Predicting Autism Spectrum Disorder (ASD) using Eye Tracking Data and Deep Learning

Suhas Nagaraj
suhas99@umd.edu

Swaraj Mundruppady Rao
swarajmr@umd.edu

FNU Koustubh
koustubh@umd.edu

Arya Sunil Gangatkar
agangadk@umd.edu

Abstract

Detecting Autism Spectrum Disorder (ASD) is a complex process due to the lack of definitive medical tests and the reliance on subjective behavioral assessments. ASD is characterized by persistent difficulties in social communication and repetitive or restricted behaviors. Among key diagnostic markers, atypical eye movement patterns have been identified as a distinguishing characteristic of people with ASD. In this work, we utilize the eye tracking data, collected from children, for ASD detection. Eye movement data, collected over time, were compressed into RGBA images, creating a dataset that encapsulates both spatial and temporal gaze dynamics [1]. Using this dataset, we employ Deep Learning techniques to classify the images and diagnose ASD. This approach offers a diagnostic tool that minimizes the need for extensive behavioral observation in clinical environments. Furthermore, our method can serve as an early screening mechanism, paving the way for more comprehensive and targeted evaluation of suspected individuals, thus improving early detection and timely intervention for ASD. In this work, we introduce a novel individual-level diagnostic approach by aggregating predictions from multiple eye-tracking images of the same participant using sequence modeling techniques such as LSTM with attention and Transformers.

Keywords:*Autism Spectrum Disorder (ASD), Individual-level approach, Eye Tracking Data, Deep Learning, CNN, LSTM, Transformers, Attention Mechanism*

1. Introduction

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition characterized by challenges in social interaction, communication, and behavior, including restrictive interests and repetitive actions. The World Health Organization estimates the global prevalence of ASD at approximately 0.76%, equating to about 16% of the global child population [3]. Despite its widespread occurrence, the diagnosis of ASD is often delayed for years after symptoms emerge, largely due to the limited availability of trained professionals and effective diagnostic tools. Early identification is critical, particularly in childhood, as timely interventions can significantly improve long-term outcomes for affected individuals [4].

Recent developments in artificial intelligence (AI) and machine learning (ML) have introduced innovative approaches to ASD diagnosis. Among these, eye-tracking technologies have shown considerable promise in capturing subtle and atypical visual attention patterns associated with ASD. These behavioral biomarkers, when integrated into machine learning models, hold potential for enhancing the accuracy and efficiency of early diagnosis [4], [5].

However, several challenges remain. Implementing these technologies in real-world clinical settings necessitates their integration with existing diagnostic frameworks, rigorous validation across diverse datasets, and addressing the concerns related to precisely collecting data from young children [2], [5]. This study seeks to enhance diagnostic capabilities by employing deep networks to analyze eye-tracking data, optimizing pre-processing and model performance to improve diagnostic accuracy [6], [7].

In this work, we propose a novel approach to ASD diagnosis by analyzing multiple eye-tracking images from individuals under evaluation. Unlike previous methods that typically rely on a single eye tracking image to classify individuals as autistic or non-autistic, our method aggregates data from a series of eye-tracking images of a subject to create a more comprehensive and dynamic profile for each individual. This approach mirrors clinical practices, where doctors observe behavioral patterns over time to inform their diagnosis. By using multiple images, we aim to capture a more complete range of visual attention behaviors, improving diagnostic accuracy.

Our methodology integrates deep learning (CNN) with sequence modeling techniques, achieving remarkable performance across various approaches. The CNN+LSTM model attained a validation accuracy of 77%, demonstrating its robust capability to capture temporal dependencies in eye-tracking data. By incorporating an attention mechanism (CNN+LSTM+Attention), the model's performance was further enhanced, achieving a validation accuracy of

88%, showcasing its ability to focus on critical temporal patterns effectively. Additionally, a Transformer-based approach leveraged self-attention mechanisms to model long-range dependencies with high efficiency, achieving a ROC AUC of 0.95. A voting-based method, serving as a robust baseline, aggregated predictions from multiple images and achieved a validation accuracy of 77% and an AUC of 0.90. These results highlight the potential of integrating CNN-based feature extraction with sequence modeling and attention mechanisms to advance ASD diagnosis, offering a scalable and accurate solution for early screening and intervention.

2. Related Work

The application of machine learning for Autism Spectrum Disorder (ASD) diagnosis has been widely studied, utilizing a variety of behavioral and physiological markers. Techniques involving functional Magnetic Resonance Imaging (fMRI) scans have also been explored, with models like ASD-DiagNet [8] achieving reported accuracies of up to 82%. Behavioral patterns, such as self-injurious behaviors [9], along with facial expression analysis using deep learning [10], have demonstrated significant success in advancing the field of ASD detection.

These methods show significant promise in developing machine learning-based classifiers; however, a key challenge lies in collecting the necessary data for accurate model predictions. For instance, recording fMRI scans often requires patients to endure lengthy sessions in an enclosed machine, which can be uncomfortable and may lead to inaccuracies in the data. In contrast, using eye-tracking data collected by displaying videos on a screen offers a more practical, quick, and user-friendly alternative, making it a highly accessible option for ASD diagnosis.

