Eyes on Autism: Utilizing Deep Learning Techniques for Classifying Autism Spectrum Disorder Through Eye Gaze and Pupil Responses

Søren Søndergaard Meiner (SM), 202205445@post.au.dk

Thorkild Kappel (TK), 202207326@post.au.dk

School of Communication and Culture, Aarhus University

Jens Chr. Skous Vej, 8000, Aarhus

Supervisor: Peter Thestrup Waade

1. Introduction (TK) 2

2. Methodology and Materials 3

2.1 Eye-tracking in Behavioural Studies (SM) 3

2.2 Eye-Tracking for Diagnosing Individuals with Autism Spectrum Disorder (SM) 3

2.3 Eye-tracking Dataset (SM) 5

2.4 Data Preprocessing (SM) 6

2.5 Pre-Modelling Data Processing: Scaling and One-Hot Encoding (SM) 8

2.6 LSTM's (TK) 9

2.7 Constructing the Model (TK) 9

2.8 Evaluation Methods and Metrics (SM) 12

3. Results 13

3.1 Model Performance and Evaluation (TK) 13

4. Discussion 15

4.1 Summary and Interpretations of Results (TK) 15

4.2 Implications: Why do our results matter? (SM) 16

4.3 Methodological limitations (SM) 17

4.4 Avenues for further studies or analyses (TK) 20

5. Conclusion (TK) 21

6. References 23

7. Appendixes 31

7.1 Scanpaths of Participants 31

7.2 Evaluation Metrics 31

7.3 Github repository for code 32

1. Introduction (TK)

Diagnosing Autism Spectrum Disorder (ASD) requires on average 13 hours with a psychiatrist in England (Galliver et al., 2017). Furthermore, in about 30% of cases, even top specialists are not confident in their diagnosis (Carroll & Herzberg, 2023). Therefore, much potential efficiency can be added to the diagnosis of the disorder both in terms of time and economic cost. In this study, we showcase the potential of eye-tracking measurements in combination with artificially intelligent models as a possible factor to include as a link in the diagnosis of the disorder.

The data set used in this article is a public data set released by a team of researchers in France on December 22 (Cilia et al., 2022). Since then, many statistical models have been created from the data set and analyzed thoroughly before our attempt. Elbattah et al. used a Convolutional Neural Network (CNN) based on the accumulated scan paths of the participants (Elbattah et al., 2023). The results showcased a model with impressive predictive power, also pointing towards the big potential of eye tracking in diagnosing children with the neurodevelopmental disorder. However, by making a model based on the accumulated scan paths alone, the model loses important information about the temporal dynamics of eye movements. Therefore, the next step in making an improved model seems to be to make a sequential model, taking the chronological temporal aspect into account.

We want to create a Long Short-Term Memory (LSTM) Neural network, as this model is ideal for learning sequential data patterns (Saxena, 2023). We aim to determine whether this method can improve the predictive power to higher accuracy than other models. Furthermore, we also seek to find whether it can be used advantageously as a possible tool in the early diagnosis of ASD. We do this because we want to show the potential of using LSTM models on eye-tracking data in the diagnosis of ASD in children

2. Methodology and Materials

2.1 Eye-tracking in Behavioural Studies (SM)

Eye-tracking offers a window into visual processing in individuals and, thereby, the underlying cognitive processes. This technology, through detailed tracking of eye movements, offers insights into how individuals engage with their visual environment. The primary application of eye-tracking is to record where an individual's gaze is directed in their visual field, indicating the focal point of their attention. In doing so, fixations and saccadic movements of the eye are being used to map the eye gaze. Fixations happen when the eye gaze is stabilized on a specific object in the visual field, whereas saccades are the rapid movements between fixations. Fixations and saccades can offer valuable insights into the underlying cognitive processes since they often represent where an individual is focusing their attention together with their scan paths in analyzing a visual scene (Franchak, 2020).

The choice between using a screen-based eye-tracking device or mobile eye-tracking glasses will depend on the research question. A mobile eye-tracking device can provide valuable insights into how an agent naturally interacts with its environment, but often with the cost of the data becoming less precise. A screen-based eye-tracking device, on the other hand, offers more precise data because of a fixed setting, and though precise, it might not encompass the full spectrum of eye movements and behaviors that occur in a more naturalistic and real-world setting. Most eye-tracking devices use near-infrared technology and utilize the pupil center corneal reflection to map the eye gaze. Some more advanced eye-tracking devices can also map pupil dilation. If the changes in pupil dilation due to light variations can be accounted for, other properties, such as emotional arousal or cognitive workload, can be extracted from pupil dilation (Farnsworth, 2022).

2.2 Eye-Tracking for Diagnosing Individuals with Autism Spectrum Disorder (SM)

Recent advancements in eye-tracking technology have opened new avenues in understanding ASD. Though there is a need for further research to fully understand the biomarkers in the visual exploration of individuals with ASD (Bast et al., 2021), studies

show that there might be differences in eye gaze, saccading and pupil dilation between TD and ASD children. Through investigation of these variations, research may uncover the fundamental biomarkers of oculomotor function in individuals with ASD (Frye et al., 2019).

Research has shown that individuals with ASD tend to exhibit atypical eye gaze patterns, and they tend to attenuate their attention to social visual stimuli (Frazier et al., 2017). Newer research also indicates that individuals with ASD seem to have saccade dysmetria, meaning the amplitude and length of saccades are typically decreased compared to individuals with Typical Development (TD). This new research also shows that saccade dysmetria happens independently of whether stimuli are videos of human/social content or a more naturalistic scene, and individuals with ASD seem to have more clustered fixations. These clustered fixations have also led researchers to propose that there might be an altered neurological difference in the pontocerebellar motor modulation, which might contribute to this atypical oculomotor coordination. However, tracking this region of the brain has previously been difficult because it is deeply situated within the brain, posing challenges for accurate tracking (Bast et al., 2021).

It is discussed in the field whether different pupil dilation features are significant only for individuals with ASD. In a meta-analysis study by de Vries et al., 2021 researchers found a significant difference in the latency of pupillary responses but no difference in the amplitude and baseline pupillary responses. In the more extensive study by Bast et al., 2021 they also explored whether pupil dilation features impacted visual exploration for individuals with ASD but found no significant difference. Other studies found significant results for children with ASD, having a larger baseline pupil size and a decreased response in pupil dilation to human faces (Anderson & Colombo, 2009). More clear and precise research in the field is needed to determine the effect of pupillary features on individuals with ASD.

