Transformer-Based Longitudinal Autism Behavioral Analysis using AQ10 Scores

Illy Hoang

Suining He, PhD

University of Connecticut, CSE Department

1 Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by persistent challenges in social environments and the presence of repetitive restricted behaviors. Understanding the development trajectory of these behaviors is critical for developing effective interventions and improving the quality of life of individuals with ASD. Longitudinal studies, which track individuals for extended periods, provide invaluable insights into the dynamic nature of these behaviors. However, contemporary research in both disciplines of neurological disorders and Machine Learning (ML) have yet to leverage state-of-the-art models to visualize these trends.

1.1 Problem Statement

Early diagnosis and intervention are critical for improving long-term outcomes for individuals with ASD. However, current diagnostic processes often rely on subjective and time-consuming clinical evaluations, which can delay access to necessary support services. Furthermore, there are significant gaps in longitudinal data collection and analysis, particularly in understanding how ASD traits evolve over time and how age-related patterns manifest in standardized screening tools such as Autism-Spectrum Quotient (AQ).

The AQ10 is a 10-item questionnaire derived from the longer 50-item AQ developed by Baron-Cohen et al. It is a widely used screening tool designed to assess traits associated with ASD in adults. AQ10 focuses on behaviors and preferences related to social interaction, communication, attention to detail, and routine, which are key characteristics of ASD. Each item is scored on a 4-point Likert scale (e.g., "Definitely agree" to "Definitely disagree"), with higher total scores indicating a greater likelihood of autistic traits. A score of 6 or higher (out of 10) is typically considered indicative of potential ASD, which warrants further clinical evaluation. AQ10, a widely used screening tool, is generally administered at a single point in time. This limits the ability to analyze how AQ10 scores and ASD traits change with age, particularly during critical developmental stages such as childhood, adolescence, and adulthood. Understanding age-related patterns in AQ10 responses could provide valuable information on the stability or variability of ASD traits over time, as well as inform personalized intervention strategies [2].

Considering the interpret-ability of data derived from the AQ10 questionnarie, ML methodologies offer a promising avenue to address gaps by leveraging existing data to uncover patterns and make predictions that would be difficult to achieve using traditional manual statistical methods.

2 Literature Review

1. ASD, as defined by the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-5), is characterized by persistent deficits in social communication and social interaction, along with restricted, repetitive patterns of behavior, interests, or activities [1]. This diagnostic framework provides the foundation for understanding the core features of ASD and serves as a reference point for research and clinical practice. While the DSM-5 provides a robust framework for diagnosis, it does not explicitly detail the developmental trajectories of specific behaviors over time. Longitudinal studies, coupled with advanced analytical techniques, are needed to further explore these trajectories. This project aims to complement the DSM-5's diagnostic criteria to potentially reveal patterns of behavioral persistence and change that are not readily apparent through traditional diagnostic assessments.

2. Longitudinal studies are crucial for elucidating the stability and change of these behaviors over time. Gotham, Pickles, and Lord demonstrated the utility of standardized observational measures, specifically the Autism Diagnostic Observation Schedule (ADOS), in tracking social interaction symptom severity from toddlerhood to adulthood. Their findings highlight the potential for quantitative assessment of ASD symptom progression, providing a foundation for our proposed analysis [5]. However, traditional statistical methods may not fully capture the complex, sequential nature of behavioral data. Recent advancements in Large Language Models (LLMs) offer a novel approach to modeling such sequences, potentially revealing previously obscured patterns in longitudinal behavioral data. This project aims to bridge the gap between established longitudinal assessment methods and deep learning techniques, exploring the feasibility of using predictive models to analyze behavioral trends in ASD.

3. Fitzmaurice, Laird, and Ware (2011) provide a comprehensive framework for statistical methodologies used in analyzing such data. Their work emphasizes the importance of accounting for within-subject correlation and between-subject variability when modeling longitudinal outcomes [4]. However, while these methods effectively capture population-level trends, they may not fully exploit the sequential dependencies inherent in individual behavioral trajectories – especially those unique communities such as being autistic. This project aims to explore the application of predictive language models. potentially offering a more granular understanding of behavioral persistence and change in ASD.