For detecting autism using the eye tracking dataset, various Machine Learning and Deep Learning based techniques have already been used. A machine learning model investigating eye-tracking datasets for Autism Spectrum Disorder (ASD) diagnosis was proposed by Akter et al. [18], achieving an accuracy of up to 74.2% with an RF classifier, when evaluated on the whole dataset. We have chosen this model as the baseline for our project.

There have been multiple studies to implement the classification based on DNN and CNN based approaches such as Ahmed and Jadhav et al. [11] suggested a DL model using CNN for categorizing individuals as either having ASD or TD. Arora et al.[12] proposed a predictive tool for Autism Spectrum Disorder (ASD) using eye gaze data, achieving an accuracy of 74.57% with ANN and 85.28% with CNN models. The ANN model applies Principal Component Analysis (PCA) to convert images into a list of floating values for evaluation, while the CNN model uses convolutions to break down images into smaller parts for training and testing.

The closest wok related to our methodology of using transfer learning is Carette et al. [13], who presented a method for visualizing eye-tracking patterns in Autism Spectrum Disorder (ASD), utilizing a pretrained Xception model for feature extraction and a stacking ensemble framework.

3. Data

The dataset utilized in this study is designed to facilitate the analysis of eye-tracking patterns in children diagnosed with Autism Spectrum Disorder (ASD) and their neurotypical counterparts. The data collection involves data from 59 children. Participants were divided into two balanced groups of children diagnosed with ASD and neurotypical children (Non-ASD), with ages ranging from approximately 3 to 13 years.

A total of 547 images were generated from eye-tracking visualizations, with 219 images representing children diagnosed with ASD and 328 images corresponding to neurotypical participants. These visualizations are derived from scanpaths, which provide a spatiotemporal mapping of the sequence of fixations and saccades made by the participants during the experimental sessions. Each participant was assigned a unique identifier to maintain anonymity, and metadata files documenting participant related information.

3.1. Data Acquisition and Equipment

Eye movement data were collected using the SMI REDm remote eye-tracker device. The eye-tracker was mounted below the display screen, which tracked the participants' gaze as they engaged with visual stimuli. Participants were seated approximately 60 cm from the screen to optimize the accuracy of the recordings, with environmental conditions controlled to minimize visual distractions.

The experimental protocol involved presenting participants with a series of carefully curated videos designed to stimulate eye movements. These videos included dynamic and engaging content such as animated objects and human presenters, tailored to capture the attention of young children. The objective was to elicit naturalistic eye movements, allowing researchers to observe and record patterns of gaze and attention under conditions approximating real-world interactions.

3.2. Visualization Methodology

The core innovation of this study lies in the transformation of raw eye-tracking data into visual representations, enabling detailed analysis of gaze dynamics. Each scanpath visualization encodes the trajectory of eye movements as a sequence of lines connecting consecutive fixations. The dynamics of these movements—velocity, acceleration, and

jerk—are captured using RGB color gradients. Velocity is represented by a gradient from black (low) to red (high). Acceleration is encoded using a gradient from black (low) to green (high). Jerk is visualized with a gradient from black (low) to blue (high). The resulting images offer a compact yet comprehensive representation of the temporal and spatial characteristics of eye movements. The vertical mirroring of images ensures that the y-axis aligns correctly with the screen layout, enhancing interpretability. To standardize the amount of information contained within each image, a threshold of 200 fixation points was applied. This threshold strikes a balance between capturing sufficient detail and avoiding excessive visual clutter.

3.3. Dataset Structure and Accessibility

The dataset is publicly available via the Figshare Data Repository (`https://figshare.com/s/5d4f93395cc49d01e2bd`) and includes two primary components. Images are stored in two subfolders, one for ASD participants and another for neurotypical participants. Metadata includes CSV and JSON files documenting participant attributes and maps the images to unique participant IDs. The images are formatted as 640×480 pixels, providing a resolution suitable for both human interpretation and computational analysis.

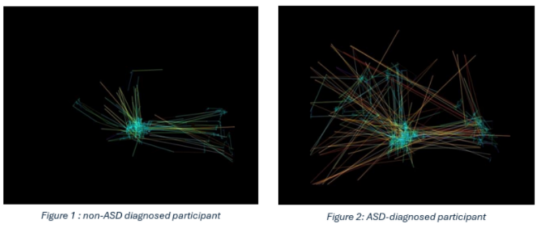

Figure 1. Comparison of eye-tracking scanpaths between a non-ASD diagnosed participant (left) and an ASD-diagnosed participant (right). The images illustrate the difference in eye movement patterns, with the ASD-diagnosed participant showing more dispersed and irregular gaze paths, indicative of atypical visual attention and fixation behavior associated with Autism Spectrum Disorder.

3.4. Exploratory Data Analysis

We conducted an exploratory data analysis to examine the structure of the dataset and identify potential biases. This analysis will help us in uncovering trends that may be relevant for diagnosing Autism Spectrum Disorder.

