# Development and validation of a social interaction based deep learning system to predict the severity of social skill in autism spectrum disorder

Joo Hyun Lee

Department of Biomedical Systems Informatics
The Graduate School, Yonsei University

# Development and validation of a social interaction based deep learning system to predict the severity of social skill in autism spectrum disorder

Directed by Professor Yu Rang Park

The Master's Thesis submitted to the Department of Biomedical Systems Informatics, and the Graduate School of Yonsei University in partial fulfillment of the requirements for the degree of Master of Science

Joo Hyun Lee

December 2023

This certifies that the Master's Thesis of
Joo Hyun Lee is approved.

___________________________________
Thesis Supervisor: Yu Rang Park

___________________________________
Thesis Committee Member I: Dukyong Yoon

___________________________________
Thesis Committee Member II: Hwajung Hong

The Graduate School
Yonsei University

December 2023

# ACKNOWLEDGEMENTS

From the moment I realized that I wanted to study the field of medical informatics, but didn't know how to do it specifically, until now, when I finished my degree, every moment has been precious.

I would like to express my deepest gratitude to my supervisor, Prof. Yu Rang Park, who has consistently and eloquently conveyed her sense of adventure in research and scholarship and her passion for teaching. Without her guidance and patience, this thesis would not have been possible.

I would also like to express my sincere gratitude to my thesis committee members, Prof. Dukyong Yoon, and Prof. Hwajung Hong. Their advice and insights helped me to complete this thesis.

I would also like to thank the members of DHLab and MCU who spent a lot of time with me and helped me with my research.

Finally, I would like to thank my family for their silent support behind the scenes.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABSTRACT

## Development and validation of a social interaction based deep learning system to predict the severity of social skill in autism spectrum disorder

Joo Hyun Lee

*Department of Biomedical Systems Informatics*
*The Graduate School, Yonsei University*

Directed by Professor by Yu Rang Park

**Background:** Children with autism spectrum disorder (ASD) have difficulty with social interactions, making social ability one of the important measures for diagnosing ASD. However, existing assessment methods for measuring social ability are costly and time-consuming and may involve examiner bias. Therefore, there is a need for an objective and standardized assessment tool to measure social ability in children with ASD.

**Objective:** We aimed to 1) develop and validate a protocol to digitize the nonverbal social communication skills of children with ASD, 2) evaluate a digitized protocol of nonverbal social communication skills and its correlation with neuropsychological test, and 3) develop a deep learning model that can predict the nonverbal social communication skills of children with ASD using video data collected through the developed protocol.

**Methods:** The study is prospective and observational study. Eligible children were assessed using the Autism Diagnostic Observation Schedule-2 (ADOS-2) and a neuropsychological test (NPT) to evaluate their social skills. A specific turn taking protocol was developed to measure and videotape the children's social interactions. This data was then used to train three different deep learning models: an RGB model, a flow model, and a combined RGB-flow late fusion model. The models were designed to predict the severity of social skills in the participating children.

**Results:** The study included data from 9 participants. The evaluation of the digitized nonverbal social communication skill measurement protocol showed significant differences in turn-taking performance between groups with mild (median (IQR): 100.0 [87.5 to 100.0]) and severe (median (IQR): 12.5 [0.0 to 66.7], p-value = 0.048) social skill impairments. The combined RGB-Flow late fusion deep learning model exhibited superior performance, achieving high accuracy (93.33%), precision (0.91), recall (1.0), F1 score (0.96), and area under the receiver operating characteristic curve (AUC, 0.99) in predicting social skill severity. The Grad-CAM algorithm was applied to these models, revealing that the models primarily focused on the child's face and toy interactions for making predictions.

**Conclusion:** To the best of our knowledge, this is first study has demonstrated the feasibility of collecting datasets for behavioral biomarkers using a standardized video data collection setup suitable for computer vision and deep learning, as well as measuring nonverbal social communication skills. According to the findings of this study, objectively measuring social-communication skills may provide objective information for ASD diagnosis or may be a good alternative for objectively measuring the effectiveness of social skill improvement interventions.

**Keywords:** Autism Spectrum Disorder; Nonverbal social communication skill; Social skill severity prediction deep learning model; Transfer leaning

# Development and validation of a social interaction based deep learning system to predict the severity of social skill in autism spectrum disorder

Joo Hyun Lee

*Department of Biomedical Systems Informatics*
*The Graduate School, Yonsei University*

Directed by Professor by Yu Rang Park

## I. INTRODUCTION

### 1.1. Background

### 1.1.1. Characteristics of Autism Spectrum Disorder (ASD)

Social interaction plays a crucial role in social relationships. Autism Spectrum Disorder (ASD) is characterized by difficulties in social interaction, verbal, and nonverbal communication, and restricted or repetitive behaviors (RRB) [1]. Children with ASD show a lack of social interactions, such as turn-taking, in contrast to typically developing children. Thus, social interaction, such as turn taking, is one of the many psychological hypotheses used to explain the psychopathology of ASD [2].

### 1.1.2. Assessing social skill in ASD

The unique characteristics of children with ASD necessitate the use of accurate and standardized tools for assessing deficits in social interactions. There are established and validated methods for evaluating behaviors related to social interactions [3]–[5]. The Autism Diagnostic Observation Schedule (ADOS), Autism Diagnostic Interview-Revised (ADI-R), and Social Responsiveness Scale (SRS) represent the most commonly utilized diagnostic tools in ASD assessment. These tools are crucial for diagnosing ASD and evaluating aspects such as communication, social interaction, and imaginative play [4], [5]. Social communication skills are important in ASD diagnosis, and standard assessments are available for this purpose. 'Turn taking' is a notable behavior that is frequently assessed in these assessments [2]–[4], [6]. These assessment tools are generally considered to be of high quality due to their structured nature and the specialized training required for their administration. Nonetheless, there is potential for bias in these standard tools. Variables that are difficult to predict can influence the interaction between the child and the examiner, and the reliability of these assessments can be dependent on the guardian’s memory and understanding. Furthermore, these tests are administered by trained professionals in specialized institutions, resulting in long waiting times and high costs [7].

### 1.1.3. Need for an interpretable predictive model

To address these issues, there has been recent development of machine learning or deep learning models for objective screening of children with ASD [8]–[11]. Nevertheless, the generalizability of these models is hampered by specialized and expensive equipment, reliance on human assessments of autistic behaviors, lack of automation, and low accuracy levels [9]–[11]. This may be due to the lack of biomarkers for the target behaviors that are important for measuring social communication skill in children with ASD or the limitations of the models in capturing the complex characteristics of children with autism. In addition, deep learning models require large amounts of data to train and label, which can be costly and labor intensive. Children's data is particularly difficult to collect. To overcome these challenges, transfer learning is employed. This approach leverages existing labeled large datasets and pre-trained models from related tasks to enhance performance and generalize model [12]–[15].  However, despite these achievements, the complexity of these models has escalated, often turning them into opaque 'black box' systems [16]. Interpretable AI enhance clarity and foster trust by offering transparent insights into their decision-making processes, a factor that is particularly critical in delicate areas like ASD. The aspect of interpretability is paramount when integrating AI technologies in healthcare settings. It is essential to comprehend the rationale behind diagnostic or screening outcomes, making this aspect as significant as the outcomes themselves [16], [17]. Therefore, there is a need for interpretable digital technologies that utilize clinically validated methods to objectively measure the social communication skills of children with ASD with a small amount of data and capture their complex characteristics.

### 1.2. Objective

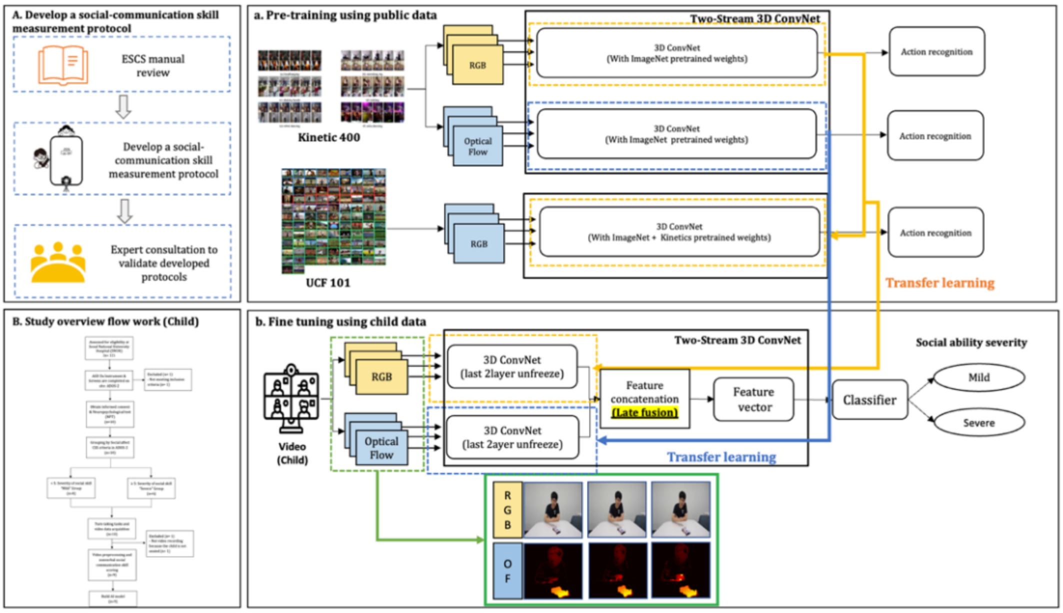
We aimed to 1) develop and validate a protocol to digitize the nonverbal social communication skills of children with ASD, 2) evaluate a digitized protocol of nonverbal social communication skills and its correlation with neuropsychological test, and 3) develop a deep learning model that can predict the nonverbal social communication skills of children with ASD using video data collected through the developed protocol.

# II. Methods

## 2.1. Study design

This prospective and observational study was designed to develop and evaluate a protocol for measuring nonverbal social communication skills in children with ASD and to develop and evaluate a deep learning model for predicting social skills in children with ASD from data collected by the protocol. Children who wish to participate in the study administered the Autism Diagnostic Observation Schedule-2 (ADOS-2) to determine if they meet criteria. They administered the Neuropsychological Test (NPT) to assess their social skills, emotional and behavioral problems, and motor coordination difficulties. We designed a turn taking protocol to measure and video record the child's social interactions. The child's social interaction behavior was observed throughout the protocol to measure their ability to engage in playful and positive interactions with others.  The test was conducted by a qualified professional examiner. During the test, the child's face and behavior captured through video recording. The data collected through the protocol was used to develop a deep learning model to predict the severity of social skills in children with ASD. The overall study design is in Figure 1. Written informed consent was obtained from all children and parents who participated in this study. This study was approved by the Seoul National University Hospital Institutional Review Board (SNUH IRB).

## A. Overview of study design


A. Develop a social-communication skill measurement protocol
ESCS manual review
Develop a social-communication skill measurement protocol
Expert consultation to validate developed protocols
B. Study overview flow work (Child)
a. Pre-training using public data
Kinetic 400
UCF 101
RGB
Optical Flow
RGB
Two-Stream 3D ConvNet
3D ConvNet (With ImageNet pretrained weights)
3D ConvNet (With ImageNet pretrained weights)
3D ConvNet (With ImageNet + Kinetics pretrained weights)
Action recognition
Action recognition
Action recognition
Transfer learning
b. Fine tuning using child data
Video (Child)
RGB
Optical Flow
Two-Stream 3D ConvNet
3D ConvNet (last 2layer unfreeze)
3D ConvNet (last 2ayer unfreeze)
Feature concatenation (Late fusion)
Feature vector
Classifier
Social ability severity
Mild
Severe
Transfer learning
R G B
O F