4. The analysis of longitudinal data necessitates models capable of capturing complex temporal dependencies. Vaswani et al. introduced the Transformer architecture, an approach in natural language

processing that relies on self-attention mechanisms to model long-range dependencies within sequential data. This innovation has revolutionized sequence modeling. Given the inherent sequential nature of longitudinal behavioral data, where past behaviors influence future patterns [10], the Transformer architecture offers a promising avenue for analysis. Unlike traditional statistical methods, Transformers can potentially identify subtle patterns of behavioral persistence and change. This project proposes to leverage the capabilities of a Transformer-based language model to analyze longitudinal survey data that will be processed into sequential representations for model learning.

5. Predictive modeling can analyze various modals of data to predict ASD diagnosis. Prediction models in past implementations have shown to consistently provide better results in terms of accuracy, specificity, sensitivity, precision and false positive rates in autism diagnosis based on screening data [7]. Implication of multi-modal analysis such as through images classification demonstrates promising classification ability with contemporary models [9]. These and numerous other studies exhibit the reducibility in reliance on lengthy clinical evaluations and provides scalable screening tools.

6. An inherent issue for longitudinal studies is garnering enough data from a sufficient population. Synthesizing data using ML techniques is a method that has been used to address gaps by generating realistic data points or filling in missing values. This enables researchers to augment datasets and analyze age-related patterns even when long-term tracking is unavailable, improving data generality as well as overall model performance [6]. Synthetic tables generated by generative adversarial networks (GANs) for training ML models have demonstrated performance statistically similar to original tables yet do not incur information leakage and also prepare the model for unknown testing cases [8]. Integration of generative artificial intelligence (GenAI) models, such as GPT-4, for synthesizing data serves as a practical means without significantly impacting data variance while overcoming limitations in data acquisition [3].

3 Project Proposal

I propose a predictive model to analyze longitudinal behavioral trends in autism using a transformer-based architecture. First, the data will be imported directly from Kaggle. I will then perform a concise exploratory analysis to highlight important features for extraction. The data will be preprocessed, transforming survey responses (binary values) into sequential representations of each individual's behavioral history, including age and time information. The data will also be augmented by replicating the sequential array with

masked survey responses for specific representative characteristics. Following this, I will train and fine-tune a Transformer-based language model, similar to BERT, framing the task as a next-token prediction problem, to model behavioral trajectories. I will then visualize and analyze the predicted trends, comparing behavioral persistence between individuals with and without ASD. To validate the model performance, I will use appropriate metrics such as standard accuracy, precision, recall, and F1-score. Finally, I will assess the feasibility of using language models for this type of longitudinal analysis and report on the identified trends. I hope to submit my work as a formal dissertation to conclude my undergraduate career at UConn.

3.1 Proposed Model

The model leverages the Transformer architecture proposed in the original paper by Vaswani et al., specifically designed to analyze questionnaire responses alongside demographic data sequentially for dual prediction tasks. The architecture includes three necessary components:

3.1.1 Feature Embedding Layer

This initial stage projects input features—comprising demographic details (age, gender) and AQ-10 questionnaire responses—into a 64-dimensional embedding space using linear transformations followed by GELU activation and layer normalization.

3.1.2 Transformer Encoder

The model consists of four Transformer encoder layers, each utilizing multi-head attention with 4 heads. This allows the model to effectively capture complex dependencies and interactions among the features through self-attention mechanisms.

3.1.3 Dual Output Heads

The final stage consists of two separate prediction heads: the **Diagnosis Prediction Head** predicts clinical diagnosis of ASD (binary classification); and the **Classification Prediction Head** predicts the patient classification based on AQ10 responses and demographic input. The model was implemented with the following configurations:

- **Model Dimension:** 64
- **Number of Transformer Encoder Layers:** 4

- **Attention Heads per Layer:** 4
- **Feedforward Network Dimension:** 256
- **Dropout Rate:** 0.2
- **Learning Rate:** 0.001
- **Batch Size:** 64
- **Epochs:** 20
- **Optimizer:** Adam optimizer was used due to its efficiency in handling noisy gradients and sparse data.
- **Regularization:** Dropout was applied at various stages within the model architecture to prevent overfitting, specifically within the embedding layers and Transformer encoder blocks. Early stopping was also implemented during training, monitoring validation loss with a patience of 3 epochs to ensure optimal model performance without overfitting.