As mentioned previously, the dataset consists of data collected from 59 children who participated, each labeled as ASD or neurotypical (non-ASD). The class labels were balanced, with 30 participants clinically diagnosed with ASD and 29 classified as neurotypical. This balance reduces the risk of model bias due to class imbalance, ensuring a fair assessment of the performance of the classifier. Key demographic observations include age and sex. The participants ranged from approximately 3 to 13 years of age, with a mean age of 7.8 years. The age distribution was normal for both the classes with a slight skew towards younger participants in the neurotypical group. The dataset comprised 38 male and 21 female participants. Among ASD-labeled participants, 76% were male, reflecting the higher prevalence of ASD in boys.

3.5. Model Output and Dimension

In CNN-Only methods the model outputs a probability score between 0 and 1 for each image, indicating the likelihood of the image belonging to an ASD diagnosed child. Whereas in CNN + Sequence based methods, the model outputs a probability score between 0 and 1 for each individual, indicating the likelihood of being classified as having ASD. A threshold of 0.5 is applied to this probability to determine the final binary classification. The model's output dimension for binary classification is a scalar value for each individual.

3.6. Loss Function and Evaluation Metric

The model employs binary cross-entropy (BCE) as the loss function to optimize the binary classification task. To evaluate performance, we utilize metrics tailored to the medical context, including accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), and the F1 Score. These metrics provide a comprehensive assessment of the model's ability to classify individuals accurately, particularly in distinguishing between ASD and non-ASD classes.

4. Method

The primary objective of this study is to leverage Deep Learning techniques to classify eye tracking images and detect Autism Spectrum Disorder (ASD). The task involves binary classification (positive: ASD, and negative: Non-ASD), and we employ a sigmoid activation in the final layer with Binary Cross-Entropy (BCE) as the loss function.

4.1. Baseline and Initial Approach: Custom CNN

Our baseline for image-level classification originates from a prior analysis on the same dataset, where Akter et al.[16] employed classical machine learning algorithms and achieved 74.2% accuracy. We aimed to improve upon this established baseline by exploring Convolutional Neural Network (CNN)-based deep learning techniques. As a first step, we implemented a custom CNN model consisting of three convolutional layers, each followed by max-pooling, and a three-layer fully connected head for classification. This model served as our initial deep learning approach for image-level classification.

Figure 2. Custom CNN model

Through iterative tuning of hyperparameters, regularization strategies (e.g., dropout, weight decay), and data augmentation, the custom CNN eventually reached a test accuracy of 76.24% with a ROC AUC of 0.83. This exceeded the classical machine learning baseline (74.2%), demonstrating the potential of deep learning techniques for this task, though there remained room for further improvement.

4.2. Exploring Transfer Learning

To potentially enhance performance beyond the custom CNN, we explored transfer learning using pre-trained models such as ResNet50, DenseNet121, and VGG19. The rationale was to leverage the robust feature extraction capabilities of models trained on large image datasets. Residual connections in ResNet50, dense connectivity in DenseNet121, and the deep architecture of VGG19 offered distinct advantages. However, as shown later, the overall improvements in AUC and accuracy over the tuned custom CNN were marginal, indicating that while transfer learning stabilized validation performance, it did not dramatically surpass the carefully tuned custom CNN model.

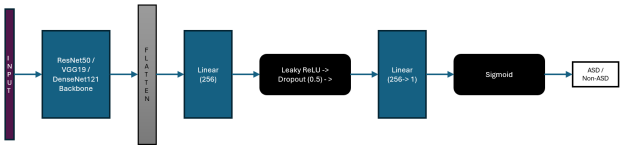

Figure 3. Transfer Learning model

4.3. Individual-Level Classification: Voting-Based Baseline

While we established a baseline for image-level classification (Akter et al.'s 74.2% and our improved CNN at 76.24%), there was no previously established baseline for individual-level classification. To address this gap, we implemented a voting-based method as our baseline approach for individual-level classification. After classifying each image from an individual using a Resnet50 based transfer learning model, we aggregated the predictions. The final individual-level prediction was determined by majority vote. This simple yet effective method served as the baseline for evaluating more advanced sequence-based models introduced subsequently.

4.4. Sequence Modeling with LSTM

We extended the modeling capabilities by incorporating a Long Short-Term Memory (LSTM) network to handle sequences of images. First, features from each image were extracted using a pre-trained CNN backbone (e.g., ResNet50 with the final classification layer removed). These features formed input sequences to the LSTM, allowing the model to capture temporal or relational patterns across multiple images of the same individual. Our aim was to surpass the voting-based baseline for individual-level classification by modeling the sequential nature of the data.

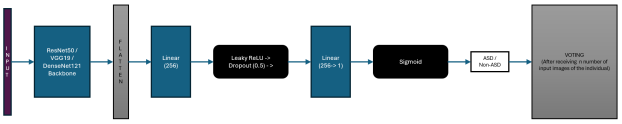

Figure 4. Voting based model

The LSTM model utilizes a hidden state dimension of 256. The architecture includes 2 stacked bidirectional LSTM layers. To prevent overfitting, a dropout rate of 0.5 is applied between the LSTM layers. The final output is generated through a fully connected layer with a sigmoid activation function for binary classification.

