1980) was employed to classify the severity of autism more precisely. The CARS method has been widely applied in the Psychology practice for describing the severity of ASD symptoms (Ozonoff et al., 2005). The scale includes various ratings on different behavioral aspects (e.g., verbal communication, activity level). Table 1 summarizes the characteristics of participants.

Table 1: Statistics of participants (Carette et al., 2018).

Number of Participants (ASD, TD)	59 (29, 30)
Gender Distribution (M, F)	38 (≈ 64%), 21 (≈ 36%)
Age (Mean, Median) years	7.88, 8.1
CARS Score (Mean, Median)	32.97, 34.50

3.2 Experimental Protocol

The experiments were conducted using an eye-tracker of the SMI Red-M model, which is a screen-based eye-tracker with 60Hz sampling rate. The eye-tracker was operated along with a 17-inch monitor. The screen resolution was 1280x1024.

The eye-tracking experiments included a set of photos and video scenarios, which were particularly designed to stimulate the eye gaze across the screen. The participants were seated at approximately 60-cm distance away from the monitor. A quiet room at the university campus was used for running our experiments. In addition, physical barriers were applied around the screen to avoid visual distractions.

The content and length of videos varied to allow for analyzing the ocular activity from different aspects and levels. We used other stimuli provided by the SMI Experiment Center Software. The Stimuli included a variety of types, typically used in the eye-tracking research. For instance, static and dynamic naturalistic scenes with and without receptive language, static face or objects and cartoons stimuli, and other joint attention stimuli. In addition, some videos included human presenters. The presenter's role was generally to turn the participant's attention to elements, which could be visible or invisible around the display area.

Eye-tracking experiments usually took about 5 minutes. The participants were inspected with respect to the quality of eye contact with the presenter, and the level of focus on other elements. A five-point scheme of calibration was applied. A set of verification procedures followed the calibration scheme. A set of 25 eye-tracking experiments was conducted to produce the output dataset. The dataset was stored in multiple CSV files, which collectively included more than 2M records.

3.3 Ethical Approval

The study received the ethical approval by the ethics committee of Rouen University (Reference: 2016-02-B). The CNIL (Commission nationale de l'informatique et des libertés) declaration number of research conformity is 2208663v0.

The study fulfils the principles and terms of the 1964 Helsinki declaration. Before starting the study, the approval was obtained from the heads of the regional and district education authorities, as well as the head and the teachers of the particpating school. The parents of participants had also given their written informed consent.

4 OUR APPROACH

The raw eye-tracking output is high-dimensional, which typically contains many variables describing the gaze position and eye movements. Therefore, the main challenge is to find out a feature representation suitable for building ML models.

In this respect, our approach aimed at developing a visual representation of eye-tracking data. The basic idea was to compactly render long-tailed eye-tracking recordings into a set of images that can visually describe the gaze behavior. In this manner, the prediction problem could be approached as an image classification task. Furthermore, the visual representation was used as a bridge to allow for applying high-performing models pretrained on large-scale image datasets.

The following sections elaborate further on the key components of our approach. First, we describe the method of transforming the raw eye-tracking data into images. Second, we refer to the pretrained models used in our experiments as described later.

4.1 Visualization of Eye-tracking Scanpaths

Scanpaths are commonly utilized in the eye-tracking context as to depict the gaze behavior in a visual manner. A scanpath basically represents a sequence of consecutive fixations and saccades as a trace through time and space (Goldberg, and Helfman, 2010). A fixation describes the brief moments while the eye gaze is paused on a particular object, which allow the brain to perform the perception process. The average duration of fixation was estimated to be around 330ms (Henderson, 2003). While saccades include a constant scanning with very rapid and short eye movements. Saccades consist of quick ballistic jumps of 2º or longer, which continue for about 30–120ms (Jacob, 1995).

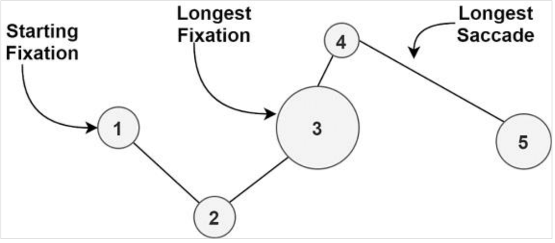

Figure 1: Eye-tracking scanpath, adapted from (Goldberg, and Helfman, 2010).