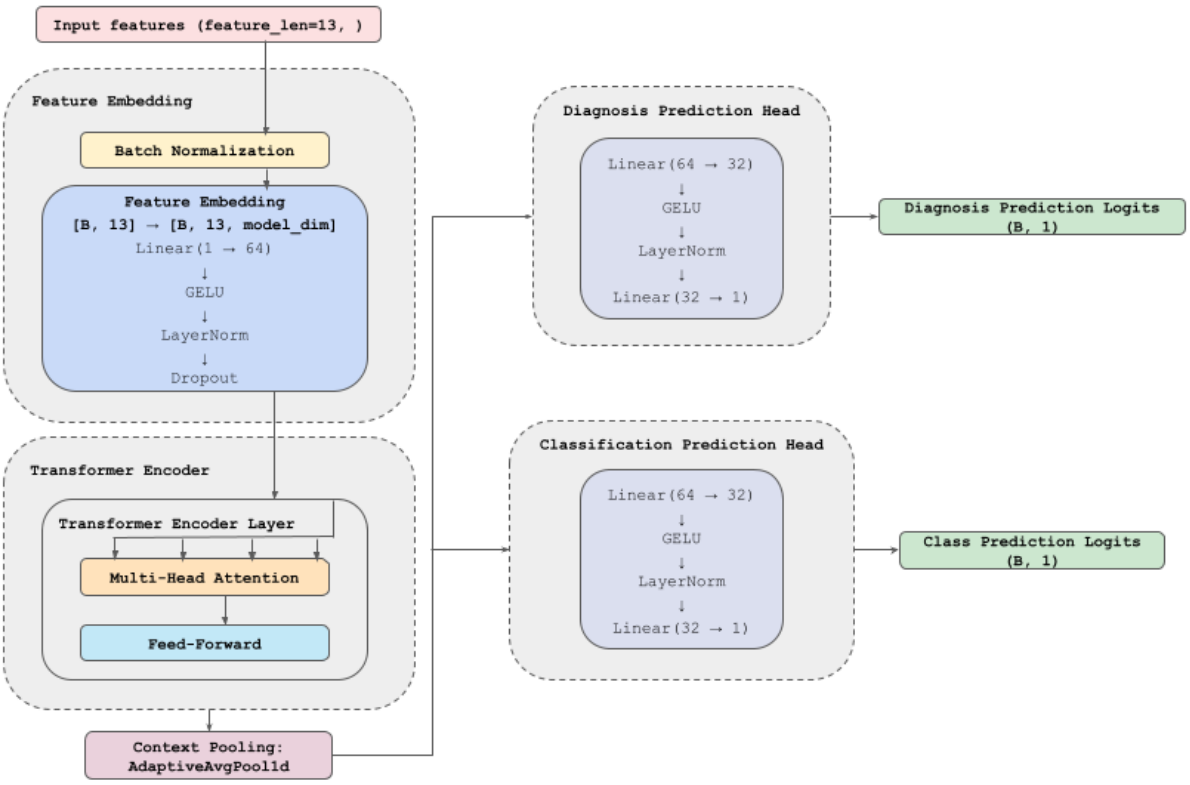

Figure 1: Architectural Design of Proposed Model.

4 Training

4.1 Dataset Description

The dataset used in this project is sourced from the publicly available "Autism Screening in Adults" dataset on Kaggle, available at www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults/data. It contains anonymized records of adults screened for Autism Spectrum Disorder (ASD). This dataset was compiled to facilitate the development of ML models for ASD prediction and analysis. The dataset contains 704 records. While it does not provide extensive coverage by recording results of individuals through various significant stages of development, the dataset provides a moderate sample size for training and validation.