## B. Overall study workflow

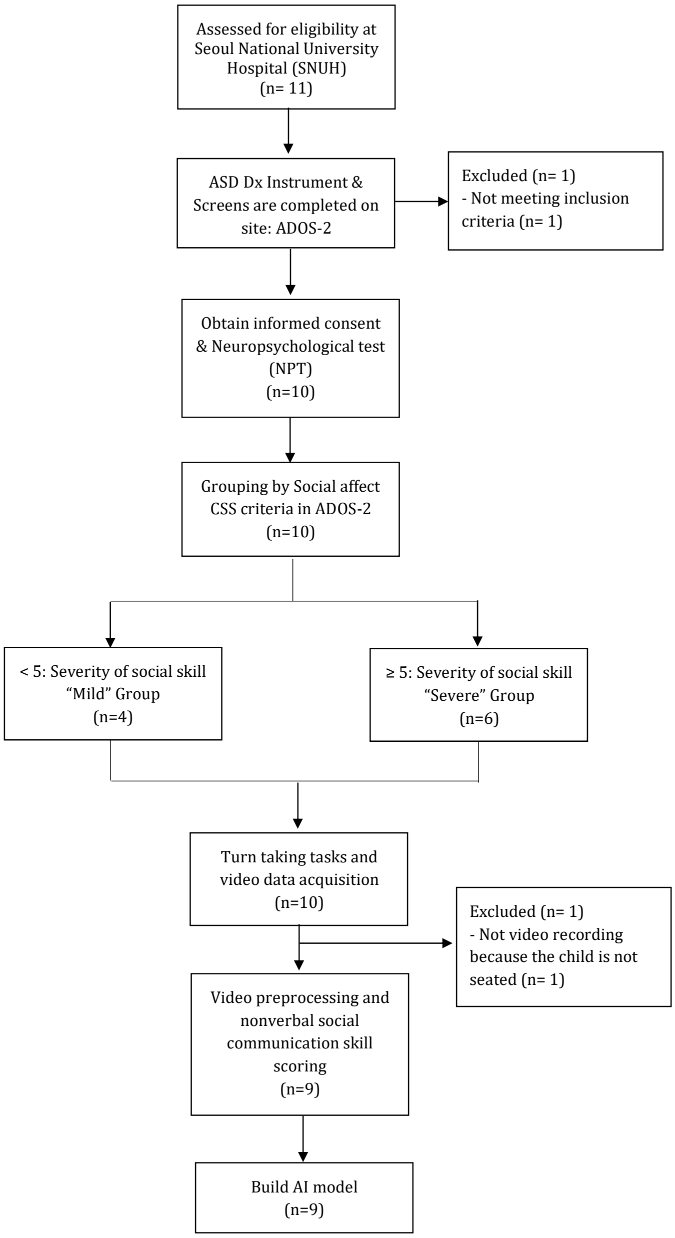

Assessed for eligibility at Seoul National University Hospital (SNUH) (n= 11)
ASD Dx Instrument & Screens are completed on site: ADOS-2
Excluded (n= 1)
- Not meeting inclusion criteria (n= 1)
Obtain informed consent & Neuropsychological test (NPT) (n=10)
Grouping by Social affect CSS criteria in ADOS-2 (n=10)
< 5: Severity of social skill "Mild" Group (n=4)
≥ 5: Severity of social skill "Severe" Group (n=6)
Turn taking tasks and video data acquisition (n=10)
Excluded (n= 1)
- Not video recording because the child is not seated (n= 1)
Video preprocessing and nonverbal social communication skill scoring (n=9)
Build AI model (n=9)

**Figure 1.** Overall study workflow (A. Overview of study design and B. Overall study workflow)

### 2.2. Participants

Children recruited from July 2023 to November 2023 from the Department of Psychiatry, Seoul National University Hospital. Inclusion criteria for children are as follows: (1) age between 30 and 71 months, (2) children diagnosed with ASD based on DSM-5 by a pediatric psychiatrist. Exclusion criteria for children are as follows: (1) the child has a history of congenital or acquired brain injury such as cerebral palsy, (2) the child has hearing or visual impairment, (3) the child has been diagnosed with neurological (motor, muscular) disorders, and (4) the child has shown adverse reactions or abnormal responses to previous sedation or anesthesia.

### 2.3. Protocol for measuring digitized nonverbal social communication skills

In this study, we designed a protocol for measuring and video recording social interaction situation, adopting methods from the ESCS manual of Mundy et al [6]. Throughout the social interaction assessment protocol, we observed the child's social interaction behavior to measure their ability to have playful, affectively positive turn-taking interactions with others. The specific tasks that can measure social interaction behavior are presented in the social interaction assessment protocol (Table 1).

**Table 1.** Turn taking task administration guidelines for measuring social interaction behavior

| Target behavior | Administration |
|---|---|
| Initiating Social Interaction (ISI) | 1) The child grabs a toy (car) in front of them.<br>2) The child gives the toy to the examiner.<br>3) The examiner gives the toy back to the child.<br>4) The child gives the toy back to the examiner again. |
| Responding Social Interaction (RSI) | 1) The examiner says to the child, "Let's play with the toy," and adopts a posture with open hands so they can catch the ball when the child throws it or stop the car when the child rolls it. (10 seconds)<br>2) If the child responds by rolling the car to the examiner, the examiner rolls the car back to the child.<br>3) Continue the activity until the child stops rolling the car or until the child has completed five turns.<br>4) If the child does not start the game, the examiner prompts the child for the toy, making a suitable noise to attract the child's attention, but does not explicitly ask for the toy. (10 seconds)<br>5) If the child does not respond after waiting 10 seconds, the examiner rolls the car to the child while making a "whirring" sound.<br>6) If the child still does not respond, the examiner repeats the action in 5). (Do not ask for the toy at this point).<br>7) After another 10 seconds without a response, the examiner explicitly asks the child for the toy. (“Give the car to the teacher.”)<br>8) If the child still does not respond after 3 seconds, repeat the action in 7) 2 more times and stop the test. |

There are two types of social interaction behavior: Initiating social Interaction (ISI) and Responding social Interaction (RSI). ISI refers to a child's ability to start turn-taking interactions and their inclination to playfully engage with the examiner. RSI relates to how often a child makes eye contact, uses gestures, and turns-taking exhibited by a child in response to turn-taking interactions initiated by the examiner. For the tasks, the examiner places a toy car within easy reach of the child. Subsequently, the examiner positions their hands on the table, prepared to catch the car if the child decides to roll it. If the child responds by rolling the car to the examiner, the examiner retrieves the car and again rolls it to the child. This turn-taking task continues until the child stops rolling the car or the child has taken 5 turns. The experimental setup for social interaction assessment is shown in Figure 2.

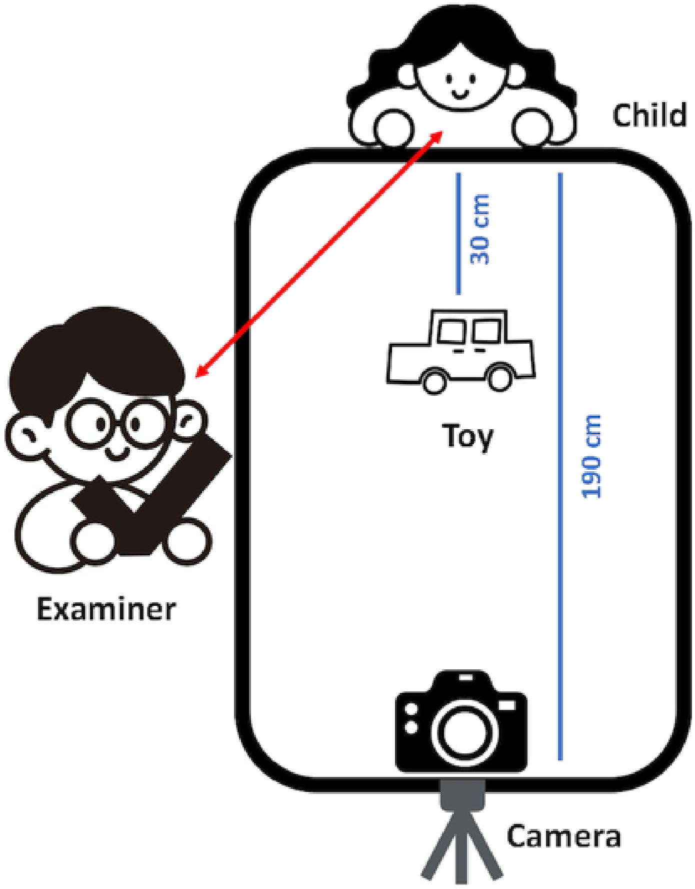


**Figure 2.** Experimental setup for nonverbal social communication skill assessment (Turn taking task)

### 2.4. Measures

#### 2.4.1. Screening and Diagnostic Assessments

Children who attended the outpatient department of pediatric psychiatry at Seoul National University Hospital who were willing to participate in this study and were suspected of having an ASD were referred to a pediatric psychiatrist for diagnostic evaluation. The ADOS-2 was administered of screen for study participation and diagnosis ASD [18]–[20]. The ADOS-2 total scores provide a measure of autism severity. These scores combine symptoms from the Social Affect (SA) and Restricted and Repetitive Behaviors (RRB) domains. In this study, we used a calibrated social affect score as a representative measure of a child's socialization and categorized into two labels (mild and severe) for training the deep learning model [19].

### 2.4.2. Cognitive Functioning Assessments

Cognitive function was assessed using a variety of measures, depending on the child's age and ability to perform demanding cognitive tasks. We defined the best estimate of IQ using the Korean Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition (K-WPPSI-IV) [21].

### 2.4.3. Neuropsychological test (NPT)

The Social Responsiveness Scale-2 (SRS-2)[3], Korean Child Behavior Checklist (K-CBCL)[22], [23], Korean Vineland-II (K-VABS-2)[24], [25], and Developmental Coordination Disorder Questionnaire (DCDQ)[26] were administered via parent questionnaires to assess children's social skills, emotional and behavioral problems, and motor coordination disorders.

### 2.4.4. Turn Taking Tasks and Video Data Acquisition

Video data collected in social interaction situations designed to measure ISI and RSI behaviors (Table 1, Figure 2). The video captures at a resolution of 1920 X 1080 and a rate of 30 frames per second, recording the children throughout the social interaction assessment. In the designated setup for this evaluation, the examiner prompts social interaction behaviors based on the Turning Tasking task protocol. A digital (RGB) camera records the child's face and upper torso in real-time during the assessment.

Assessment of task performance is typically conducted via observation made from recorded video. Primary coding consisted of recording the occurrence of social interaction behaviors. In looking at an interaction, the coder should, first classified the function, and second, determined who initiated the function (to determine if the child's behavior was an initiation or response). This sequence of judgments is important to note as individual behavioral forms (e.g., "points") are rated by behavioral function rather than just behavioral topography. In each trial, the scores for ISI and RSI are calculated. The scores for these two behaviors are summed and defined as "success" for a score of 2 and "fail" for a score of 1 or 0 (Table 2). Coding was performed by two people: one pediatric psychiatrist and one registered nurse.

**Table 2.** Social interaction score guideline (a. Social interaction assessment criteria, b. Social interaction score for each behavior, c. Social interaction total score, d. Social interaction score formula)

a. Social interaction assessment criteria

| | Social interaction assessment criteria |
|---|---|
| ISI | - Child rolls car to examiner (first time)<br>- Child returns car to the examiner (after the examiner returns the car to the child) |
| RSI | - The child exchanges cars with the examiner during a sequence.<br>(Refers to a sequence of turn taking in which the child rolls the car to the examiner) |

b. Social interaction score for each behavior

| Score | Behavior type | |
|---|---|---|
| | ISI | RSI |
| 0 | Initiate social interaction fail | Response social interaction fail |

| | | |
|---|---|---|
| 1 | Initiate social interaction success | Response social interaction success |

c. Social interaction total score

| | Total social interaction label |
|---|---|
| Success | Combined score of ISI and RSI equals 2 |
| Fail | Combined score of ISI and RSI equals 1 or 0 |

d. Social interaction score formula

**1. Definition**

$ISI_{score}$: Score for Initiating Social Interaction

$RSI_{score}$: Score for Responding Social Interaction

$Total_{score}$: Combined score of $ISI_{score}$ and $RSI_{score}$

**2. Initialization**

$ISI_{score} = 0$

$RSI_{score} = 0$

$Total_{score} = 0$

**3. Evaluation scoring**

```
1) Evaluate Social Interaction ($ISI_{behavior}, RSI_{behavior}$):
        $if\ ISI_{behavior}$ = ‘Success’
        $ISI_{score} = 1$,
        $else\ ISI_{behavior}$ = ‘Fail’
        $ISI_{score}$ = 0,

        $if\ RSI_{behavior}$ = ‘Success’
        $RSI_{score} = 1$,
        $else\ RSI_{behavior}$ = ‘Fail’
        $RSI_{score} = 0$
```

```
4. Calculate Total Score
$Total_{score} = ISI_{score} + RSI_{score}$

5. Determine Outcome
        $if\ \ Total_{score} = 2$,
        Outcome = ‘Success’
        $else\ Total_{score} = 1$ or $Total_{score} = 0$
        Outcome = ‘Fail’
```

*ISI: Initiate Social Interaction, RSI: Response to Social Interaction*

## 2.5. Developing a deep learning model to assess the severity of social skill in ASD

### 2.5.1. Pre-training with public data

To overcome the limitation of having insufficient child data for predicting the social ability of children with ASD, we pre-trained our model using public data. This strategy aligns with the findings that models pre-trained on extensive video data (such as Kinetics [27], UCF-101[28]) exhibit enhanced performance and better generalization [13]–[15], [29]. We trained the models use ImageNet and Kinetics pre-trained Inflated 3D ConvNets (I3D) model, and UCF 101 dataset. UCF101 is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories with 13,320 videos. For all architectures, we follow each convolutional layer with a batch normalization layer and a ReLU activation function, except for the final convolutional layers, which produce the class scores for each network.