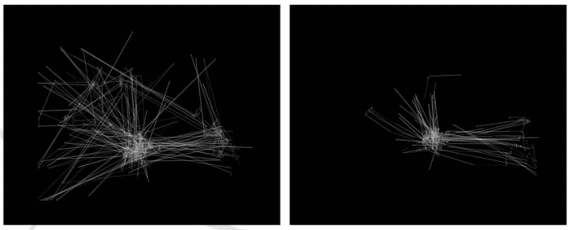

Figure 2: Samples of the scanpath visualization (Carette et al., 2018). The left-sided image relates to an ASD participant, while the right-sided image is for a TD.

Figure 1 gives a basic scanpath example, which includes a few fixations and saccades. As it appears, the fixations are shown as circles, and the saccades as lines connecting those fixations. The diameter of a fixation indicates the duration, while the line length represents the continuation of saccades.

Our approach was based on transforming the eye-tracking scanpaths into an image-based format. Our representation of scanpaths follows on the core idea of visualizing fixations and saccades. Further, it was aimed to visually encode the dynamics of gaze using color gradients. Given the coordinates and timing of eye-tracking records, the velocity of gaze movement could be calculated. Using a grayscale spectrum, the color values were tuned based on the magnitude of velocity with respect to time. The visualizations were produced using Python with the Matplotlib library (Hunter, 2007). A more comprehensive presentation of that part is elaborated in our earlier work (Carette et al., 2018).

The outcome of data transformation process was an image dataset containing more than 500 images. Specifically, 328 images related to the TD participants, and another 219 images for the ASD-diagnosed. The default image dimensions were set as 640×480. The dataset along with metadata files have been made publicly available on the Figshare repository (Figshare, 2019). Figure 2 presents sample images from the dataset.

4.2 Transfer Learning

TL is a rapidly developing approach in the arena of Deep Learning, and AI in general. TL allows for extracting knowledge from one or more source tasks towards the application to another target task (Pan, and Yang, 2009). Thanks to the visual representation of eye-tracking data, the source task could be linked to image classification in a broad sense. As such, we were able to consider the state-of-the-art models of Computer Vision.

A set of popular TL models were experimented. The models were originally pretrained using the ImageNet dataset (Deng et al., 2009). Specifically, our experiments included the following models:

i) VGG-16: The VGG-16 model is a deep CNN architecture developed by a group of researchers from the University of Oxford (Simonyan, and Zisserman, 2014). The model could achieve an excellent accuracy of 92.7% with ImageNet data, over 14M images organized into 1000 categories. As its name suggests, the VGG-16 architecture includes 16 layers. More specifically, the model includes a couple of blocks of 2 convolutional layers followed by a max-pooling operation. Subsequently, it has a sequence of 3 blocks including 3 convolutional layers followed by max-pooling. Eventually, the output is connected to a number of fully connected dense layers.

ii) ResNet: Residual Networks (ResNet) (He et al., 2016) have become the mainstream architecture for many tasks of Computer Vision. The fundamental breakthrough made by the ResNet architecture was to allow for training extremely deep neural networks. The key idea of ResNet was the so-called “skip connection”, which can add the original input to the output of the convolution block. This was a major contribution to address the problem of vanishing gradients (Hochreiter, 1998). There are variants of the ResNet architectures (e.g. ResNet50, ResNet101). In our case, the ResNet50 model was applied only. The ResNet50 model includes 48 convolutional layers along with a max-pooling layer, and another average-pooling layer.

iii) DenseNet: The Densely Connected Convolutional Network (DenseNet) architecture (Huang et al., 2017) can be conceived as a further extension of the ResNet. The basic idea of the DenseNet architecture is that each layer is connected to each other layer. As such, the feature maps of all the preceding layers are used as inputs for each layer, and its own feature maps are used as input for each subsequent layer. There are variants of the DenseNet architectures including the DenseNet-121, DenseNet-160, and DenseNet-201. The number denotes the number of layers in the model (i.e. 121 layers, 160 layers, etc.). In our case, we only applied the DenseNet-121 architecture.