4.2 Feature Preparation

Input features included normalized demographic data (age, gender) and binary responses from the AQ-10 questionnaire. Features were standardized to ensure optimal performance and comparability.

statistic	A1-10_Score	age (year)	gender	result	diagnosis	classification
count	702	702	702	702	702	702
mean	*	29.698006	*	4.883191	*	*
std	*	16.507465	*	2.498051	*	*
min	*	17	*	0	*	*
max	*	383**	*	10	*	*

Table 1: Summary of relevant statistics for working DataFrame.
*Omitted values as they hold no significant value to scope of study
**Seems unrealistic relative to human age. This entry was still used in model training.

4.2.1 AQ Question Score

Each A1_Score to A10_Score corresponds to a specific behavioral question (e.g., social preferences, attention to detail) according to the AQ10 questionnaire developed by Baron-Cohen et al. As mentioned prior, each item is scored on a 4-point Likert scale with higher total scores indicating a greater likelihood of autistic traits. The dataset scores patient responses in binary values with 1 signifying either "definitely agree" or "slightly agree" and 0 signifying "slightly disagree" or "definitely disagree." A score of 6 or higher (out of 10) is typically considered indicative of potential ASD.

4.2.2 Age

This column indicates the age of the participant (continuous value). The values in this column are normalized to the range [0, 1] to improve model training stability using

$$\frac{\text{value} - \min(\text{column})}{\max(\text{column}) - \min(\text{column})}, \tag{1}$$

to normalize each entry in this column during data preprocessing. There were also two entries with null age values which were removed during preprocessing.

4.2.3 Gender

This column indicates the gender of the participant, with 1 signifying the a male participant and 0 signifying a female participant.

4.2.4 Result

This column indicates the total AQ10 scored (sum of A1_Score to A10_Score), ranging from 0 to 1. As used for the values under the Age column, the values in this column were normalized using Equation (1) during preprocessing.

4.2.5 Autism

This column is the clinical diagnosis of the participant. The values in this column were binary encoded, renamed under the alias "Diagnosis", with 1 indicating a positive diagnosis for ASD and 0 indicating a negative diagnosis, during preprocessing.

4.2.6 Classification

This column is the screening classification of the participant based on their AQ10 result. This column was also binary encoded during preprocessing with 1 signifying an AQ10 score ≥ 6 and 0 signifying an AQ10 score < 6.

4.3 Baseline Models

I selected a variety of traditional ML algorithms commonly used for classification tasks to establish a benchmark for evaluating the performance of the proposed model. The chosen baseline models include Logistic

Regression, Decision Trees, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, MultiLayer Perceptron (MLP), eXtreme Gradient Boost (XGBoost), and Light Gradient-Boosting Machine (LightGBM).

4.4 Model Training

Each model was individually trained using the training dataset, employing appropriate hyperparameters. Specifically, the Decision Tree classifier was limited to a depth of 5, Random Forest utilized 100 estimators, Logistic Regression was trained with a maximum of 1000 iterations, and SVM applied an RBF kernel.

5 Evaluation

The trained models were evaluated using the test dataset on both tasks to measure predictive performance. The evaluation metrics included accuracy, precision, recall, and F1-score.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{2}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The baseline models' performance metrics were systematically recorded, and each model was saved for reproducibility and comparative analysis against the proposed model. The collected metrics established baseline performance levels for further validating improvements achieved by the proposed approach.

5.1 Preliminary Results

After training both the baseline and preliminary models, each model was evaluated on predicting a participant's classification status (based on whether their AQ10 score is ≥ 6 and considering demographic data) and predicting the diagnosis label.