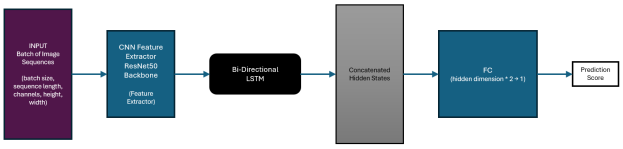

Figure 5. CNN+LSTM model

4.5. Attention-Augmented LSTM

To further enhance sequence modeling, we integrated an attention mechanism. This allowed the model to focus selectively on the most informative parts of the input sequence. By assigning varying importance to different images, the attention-augmented LSTM could highlight critical patterns, potentially improving both accuracy and interpretability compared to the baseline methods.

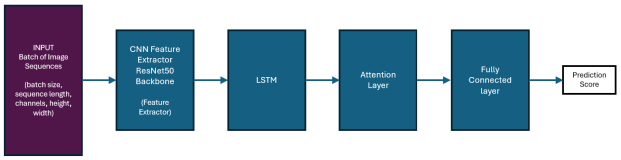

Figure 6. CNN + LSTM + Attention model

4.6. Transformer-Based Approach

Finally, we explored a Transformer-based approach to replace the recurrent architecture. The Transformer leverages self-attention to efficiently capture long-range dependencies and relationships within the sequence of image embeddings. This architecture scales well and can handle

variable-length inputs in parallel, potentially providing a more powerful and faster alternative to the LSTM-based models.

The transformer model consists of 512 hidden dimensions and 4 encoder layers. Each encoder utilized 8 attention heads and a feed forward network with a dimension of 2048. A dropout rate of 0.5 ensured regularization. A fully connected layer with a sigmoid activation was employed for binary classification.

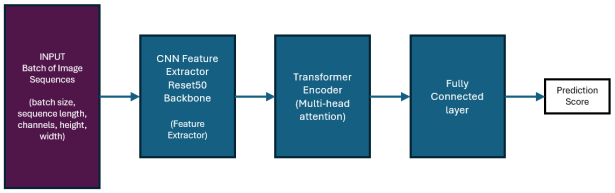

Figure 7. CNN + Transformer model

5. Experiments

5.1. Experimental Setup and Evaluation Metrics

To prevent data leakage, the dataset was split at the individual level rather than the image level, ensuring that images from the same individual did not appear across training, validation, and test sets. Participants were divided into 70% training, 15% validation, and 15% testing subsets, maintaining an equal ratio of positive and negative data points in each subset. This stratified approach preserved class balance and enhanced the reliability of model evaluation.

Data augmentation was applied exclusively to the training set to enhance generalization. Each participant's data was augmented 10 times using techniques such as random rotation, shearing, zooming, and filling. Cross-validation was avoided to prevent data leakage from augmented data inflating validation accuracy. Instead, a fixed validation set created during the initial split was consistently used across all models.

Preprocessing involved resizing images to a uniform size of (w,h) = (224,224) and normalizing pixel values using the mean and standard deviation computed from the training set. This normalization stabilized training, mitigated vanishing gradients, and ensured that all input data had a similar scale, facilitating faster convergence and improved performance.

Hyperparameter tuning was performed using grid search to maximize validation accuracy and model generalization. The key parameters tuned included the number of layers, layer dimensions, kernel sizes, learning rate, weight decay, dropout, batch size, and optimizer type. Advanced configurations such as the number of transformer heads, sequence model dimensions, and fine-tuning of transfer learning layers were also explored. A learning rate scheduler with a "decay on plateau" strategy was employed to reduce the learning rate when validation loss plateaued, helping the model escape suboptimal minima and achieve better overall performance.

Additionally, we explored a model checkpoint strategy to store the best model at different epochs during training. This approach ensured early stopping, allowing us to select the best model based on validation performance and use it for testing, avoiding overfitting and optimizing model generalization.

The models described in the 'Method' section were trained and evaluated using a dataset of eye-tracking images. We measured performance using metrics such as accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), F1 Score, and confusion matrices. Given the medical context, generalization and recall were emphasized. Experiments were conducted at both the image level (to improve upon the Akter et al. baseline of 74.2%) and the individual level. For individual-level classification, our voting-based method served as the baseline, against which we compared sequence-based models (LSTM, attention-augmented LSTM, and Transformers).

5.2. Custom CNN and Transfer Learning Results

The custom CNN initially faced issues with generalization and overfitting. However, after careful tuning, it achieved a test accuracy of 76.24% (improving upon the 74.2% baseline of Akter et al.[16]) and a ROC AUC of 0.83. This demonstrated the feasibility of surpassing the classical machine learning baseline using a deep learning approach.

We then explored transfer learning. Among the pre-trained models, ResNet50 achieved an AUC of 0.846, DenseNet121 achieved 0.834, and VGG19 achieved 0.802. While these performances were in a similar range as the tuned custom CNN, the margin of improvement was not substantial. The tuned custom CNN's competitive results indicate that a domain-specific, carefully optimized model can perform comparably to more complex pre-trained architectures.