The insights from these studies shape the foundation for our decision to utilize eye-tracking technology in the creation of a model for ASD classification. Understanding the perception of individuals with ASD provides insights into how they act and is essential for making inferences in diagnosing and treating individuals with the disorder, as analyzing

both their perceptions and actions offers a window into their cognitive processes (Knoblich & Sebanz, 2006)

2.3 Eye-tracking Dataset (SM)

In this research, we utilized a comprehensive eye-tracking dataset developed by Cilia et al., 2022, which provides the foundation of our research. The researchers recruited 59 participants from schools in the Hauts-de-France region. The participants ranged from 3 to 12 years old and were at an early stage of development. The participants were categorized in binary groupings of children with ASD and TD. Additionally, the researchers collected Childhood Autism Rating Scale (CARS) scores from participants with ASD (Schopler et al., 1980), a scale used to assess the degree of ASD.

The participants were shown various stimuli, including photos and videos, in varying lengths to achieve diverse data. The content of the videos and photos typically included visual items particularly attractive to children, and a human-like presenter tried to draw the kids' attention to the object. Unfortunately, we could not access the stimuli presented in the dataset by Cilia et al., which could have provided broader insights into how the data was collected. The dataset encapsulates various eye-tracking metrics such as fixations, saccades, pupil diameter and point of regard/eye gaze. For their experiment, they used an SMI-Red-M eye tracker running at a 60hz sampling rate and the screen size they used was 1280x1024.

An example of a gaze plot from one stimulus of participants with ASD and TD can be seen in Appendix 1. Figure 1 provides a visual representation of our workflow, depicting and offering a comprehensive overview of the model creation process, which is detailed later in the paper.

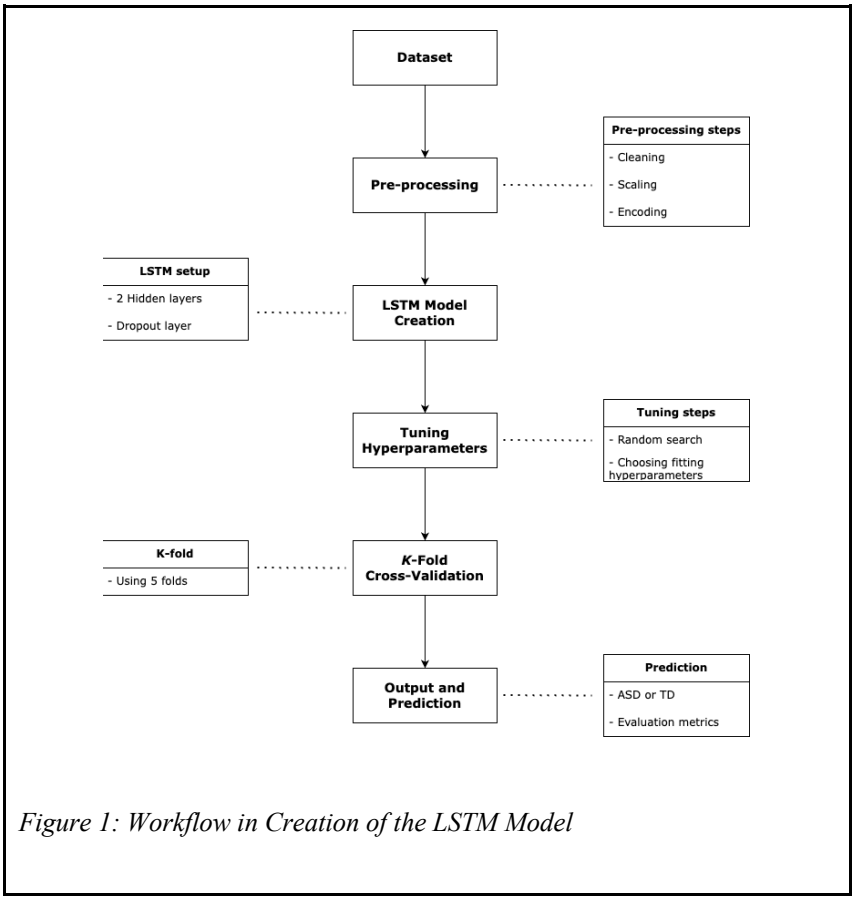


Figure 1: Workflow in Creation of the LSTM Model

2.4 Data Preprocessing (SM)

The initial step in the analysis involved a comprehensive preprocessing of the eye-tracking dataset. The 25 CSV files from the original dataset were concatenated along with the participants' metadata and saved into a single CSV file. After this step, an explorative analysis of the data was done to understand the structure of the dataset, including which columns were duplicates and which included missing values and had to be removed.

For the general data analysis, Python 3 was utilized (Van Rossum & Drake, 2009), and for data preprocessing steps, the Python library pandas was utilized (McKinney & others, 2010). After an overview of the data was achieved, the cleaning of data began, which involved several different steps. The ‘Time of Day’ column was in a format of (h:m:s:ms) which had to be converted into a timedelta format to help facilitate the normalization steps for this value later on. The variables that would be the basis of the model were then sorted, and in this step, unnecessary variables for the model were also removed. A custom imputation algorithm was then implemented, replacing missing numerical values with the

mean of their adjacent values, which was possible because the data was continuously linear, which assisted in preserving as much data as possible. After this step, additional rows containing missing values of other variable types were removed. Unfortunately, the 'Participant' column contained unidentified participants, which was removed, causing some lost data. The dataset was then sorted by 'Participant', 'Trial' and 'Stimulus' to maintain consistency in the dataset, which is crucial for any time series model and saved the new cleaned dataset.

After cleaning the dataset, some general descriptive measures were achieved, as seen in Table 1. After cleaning, the dataset included 27 participants with ASD and 29 TD, summing up to 56 participants in total. The dataset had a disparity in gender, with 36 male and 20 female participants. Participants had a mean age of 7.8 years. The mean CARS score for participants with ASD was 32.65, with a standard deviation of 6.68, indicating a range of scores within the group. The participants with TD did not have a CARS score since it is an assessment of the level of autism. The number of unique stimuli was 114.

Summary Statistics of Eye-tracking Dataset

STATISTIC	VALUE
Total Unique Participants	56
Class Distribution - ASD	27
Class Distribution - TD	29
Gender Distribution - Female	F 20 (ASD: 4)
Gender Distribution - Male	M 36 (ASD: 23)
Mean Age of Participants	7.80 (SD: 2.82) years
Mean CARS Score for ASD	32.65 (SD: 6.68)
Number of Unique Stimuli	114

Table 1: Descriptive Statistics of the Preprocessed Data

2.5 Pre-Modelling Data Processing: Scaling and One-Hot Encoding (SM)

An essential step in preparing our dataset for deep learning required some data processing techniques: scaling, one-hot encoding and splitting the data into a training and a test set. These techniques are crucial for optimizing the model's performance by ensuring that the input data is structured and standardized (Ghosh, 2022).