5 EXPERIMENTS

5.1 Preprocessing and Augmentation

At the outset, a set of image processing techniques was applied as follows. First, the blank background was cropped out as far as possible. The cropping was implemented using the OpenCV library (Bradski, G., 2000). Second, all images were consistently scaled down to 256x256 dimensions.

Further, data augmentation was applied to enlarge the dataset, and increase the diversity of images. Augmentation is a prevalent approach to help models generalize better and reduce the risk of overfitting. The ML literature largely reported that augmentation could improve the performance of classification models (e.g., Xu et al. 2016; Perez, and Wang, 2017).

In our case, each scanpath image in the dataset was augmented with five synthetic samples. The synthetic samples were created using a random set of image transformations such as rotation, and shearing. The implementation of data augmentation was largely facilitated thanks to the Keras library (Chollet, 2015), which comes with built-in functionalities for that purpose. The synthetic set was added exclusively to the train data during the model training as explained in the following part.

5.2 Data Splitting

The dataset was partitioned into train and test sets based on a 3-fold cross-validation procedure. The split process was conducted over two stages to ensure that images of a particular participant would exist either in the train or test sets. Specifically, the dataset was split using the following stepwise procedures:

1. **Split Participants:** Initially, the group of 59 participants were randomly split into two independent sets (i.e., train and test).
2. **Match Images:** Based on the IDs of participants, the scanpath images were matched and loaded into the train and test sets. The images produced earlier by the

augmentation were used exclusively in the train set based on participant IDs as well.

3. **Repeat:** Step #1 and Step #2 were repeated for each round of the cross-validation.

5.3 Classification Models

Our experiments included a set of three classifiers based on the pretrained models mentioned earlier. Each classification model consisted of two main parts as follows:

- **Base Model:** The pretrained convolutional model for performing the feature extraction.
- **Classifier Model:** The model used to classify the input image based on the features extracted by the convolutional base.

It is also important to note that those pretrained models were originally trained against the classes of ImageNet, which consisted of 1000 labels. Therefore, we removed the final layers that perform the classification task. Instead, we added a simple MLP module on top of each base model.

The MLP module consisted of three layers including 256, and 128 neurons, along with the final output layer. On the one hand, the MLP weights were trainable during the model training. On the other hand, the base model weights were fully transferred. The ReLU activation function was used at the MLP front layers. While the output layer used a sigmoid function, which was sufficient to perform the binary classification task (i.e., TD or ASD).

5.4 Performance Evaluation

Three rounds of cross-validation were applied for evaluating the performance of each model. Figure 3 compares the ROC curves. As it appears, the VGG-16 model could achieve the best performance with ROC-AUC≈0.78. However, the ResNet and DenseNet models could achieve quite comparable performance with ROC-AUC of about 0.77 and 0.76, respectively. Table 2 analyzes the performance with respect to precision and recall. Likewise, the VGG-16 had the highest values.

The experiments were run on the Google Cloud platform using a VM including a single GPU (P-100 Nvidia), and 25GB RAM. All classification models were implemented using Keras (Chollet, 2015) along with TensorFlow backend (Abadi et al., 2016).

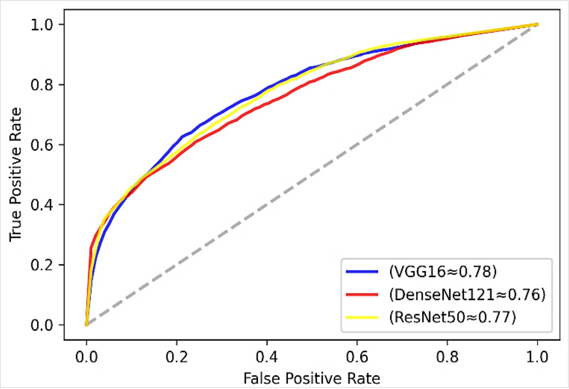

Figure 3: ROC curves.

Table 2: Performance analysis of models.