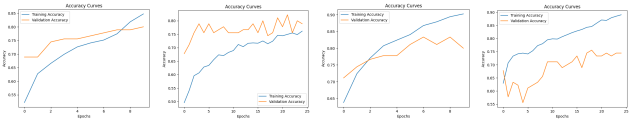

Figure 8. Training and validation accuracy curves for ResNet50, DenseNet121, VGG19 and Custom CNN models (left to right).

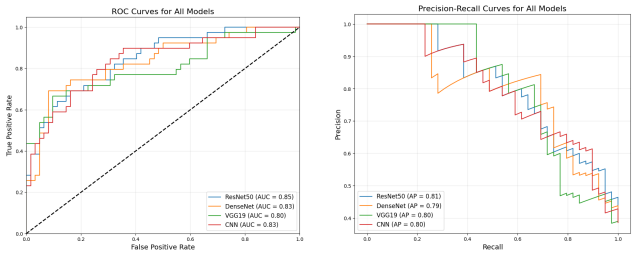

Figure 9. ROC and PR curves of ResNet50, DenseNet121, and VGG19 and Custom CNN models

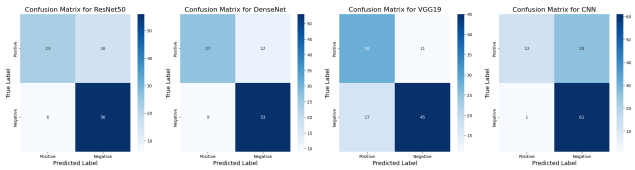

Figure 10. Confusion matrices for ResNet50, Densenet121, VGG19 and custom CNN (lest to right). The tuned custom CNN is competitive with transfer learning models, illustrating that careful optimization can yield strong baseline performance.

Table 1. Metrics/Performance of Custom CNN (A), ResNet (B), DenseNet (C), and VGG19 (D).

Met/Mod	Base	A	B	C	D
Accuracy	0.742	0.7624	0.7822	0.7921	0.7228
AUC	0.715	0.8346	0.8465	0.8341	0.8023
F1-Score	0.736	0.49056	0.6764	0.72	0.6666

Upon observing Figure 8, ResNet50 exhibits a steady improvement in both training and validation accuracy, indicating robust generalization and minimal overfitting. DenseNet121, while showing effective training progression, has minor fluctuations in validation accuracy, suggesting some sensitivity to the dataset. VGG19's performance, with its larger gap between training and validation accuracies. The Custom CNN model exhibited a huge gap between the training and validation accuracies. These observations indicate that ResNet50 performs better and is able to generalize better among the models.

Figure 9 shows the ROC and PR curves, where we can see that the ResNet and the DenseNet demonstrate superior discriminatory power and better precision-recall balance. The Figure 10 shows the confusion matrices where DenseNet achieves the best balance of sensitivity and specificity, while VGG19 underperformed with a higher misclassification rate, particularly in false negatives.

From Table 1 we can summarize that DenseNet achieved the highest accuracy(79.21%) and F1-score (0.72), highlighting its strong overall performance in binary classification. ResNet attained the highest AUC(0.8465), demonstrating its ability to distinguish between classes effectively. The custom CNN performed competitively with an accuracy of 76.42% and AUC of 0.8346, demonstrating the effectiveness of domain specific training. The VGG19 model lagged behind, indicating it may be less suitable for this task.

5.3. Individual-Level Classification: Voting-Based Baseline Results

For individual-level classification, no established baseline existed. By adopting a voting-based approach (aggregating predictions from multiple images of the same individual), we established a baseline to evaluate sequence-based models. The voting-based method improved stability at the individual level and reached an AUC of 0.90, indicating the value of combining multiple samples per individual.

5.4. LSTM and Attention-Augmented LSTM Results

By introducing LSTM-based models, we aimed to surpass the voting-based baseline. The LSTM incorporated sequence modeling, capturing temporal dependencies and improving classification beyond simple aggregation. Adding an attention mechanism further refined this approach, allowing the model to emphasize crucial images. This resulted in near-perfect performance (AUC = 1.0) on validation data, underscoring the advantages of sequence modeling and informed attention in boosting individual-level classification accuracy.

5.5. Transformer-Based Model Results

Transitioning to a Transformer architecture facilitated more efficient and potentially more powerful sequence modeling. The Transformer-based model achieved over 85% validation accuracy and a ROC AUC of 0.95, illustrating strong discriminative capability while handling variable-length inputs in parallel. These results confirm that more advanced sequence modeling architectures can significantly outperform the voting-based baseline for individual-level classification.