### 2.5.2. Fine-tuning using child data

In our study, we employed a late fusion deep learning along with RGB and Optical Flow data. The foundational architecture of our model relies on the well-established "Inception-v1 I3D" architecture (Figure 3) [29]. This architecture incorporates "Inception" modules that introduce parallel paths to process the given inputs and then concatenation the outputs of each path to the respective module's output. These "Inception" blocks are reiterated multiple times at specific points, performing max-pooling operations and intermittent down-sampling. In our work, we utilized RGB video data and optical flow based on the RGB data as individual modalities. Both of these modalities are trained using Inception-v1 I3D, and to accommodate the characteristics of our dataset, the last two module was unfrozen during the learning phase. After the training, the features are concatenated (late fusion) following the last dense layer.

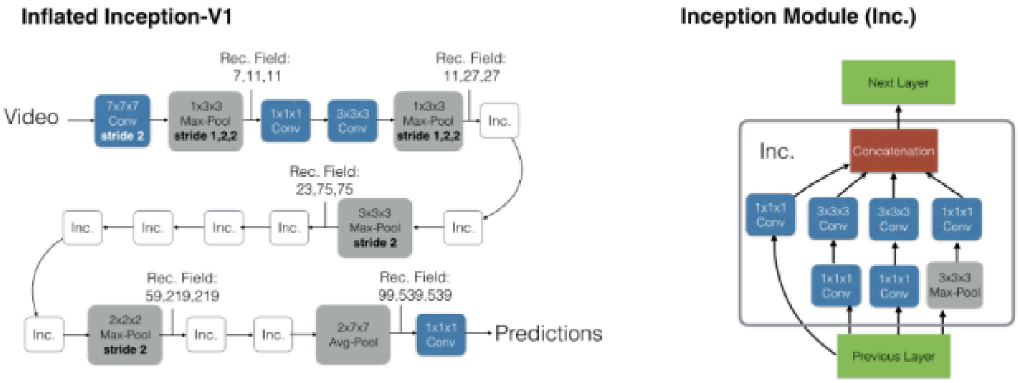


**Figure 3.** Layout of the Inception-v1 I3D model (adapted from [29])

We trained to predict the social abilities of children with ASD by transfer learning a model that we pre-trained using late fusion of video data. We use the same model architecture "Inception-v1 I3D" for the children's video data as the model we pre-trained with public data. The video data is extracted into RGB video and optical flow[30] based on the RGB data, and each individual modality is trained using "Inception-v1 I3D", and the features of each individual modality are concatenated (late fusion) to extract a feature map representing the learned features. The fusion step is to connect the feature maps generated from the RGB data and the optical flow data to generate a vector feature map from all modalities (late fusion), followed by a classifier to model the entire multimodal feature representation to predict social ability. The size of the input data for deep learning is 224*224*100. (The average number of video frames was 274, so we sampled uniformly every 3 frames to extract 100 frames. If the length is less than 100 frames, we perform zero padding.) Data augmentation is known to be of crucial importance for the performance of deep learning model. During the training we used random horizontal flipping (left-right flipping) with a 50% probability. To address the imbalance in the data set, class weights were calculated to ensure balanced learning[31]. Given two classes, ‘Mild’ and ‘Severe’, the number of samples for each class $\boldsymbol{N_{Mild}}$ and $\boldsymbol{N_{Severe}}$ were first determined. The weights for each class were then computed as follows:

$$W_{Mild} = \frac{N_{Severe}}{N_{Mild}}$$

$$W_{severe} = \frac{N_{Severe}}{N_{Severe}}$$

These weights were applied to the loss function to mitigate the impact of class imbalance during the training process. The loss function used for training the I3D model is the Cross-Entropy Loss, modified to incorporate class weights to handle class imbalances. Given the class weights $\boldsymbol{W_{Mild}}$ and $\boldsymbol{W_{severe}}$ calculated previously, the weighted Cross-Entropy Loss $\boldsymbol{L}$ for a single prediction $\boldsymbol{y}$ with the true label $\boldsymbol{t}$ is given by:

$$L(y,t) = -\sum_{c=1}^{C} w_c \cdot t_c \cdot \log\left(softmax(y)_c\right)$$

Where, $\boldsymbol{C}$ is the number of classes, $\boldsymbol{w_c}$ is the weight for class $\boldsymbol{c}$, which is $\boldsymbol{W_{Mild}}$ for the ‘Mild’ class and $\boldsymbol{W_{severe}}$ for ‘Severe’ class. $\boldsymbol{t_c}$ is 1 if the true class is $\boldsymbol{c}$ and 0 otherwise. $\boldsymbol{softmax(y)_c}$ is the predicted probability of class $\boldsymbol{c}$ after applying the softmax function to the output $\boldsymbol{y}$.

For model training and evaluation, the dataset was divided into training and test subsets. We used 9-fold group-wise (by individual) cross-validation for the development of social ability severity prediction model. The model was trained using a training subset while monitoring performance metrics such as accuracy, precision, recall, and F1-score. To optimize performance, model hyperparameters (e.g., learning rate, batch size, and normalization) were fine-tuned. The overall model architecture for predicting social ability in children with ASD by transfer learning a pre-trained model based on adult data is shown in Figure 4.

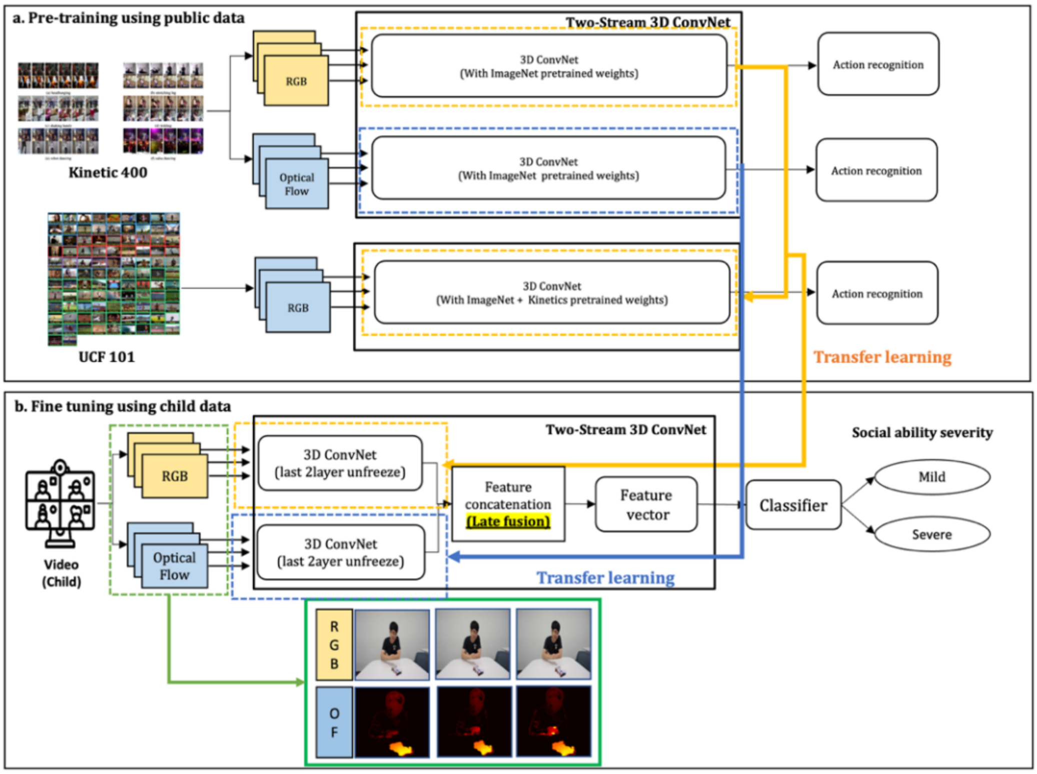


**Figure 4.** The proposed framework of the fine-tuning model based on a deep learning technique to predict the social skill severity of ASD children into two level: mild, and severe

## 2.6. Interpretability of the deep learning model: Use of Class Activation Map

The gradient-weighted class activation mapping (Grad-CAM) technique was used to generate a visual description of how the system makes predictions by superimposing a visualization layer on top of the CNN model [32]. This method uses the gradients of all target concepts accumulated in the final CNN layer to generate a localization heatmap that highlights key areas of the image for concept prediction. At the dataset level, we drew a heatmap on the test dataset for each turn taking task to validate the consistency of the DL system's decisions by visualizing how the DL system discriminates between the severity of social ability in ASD.

### 2.7. Statistical analysis

In this study, non-parametric analysis methods were chosen due to the non-normal distribution and limited size of the data set. Participant demographics, ADOS-2 scores, cognitive assessments, and social communication skills were summarized based on different levels of social ability, using median and IQR for continuous data and count and percentage for categorical data. The study compared ADOS-2, K-WPPSI-IV, NPT, and turn-taking success scores between mild and severe social ability severity groups, using median and IQR were tested with the Wilcoxon signed-rank test. The significance threshold was set at $P < .05$, and CIs were estimated using the Hanley and McNeil method. Pearson correlation analysis was used in this study to examine the relationships between turn-taking success and failure rates and three key variables: ADOS-2 score, cognitive function, and NPT scores. The Pearson correlation coefficient (r) was calculated to quantify the strength and direction of the linear relationships. The coefficient ranges from -1 to +1, with +1 indicating a perfect positive linear relationship, -1 indicating a perfect negative linear relationship, and 0 indicating no linear relationship. The significance of each correlation was determined using p-values, with a significance level of $p<0.05$. Simple linear regression analyses were performed to examine the relationship between the turn-taking success rate and several variables, including ADOS-2 scores, cognitive function, and NPT scores. Each variable was evaluated separately for its linear association with the turn-taking success score rate. The coefficient of determination ($R^2$), which explains the proportion of variance in the dependent variable that can be predicted by the independent variable, was used to quantify the strength and direction of these associations. The resulting $R^2$ values indicate the extent to which variations in ADOS-2 scores, cognitive function, and

NPT can be linearly related to variations in the turn-taking success rate. Statistical significance was determined using p-values at a significance level of 0.05.

The area under the receiver operating characteristic (AUROC), accuracy, recall, and precision were computed to evaluate the performance of the prediction models. Data analysis and training of the deep learning models were performed using R (version 4.1.0), Python (versions 3.6.8 and 3.9.12), and Pytorch (versions 1.13.1).

## III. Result

### 3.1. Demographic characteristics

Between July 2023 and November 2023, 10 participants who met our inclusion criteria were enrolled and data were collected. One participant who did not complete the social communication skills protocol was excluded from the analysis. Finally, a total of 9 participants were included in the data analysis and trained prediction model (Figure.1-B).

Demographic and baseline variables were compared between the mild and severe social ability groups. Of the nine children, three (33%) were in the mild group and six (67%) were in the severe group. There was no significant difference in sex between the groups; 100% of the participants in both groups were boys. There was no significant difference between the mild and severe groups in demographic variables, including age and K-WISC-IV. However, there was a significant difference between the two groups on the ADOS-2 test. The total Calibrated Severity Score (CSS) in the mild group was a median of 3.0 (IQR; 3.0 to 3.5), while the total CSS in the severe group was 7.5 (6.0 to 9.0). The demographic and baseline variables of both groups are summarized in Table 3, Appendix Table S1.