Given the sensitivity of neural networks to the input data scale, numerical variables had to be normalized. This was a crucial step because some variables had a higher magnitude than others, and if not scaled, these variables would have a more significant bias compared to other variables (Brownlee, 2019). To do this, a Min-Max scaling technique was utilized from the Python library Scikit-learn (Pedregosa et al., 2011), making the variables range in scale from 0 to 1. Consequently, the point of regard before scaling ranged from 1280x1024, but after the scaling, both the x and y-coordinates were now transformed into numerical values ranging from 0 to 1.

One-hot encoding was an essential technique to employ due to the nature of neural networks requiring a numerical input. This method effectively addresses the challenge of categorical data, which is not numerical, particularly when the categories do not have an intrinsic ordinal relationship. The process involves generating a new binary column for each category level and assigning the value 1 if true and 0 if false (Brownlee, 2017b). For hot-one encoding, we used the Python library pandas (McKinney & others, 2010).

We then converted the data type objects to float32 because the neural network requires a uniform input in data type. The reason for choosing float32 compared to float64 was based on float32 being computationally less expensive and the extra precision from using float64 not being needed (Turner-Trauring, 2023).

Due to the nature of LSTM models, we chose to sequence our data into snippets of 2.4 seconds of data. By both estimation and trial, snippets of 2.4 seconds seemed fitting since the LSTM would be able to encapsulate an action happening in the stimulus and store this sequence in its short-term memory while also being able to attenuate irrelevant information.

2.6 LSTM's (TK)

An LSTM model or a Long Short Term Memory Neural Network is a type of Deep Neural Network (DNN). Deep Neural Networks are Deep learning algorithms that, to some degree, are made to mimic the mechanisms of the human brain. They are composed of multiple layers of neurons, and these layers are connected through weights and biases that, prior to training, are set to random values. These weights and biases are then adjusted according to an error function that is defined in terms of how close the current weights and biases are in outputting the true values when running the data through them. Gradient descent, also called backpropagation, is applied to the error function in order to minimize the error (Kostadinov, 2019). The process is then repeatedly conducted across the entire dataset multiple times. LSTM models have a short-term memory, meaning they take the output of the previous sequence into account and the sequence fed into the network while training the model. Apart from this, layers have a long-term factor that is trained and updated from all the sequences or inputs in the network, independent of everything else in the model. These factors are then used in the computations together with the neuron's respective outputs throughout the network. Thus, it has a long-term memory independent of the weights and biases. The model is also able to adjust how much information is let in from the input in the classification processes, which is also an essential factor for attenuating information that is not valuable for classification which can prevent potential overfitting. The effectiveness of handling sequential data in LSTM models, such as eye-tracking, has shown promising results, exemplified by Mazzeo et al. in 2021.

2.7 Constructing the Model (TK)

After the original data set was released on 2022/12/31 (Cilia et al., 2022), researchers trained different models from the dataset. These ranged from lower to higher in complexity. Other neural network techniques have been used on the data set, like a CNN based on the images of the scan paths of the participants' eye gaze (Elbattah et al., 2023). They attained a relatively high accuracy with an ROC-AUC score of 0.90 from a 3-fold cross-validation. It appears from their accuracy plot, with an 80/20 train-test split, that they achieved about 84-86% validation accuracy. However, researchers have argued that it

might be a problem that CNN-models do not capture the temporal aspect of the eye-tracking data (Carette, Elbattah, et al., 2018). Consequently, a neural network based on long short-term memory could prove to be a more suitable method. Given their enhanced capability to encapture temporal features of the data, they might be a more appropriate method for making predictions using eye-tracking data.

The creation of the model required thoughtful engineering and the acquisition of knowledge in the literature regarding eye-tracking and ASD. The relevant factors from the eye tracking data used to construct the final model include the specific stimulus given to the participant, the coordinates of their right eye gaze, the right eye pupil size, the time passed within a given stimulus, the tracking ratio percentage, the category of the current eye-action and the age of the participant.

In selecting the predictors, it was thoroughly investigated that each predictor incorporated, had its grounding in empirical evidence to ensure the model was not created from random correlational patterns. The fact that Elbattah et al., 2023 found their CNN model to have a ROC-AUC score of 0.90, only based on the accumulated gaze plots of the participants over a full stimulus from the same data set, shows us that the gaze of the participants' eyes carries significant information as well. This effect enhances our model's predictive power, but it might further do so when paired with LSTM models' sequential properties, as there might be chronological patterns in the data that a CNN model cannot capture.
As described earlier, there is contradicting research pointing towards whether pupil dilation could be an important factor in distinguishing children with ASD. Pupil dilation could enhance the predictive power, but the model will automatically attenuate the effect to some degree if there is no correlation, and thereby account for the contradicting findings on the subject.
The stimuli are also an important predictor, as certain neurons in the model could be taught to be more or less activated after the specific stimulus the model is presented with. This is logical as there will be different eye gaze patterns according to the stimulus the participant is presented with.

Age was also included as a predictor as this might also have some effect on the participant's eye gaze or their ability to concentrate efficiently on the task (Ridderinkhof & Van Der Stelt, 2000).
Tracking ratio was also included which is a metric for how precise the data was recorded. In including this metric, hopes were that it would help the model attenuate inaccurate data. Hereby, as much data as possible were kept intact, without the need for removing data with less accuracy.
Gender was also included, as it has been argued that girls are better at keeping attention (Pascualvaca et al., 1997). Cultural or social expectations might differ between genders, as argued by Freeman, 2004. Thus, gender as a variable might also carry an amount of predictive power.

Once all the variables were prepared, the data was split into training and testing tensors, a package of data in 3 dimensions provided by Tensorflow (Abadi et al., 2015), containing sequences of data over 2.4 seconds. The data was split such that it consisted of 36 training participants and 20 validation participants for the purpose of selecting the best hyperparameters for later cross-validation. The Python package Numpy (Harris et al., 2020) was also utilized to format and combine different data arrays.

The model was built using Keras (Chollet & others, 2015), consisting of two output neurons instead of one, as this seemed to encapsulate more complexities and lead to higher accuracy based on testing the model using different parameters. For the creation of the model, two hidden layers of neurons and a dropout layer were chosen, as testing on the train-test split pointed towards this being the best balance between prediction and not overfitting. A dropout layer is a layer that randomly sets specific neuron inputs to zero such that the model is trained to predict independently of specific neurons and does not overfit. The number of epochs had to be set as well. Epochs are the number of times the data set is run through the training algorithm.