Model	Recall (~)	Precision (~)
VGG-16	**0.56**	**0.67**
ResNet	0.54	0.65
DenseNet	0.55	0.65

6 CONCLUSIONS

This paper presented a vision-based approach for detecting autism using Transfer Learning (TL) and eye-tracking. The key idea presented here was to utilize a visual representation of eye-tracking recordings as an enabler to approach sophisticated models pretrained on image classification tasks.

Interestingly, popular vision models such as VGG-16, ResNet, and DenseNet could achieve a quite promising performance. It turned out that the TL approach was largely applicable, even though the source dataset (i.e., ImageNet) is assumed to have included quite different types of images. This could translate into the suitability of synthetic visual representations. Compared to earlier work (e.g., Carette et al., 2019), it is not claimed at all that the TL approach could provide superior performance. However, it is conceived that the scarcity or imbalance of datasets could make such TL approaches attractive for further investigation.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. *In Proceedings of the 12th {USENIX} Symposium on Operating Systems Design and Implementation* (pp. 265-283).

Ahmed, Z. A. T., & Jadhav, M. E. (2020). A Review of Early Detection of Autism Based on Eye-Tracking and Sensing Technology. *In Proceedings of the International Conference on Inventive Computation Technologies (ICICT)* (pp. 160-166). IEEE.

Ahuja, S., Panigrahi, B. K., Dey, N., Rajinikanth, V., & Gandhi, T. K. (2021). Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices. *Applied Intelligence*, 51(1), 571-585.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (DSM-5®). American Psychiatric Pub.

Bradski, G. (2000). The OpenCV library. *Dr Dobb's Journal of Software Tools*, 25, 120-125.

Buescher, A. V., Cidav, Z., Knapp, M., & Mandell, D. S. (2014). Costs of autism spectrum disorders in the United Kingdom and the United States. *JAMA Pediatrics*, 168(8), 721-728.

Carette, R., Elbattah, M., Dequen, G., Guérin J.L., & Cilia, F. (2018). Visualization of eye-tracking patterns in autism spectrum disorder: Method and dataset. *In Proceedings of the 13th International Conference on Digital Information Management (ICDIM 2018)*. IEEE.

Carette, R., Elbattah, M., Cilia, F., Dequen, G., Guérin, J.L., & Bosche, J. (2019). Learning to predict autism spectrum disorder based on the visual patterns of eye-tracking scanpaths. *In Proceedings of the 12th International Conference on Health Informatics (HEALTHINF).*

Chen, S., & Zhao, Q. (2019). Attention-based autism spectrum disorder screening with privileged modality. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1181-1190).

Chollet, F. (2015). Keras. Github Repository: https://github.com/fchollet/keras

Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R. M., ... & Rehg, J. M. (2017). Detecting gaze towards eyes in natural social interactions and its use in child assessment. *In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3), 1-20.

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., ... & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics*, 125(1), e17-e23.

de Belen, R. A. J., Bednarz, T., Sowmya, A., & Del Favero, D. (2020). Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational Psychiatry*, 10(1), 1-20.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248-255). IEEE.

Elbattah, M., Guérin, J. L., Carette, R., Cilia, F., & Dequen, G. (2020, December). NLP-Based Approach to Detect Autism Spectrum Disorder in Saccadic Eye Movement. *In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1581-1587). IEEE.

Eraslan, S., Yaneva, V., Yesilada, Y., & Harper, S. (2019). Web users with autism: eye tracking evidence for differences. *Behaviour & Information Technology*, 38(7), 678-700.

Eslami, T., & Saeed, F. (2019). Auto-ASD-network: a technique based on deep learning and support vector machines for diagnosing autism spectrum disorder using fMRI data. *In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (pp. 646-651).

Estes, A., Munson, J., Rogers, S. J., Greenson, J., Winter, J., & Dawson, G. (2015). Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 54(7), 580-587.

Figshare. (2019). Visualization of Eye-Tracking Scanpaths in Autism Spectrum Disorder: Image Dataset. Available online: https://figshare.com/s/5d4f93395cc49d01e2bd

Fukui, T., Chakrabarty, M., Sano, M., Tanaka, A., Suzuki, M., Kim, S., ... & Wada, M. (2021). Enhanced use of gaze cue in a face-following task after brief trial experience in individuals with autism spectrum disorder. *Scientific Reports*, 11(1), 1-10.