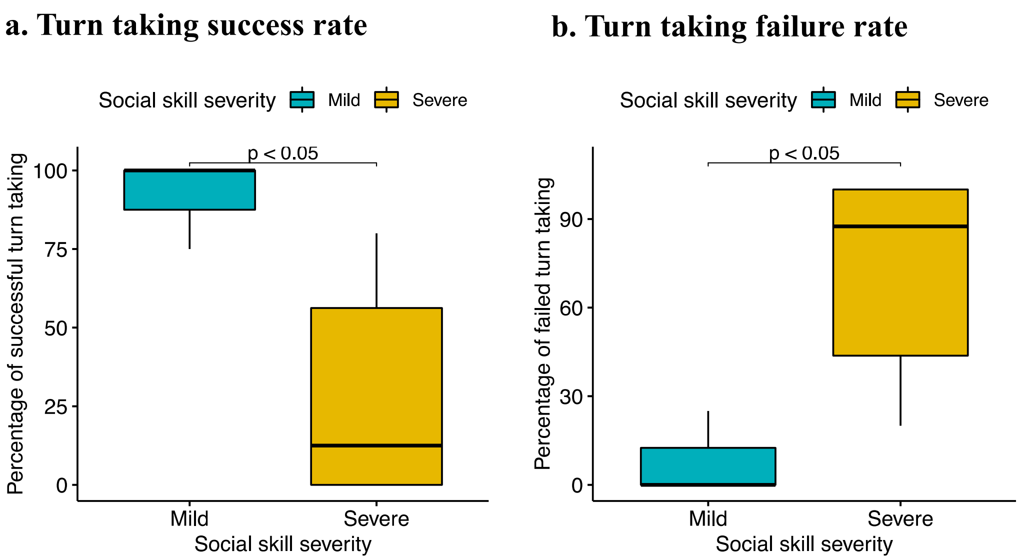
**Table 3.** Participant's characteristics

| Characteristics | Severity of social skill (ADOS-SA CSS) | | p-value |
|---|---|---|---|
| | Mild<br>(n=3) | Severe<br>(n=6) | |
| Age (years), median (IQR) | 5.0 [ 4.5; 5.0] | 4.5 [ 4.0; 5.0] | 0.665 |
| Sex, n (%) | | | |
| - Male | 3 (100%) | 6 (100%) | |
| K-WPPSI-IV, median (IQR) | | | |
| - Full scale IQ | 85.5 [76.0;95.0] | 67.0 [53.0;96.0] | 0.857 |
| - VCI | 82.0 [59.0;105.0] | 84.0 [53.0;95.0] | 0.857 |
| - VSI | 92.0 [85.0;99.0] | 67.0 [64.0;105.0] | 0.857 |
| - FRI | 88.5 [81.0;96.0] | 83.5 [52.5;122.0] | 1.000 |
| - WMI | 99.5 [86.0;113.0] | 52.0 [52.0;92.0] | 0.430 |
| - PSI | 92.0 [92.0;92.0] | 82.0 [50.0;116.0] | 1.000 |
| ADOS-2, median (IQR) | | | |
| - module, n (%) | | | 0.236 |
| - module1 | 0 (0.0%) | 4 (66.7%) | |
| - module 2 | 3 (100.0%) | 2 (33.3%) | |
| - Total CSS | 3.0 [ 3.0; 3.5] | 7.5 [ 6.0; 9.0] | 0.026 |
| - RRB CSS | 4.0 [ 2.5; 5.0] | 7.0 [ 7.0; 7.0] | 0.030 |
| - Social affect CSS | 4.0 [ 3.5; 4.0] | 8.0 [ 7.0;10.0] | 0.026 |
| -- Community | 2.0 [ 1.5; 2.0] | 4.0 [ 3.0; 6.0] | 0.026 |
| -- Social interaction | 4.0 [ 3.0; 4.5] | 11.0 [ 8.0;12.0] | 0.027 |

*ADOS-SA-CSS: The Autism Diagnostic Observation Schedule Social Affect Calibrated Severity Score, IQR: InterQuartile Range, K-WPPSI-IV: The Korean Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition, IQ: Intelligence quotient, VCI: Visual Comprehension IQ, VSI: Visual Spatial IQ, FRI: Fluid Reasoning IQ, WMI: Working Memory IQ, PSI: Processing Speed IQ, ADOS-2: The Autism Diagnostic Observation Schedule-2, Total CSS: Total Calibrated Severity Score, RRB CSS: Restricted and Repetitive Behaviors Calibrated Severity Score*

### 3.2. Evaluating the validity of the digitized nonverbal social communication skill measurement protocol

To evaluate the validity of the digitized nonverbal social communication skill measurement protocol we developed, we compared the performance rate of the turn taking task in the social skill severity "Mild" and "Severe" groups (Figure 5, Appendix Table S2). Using our social interaction scoring metric, we found that the turn taking success rate was statistically significantly different between the social skill severity "Mild"(median (IQR): 100.0 [87.5 to 100.0]) and "Severe"(median (IQR): 12.5 [0.0 to 66.7]) groups (p-value = 0.048). The turn taking failure rate was also significantly different between the two groups (p-value = 0.048).

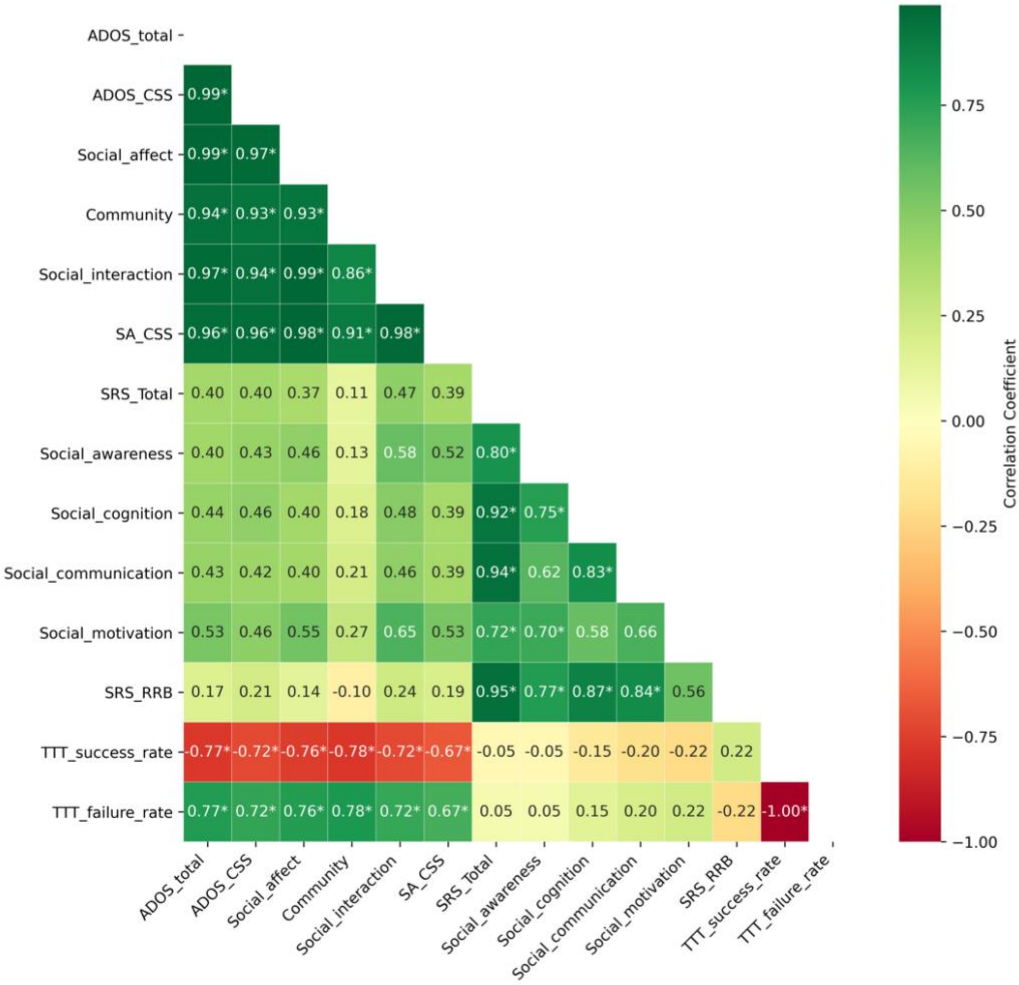


a. Turn taking success rate

b. Turn taking failure rate

**Figure 5.** Boxplot of compare turn taking task success and fail rates by social skill severity

The study used Pearson correlation analysis to examine the relationship between turn-taking success and failure rates and ADOS-2 score, cognitive function, and NPT scores. The results showed a negative correlation between the turn taking success rate and the total ADOS-2 score (r = -0.77, p-value = 0.02). This negative correlation was also observed with the ADOS-2 Community (r = -0.78, p-value = 0.01) and Social Affect (r = -0.76, p-value = 0.02) subscales. Conversely, a positive correlation was found between the turn-taking success rate and the failure rate. However, none of the variables in the SRS-2 were, a questionnaire assessed by caregivers, significantly correlated with the success or failure rate of turn taking. Cognitive function analysis revealed a significant positive correlation between the turn taking success rate and the Visual Spatial Index (VSI) with a correlation coefficient of r = 0.82 and a p-value < 0.05. And it was also a significant positive correlation between turn taking success rate and Working Memory Index (WMI) with r = 0.92 and p value < 0.05. It was not statistically significantly correlated with the rest of the NPT scores, but it was positively correlated with the K-Vineland-II subscales: socialization and turn taking success rate (r=0.72, p-value = 0.07). For more information, see Figure 6 and Appendix Table S3, Figure S1.



| | ADOS_total | ADOS_CSS | Social_affect | Community | Social_interaction | SA_CSS | SRS_Total | Social_awareness | Social_cognition | Social_communication | Social_motivation | SRS_RRB | TTT_success_rate | TTT_failure_rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ADOS_total | | | | | | | | | | | | | | |
| ADOS_CSS | 0.99* | | | | | | | | | | | | | |
| Social_affect | 0.99* | 0.97* | | | | | | | | | | | | |
| Community | 0.94* | 0.93* | 0.93* | | | | | | | | | | | |
| Social_interaction | 0.97* | 0.94* | 0.99* | 0.86* | | | | | | | | | | |
| SA_CSS | 0.96* | 0.96* | 0.98* | 0.91* | 0.98* | | | | | | | | | |
| SRS_Total | 0.40 | 0.40 | 0.37 | 0.11 | 0.47 | 0.39 | | | | | | | | |
| Social_awareness | 0.40 | 0.43 | 0.46 | 0.13 | 0.58 | 0.52 | 0.80* | | | | | | | |
| Social_cognition | 0.44 | 0.46 | 0.40 | 0.18 | 0.48 | 0.39 | 0.92* | 0.75* | | | | | | |
| Social_communication | 0.43 | 0.42 | 0.40 | 0.21 | 0.46 | 0.39 | 0.94* | 0.62 | 0.83* | | | | | |
| Social_motivation | 0.53 | 0.46 | 0.55 | 0.27 | 0.65 | 0.53 | 0.72* | 0.70* | 0.58 | 0.66 | | | | |
| SRS_RRB | 0.17 | 0.21 | 0.14 | -0.10 | 0.24 | 0.19 | 0.95* | 0.77* | 0.87* | 0.84* | 0.56 | | | |
| TTT_success_rate | -0.77* | -0.72* | -0.76* | -0.78* | -0.72* | -0.67* | -0.05 | -0.05 | -0.15 | -0.20 | -0.22 | 0.22 | | |
| TTT_failure_rate | 0.77* | 0.72* | 0.76* | 0.78* | 0.72* | 0.67* | 0.05 | 0.05 | 0.15 | 0.20 | 0.22 | -0.22 | -1.00* | |

**Figure 6.** Heatmap of the correlation matrix generated by the Pearson correlation coefficient for both turn taking task recording video and ADOS-2 and SRS-2

**: p-value < 0.05*
*ADOS: The Autism Diagnostic Observation Schedule-2, ADOS CSS: The Autism Diagnostic Observation Schedule Calibrated Severity Score, SA CSS: Social Affect Calibrated Severity Score, SRS: Social Responsiveness Scale-2, SRS RRB: Social Responsiveness Scale Restricted and Repetitive Behaviors, TTT: Turn Taking Task*

The regression analysis shows a variety of negative correlations between various ADOS-2 subscale scores and the success rate of turn taking. The total ADOS-2 score has a significant negative correlation with the turn taking success rate ($R^2$ = 0.59, p = 0.02), indicating that higher ADOS-2 scores are associated with a lower turn taking success rate. The ADOS-2 CSS ($R^2$ = 0.52, p = 0.03), Social Affect ($R^2$ = 0.57, p = 0.02), and the Community subscale ($R^2$ = 0.61, p = 0.01) all show similar negative correlations. The Social Interaction subscale ($R^2$ = 0.51, p = 0.03), in addition to the social affection CSS subscale ($R^2$ = 0.45, p = 0.05), show a significant negative correlation. These findings indicate that difficulties in social affect and social interaction, as measured by ADOS-2, could affect the ability to effectively participate in turn taking interactions. In contrast, the RRB subscale, while showing a negative correlation ($R^2$ = 0.47, p = 0.04), indicates a weaker relationship compared to other subscales (Figure 7). The results of the relationship between the cognitive function and the NPT score can be seen in Figure S2 in the Appendix.