For the prediction, a custom prediction function had to be created. This is because the model has only been trained to predict over a sequence of 2.4 seconds at a time, but it is

desirable to predict a child's diagnosis from all of the sequences. Thus, a prediction function that groups together the data of one child and then makes predictions was made for each sequence of that child's data. The output probabilities, generated by two output nodes using a softmax activation function, were summed up for each possible outcome. The prediction of a binary classification for a participant was then made using these accumulated outputs.

To select the best hyperparameters for the model, a hyperparameter tuning engine was created based on a random search. This method essentially tests random hyperparameters in combination with intervals that are specified. Intervals of potential hyperparameters that were found fitting based on earlier testing and evaluation were selected. For back-propagation, the Adam-optimization algorithm was used, which is able to adapt learning rates to the specific parameters when using gradient descent (Brownlee, 2017a). Then, the tuning engine ran ten models with different combinations of hyperparameters, whereof the model that had the best accuracy on the validation set was saved. One could argue whether this approach could lead to overfitting on the particular train-test split that was set. This could very well be, but it might also translate into higher predictive power when training on other splits. As an example the correct amount of neurons or complexity will also possibly translate into other splits where the complexity of the data will probably be somewhat the same. Neural network training is computationally expensive, so more thorough parameter tuning methods would be too computationally expensive.

2.8 Evaluation Methods and Metrics (SM)

The k-fold cross-validation method was utilized to validate the model. Cross-validating the dataset using a *k*-fold is often a more robust method as one might hit a non-representative random seed by only using a training and test split, especially because of the random factors in deep neural networks (scikit-learn, 2007). A *k*-fold cross-validation splits the training dataset into *k* equal amounts, and afterward, the model is trained on $k - 1$ of the folds. Consequently, when cross-validating the model, it is validated on *k* amounts of different training sessions. Since *k*-folding can be a computationally heavy task, we chose

for our model to fold five times while still ensuring the robustness of the model performing well on a diverse dataset. For folding our dataset, we utilized the Python package scikit-learn (Pedregosa et al., 2011).

Standard evaluation metrics were utilized to assess the performance of the model and consisted of accuracy, sensitivity/recall, specificity and F1 scores. The accuracy score is used to assess how accurately the model performed by dividing the correct number of predictions by the total number of predictions. The sensitivity score is used to assess the ratio of true positives to actual positives. On the other hand, the specificity score is used to assess the ratio of true negatives to the total actual negatives. The F1 score is the harmonic mean between sensitivity and precision (ratio of true positives to all predicted positives) and holds more information compared to the accuracy score but may be more difficult to interpret. The equations for the evaluation metrics can be seen in Appendix 2. A confusion matrix was also established using ggplot2 (Wickham, 2016) and R (R Core Team, 2022), to provide a visualization of the ground truth versus the predicted values (Bajaj, 2022). The repository for creation of the model can be found Appendix 3.

3. Results

3.1 Model Performance and Evaluation (TK)

The 5-fold cross-validation with the chosen hyperparameters yielded the following aggregated results: Accuracy score of 83.94% (95% CI: 0.71 - 0.97), F1 score of 85.69% (95% CI: 0.75 - 0.96), Sensitivity of 86.48% (95% CI: 0.78 - 0.95), and Specificity of 80.00% (95% CI: 0.56 - 1.00). Model performance results are also presented in Table 2.

Summary of Model Performance Scores

METRIC	SCORE
Accuracy	83.94% (95% CI: 0.71 - 0.97)
F1 Score	85.69% (95% CI: 0.75 - 0.96)
Sensitivity	86.48% (95% CI: 0.78 - 0.95)
Specificity	80.00% (95% CI: 0.56 - 1.00)

Table 2: Model Performance Scores

The amount of correctly versus incorrectly predicted participants accumulated through folds is visualized in the confusion matrix as seen in Figure 2.

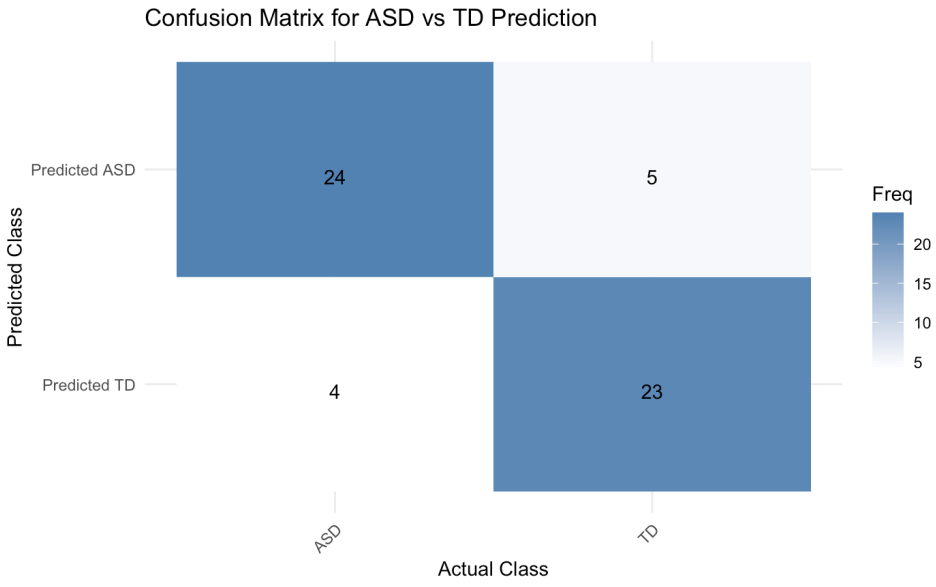


Figure 2: Confusion Matrix for Predicted and Actual Values

In the random search, the best-performing configuration consisted of 70 neurons in the first layer and 50 in the second. Additionally, the best-performing dropout rate was found

to be 0.05 and the learning rate was 0.001. The number of epochs in the training was set to five. The amount of k folds was set to five. Training a model took slightly more than a minute using the m1 Pro processor (Apple, 2021).