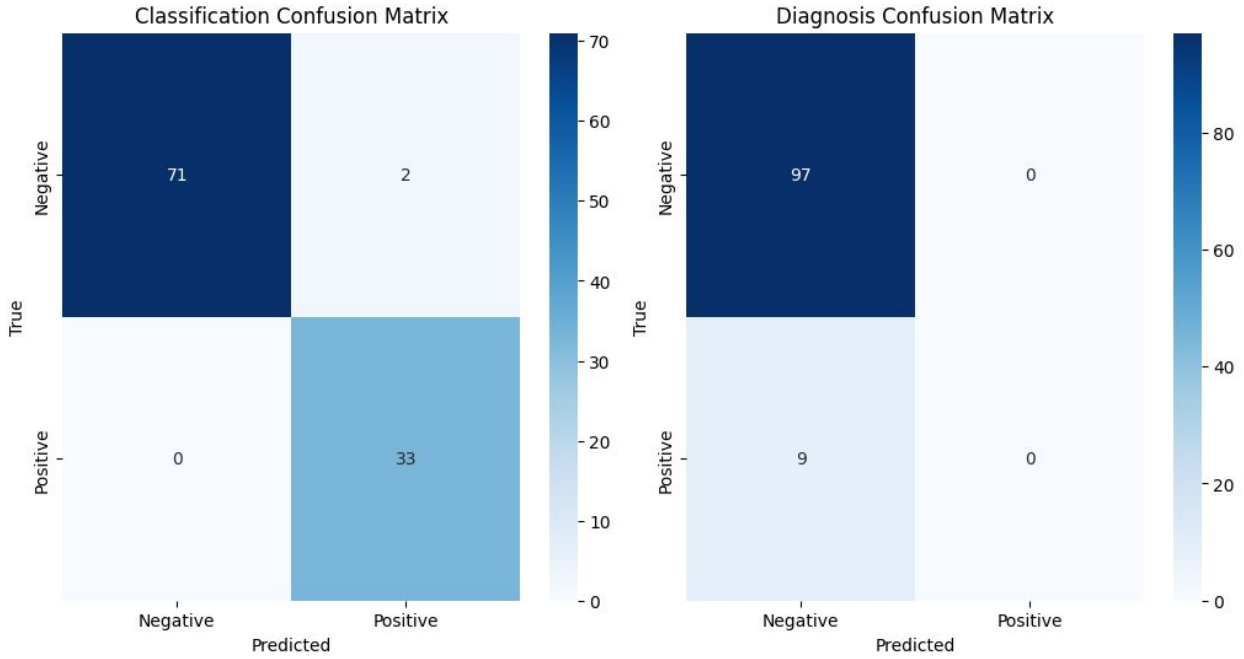

Figure 2: Confusion Matrices of Classification (left) and Diagnosis (right) Predictions of Preliminary Model.

For the classification task, the model exhibits near-perfect performance, with almost all positive and negative cases correctly classified. By contrast, in the diagnosis task's confusion matrix, the model displays a bias toward the negative class—leading to zero true positives and a failure to detect any actual positives. This discrepancy suggests that the dataset's distribution for diagnosis is more imbalanced or less informative than the classification task, requiring additional tuning or data augmentation to capture true positives effectively.