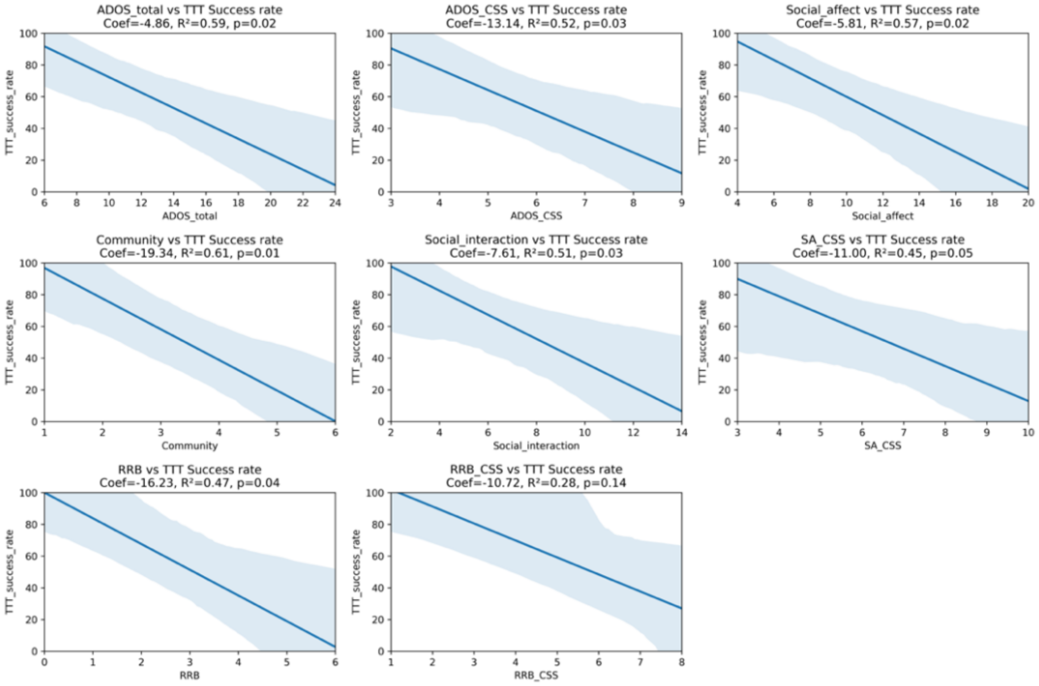


**Figure 7.** Linear regression analysis between turn taking success rate and ADOS-2

*ADOS: The Autism Diagnostic Observation Schedule-2, TTT: Turn Taking Task, ADOS CSS: The Autism Diagnostic Observation Schedule Calibrated Severity Score, SA CSS: Social Affect Calibra ted Severity Score, RRB: Restricted and Repetitive Behaviors, RRB CSS: Restricted and Repetitive Behaviors Calibrated Severity Score*

### 3.3. Turn taking based Deep learning Model for predcition of ASD social skill severity

The performance of three deep learning models - the RGB model, the Flow model, and a combined RGB-Flow late fusion model - in predicting the severity of social skills in children with ASD, predicting mild and severe cases, was compared. The RGB-Flow late fusion model outperformed the other two models in all performance metrics, achieving the highest accuracy (93.33%), precision (0.91), recall (1.0), F1 score (0.96), and area under the receiver operating characteristic curve (AUC) (0.99), as shown in Table 4, Figure 8-a. Although less accurate than the late fusion model, the optical flow model had the highest recall (0.97) but the lowest AUC (0.52). The RGB model had an accuracy of 75.56%, a precision of 0.82, a recall of 0.84, an F1 score of 0.83, and an AUC of 0.85.

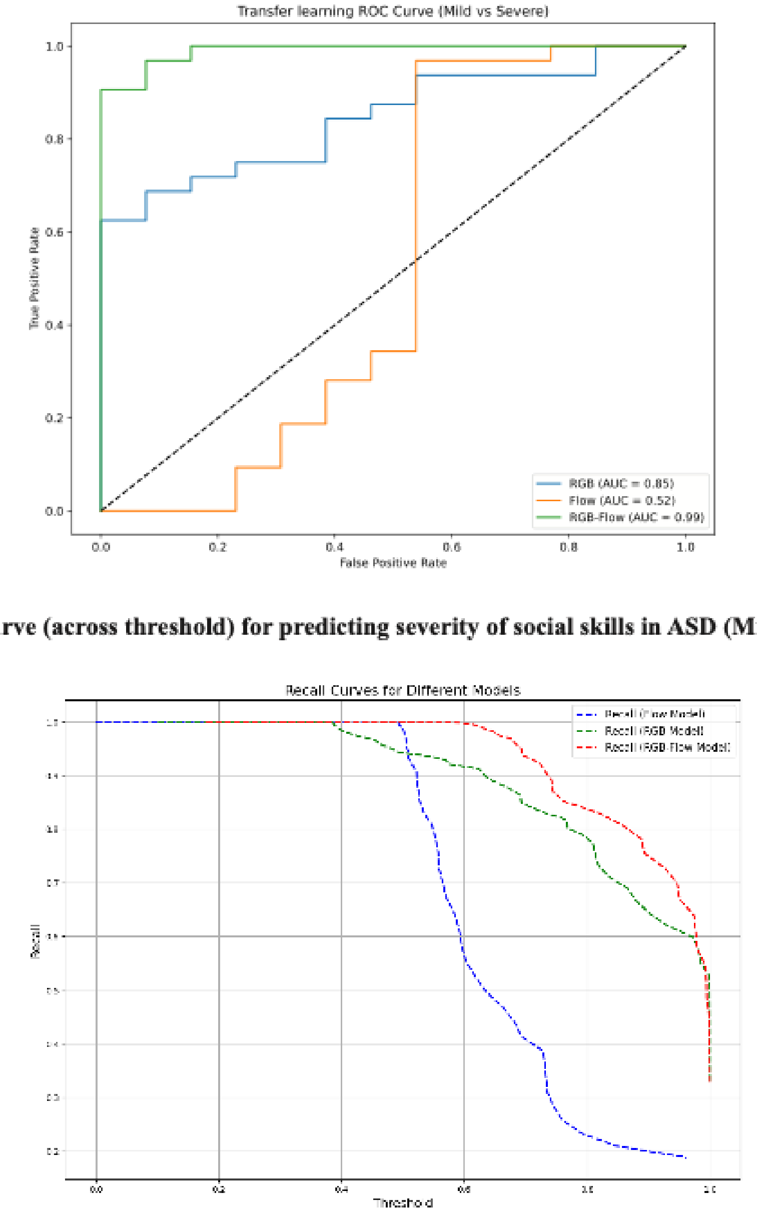
**Table 4.** The performance of deep learning models predicting social skill severity in ASD (Mild vs Severe)

| | Accuracy (%) | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| RGB | 75.56 | 0.82 | 0.84 | 0.83 | 0.85 |
| Optical Flow | 82.22 | 0.82 | 0.97 | 0.89 | 0.52 |
| RGB + Optical flow<br>(Late fusion) | 93.33 | 0.91 | 1.0 | 0.96 | 0.99 |

*AUC: Area Under the ROC Curve, ROC: Receiver Operating Characteristic*

Figure 8-b shows the performance of three different models in terms of their ability to correctly identify positive instances (true positives) at different threshold settings: RGB model, Flow model, and RGB-Flow model. The recall of the RGB model decreases sharply as the threshold increases, indicating that its ability to detect true positives is highly dependent on the threshold setting and that it is more prone to miss true positives at higher thresholds. The Flow model shows a gradual decrease in recall as the threshold is increased, indicating that it is moderately sensitive to the threshold setting, but not as sensitive as the RGB model. The RGB-Flow model maintains a recall of 1.0 over a wide range of thresholds before decreasing as the threshold approaches one. This demonstrates a strong ability to identify true positives with few false negatives at different threshold levels.

**a. Receiver Operating Characteristic (ROC) for predicting severity of social skills in ASD (Mild vs. Severe)**



**b. Recall curve (across threshold) for predicting severity of social skills in ASD (Mild vs. Severe)**

**Figure 8.** Receiver Operating Characteristic (ROC) and recall curve (across thresholds) for predicting severity of social skills in ASD (Mild vs. Severe)

 Figure 9 shows a comparison of the error rates of the three models. The Flow model, which uses temporal information, performed better (17.78%) than the RGB model, which relies on static visual cues (24.44%). The combined RGB-Flow late fusion model, on the other hand, showed a significantly lower error rate (6.67%) by using a late fusion technique that integrates both spatial and temporal features.

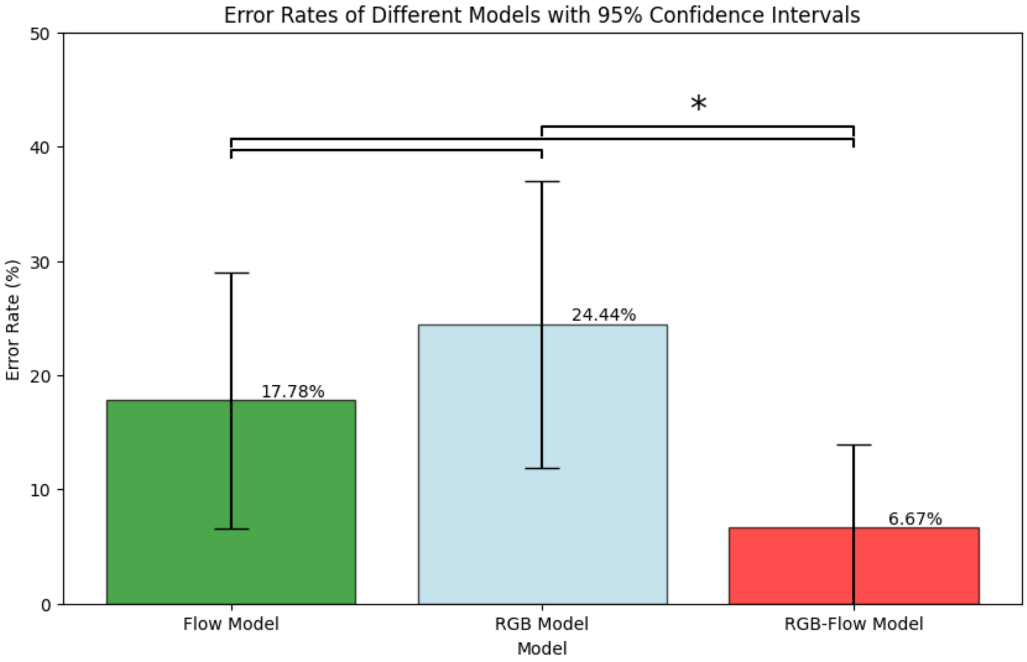


**Figure 9.** Comparing the error rates of RGB, optical flow, and RGB + optical flow models.

**: p-value < 0.05*

### 3.4. Interpreting the deep learning prediction premise: Grad-CAM

The Grad-CAM algorithm was applied to identify the critical regions of the videos that the deep learning model focuses on for prediction. Figure 10 shows the Grad-CAM results for "Mild" and "Severe" social skill severity. For a quantitative understanding of the video regions, we examined the spatial distribution of gradient weights across all RGB images. Gradient weights were significantly concentrated in the central region of the image, corresponding to the locations of the child's face and toy. These results suggest that the model learns to interact with the examiner by responding to the child's gaze or the movement of the toy.