4. Discussion

4.1 Summary and Interpretations of Results (TK)

Based on these results, the model seems to perform similarly to the best models currently trained on this particular data set. As an example, the CNN-based model had a roc-auc score of 0.9 based on 3-fold cross-validation (Elbattah et al., 2023). Employing an LSTM model showed great performance in classifying ASD from eye-tracking data when compared with the CNN model. However, it is important to consider that neural network-based models are always set to a random initial state and then converge towards less prediction error. This will inevitably yield different models each time they are trained, even on the same data. For this reason, it is not certain that the 5-fold cross-validation will result in the same values of the precision metrics when it is trained multiple times. As the evaluation metrics are aggregated across only 5 folds, somewhat big 95% confidence intervals are to be expected. However, it is worth noting that the specificity seems to vary a lot, leading to a wide confidence interval. This means that for specific folds, the model's uncertainty in classifying TD children has varied. This might be an effect of some of the TD children having some degree of oculomotor behavior similar to those observed in children with ASD.
Another study has shown promising results (83% accuracy) of using an LSTM with eye tracking data for distinguishing neurotypical children from children with Autism Spectrum Disorder (ASD) in the age of 8-10 (Carette, Cilia, et al., 2018). However, this study is based on a sample size of only 17 participants and a validation data size of only six participants. For this reason, our model is a more stable predictive model.

K-fold cross-validation was used because it takes multiple models trained on different data splits into account, making the model less prone to overfitting and giving a better insight into the average performance of the model. The number of k-folds was set to five

as this provided a good overview of the average performance of the model, and a higher value of *k* would not necessarily lead to a more accurate model. We could have added the mean ROC-AUC score in order to make our model more comparable with the already constructed CNN model (Elbattah et al., 2023), but complications arise when our prediction function is based on the sum of the outputs of the model, as well the model's output being binary and not being probabilistically calibrated. This could have been done by restructuring the output of the model and the way the prediction function works. However, the validation metrics already selected provide a comprehensive overview of the model's performance, thus eliminating the need for additional measures.

Calibrating the model's output to represent the true probability of its confidence more accurately could enhance its applicability in real diagnostic scenarios. To assess the actual probability of ASD based on the model's output, it would be necessary to apply a Bayes factor to obtain the true probability of having ASD (Joyce, 2021).
Our model seemed to train faster (≈ 1.2 minutes) compared to, eg. a gaze-plot-based CNN which took around three minutes (Elbattah et al., 2023), although they were using a slower processor. Their method also required more heavy preprocessing. While the difference might not seem of big importance, it could become significant when working with a larger data set.

4.2 Implications: Why do our results matter? (SM)

According to the Centers for Disease Control and Prevention, in 2018, the average age of children getting their diagnosis was 4.5 years in the US. However, 85% of those diagnosed with ASD had concerns about development already at three years of age (CDC, 2018). Early identification of the diagnosis, especially before the age of three, is vital due to brain plasticity being very high up until this age (Huttenlocher, 1994). It is hypothesized that if autism is diagnosed before the age of three, interventions might help with impairments such as facial recognition because children before the age of three with autism fail to attend to faces. However, with professional assistance, they will get a more developed fusiform gyro, which is a center in the brain responisble for the processing of faces and expertise (Dawson & Zanolli, 2003). One of the first behavioral characteristics

regarding social responsiveness of children with ASD that emerges is the infrequent look at others' faces, gaze aversion and eye contact (Landa, 2008). The research by Landa, 2008 also suggests that a diagnosis of ASD becomes possible at 14 months of age, but the diagnosis might be unstable up until the age of 30 months, why Landa suggests a screening from the age of 18 months and a repeated screening at the age of 24 to 36 months of age.
New research has shown that such a screening technique using eye-tracking has proved to be successful in aid of diagnosing children at the age of 16 to 30 months as accurately as a specialist would. The research was a large study based on 499 children in the US with different backgrounds (Jones et al., 2023; Smith, 2023).

Currently, diagnosing autism is expensive, and in a study done in the UK, they found that the median professional time involved in diagnosing autism was 13 hours (Galliver et al., 2017). This is why a good model is needed for screening to reduce resources spent on diagnosing, making the screening process less expensive and more accessible. Our study showed promising results in using an LSTM model, and each step closer to a better model leads to reducing the costs for diagnosing and a potentially more accurate diagnosis. The current way screening is done today is mainly through questionnaires from parents and a few involve looking at how the children interact in play. These screening techniques already scored well on specificity and sensitivity, with scores greater than 70% (CDC, 2022). However, these techniques, compared to eye-tracking screening, reaching specificity and sensitivity scores greater than 80%, may seem outdated.

4.3 Methodological limitations (SM)

One of the critical limitations arose from not collecting the data ourselves. As an example, it would have been optimal to have known the different stimuli in the choice of sequencing our data into timeframes to make this represent the duration of an action in the stimuli. Valuable insights from conducting the experience could also have been gained, such as asking the participants themselves how they analyzed the stimuli, which could be quite

advantageous in preprocessing of the data. Even though the description of both data gathering and stimuli is well described in the paper by Cilia et al., 2022, the experience of having been there could have provided valuable insights into data handling. For example, it would have been convenient to know whether it would be suitable to trim the stimuli, as there could have been buffers in the recording process (such as waiting a couple of seconds before beginning), which would result in worse predicting. In the preprocessing of the data, potentially valuable information could also have been lost, and two participants were sorted out in the process of cleaning the data. Some of the eye-tracking data themselves were imprecise, but fortunately, the dataset included this metric, which was also included in the training of the model, which could have helped the model in making more accurate predictions, but also a potential in some overfitting of this noise.

Another potential issue is the sample size of the dataset. Though the dataset is diverse and contains children in ages ranging from 3-12 years, and after cleaning 56 unique participants and an almost equal amount of children with ASD and TD, some concerns still arise. The first concern is regarding the sample size of the dataset. Due to autism being a spectrum of many dimensions, this allows for much variance within the disorder, and even though the CARS score was provided in the dataset, 27 participants with ASD may not be sufficiently large enough to fully encapsulate the diversity of ASD (Jaarsma & Welin, 2012). This is because ASD diagnosed children might have differing symptoms and these could manifest themselves very differently. The CARS score was not added to the model as the TD children did not have such a score and the model would be able to classify a diagnosis from this score alone.

One issue regarding the accuracy of the model could also be the misdiagnosis of the participant children in the first place. Children with TD might also exhibit some degree of symptoms that would have some characteristics of ASD, but not enough to be labeled with ASD by the French psychiatric standards. This arises from the nature of autism being a multidimensional spectrum with a variety of symptoms, but it is necessary to establish a diagnostic threshold at some point.

Another concern is regarding the lower sample of females included in the dataset, primarily because ASD manifests differently across genders. Females with ASD often exhibit different symptoms and behavioral patterns compared to males with ASD, which might

not be adequately captured by a model primarily trained on male participants (Manjra & Masic, 2022).

There is also a concern regarding cultural differences, mainly because the dataset only includes participants from a specific region in France. There is no unison standard for diagnosing autism across cultures, and autism is socially constructed differently across cultures (Kim, 2012). ASD itself also varies in its expression across different cultures, and a model that encompasses this would need an immense sample size with many different cultures represented (Ennis-Cole et al., 2013). Therefore, our trained model can not be used in other cultural paradigms in diagnosing ASD.