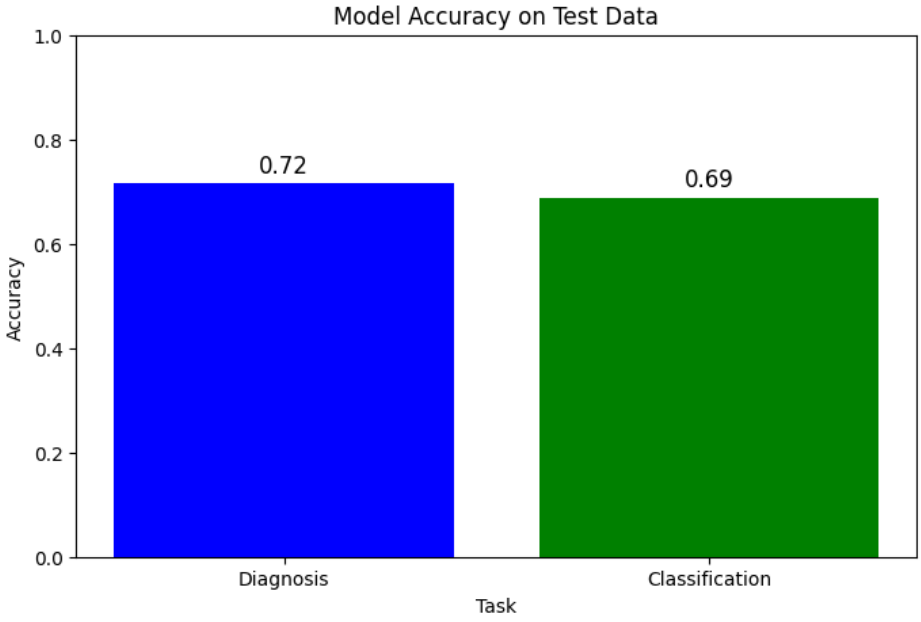

Figure 3: Bar Plot of Preliminary Model Accuracy.

While accuracy reaches approximately 0.98 (98.1%) on classification, it is around 0.92 (91.5%) for diagnosis. At first glance, a 91.5% accuracy for diagnosis might seem strong, but the confusion matrix (Figure 2) reveals that this improvement is coming from predicting "negative" for nearly the entire test set. Because most participants in this dataset are negative for a clinical ASD diagnosis, defaulting to the negative class yields high accuracy but leads to very poor precision, recall, and F1.

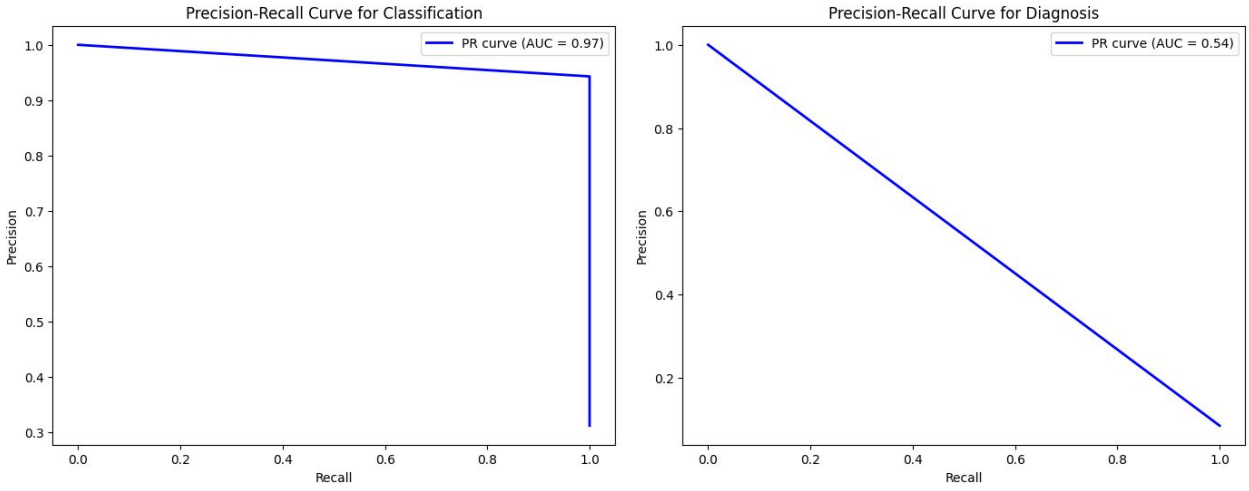

Figure 4: Precision-Recall Curve of Preliminary Classification (left) and Diagnosis (right) Predictions.

In Figure 4, the left chart (classification) shows a higher area under the curve, indicating that the preliminary model can maintain high precision for a broad range of recall values. On the right chart (diagnosis), however, the curve collapses, reflecting the model's difficulty in identifying the positive class. The near-zero precision and recall confirm that it often fails to predict true positives.