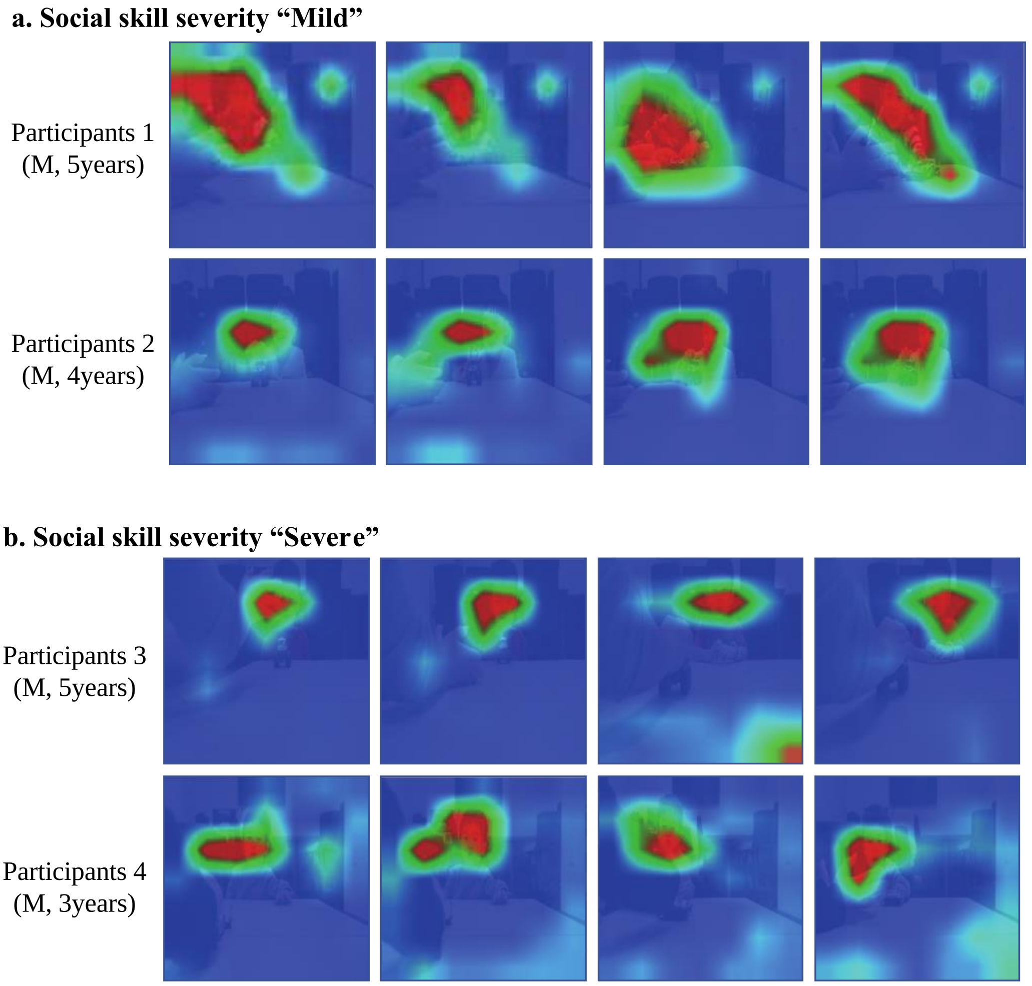


**Figure 10.** Gradient-Weighted Class Activation Maps (GradCAM) of nonverbal social communication skill video

### 3.5. Compare model performance with and without pre-trained weight

Table 5 and Appendix Figure S3 show a comparison of the performance of the I3D models with and without pre-trained weights. The use of pre-trained weights improves the accuracy from 48.89% to 75.56%, the precision from 0.76 to 0.82, and the recall from 0.41 to 0.84 in the RGB model. In addition, the F1 score increases from 0.53 to 0.83 and the AUC increases from 0.72 to 0.85 as a result of the improvement. The pre-trained weights improve the accuracy of the optical flow model from 73.33% to 82.22%, while the recall improves from 0.91 to 0.97. The F1 score increases from 0.83 to 0.89, but the AUC decreases from 0.83 to 0.52. In RGB + Optical Flow (Late Fusion), the combined model with pre-trained weights shows the most significant improvements, with accuracy increasing from 40.00% to 93.33%, precision decreasing from 1.00 to 0.91, but recall increasing from 0.16 to 1.0. The F1 score increases significantly from 0.27 to 0.96, the AUC increases significantly from 0.75 to 0.99.

**Table 5.** Comparing the performance of deep learning models predicting social skill severity in ASD (Mild vs Severe); a. I3D model (without pretrained weight), b. Transfer learning model (I3D model with pretrained weight)

a. I3D model (without pretrained weight)

| | Accuracy (%) | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| RGB | 48.89 | 0.76 | 0.41 | 0.53 | 0.72 |
| Optical Flow | 73.33 | 0.76 | 0.91 | 0.83 | 0.83 |
| RGB + Optical flow (Late fusion) | 40.00 | 1.00 | 0.16 | 0.27 | 0.75 |

b. Transfer learning model (our model, I3D model with pretrained weight)

| | Accuracy (%) | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| RGB | 75.56 | 0.82 | 0.84 | 0.83 | 0.85 |
| Optical Flow | 82.22 | 0.82 | 0.97 | 0.89 | 0.52 |
| RGB + Optical flow (Late fusion) | 93.33 | 0.91 | 1.0 | 0.96 | 0.99 |

*AUC: Area Under the ROC Curve, ROC: Receiver Operating Characteristic*

## IV. Discussion

To the best of our knowledge, no previous research has demonstrated that a behavioral biomarker such as a turn-taking task can be used to develop a digited nonverbal social skill assessment protocol and deep learning model for assessing nonverbal social communication skills in children with ASD using a video dataset. Our study makes a significant contribution by digitizing these complex nonverbal social communication behaviors and developing a protocol that not only demonstrates, but also withstands validation. Previous studies of ASD have focused primarily on diagnosing or screening children [8]–[11], [33]. While diagnosis is critical, it is also important to recognize that children with ASD face many difficulties in social communication and reciprocity[34]–[36]. The significant variability in symptom expression and functional abilities observed throughout the spectrum demonstrates the complexity of ASD. Because of this diversity, assessment strategies must be complex and individualized.

The results of the turn taking task across severity groups show statistically significant differences in performance, confirming the sensitivity of the protocol to different levels of social skills. The observed negative correlations between turn taking success rate and ADOS-2 scores, particularly within the Community and Social Affect subscales, support the validity of the protocol in reflecting ASD social skills symptomatology[2], [6]. In addition, the positive correlations with cognitive function indices such as the Visual Spatial Index (VSI) and the Working Memory Index (WMI) highlight the multifaceted nature of social communication in ASD. There was no significant correlation with caregiver-measured NPT scores, which is consistent with previous research that has raised concerns about caregiver reliability when administering these tests and suggests a potential bias that could influence results[37]. These observations highlight the need for more objective and standardized measures of social skills in children with autism.

In this study, we systematically collected input video data for a deep learning model and obtained prediction results with high accuracy and precision. The results of the RGB-flow late fusion model demonstrate the power of combining spatial and temporal data. In addition, the Explainable AI tool showed significant results, demonstrating that the deep learning model's predictions were based on observable behavioral differences, such as gaze direction and interaction with toys, and were similar to how trained professionals make judgments[2], [20], [34]. These results, which are consistent with expert judgment, reinforce the potential for deep learning models to replicate and augment the nuanced diagnostic process of ASD. This not only validates the predictive power of the model, but also sheds light on the behavioral patterns associated with different levels of ASD nonverbal social communication skills symptoms.

When we compared I3D models with and without pre-trained weights, we found that pre-trained weights significantly improved the accuracy and predictive power of the model. This suggests that in cases where data collection is difficult, such as with children, using pre-trained weights to train a model with a small amount of data can significantly improve accuracy, precision, recall, and AUC[12]–[15].

## V. Limitation

This study had several limitations. First, we recruited only males, so we could not observe the performance of the model on females. This is a limitation due to the higher prevalence of autism in males than females in Korea[38]. Future studies should consider gender balance when recruiting female subjects.

Second, to learn the spectrum of features of ASD, we need a model that takes into account multiple behaviors. However, the protocol we developed is a simple way to measure a single behavior, but we tested the performance of the model by capturing the broad features of social skills in children with ASD using a turn-taking task.

## VI. Conclusion

This study demonstrated the feasibility of collecting datasets for behavioral biomarkers using a standardized video data collection setup suitable for computer vision and deep learning, as well as measuring nonverbal social communication skills. According to the findings of this study, objectively measuring social-communication skills may provide objective information for ASD diagnosis or may be a good alternative for objectively measuring the effectiveness of social skill improvement interventions.

# REFERENCES

[1] A. Masi, M. M. DeMayo, N. Glozier, and A. J. Guastella, “An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options,” *Neurosci Bull*, vol. 33, no. 2, pp. 183–193, Apr. 2017, doi: 10.1007/s12264-017-0100-y.

[2] C.-H. Chiang, W.-T. Soong, T.-L. Lin, and S. J. Rogers, “Nonverbal Communication Skills in Young Children with Autism,” *J Autism Dev Disord*, vol. 38, no. 10, pp. 1898–1906, Nov. 2008, doi: 10.1007/s10803-008-0586-2.

[3] J. N. , & G. C. P. Constantino, *Social Responsiveness Scale–Second Edition (SRS-2)*. Torrance, CA: Western Psychological Services., 2012.

[4] Catherine Lord and Michael Rutter, *Autism diagnostic observation schedule*, Second edition. Torrance, CA: Western Psychological Services, 2012.

[5] C. Lord, M. Rutter, and A. Le Couteur, “Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders,” *J Autism Dev Disord*, vol. 24, no. 5, pp. 659–685, Oct. 1994, doi: 10.1007/BF02172145.

[6] P. Mundy *et al.*, “DRAFT A Manual for the EARLY SOCIAL COMMUNICATION SCALES (ESCS),” 2003. [Online]. Available: http://edscholars.ucdavis.edu/vrlab/home

[7] M. Randall *et al.*, “Diagnostic tests for autism spectrum disorder (ASD) in preschool children,” *Cochrane Database of Systematic Reviews*, vol. 2018, no. 7, Jul. 2018, doi: 10.1002/14651858.CD009044.pub2.

[8] H. Abbas, F. Garberson, S. Liu-Mayo, E. Glover, and D. P. Wall, “Multi-modular AI Approach to Streamline Autism Diagnosis in Young Children,”

*Sci Rep*, vol. 10, no. 1, p. 5014, Mar. 2020, doi: 10.1038/s41598-020-61213-w.

[9] Q. Tariq, J. Daniels, J. N. Schwartz, P. Washington, H. Kalantarian, and D. P. Wall, “Mobile detection of autism through machine learning on home video: A development and prospective validation study,” *PLoS Med*, vol. 15, no. 11, p. e1002705, Nov. 2018, doi: 10.1371/journal.pmed.1002705.

[10] N. Kojovic, S. Natraj, S. P. Mohanty, T. Maillart, and M. Schaer, “Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children,” *Sci Rep*, vol. 11, no. 1, p. 15069, Jul. 2021, doi: 10.1038/s41598-021-94378-z.

[11] H. Drimalla *et al.*, “Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (SIT),” *NPJ Digit Med*, vol. 3, no. 1, p. 25, Dec. 2020, doi: 10.1038/s41746-020-0227-5.

[12] K. Santosh, A. Goyal, D. Aouada, A. Makkar, Y.-Y. Chiang, and S. K. Singh, Eds., *Recent Trends in Image Processing and Pattern Recognition*, vol. 1704. in Communications in Computer and Information Science, vol. 1704. Cham: Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-23599-3.

[13] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10. pp. 1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?”

[15] M. F. Rabbi *et al.*, “Autism Spectrum Disorder Detection Using Transfer Learning with VGG 19, Inception V3 and DenseNet 201,” in

*Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 190–204. doi: 10.1007/978-3-031-23599-3_14.

[16] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[17] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective,” *BMC Med Inform Decis Mak*, vol. 20, no. 1, p. 310, Dec. 2020, doi: 10.1186/s12911-020-01332-6.

[18] A. N. Esler, V. H. Bal, W. Guthrie, A. Wetherby, S. E. Weismer, and C. Lord, “The Autism Diagnostic Observation Schedule, Toddler Module: Standardized Severity Scores,” *J Autism Dev Disord*, vol. 45, no. 9, pp. 2704–2720, Sep. 2015, doi: 10.1007/s10803-015-2432-7.

[19] V. Hus, K. Gotham, and C. Lord, “Standardizing ADOS Domain Scores: Separating Severity of Social Affect and Restricted and Repetitive Behaviors,” *J Autism Dev Disord*, vol. 44, no. 10, pp. 2400–2412, Oct. 2014, doi: 10.1007/s10803-012-1719-1.

[20] D. Hedley, R. Nevill, M. Uljarević, E. Butter, and J. A. Mulick, “ADOS-2 Toddler and Module 1 standardized severity scores as used by community practitioners,” *Res Autism Spectr Disord*, vol. 32, pp. 84–95, Dec. 2016, doi: 10.1016/j.rasd.2016.09.005.

[21] H. Park, Y. Seo, and J. Lee, “A Study of Concurrent Validities of K-WPPSI-IV,” *Korean Journal of Child Studies*, vol. 36, no. 1, pp. 65–83, Feb. 2015, doi: 10.5723/KJCS.2015.36.1.65.

[22] Thomas M. Achenbach, *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Burlington, VT: Department of Psychiatry, University of Vermont, 1991.

[23] Lee Helen, Oh Kyung Ja, Hong Kang-E, and Ha Eun Hye, "CLINICAL VALIDITY STUDY OF KOREAN CBCL THROUGH ITEM ANALYSIS".

[24] S.-T. Hwang, J.-H. Kim, S.-H. Hong, S.-H. Bae, and S.-W. Jo, "Standardization Study of the Korean Vineland Adaptive Behavior Scales-(K-Vineland-II)," *Korean Journal of Clinical Psychology*, vol. 34, no. 4, pp. 851–876, 2015, [Online]. Available: http://www.dbpia.co.kr/journal/articleDetail?