Though using an LSTM model is suitable and provides excellent results in predicting ASD, it also has its limitations. One of the most significant limitations here lies in the fact that we do not know which factors the model finds essential. As an example, the weights and biases assigned to pupil dilation would have provided valuable insights into how significant this effect is for predicting ASD. This limitation stems from the innate complexity of neural networks, which are often challenging to interpret internal workings and often described as “black boxes” (Blazek, 2022). One way to assess the importance of the different variables could have been to retrain the model either without them or manipulate specific variables and assess the difference in the performance in prediction.

A critical issue in using LSTM models is the possibility of overfitting. In training the model, there is a possibility of the model becoming too aligned with the specific characteristics of the training data, impairing its ability to generalize and perform accurately on new data. To counteract and mitigate this issue, dropout techniques were utilized as well as training the model on the proposed amount of epochs. The number of epochs selected was based on tests where the validation accuracy and accuracy appeared to be converging, which is significant in the case of a non-overfitting model.

Another issue with our model is that its output layer consists of two neurons which as discussed in the methodology section adds to the complexity of the model, but also makes the model perform better in our case. The outputs, however, have not been calibrated or tested. Therefore these outputs might not correctly reflect the actual probability of the model's classification being correct. This calibration could also potentially have improved

the performance of the model. Also if we had used a single output neuron with a sigmoid activation function, the output would have been more directly interpretable.

4.4 Avenues for further studies or analyses (TK)

In the future, we hope for more extensive research on which particular factors of the eye-tracking data are the most important, particularly for LSTM-models performance in their ability to diagnose children with ASD from eye-tracking data. This could be done through different tools and packages available for the analysis of neural network-models like doing a visualization with TensorBoard (Abadi et al., 2015). Further calibration of the model could also increase performance as well as making it more generally applicable for diagnosis of ASD. Parameters could also be further tuned for a more generalizable model across different data splits, but computational cost is a big factor to consider here. Further feature engineering could also be applied. Eg. if we had split the eye gaze up into more variables such that the y axis is split up into 50 or more sections, with a sequence of neurons for each of the 50 sections. This might have improved the model further as certain areas might be more interesting to look at over time. The model is not currently able to understand specific areas of the screen as the input neurons only understand a granular decrease or increase in the eye-gaze axes and the variation hereof. This is something that a CNN is better at capturing because it has a neuron for every pixel, and probably a big reason why the CNN model had great performance as well. This preprocessing technique could be implemented in future LSTM models.
Another model that could possibly enhance performance is a transformer neural network which was first presented in the article “Attention is all you need” (Vaswani et al., 2017). In these neural networks, attention between the different layers of a sequence is implemented as a factor. Thus, these types of neural networks could be beneficial in this type of data, as each layer in a sequence would have further “knowledge” about its relative position in a sequence to the other layers. This new factor to the network could potentially increase the accuracy, although the complexity of further analysis would increase.

If ASD diagnosis with eye tracking continues to develop as a method, it could potentially be used as a tool in the actual diagnosis of children in the future for identifying symptoms

of ASD associated with oculomotor control. However, this would also require testing of different types of data collection, as there is most probably a method used for data collection that could be better at distinguishing children with ASD from TD children. For example, one stimulus might provide a more distinguishable data set in classification.

The eye-tracking method could be used as a standalone method in the process of screening, but in diagnosing, we potentially foresee a multimodal model that encaptures additional dimensions of autism, such as social capabilities, motor capabilities and repetitive behaviors, among others (Faras et al., 2010; Mahmood, 2023). Analyzing these behaviors could also improve the precision of diagnosis so that psychiatrists can capture more dimensions of the diagnosis and offer an in-depth diagnosis with more dimensions rather than a binary classification, which does not hold much information about specific symptoms.
Multimodal models have already been tested in the screening process by combining electroencephalogram (EEG) data with eye-tracking data in a study by Han et al., 2022. Utilizing this approach showed auspicious results for an accurate screening model with an accuracy of 95.56%. Finding a well-performing eye-tracking screening technique might be sufficient in the screening process alone and be more cost-effective than including an EEG. Even if cost is not a problem, a high-performing model for each technique remains essential in order to construct the most accurate predicting system.

5. Conclusion (TK)

In this paper, we have explored the use of LSTM-based neural networks on sequential eye-tracking data, in classifying children in the age of 3-12 years with ASD. From our results, it is evident that using this method shows promising results in terms of the classification of ASD in children in the age group when compared to other models trained on the same data set, presented in the available literature.
An evaluation of the LSTM-based neural network method has been presented in the use of classifying ASD from sequential eye tracking data. We have laid out our method in terms of data preprocessing, like the different encoding approaches of the categorical variables, and our thoughts regarding these methods.

The advantages and disadvantages of using an LSTM model in this use have been discussed, as well as what improvements could have been made in terms of the creation of such a model. Furthermore, the results from the 5-fold cross-validation of our model and their implications have been revealed and discussed in terms of their implication. Looking forward, a roadmap for future research has been outlined, and the next steps in advancing this specific method, as we suggest it might have big potential.

6. References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*.

Anderson, C. J., & Colombo, J. (2009). Larger Tonic Pupil Size in Young Children With Autism Spectrum Disorder. *Developmental Psychobiology*, *51*(2), 207–211. https://doi.org/10.1002/dev.20352

Apple. (2021). *Vi introducerer M1 Pro og M1 Max: De mest effektive chips, Apple nogensinde har udviklet*. Apple Newsroom (Danmark). https://www.apple.com/dk/newsroom/2021/10/introducing-m1-pro-and-m1-max-the-most-powerful-chips-apple-has-ever-built/

Bajaj, A. (2022, July 21). *Performance Metrics in Machine Learning [Complete Guide]*. Neptune.Ai. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

Bast, N., Mason, L., Freitag, C. M., Smith, T., Portugal, A. M., Poustka, L., Banaschewski, T., Johnson, M., & Group, T. E.-A. L. (2021). Saccade dysmetria indicates attenuated visual exploration in autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *62*(2), 149–159. https://doi.org/10.1111/jcpp.13267

Blazek, P. J. (2022, March 2). *Why we will never open deep learning's black box*. Medium. https://towardsdatascience.com/why-we-will-never-open-deep-learnings-black-box-4c27cd335118

Brownlee, J. (2017a, July 2). Gentle Introduction to the Adam Optimization Algorithm for Deep Learning. *MachineLearningMastery.Com*. https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