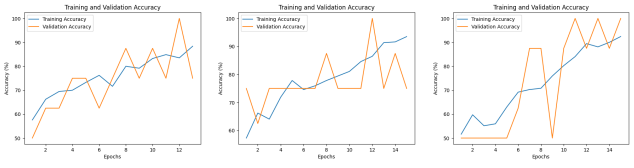

Figure 11. Training and validation accuracies for the LSTM (left), Attention Augmented LSTM (centre) and Transformer based (right) models

The transformer based model shows steady and consistent improvement during both training and validation,

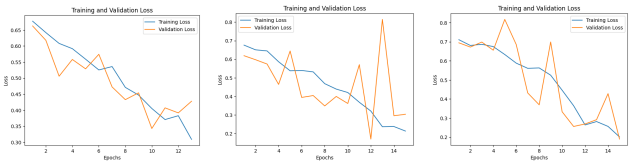

Figure 12. Training and validation Loss for the LSTM (left), Attention Augmented LSTM (centre) and Transformer based (right) models

as seen in the figure above. This highlights its ability to generalize better compared to the LSTM and Attention-Augmented LSTM models, which display more fluctuations in validation accuracy, suggesting a sensitivity to the complexity of the data.

The Attention-Augmented LSTM performs better than the basic LSTM, achieving higher accuracy and lower loss. This improvement is due to the attention mechanism, which allows the model to focus on the most important parts of the input sequence and learn relevant features more effectively. However, despite this enhancement, it still lacks the stability seen in the Transformer model.

The Transformer model stands out by showing minimal overfitting, with a smaller gap between training and validation performance. Its loss graphs also demonstrate faster and more consistent convergence to a lower validation loss, indicating that it not only learns more quickly but also better captures the underlying patterns in the data. While the Attention-Augmented LSTM improves upon the basic LSTM, it cannot match the efficiency and stability of the Transformer model.

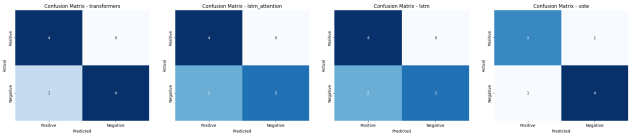

Figure 13. Confusion matrices for CNN + Transformer, CNN + LSTM + Attention, CNN + LSTM and CNN + Voting models (left to right). It can be observed that Transformer based model has the best performance

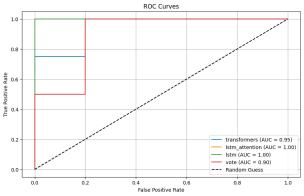

Figure 14. ROC Curves for individual-level classification models.

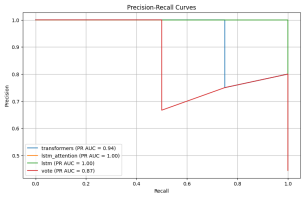

Figure 15. PR Curves for individual-level classification models.

Table 2. Performance Metrics for CNN + Voting (A), CNN + LSTM (B), CNN + LSTM + Attention (C), and CNN + Transformer (D).

Probability / Confidence Score Table				
Metrics/Model	A	B	C	D
Accuracy	0.7778	0.7778	0.7778	0.8889
AUC	0.9000	1.0000	1.0000	0.9500
F1-Score	0.6153	0.8000	0.8000	0.8888

5.6. How Confident Are the Models in Prediction?

To assess how confident the models are in their predictions and how the sequence length (i.e., the number of images for an individual) affects model prediction confidence, we conducted an experiment where the sequence length for individuals in the test set was incrementally increased. Figure 16 illustrates the results for the Transformer-based model, showing the predicted probability for Participant 19, a positive ASD sample, as the number of input images increases from 1 to 16. The plot demonstrates that the confidence of the Transformer model steadily increases as it is exposed to more images, highlighting its ability to utilize additional information to make more confident predictions.

Table 3 presents the confidence scores for various test set individuals as predicted by different models, including CNN + Voting, CNN + LSTM, CNN + LSTM + Attention, and the Transformer-based model. The Transformer model consistently achieves higher confidence in its predictions compared to the other models, underscoring its superior effectiveness in aggregating and analyzing sequence-level information. For Participant 19, the Transformer model demonstrates high confidence with a probability close to 1.0 as more images are included, confirming its robustness in classifying positive ASD cases.

From these results, it is evident that increasing the number of images significantly enhances prediction confidence, particularly for the Transformer model. This highlights the advantages of Transformer-based architectures in capturing patterns across multiple input images, making them well-suited for this task. In contrast, the other models, while capable, demonstrate less confidence in their predictions, further emphasizing the robustness and reliability of the Transformer-based approach.

Nevertheless, certain misclassifications (e.g., Participant 48 being misclassified as autistic) suggest that data noise or outliers can mislead even advanced models. Such cases underscore the need for careful data quality checks, potential domain-specific preprocessing, or incorporating clinical insights.