[25] S. S. Sparrow and D. V Cicchetti, "Diagnostic Uses of the Vineland Adaptive Behavior Scales," 1985. [Online]. Available: https://academic.oup.com/jpepsy/article/10/2/215/987476

[26] J. Ko, W. Lee, J. Woon, and Y. Kim, "Development of Korean Version of Developmental Coordination Disorder Questionnaire (DCDQ-K)," *The Journal of Korean Physical Therapy*, vol. 32, no. 1, pp. 44–51, Feb. 2020, doi: 10.18857/jkpt.2020.32.1.44.

[27] W. Kay *et al.*, "The Kinetics Human Action Video Dataset," May 2017, [Online]. Available: http://arxiv.org/abs/1705.06950

[28] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," Dec. 2012, [Online]. Available: http://arxiv.org/abs/1212.0402

[29] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," May 2017, [Online]. Available: http://arxiv.org/abs/1705.07750

[30] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion." [Online]. Available: http://www.isy.liu.se/cvl/

[31] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples."

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[33] C. Ko, J.-H. Lim, J. Hong, S.-B. Hong, and Y. R. Park, "Development and Validation of a Joint Attention–Based Deep Learning System for Detection and Symptom Severity Assessment of Autism Spectrum Disorder," *JAMA Netw Open*, vol. 6, no. 5, p. e2315174, May 2023, doi: 10.1001/jamanetworkopen.2023.15174.

[34] E. Rotheram-Fuller, "Social skills assessments for children with autism spectrum disorders," *Autism Open Access*, vol. 04, no. 02, 2014, doi: 10.4172/2165-7890.1000122.

[35] J. S. Beighley and J. L. Matson, "Comparing Social Skills in Children Diagnosed with Autism Spectrum Disorder According to the DSM-IV-TR and the DSM-5," *J Dev Phys Disabil*, vol. 26, no. 6, pp. 689–701, Oct. 2014, doi: 10.1007/s10882-014-9382-4.

[36] H. T. Wang, S. R. Sandall, C. A. Davis, and C. J. Thomas, "Social skills assessment in young children with autism: A comparison evaluation of the SSRS and PKBS," *J Autism Dev Disord*, vol. 41, no. 11, pp. 1487–1495, Nov. 2011, doi: 10.1007/s10803-010-1175-8.

[37] J. Duvekot, J. van der Ende, F. C. Verhulst, and K. Greaves-Lord, "The Screening Accuracy of the Parent and Teacher-Reported Social

Responsiveness Scale (SRS): Comparison with the 3Di and ADOS," *J Autism Dev Disord*, vol. 45, no. 6, pp. 1658–1672, Jun. 2015, doi: 10.1007/s10803-014-2323-3.

[38] M. Hong, S. M. Lee, S. Park, S. J. Yoon, Y. E. Kim, and I. H. Oh, "Prevalence and Economic Burden of Autism Spectrum Disorder in South Korea Using National Health Insurance Data from 2008 to 2015," *J Autism Dev Disord*, vol. 50, no. 1, pp. 333–339, Jan. 2020, doi: 10.1007/s10803-019-04255-y.

# APPENDIX

**Table S1.** Participant's characteristics

| Characteristics | Severity of social skill (ADOS-SA-CSS) | | p-value |
|---|---|---|---|
| | Mild<br>(n=3) | Severe<br>(n=6) | |
| Age (years), median (IQR) | 5.0 [ 4.5; 5.0] | 4.5 [ 4.0; 5.0] | 0.665 |
| Sex, n (%) | | | |
| - Male | 3 (100%) | 6 (100%) | |
| K-WPPSI-IV, median (IQR) | | | |
| - Full scale IQ | 85.5 [76.0;95.0] | 67.0 [53.0;96.0] | 0.857 |
| - VCI | 82.0 [59.0;105.0] | 84.0 [53.0;95.0] | 0.857 |
| - VSI | 92.0 [85.0;99.0] | 67.0 [64.0;105.0] | 0.857 |
| - FRI | 88.5 [81.0;96.0] | 83.5 [52.5;122.0] | 1.000 |
| - WMI | 99.5 [86.0;113.0] | 52.0 [52.0;92.0] | 0.430 |
| - PSI | 92.0 [92.0;92.0] | 82.0 [50.0;116.0] | 1.000 |
| ADOS-2, median (IQR) | | | |
| - module, n (%) | | | 0.236 |
| - module1 | 0 (0.0%) | 4 (66.7%) | |
| - module 2 | 3 (100.0%) | 2 (33.3%) | |
| - Total CSS | 3.0 [ 3.0; 3.5] | 7.5 [ 6.0; 9.0] | 0.026 |
| - RRB CSS | 4.0 [ 2.5; 5.0] | 7.0 [ 7.0; 7.0] | 0.030 |
| - Social affect CSS | 4.0 [ 3.5; 4.0] | 8.0 [ 7.0;10.0] | 0.026 |
| -- Community | 2.0 [ 1.5; 2.0] | 4.0 [ 3.0; 6.0] | 0.026 |
| -- Social interaction | 4.0 [ 3.0; 4.5] | 11.0 [ 8.0;12.0] | 0.027 |
| SRS-2 Total | 76.0 [74.5;84.5] | 88.0 [81.0;122.0] | 0.437 |
| - Social awareness | 5.0 [ 4.5; 6.5] | 7.0 [ 7.0; 9.0] | 0.429 |
| - Social cognition | 16.0 [14.0;16.5] | 17.0 [14.0;23.0] | 0.427 |
| - Social communication | 26.0 [25.5;28.5] | 35.0 [26.0;41.0] | 0.298 |

| | | | |
|---|---|---|---|
| - Social motivation | 11.0 [ 9.5;11.5] | 13.5 [12.0;14.0] | 0.118 |
| - RRB | 22.0 [20.0;23.0] | 20.5 [17.0;33.0] | 1.000 |
| K-CBCL Total | 55.5 [54.0;57.0] | 58.0 [55.0;65.0] | 0.571 |
| - Externalizing | 61.5 [59.0;64.0] | 58.0 [58.0;59.0] | 0.430 |
| - Internalizing | 53.0 [52.0;54.0] | 58.0 [56.0;61.0] | 0.118 |
| - Aggression | 59.5 [58.0;61.0] | 55.0 [55.0;60.0] | 0.558 |
| - Attention | 63.5 [59.0;68.0] | 64.0 [50.0;64.0] | 0.691 |
| - Sleep | 54.5 [50.0;59.0] | 55.0 [50.0;55.0] | 1.000 |
| - Withdrawn | 54.0 [54.0;54.0] | 62.0 [58.0;65.0] | 0.160 |
| - Somatic | 56.5 [50.0;63.0] | 59.0 [54.0;63.0] | 0.693 |
| - Anxious | 53.0 [50.0;56.0] | 50.0 [50.0;52.0] | 1.000 |
| - Emotional | 51.0 [51.0;51.0] | 58.0 [51.0;66.0] | 0.285 |
| - Affective | 50.0 [50.0;50.0] | 55.0 [55.0;55.0] | 0.143 |
| - Anxiety | 50.0 [50.0;50.0] | 53.0 [50.0;56.0] | 0.285 |
| - PDD | 54.0 [50.0;58.0] | 70.0 [66.0;72.0] | 0.079 |
| - ADHD | 66.5 [63.0;70.0] | 59.0 [52.0;63.0] | 0.241 |
| - ODD | 57.0 [51.0;63.0] | 63.0 [50.0;63.0] | 1.000 |
| K-Vineland | | | |
| - Communication | 70.0 [69.0;71.0] | 97.0 [71.0;104.0] | 0.434 |
| - Daily living skills | 79.5 [74.0;85.0] | 80.0 [80.0;89.0] | 0.845 |
| - Socialization | 97.0 [92.0;102.0] | 81.0 [52.0;83.0] | 0.190 |
| DCDQ | | | |
| - Control during movement | 17.0 [13.0;21.0] | 19.0 [13.0;20.0] | 1.000 |
| - Fine motor/Handwriting | 8.0 [ 6.0;10.0] | 8.0 [ 8.0;17.0] | 0.844 |
| - General coordination | 13.5 [10.0;17.0] | 16.0 [11.0;17.0] | 0.693 |

*IQR: InterQuartile Range, K-WPPSI-IV: The Korean Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition, IQ: Intelligence quotient, VCI: Visual Comprehension IQ, VSI: Visual Spatial IQ, FRI: Fluid Reasoning IQ, WMI: Working Memory IQ, PSI: Processing Speed IQ, ADOS-2: The Autism Diagnostic Observation Schedule-2, Total CSS: Total Calibrated Severity Score, RRB CSS: Restricted and Repetitive Behaviors Calibrated Severity Score, SRS-2: Social Responsiveness Scale-2, RRB: Restricted and Repetitive Behaviors, K-CBCL: Korean Child Behavior Checklist,* *PDD: Pervasive Developmental Disorder, ADHD: Attention Deficit Hyperactivity Disorder, ODD: Oppositional Defiant Disorder, DCDQ: Developmental Coordination Disorder Questionnaire*

**Table S2.** Compare turn taking task success and fail rates by social ability severity

| Characteristics | Severity of social skill (ADOS-SA-CSS) | | p-value |
|---|---|---|---|
| | Mild<br>(n =3) | Severe<br>(n =6) | |
| Turn Tasking Task, median (IQR) | | | |
| - Success count | 4.0 [ 3.5; 4.5] | 1.0 [ 0.0; 4.0] | 0.142 |
| - Percentage of success | 100.0 [87.5;100.0] | 12.5 [ 0.0;66.7] | 0.048 |
| - Fail count | 0.0 [ 0.0; 0.5] | 4.0 [ 2.0; 5.0] | 0.036 |
| - Percentage of fail | 0.0 [ 0.0;12.5] | 87.5 [33.3;100.0] | 0.048 |

*ADOS-SA-CSS: The Autism Diagnostic Observation Schedule Social Affect Calibrated Severity Score*

**Table S3.** Pearson correlation coefficient (upper value) and p-value (lower value) of turn taking task recording video and ADOS-2