Brownlee, J. (2017b, July 27). Why One-Hot Encode Data in Machine Learning? *MachineLearningMastery.Com*. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

Brownlee, J. (2019, February 3). How to use Data Scaling Improve Deep Learning Model Stability and Performance. *MachineLearningMastery.Com*. https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/

Carette, R., Cilia, F., Dequen, G., Bosche, J., Guerin, J.-L., & Vandromme, L. (2018). Automatic Autism Spectrum Disorder Detection Thanks to Eye-Tracking and Neural Network-Based Approach. In M. U. Ahmed, S. Begum, & J.-B. Fasquel (Eds.), *Internet of Things (IoT) Technologies for HealthCare* (pp. 75–81). Springer International Publishing. https://doi.org/10.1007/978-3-319-76213-5_11

Carette, R., Elbattah, M., Dequen, G., Guerin, J.-L., & Cilia, F. (2018). Visualization of Eye-Tracking Patterns in Autism Spectrum Disorder: Method and Dataset. *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, 248–253. https://doi.org/10.1109/ICDIM.2018.8846967

Carroll, L., & Herzberg, J. (2023, September 5). *Autism may be identified early with eye-tracking device, studies show*. NBC News. https://www.nbcnews.com/health/kids-health/autism-identified-early-eye-tracking-device-diagnosis-studies-rcna103361

CDC. (2018, April 26). *Spotlight On: Delay Between First Concern to Accessing Services*. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/addm-community-report/delay-to-accessing-services.html

CDC. (2022, December 6). *Healthcare Providers | Autism Spectrum Disorder (ASD) | NCBDDD | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/hcp-screening.html

Chollet, F. & others. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Cilia, F., Carette, R., Elbattah, M., Guérin, J.-L., & Dequen, G. (2022). *Eye-tracking dataset to support the research on autism spectrum disorder*. *1*. https://doi.org/10.5220/0011540900003523

Dawson, G., & Zanolli, K. (2003). *Early Intervention and Brain Plasticity in Autism*. *251*, 266–280. https://doi.org/10.1002/0470869380.ch16

de Vries, L., Fouquaet, I., Boets, B., Naulaers, G., & Steyaert, J. (2021). Autism spectrum disorder and pupillometry: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *120*, 479–508. https://doi.org/10.1016/j.neubiorev.2020.09.032

Elbattah, M., Carette, R., Cilia, F., Guerin, J.-L., & Dequen, G. (2023). *Applications of machine learning methods to assist the diagnosis of autism spectrum disorder* (pp. 99–119). https://doi.org/10.1016/B978-0-12-824421-0.00013-8

Ennis-Cole, D., Durodoye, B. A., & Harris, H. L. (2013). The Impact of Culture on Autism Diagnosis and Treatment: Considerations for Counselors and Other Professionals. *The Family Journal*, *21*(3), 279–287.

https://doi.org/10.1177/1066480713476834

Faras, H., Al Ateeqi, N., & Tidmarsh, L. (2010). Autism spectrum disorders. *Annals of Saudi Medicine*, *30*(4), 295–300. https://doi.org/10.4103/0256-4947.65261

Farnsworth, B. (2022, August 2). *Eye Tracking: The Complete Pocket Guide - iMotions*. https://imotions.com/blog/learning/best-practice/eye-tracking/

Franchak, J. M. (2020). Chapter Three—Visual exploratory behavior and its development. In K. D. Federmeier & E. R. Schotter (Eds.), *Psychology of Learning and Motivation* (Vol. 73, pp. 59–94). Academic Press. https://doi.org/10.1016/bs.plm.2020.07.001

Frazier, T. W., Strauss, M., Klingemier, E. W., Zetzer, E. E., Hardan, A. Y., Eng, C., & Youngstrom, E. A. (2017). A Meta-Analysis of Gaze Differences to Social and Nonsocial Information Between Individuals With and Without Autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(7), 546–555. https://doi.org/10.1016/j.jaac.2017.05.005

Freeman, J. (2004). Cultural influences on gifted gender achievement. *High Ability Studies*, *15*(1), 7–23. https://doi.org/10.1080/1359813042000225311

Frye, R. E., Vassall, S., Kaur, G., Lewis, C., Karim, M., & Rossignol, D. (2019). Emerging biomarkers in autism spectrum disorder: A systematic review. *Annals of Translational Medicine*, *7*(23), 792. https://doi.org/10.21037/atm.2019.11.53

Galliver, M., Gowling, E., Farr, W., Gain, A., & Male, I. (2017). Cost of assessing a child for possible autism spectrum disorder? An observational study of current practice in child development centres in the UK. *BMJ Paediatrics Open*, *1*(1), e000052. https://doi.org/10.1136/bmjpo-2017-000052

Ghosh, S. (2022, July 21). *A Comprehensive Guide to Data Preprocessing*. Neptune.Ai. https://neptune.ai/blog/data-preprocessing-guide

Han, J., Jiang, G., Ouyang, G., & Li, X. (2022). A Multimodal Approach for Identifying Autism Spectrum Disorders in Children. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *30*, 2003–2011. https://doi.org/10.1109/TNSRE.2022.3192431

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Huttenlocher, P. R. (1994). Synaptogenesis in human cerebral cortex. In *Human behavior and the developing brain* (pp. 137–152). The Guilford Press.

Jaarsma, P., & Welin, S. (2012). Autism as a Natural Human Variation: Reflections on the Claims of the Neurodiversity Movement. *Health Care Analysis*, *20*(1), 20–30. https://doi.org/10.1007/s10728-011-0169-9

Jones, W., Klaiman, C., Richardson, S., Aoki, C., Smith, C., Minjarez, M., Bernier, R., Pedapati, E., Bishop, S., Ence, W., Wainer, A., Moriuchi, J., Tay, S.-W., & Klin, A. (2023). Eye-Tracking–Based Measurement of Social Visual Engagement Compared With Expert Clinical Diagnosis of Autism. *JAMA*, *330*(9), 854–865. https://doi.org/10.1001/jama.2023.13295

Joyce, J. (2021). Bayes' Theorem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University.

https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem/

Kim, H. U. (2012). Autism across cultures: Rethinking autism. *Disability & Society*, *27*(4), 535–545. https://doi.org/10.1080/09687599.2012.659463

Knoblich, G., & Sebanz, N. (2006). The Social Nature of Perception and Action. *Current Directions in Psychological Science*, *15*(3), 99–104. https://doi.org/10.1111/j.0963-7214.2006.00415.x

Kostadinov, S. (2019, August 12). *Understanding Backpropagation Algorithm*. Medium. https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd

Landa, R. J. (2008). Diagnosis of autism spectrum disorders in the first 3 years of life. *Nature Clinical Practice Neurology*, *4*(3), 138–147. https://doi.org/10.1038/ncpneuro0731

Mahmood, O. (2023, October 16). *What are Multimodal models?* Medium. https://towardsdatascience.com/what-are-multimodal-models-fe118f3ef963

Manjra, I. I., & Masic, U. (2022). Gender diversity and autism spectrum conditions in children and adolescents: A narrative review of the methodologies used by quantitative studies. *Journal of Clinical Psychology*, *78*(4), 485–502. https://doi.org/10.1002/jclp.23249

Mazzeo, P. L., D'Amico, D., Spagnolo, P., & Distante, C. (2021). Deep Learning based Eye gaze estimation and prediction. *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*, 1–6. https://doi.org/10.23919/SpliTech52315.2021.9566413

McKinney, W. & others. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Pascualvaca, D. M., Anthony, B. J., Arnold, L. E., Rebok, G. W., Ahearn, M. B., Kellam, S. G., & Mirsky, A. F. (1997). Attention performance in an epidemiological sample of urban children: The role of gender and verbal intelligence. *Child Neuropsychology*, *3*(1), 13–27. https://doi.org/10.1080/09297049708401365

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ridderinkhof, K. R., & Van Der Stelt, O. (2000). Attention and selection in the growing child: Views derived from developmental psychophysiology. *Biological Psychology*, *54*(1–3), 55–106. https://doi.org/10.1016/S0301-0511(00)00053-3

Saxena, A. (2023, March 17). Introduction to Long Short-Term Memory (LSTM). *Analytics Vidhya*. https://medium.com/analytics-vidhya/introduction-to-long-short-term-memory-lstm-a8052cd0d4cd

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. https://doi.org/10.1007/BF02408436

scikit-learn. (2007). *3.1. Cross-validation: Evaluating estimator performance*. Scikit-

Learn. https://scikit-learn/stable/modules/cross_validation.html

Smith, S. (2023, September 5). *JAMA Publishes Two Large Studies Demonstrating the Diagnostic Accuracy of EarliTec's Evaluation for Autism Spectrum Disorder in Children as Young as 16 Months*. https://www.business-wire.com/news/home/20230905447529/en/JAMA-Publishes-Two-Large-Studies-Demonstrating-the-Diagnostic-Accuracy-of-EarliTec%E2%80%99s-Evaluation-for-Autism-Spectrum-Disorder-in-Children-as-Young-as-16-Months

Turner-Trauring, I. (2023, January 27). *The problem with float32: You only get 16 million values*. Python⇒Speed. https://pythonspeed.com/articles/float64-float32-precision/

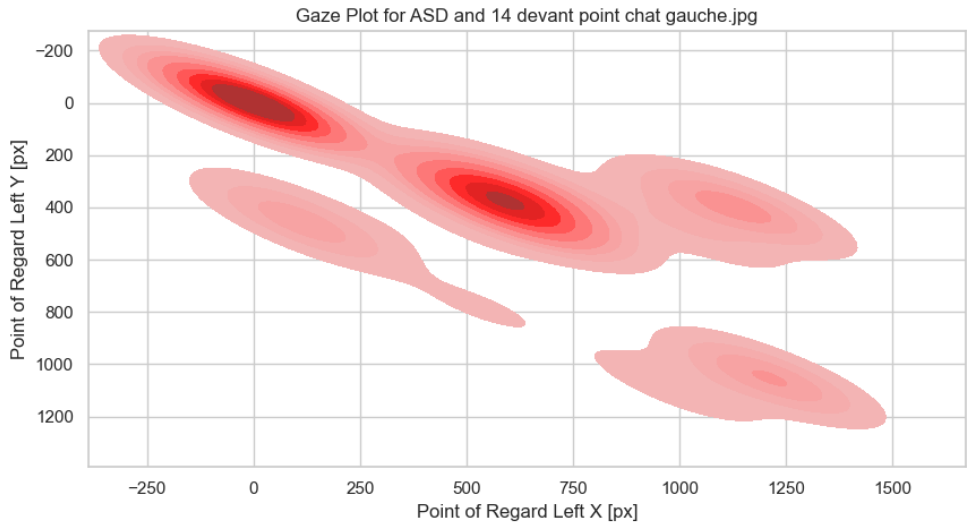
Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

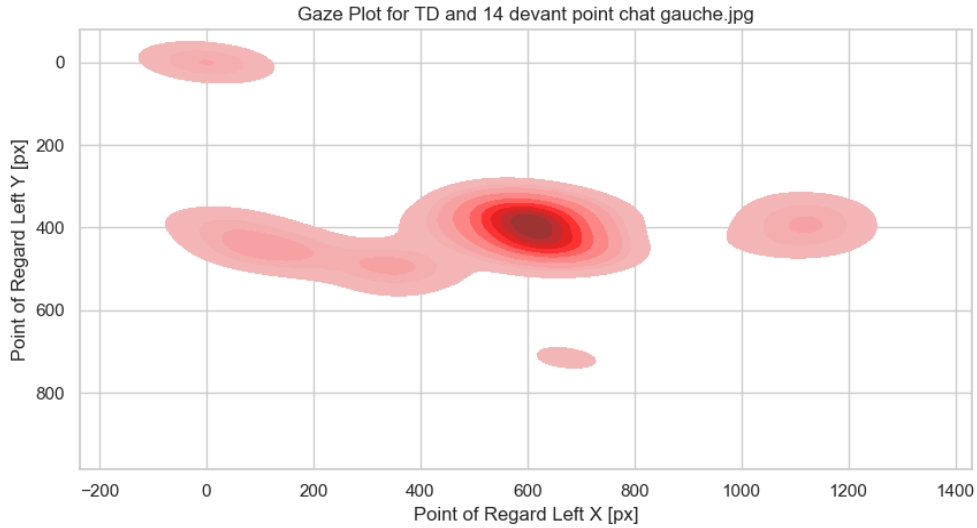
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

7. Appendixes

7.1 Scanpaths of Participants





7.2 Evaluation Metrics

Accuracy score presents the percentage of correct predictions made:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Specificity also known as true negative rate:

$$\mathbf{specificty} = \frac{TN}{TN+FP}$$

Sensitivity/recall or also known as true positive rate:

$$\mathbf{sensitivity} = \frac{TP}{TP+FN}$$

Precision:

$$Precision = TP/(TP + FP)$$

F1 score, a score to measure the model's accuracy:

$$\mathbf{F1} = \frac{2}{precision^{-1}+recall^{-1}}$$

7.3 Github repository for code

https://github.com/asw615/PercAct/tree/main