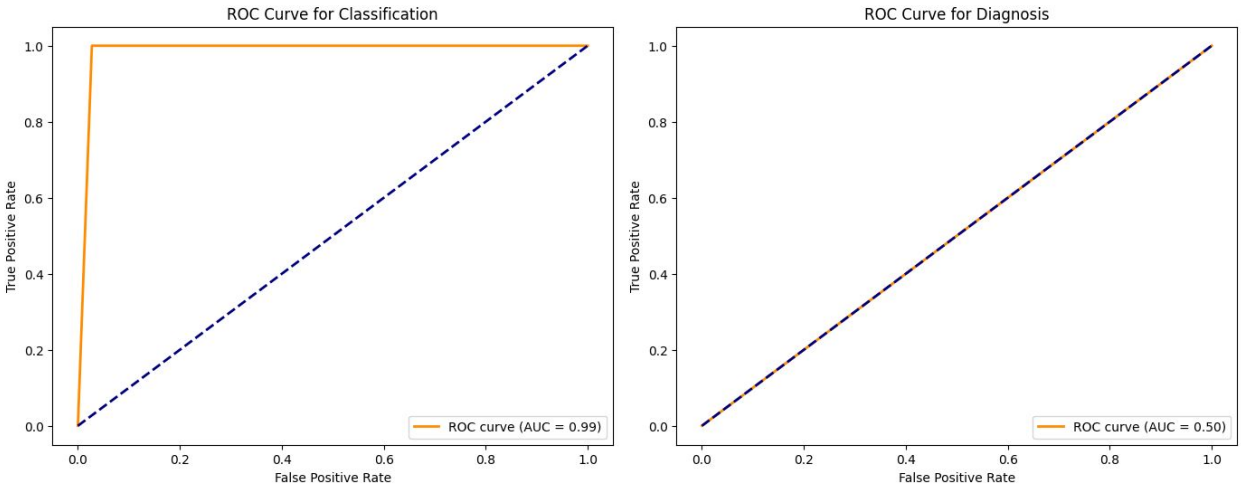

Figure 5: ROC of Preliminary Classification (left) and Diagnosis (right) Predictions.

In Figure 5, for classification, the curve remains near the top-left corner, implying strong discriminative power with a high true positive rate across thresholds. For diagnosis, the curve highlights minimal sensitivity (true positive rate), reinforcing the observation that the model rarely predicts positive for ASD diagnosis under current settings.

Tables 2 and 3 summarize how the baseline models and the preliminary model performed on classification and diagnosis, respectively.

model_name	accuracy	precision	recall	f1_score
svm	0.943	1.000	0.818	0.900
decision_tree	1.000	1.000	1.000	1.000
random_forest	1.000	1.000	1.000	1.000
log_regression	1.000	1.000	1.000	1.000
knn	0.924	0.931	0.818	0.871
naive_bayes	0.774	1.000	0.273	0.428
mlp	0.924	1.000	0.758	0.862
xgboost	1.000	1.000	1.000	1.000
lightgbm	1.000	1.000	1.000	1.000
prelim_model	0.981	0.943	1.000	0.970

Table 2: Summary Table of Preliminary Performance of Classification Task against Baseline Models.

Table 2 shows that several baseline methods (Decision Tree, Random Forest, Logistic Regression, XGBoost, LightGBM) achieved perfect metrics—100% accuracy, precision, recall, and F1—on the classification task. In contrast, some baselines like Naïve Bayes or KNN had lower scores. The preliminary Transformer model scored 98.1% accuracy, 94.3% precision, 100% recall, and 97.0% F1. Although it did not exceed certain baselines that achieved perfect results, it was still extremely competitive. The near-perfect performance across so many baselines suggests potential overfitting or data leakage for the classification label. Further cross-validation or holdout subsets may be warranted to ensure these scores truly generalize.

model_name	accuracy	precision	recall	f1_score
svm	0.755	0.185	0.556	0.278
decision_tree	0.736	0.212	0.778	0.333
random_forest	0.736	0.212	0.778	0.333
log_regression	0.736	0.212	0.778	0.333
knn	0.736	0.172	0.556	0.263
naive_bayes	0.849	0.111	0.111	0.111
mlp	0.792	0.240	0.667	0.353
xgboost	0.736	0.212	0.778	0.333
lightgbm	0.736	0.212	0.778	0.333
prelim_model	0.915	0.000	0.000	0.000

Table 3: Summary Table of Preliminary Performance of Diagnosis Task against Baseline Models.