5.7. Result Inference

The comparison of different models highlights key insights into their effectiveness for diagnosing Autism Spec-

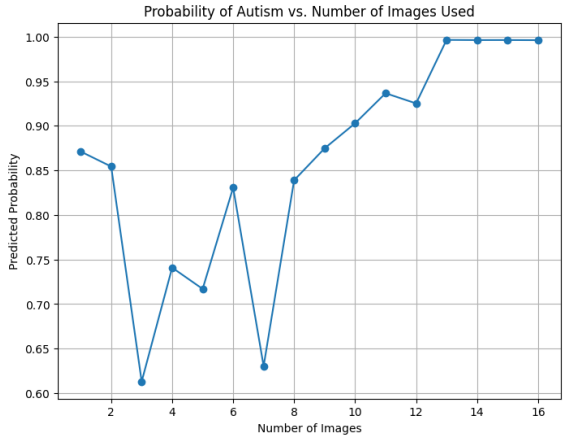

Figure 16. Participant 19 data on Transformer based model: Increasing the number of images per individual leads to more confident predictions, emphasizing the value of richer participant-specific data.

Table 3. Probability / Confidence Score Table for CNN + Voting (A), CNN + LSTM (B), CNN + LSTM + Attention (C), and CNN + Transformer (D).

Probability / Confidence Score Table				
ID / Model	A	B	C	D
TC-49	0.0000	0.6314	0.5582	0.1687
TC-31	0.0000	0.2026	0.1138	0.0330
TC-33	0.0526	0.2110	0.0445	0.0158
TC-58	0.0000	0.4207	0.1209	0.0113
TC-48	0.1250	0.6472	0.6154	0.9831
TS-9	0.7500	0.8828	0.8216	0.9978
TS-4	0.0714	0.8915	0.9068	0.9969
TS-19	0.6250	0.9239	0.9302	0.9961
TS-8	0.6000	0.8496	0.7683	0.9931

trum Disorder (ASD) using eye-tracking data. The Custom CNN performed well as a baseline model, achieving an accuracy of 76.24% and an AUC of 0.83. Transfer learning models like ResNet50 and DenseNet121 offered slight improvements, with ResNet50 achieving the highest AUC of 0.846 and an accuracy of 78.22%. This suggests that transfer learning provides modest benefits, but a well-tuned, task-specific model like the Custom CNN can perform nearly as well.

Sequence-based models demonstrated a clear advantage over simpler methods like the voting-based approach. While the voting-based model aggregated predictions effectively, achieving an AUC of 0.90, it lacked the ability to analyze relationships between images. Models like LSTM and Attention-Augmented LSTM outperformed the voting method by incorporating temporal dependencies, achieving near-perfect AUC scores of 1.0. The attention mechanism further improved performance by focusing on the most critical parts of the data, resulting in greater prediction stability.

The Transformer model stood out as the best-performing approach, achieving an accuracy of 88.89% and an AUC of 0.95. Its ability to process sequences of varying lengths and capture complex relationships between inputs made it particularly effective. The experiment with Participant 19, a positive ASD case, showed that the Transformer's prediction confidence increased steadily as more images were provided, highlighting the importance of richer datasets. This underscores the Transformer model's strength in making reliable predictions when given more information.

6. Limitations

The main limitation of our method is the limited dataset size. With data from only 59 participants, it becomes challenging to ensure robust generalization across diverse populations. While satisfactory performance was achieved using deep learning techniques and careful parameter tuning, the generalizability of our approach would likely improve with a larger dataset that captures more variation in the target population.

Another limitation is the presence of noise in the dataset. Variability and outliers, such as image data from individuals diagnosed as ASD-positive but displaying non-ASD characteristics (and vice versa), impact the consistency of model performance. These inconsistencies reduce the accuracy of predictions and highlight a need for better data curation.

A minor limitation arises from the specialized equipment required for data collection, particularly the need for eye-tracking devices. This dependency increases the difficulty of gathering new, high-quality data, potentially slowing future improvements to the model.

Despite these limitations, our method has demonstrated strong performance, particularly with the transformer-based model, which excels in identifying patterns and classifying data with high precision. We observed that increasing the number of images per individual further enhances prediction confidence, suggesting the value of richer datasets in improving outcomes.

7. Conclusion

This study explored how eye-tracking data, combined with advanced deep learning techniques, can help diagnose Autism Spectrum Disorder (ASD) more effectively. By using models like CNN-LSTM with attention and Transformers. We found that incorporating temporal and relational patterns in the data through sequence modeling, along with leveraging attention mechanisms, significantly improved the accuracy and reliability of our predictions. Moreover, using multiple images per participant allowed us to create a more detailed and robust profile for each individual, reinforcing the importance of richer datasets in achieving accurate diagnoses. From this project, we learned that

aggregating data is essential for stronger individual-level predictions. Sequence-based models, especially those enhanced with attention mechanisms, provided a clear advantage by focusing on the most important parts of the data, offering both precision and insights into how the models arrived at their decisions. Transformers, in particular, stood out as its ability to handle complex, variable-length data efficiently, makes it ideal for scaling up this approach in real-world settings.

7.1. Future Directions

We believe that there is plenty of potential to build on this work. One of the biggest next steps is to expand the dataset. With more diverse participants, including different age groups, the models can become even more generalizable and effective across a wider range of populations and can also reduce the noise. Another potential option is combining eye-tracking data with other information, like speech patterns, facial expressions, or physiological data, to create a more complete and nuanced diagnostic method for a more accurate prediction.