Greene, R. K., Parish-Morris, J., Sullivan, M., Kinard, J. L., Mosner, M. G., Turner-Brown, L. M., ... & Dichter, G. S. (2021). Dynamic Eye Tracking as a Predictor and Outcome Measure of Social Skills Intervention in Adolescents and Adults with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 51(4), 1173-1187.

Goldberg, J. H., & Helfman, J. I. (2010). Visual scanpath representation. *In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 203-210). ACM.

Habayeb, S., Tsang, T., Saulnier, C., Klaiman, C., Jones, W., Klin, A., & Edwards, L. A. (2021). Visual traces of language acquisition in toddlers with autism spectrum disorder during the second year of life. *Journal of Autism and Developmental Disorders*, 51(7), 2519-2530.

Han, J., Li, X., Xie, L., Liu, J., Wang, F., & Wang, Z. (2018). Affective Computing of Childern with Authism Based on Feature Transfer. *In Proceedings of the 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* (pp. 845-849). IEEE.

Harrison, A. J., & Slane, M. M. (2020). Examining how types of object distractors distinctly compete for facial attention in autism spectrum disorder using eye tracking. *Journal of Autism and Developmental Disorders*, 50(3), 924-934.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778).

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17, 16-23.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11), 498-504.

Hobson, R. P. (1986). The autistic child's appraisal of expressions of emotion. *Journal of Child Psychology and Psychiatry*, 27(3), 321-342.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02), 107-116.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4700-4708).

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90-95.

Jacob, R. (1995). Eye tracking in advanced interface design. In W. Barfield W, T.A. Furness (eds), *Virtual Environments and Advanced Interface Design*, pp. 258–288. New York: Oxford University Press.

Jiang, M., & Zhao, Q. (2017). Learning visual attention to identify people with autism spectrum disorder. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 3267-3276).

Khan, S., Islam, N., Jan, Z., Din, I. U., & Rodrigues, J. J. C. (2019). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125, 1-6.

Li, B., Mehta, S., Aneja, D., Foster, C., Ventola, P., Shic, F., & Shapiro, L. (2019). A facial affect analysis system for autism spectrum disorder. *In Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 4549-4553). IEEE.

Liu, W., Yu, X., Raj, B., Yi, L., Zou, X., & Li, M. (2015). Efficient autism spectrum disorder prediction with eye movement: A machine learning framework. *In Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 649-655). IEEE.

Liu, W., Li, M., & Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9(8), 888-898.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19(2), 185-212.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5), 659-685.

Loth, E., Garrido, L., Ahmad, J., Watson, E., Duff, A., & Duchaine, B. (2018). Facial expression recognition as a candidate marker for autism spectrum disorder: how frequent and severe are deficits?. *Molecular Autism*, 9(1), 1-11.

Ozonoff, S., Goodlin-Jones, B. L., & Solomon, M. (2005). Evidence-based assessment of autism spectrum disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, 34(3), 523-540.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Schopler, E., Reichler, R. J., DeVellis, R. F., & Daly, K. (1980). Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *Journal of Autism and Developmental Disorders*, 10(1), 91-103.

Shukla, P., Gupta, T., Saini, A., Singh, P., & Balasubramanian, R. (2017). A deep learning frame-work for recognizing developmental disorders. *In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 705-714). IEEE.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556.

Smith, T., Groen, A. D., & Wynn, J. W. (2000). Randomized trial of intensive early intervention for children with pervasive developmental disorder. *American Journal on Mental Retardation*, 105(4), 269-285.

Tang, C., Zheng, W., Zong, Y., Cui, Z., Qiu, N., Yan, S., & Ke, X. (2018). Automatic smile detection of infants in mother-infant interaction via CNN-based feature learning. *In Proceedings of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data* (pp. 35-40).

Weeks, S. J., & Hobson, R. P. (1987). The salience of facial expression for autistic children. *Journal of Child Psychology and Psychiatry*, 28(1), 137-152.

Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., & Jin, Z. (2016). Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv preprint arXiv*:1601.03651.