Table 3 reveals a stark contrast in performance for the diagnosis label. Most baseline models (Decision Tree, Random Forest, Logistic Regression, XGBoost, LightGBM) cluster around 73–75% accuracy, with modest precision and recall. Naïve Bayes stands out at 84.9% accuracy but yields very low precision and recall (0.111 each). Meanwhile, the preliminary Transformer model reports a 91.5% accuracy but 0.0 precision, 0.0 recall, and 0.0 F1. This result aligns with the confusion matrix in Figure 2, indicating that the model is defaulting to a negative prediction for the diagnosis class. Consequently, while the reported accuracy is high, the model is not actually capturing any true positives (i.e., individuals truly diagnosed with ASD).

6 Discussion

6.1 Preliminary Findings

The significant drop in recall (and precision) for diagnosis, compared to classification, strongly suggests label imbalance or insufficient discriminative features. If only a small fraction of participants are truly diagnosed with ASD, models can inflate accuracy by predicting everyone as negative. This inflates "accuracy" while

yielding 0 for precision and recall.

The ease with which several baseline models (Decision Tree, Random Forest, Logistic Regression, XGBoost, LightGBM) achieve 100% classification scores may indicate overfitting or potential data leakage. Despite rigorous cross-validation attempts, such perfect metrics in real-world data are uncommon. Additional verification procedures—like a larger holdout set, different random seeds, or cross-site data—may be needed.

The preliminary architecture matches or nearly matches the best baselines for classification, confirming its capability to handle the questionnaire features effectively. But it's shortcoming on the diagnosis task highlights the complexity and sparsity of clinically diagnosed ASD data. Future work might focus on addressing data imbalance (via re-weighting the loss function, oversampling minority classes, or generating synthetic samples for the diagnosis-positive class).

Given the starkly different performance across these two labels, additional experiments—adjusting hyperparameters (e.g., class weighting) or employing data augmentation—are vital for improving recall on the diagnosis task. Interpreting the results of advanced metrics beyond accuracy is also critical to ensure that truly positive cases of ASD are not being missed.

6.1.1 Next Steps

The classification task proved relatively straightforward for both baselines and the preliminary Transformer model, achieving near-perfect accuracy and F1. However, the diagnosis task (verifying a clinically confirmed ASD diagnosis) emerged as more challenging and imbalanced. While the Transformer achieved higher raw accuracy (91.5%) than many baselines, it essentially captured no positive ASD cases in its predictions.

Moving forward, balancing the dataset, tuning hyperparameters, and modifying the loss function (e.g., focal loss, weighted cross-entropy) are recommended to improve the Transformer's sensitivity to true positive diagnoses. Additional evaluation strategies (such as a more stringent cross-validation or a larger external dataset) can also help confirm whether the classification success is robust or if overfitting is inflating the results.

References

[1] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th, 2013.

[2] Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders*, 31:5–17, 2001.

[3] Chung Yin Chan. Data-driven innovation: The potential of synthetic data through generative ai. 2024.

[4] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied longitudinal analysis*. John Wiley & Sons, 2012.

[5] Katherine Gotham, Andrew Pickles, and Catherine Lord. Standardizing ados scores for a measure of severity in autism spectrum disorders. *Journal of autism and developmental disorders*, 39:693–705, 2009.

[6] Cherry Khosla and Baljit Singh Saini. Enhancing performance of deep learning models with different data augmentation techniques: A survey. In *2020 international conference on intelligent engineering and management (ICIEM)*, pages 79–85. IEEE, 2020.

[7] Kazi Shahrukh Omar, Prodipta Mondal, Nabila Shahnaz Khan, Md Rezaul Karim Rizvi, and Md Nazrul Islam. A machine learning approach to predict autism spectrum disorder. In *2019 International conference on electrical, computer and communication engineering (ECCE)*, pages 1–6. IEEE, 2019.

[8] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.

[9] Bhaskar Sen, Neil C Borle, Russell Greiner, and Matthew RG Brown. A general prediction model for the detection of adhd and autism using structural and functional mri. *PloS one*, 13(4):e0194856, 2018.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.