Beyond ASD, this methodology could also be applied to other conditions where eye-tracking data can provide insights, such as ADHD or anxiety disorders. With further development, this work has the potential to become a powerful, non-invasive tool for early diagnosis, improving the lives of those affected and helping clinicians make more informed decisions.

In addition to these, future work could explore the use of advanced techniques like Visual Transformers, which are particularly suited for handling spatial information in image-based tasks. These models may offer improved performance in understanding complex visual patterns in eye-tracking data. Other advanced techniques, such as self-supervised learning, few-shot learning, or ensemble methods, could further enhance the robustness and accuracy of the models.

7.2. Code

The code for this project can be found at the following link: GITHUB

References

[1] Figshare, "Dataset for Autism Spectrum Disorder Detection Using Eye Tracking Data," [Online]. Available: `https://figshare.com/s/5d4f93395cc49d01e2bd`.

[2] A. Hussaini, et al., "A Systematic Literature Review on the Application of Machine-Learning Models in Behavioral Assessment of Autism Spectrum Disorder," *Journal of Personalized Medicine*, vol. 11, no. 4, Apr. 2021.

[3] Hodges, H., Fealko, C., Soares, N., "Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation," *Transl Pediatr.*, vol. 9, Suppl 1, pp. S55–65, 2020.

[4] Raj, S., Masood, S., "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," *Procedia Computer Science*, 2020. DOI: 10.1016/j.procs.2020.03.399.

[5] Crippa, A., Salvatore, C., et al., "Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities," *Journal of Autism and Developmental Disorders*, 2015. DOI: 10.1007/s10803-015-2379-8.

[6] Tao, Y., Shyu, M.L., "SP-ASDNet: CNN-LSTM Based ASD Classification Model Using Observer Scanpaths," *2019 IEEE ICME Workshops*. DOI: 10.1109/ICMEW.2019.00062.

[7] Alam, M.E., et al., "An IoT-Belief Rule Base Smart System to Assess Autism," *Proceedings of iCEEiCT 2018*. DOI: 10.1109/ICEEICT.2018.8646020.

[8] Eslami T, Mirjalili V, Fong A, Laird AR, Saeed F. ASD-DiagNet: A Hybrid Learning Approach for Detection of Autism Spectrum Disorder Using fMRI Data. Front Neuroinform. 2019;13:70. Published 2019 Nov 27. doi:10.3389/fninf.2019.00070

[9] Cantin-Garside, K.D., Kong, Z., White, S.W. et al. Detecting and Classifying Self-injurious Behavior in Autism Spectrum Disorder Using Machine Learning Techniques. J Autism Dev Disord 50, 4039–4052 (2020). https://doi.org/10.1007/s10803-020-04463-x

[10] Beary, M., Hadsell, A., Messersmith, R., & Hosseini, M. (2020). Diagnosis of Autism in Children using Facial Analysis and Deep Learning. ArXiv, abs/2008.02890.

[11] Z. A. T. Ahmed, "Convolutional neural network for prediction of autism based on eye-tracking scanpaths", Int. J. Psychosocial Rehabil., vol. 24, no. 5, pp. 2683-2689, Apr. 2020.

[12] E. Arora, H. S. Jolly and A. Rehalia, "Prediction of Autism Spectrum Disorder Using ANN and CNN," 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023, pp. 457-461, doi: 10.1109/ICTACS59847.2023.10390530. keywords: Deep learning;Autism;Image analysis;Predictive models;Data models;Behavioral sciences;Reliability;ASD;ANN;CNN;Cognitive;Eye-tracking;Eyeblink;Eye-fixation,

[13] Carette et al. [1] presented a method for visualizing eye-tracking patterns in Autism Spectrum Disorder (ASD), utilizing a pretrained Xception model for feature extraction and a stacking ensemble framework

[14] Loth, E., Charman, T., Mason, L., Tillmann, J., Jones, E. J. H., Wooldridge, C., Ahmad, J., Auyeung, B., Brogna, C., Ambrosino, S., et al., "The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders," *Molecular Autism*, vol. 8, pp. 1–19, 2017.

[15] Guillon, Q.; Hadjikhani, N.; Baduel, S.; Roge, B. "Visual social attention in autism spectrum disorder: Insights from eye tracking studies," *Neurosci. Biobehav. Rev.*, 2014, 42, 279–297.

[16] Lord, C.; Risi, S.; DiLavore, P.S.; Shulman, C.; Thurm, A.; Pickles, A. "Autism from 2 to 9 years of age," *Arch. Gen. Psychiatry*, 2006, 63, 694–701.

[17] Centers for Disease Control and Prevention, "Diagnosing Autism Spectrum Disorder (ASD)," [Online]. Available: `https://www.cdc.gov/autism/diagnosis/index.html`.

[18] T. Akter, M. H. Ali, M. I. Khan, M. S. Satu, and M. A. Moni, "Machine Learning Model To Predict Autism Investigating Eye-Tracking Dataset," *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, Dhaka, Bangladesh, 2021, pp. 383-387, doi: 10.1109/ICREST51555.2021.9331152.