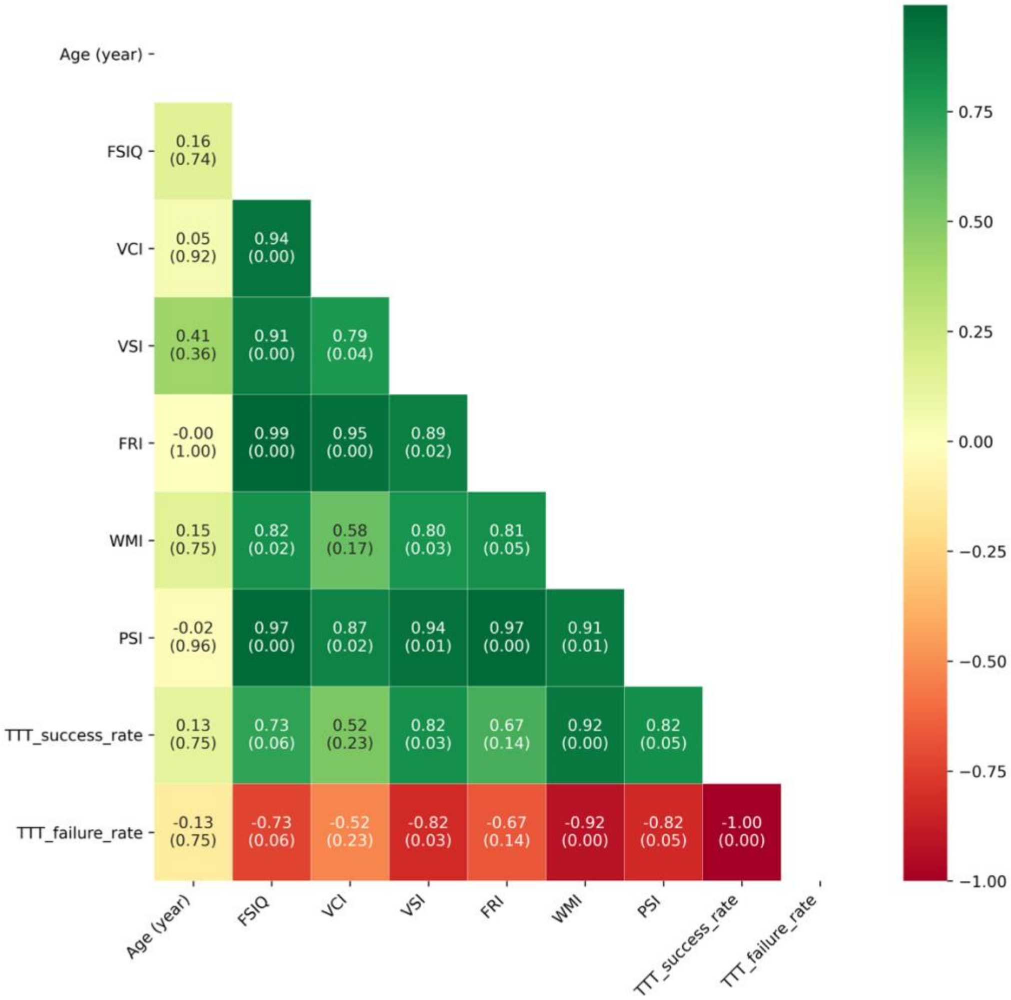
| | **Age (year)** | **ADOS total** | **ADOS CSS** | **Social Affect** | **Comm unity** | **Social interact ion** | **SA CSS** | **RRB** | **RRB CSS** | **Good video** | **Bad video** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Age (year)** | 1.00<br>(0.00) | -0.07<br>(0.86) | -0.18<br>(0.65) | -0.00<br>(0.99) | 0.02<br>(0.96) | -0.01<br>(0.97) | -0.04<br>(0.91) | -0.14<br>(0.73) | -0.45<br>(0.23) | 0.13<br>(0.75) | -0.13<br>(0.75) |
| **ADOS total** | -0.07<br>(0.86) | 1.00<br>(0.00) | 0.99<br>(0.00) | 0.99<br>(0.00) | 0.94<br>(0.00) | 0.97<br>(0.00) | 0.96<br>(0.00) | 0.71<br>(0.03) | 0.74<br>(0.02) | -0.77<br>(0.02) | 0.77<br>(0.02) |
| **ADOS -2 CSS** | -0.18<br>(0.65) | 0.99<br>(0.00) | 1.00<br>(0.00) | 0.97<br>(0.00) | 0.93<br>(0.00) | 0.94<br>(0.00) | 0.96<br>(0.00) | 0.69<br>(0.04) | 0.76<br>(0.02) | -0.72<br>(0.03) | 0.72<br>(0.03) |
| **Social Affect** | -0.00<br>(0.99) | 0.99<br>(0.00) | 0.97<br>(0.00) | 1.00<br>(0.00) | 0.93<br>(0.00) | 0.99<br>(0.00) | 0.98<br>(0.00) | 0.61<br>(0.08) | 0.64<br>(0.06) | -0.76<br>(0.02) | 0.76<br>(0.02) |
| **Comm unity** | 0.02<br>(0.96) | 0.94<br>(0.00) | 0.93<br>(0.00) | 0.93<br>(0.00) | 1.00<br>(0.00) | 0.86<br>(0.00) | 0.91<br>(0.00) | 0.69<br>(0.04) | 0.68<br>(0.04) | -0.78<br>(0.01) | 0.78<br>(0.01) |
| **Social interac tion** | -0.01<br>(0.97) | 0.97<br>(0.00) | 0.94<br>(0.00) | 0.99<br>(0.00) | 0.86<br>(0.00) | 1.00<br>(0.00) | 0.98<br>(0.00) | 0.55<br>(0.13) | 0.59<br>(0.09) | -0.72<br>(0.03) | 0.72<br>(0.03) |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **SA CSS** | -0.04<br>(0.91) | 0.96<br>(0.00) | 0.96<br>(0.00) | 0.98<br>(0.00) | 0.91<br>(0.00) | 0.98<br>(0.00) | 1.00<br>(0.00) | 0.52<br>(0.15) | 0.61<br>(0.08) | -0.67<br>(0.05) | 0.67<br>(0.05) |
| **RRB** | -0.14<br>(0.73) | 0.71<br>(0.03) | 0.69<br>(0.04) | 0.61<br>(0.08) | 0.69<br>(0.04) | 0.55<br>(0.13) | 0.52<br>(0.15) | 1.00<br>(0.00) | 0.86<br>(0.00) | -0.69<br>(0.04) | 0.69<br>(0.04) |
| **RRB CSS** | -0.45<br>(0.23) | 0.74<br>(0.02) | 0.76<br>(0.02) | 0.64<br>(0.06) | 0.68<br>(0.04) | 0.59<br>(0.09) | 0.61<br>(0.08) | 0.86<br>(0.00) | 1.00<br>(0.00) | -0.53<br>(0.14) | 0.53<br>(0.14) |
| **TTT success rate** | 0.13<br>(0.75) | -0.77<br>(0.02) | -0.72<br>(0.03) | -0.76<br>(0.02) | -0.78<br>(0.01) | -0.72<br>(0.03) | -0.67<br>(0.05) | -0.69<br>(0.04) | -0.53<br>(0.14) | 1.00<br>(0.00) | -1.00<br>(0.00) |
| **TTT failure rate** | -0.13<br>(0.75) | 0.77<br>(0.02) | 0.72<br>(0.03) | 0.76<br>(0.02) | 0.78<br>(0.01) | 0.72<br>(0.03) | 0.67<br>(0.05) | 0.69<br>(0.04) | 0.53<br>(0.14) | -1.00<br>(0.00) | 1.00<br>(0.00) |

*ADOS-2: The Autism Diagnostic Observation Schedule-2, SA CSS: Social Affect Calibrated Severity Score, RRB: Restricted and Repetitive Behaviors, RRB CSS: Restricted and Repetitive Behaviors Calibrated Severity Score, TTT: Turn Taking Task*

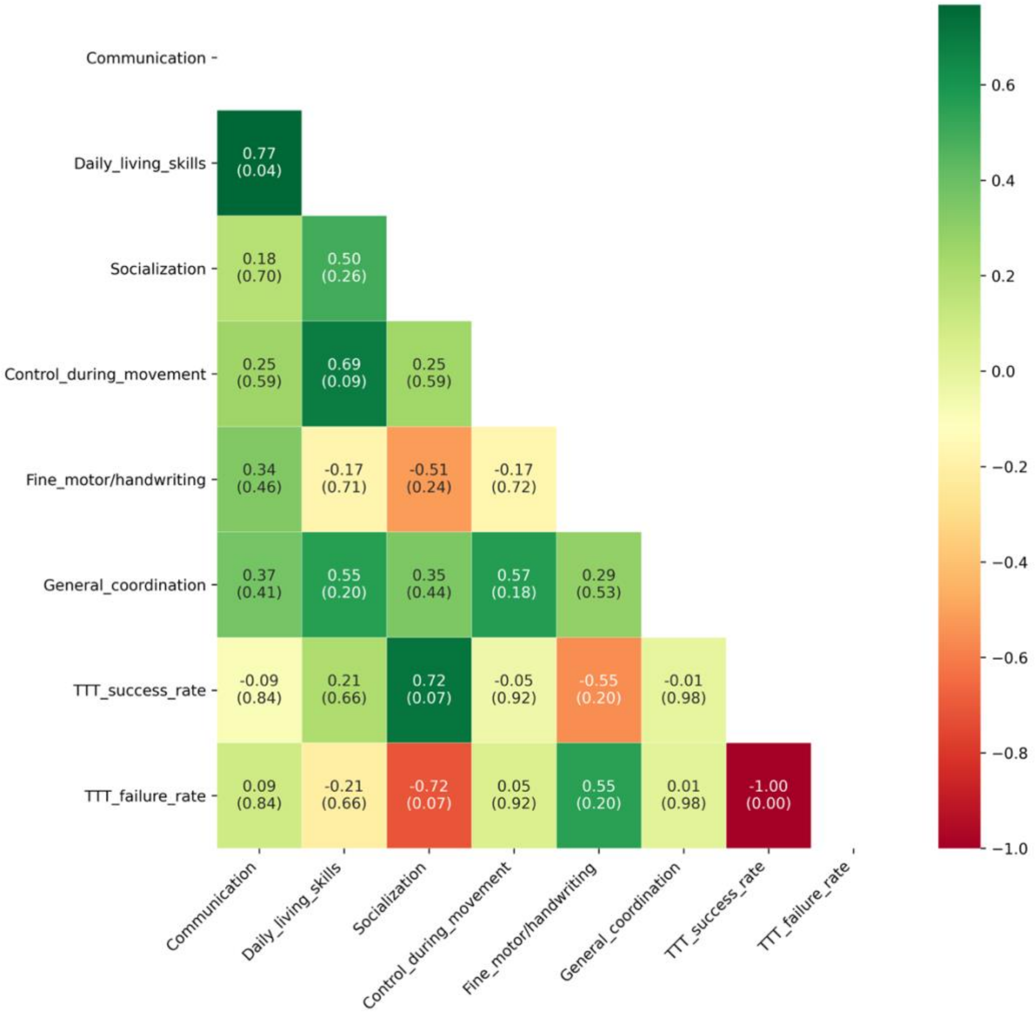
**Figure S1.** Heatmap of the correlation matrix generated by the Pearson correlation coefficient for both turn taking task recording video and cognitive function, NPT scores (a. Cognitive function, b. K-Vineland-II and DCDQ, c. K-CBCL, d. SRS-2)

a. Cognitive function test



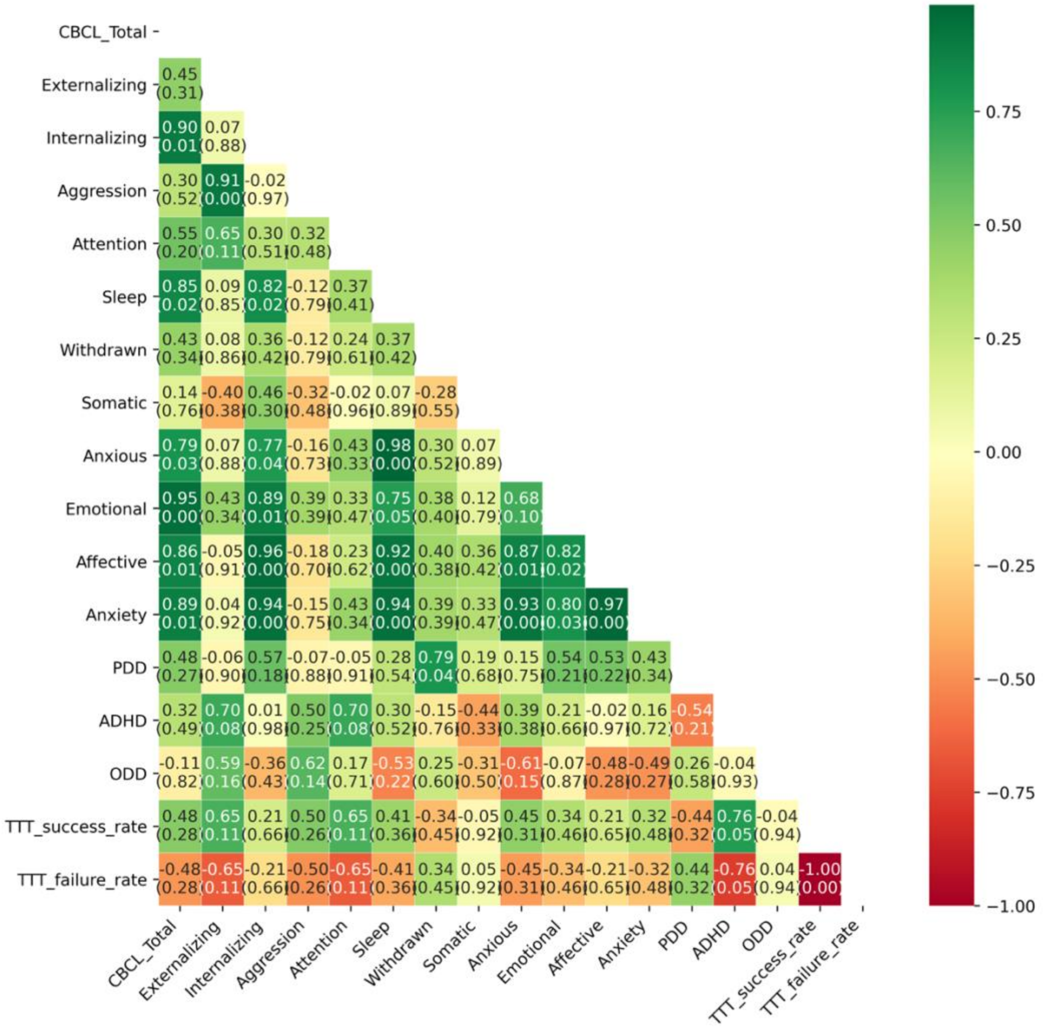
*FSIQ: Full Scale Intelligence quotient, VCI: Visual Comprehension IQ, VSI: Visual Spatial IQ, FRI: Fluid Reasoning IQ, WMI: Working Memory IQ, PSI: Processing Speed IQ, TTT: Turn Taking Task*

b. K-Vineland-II, DCDQ



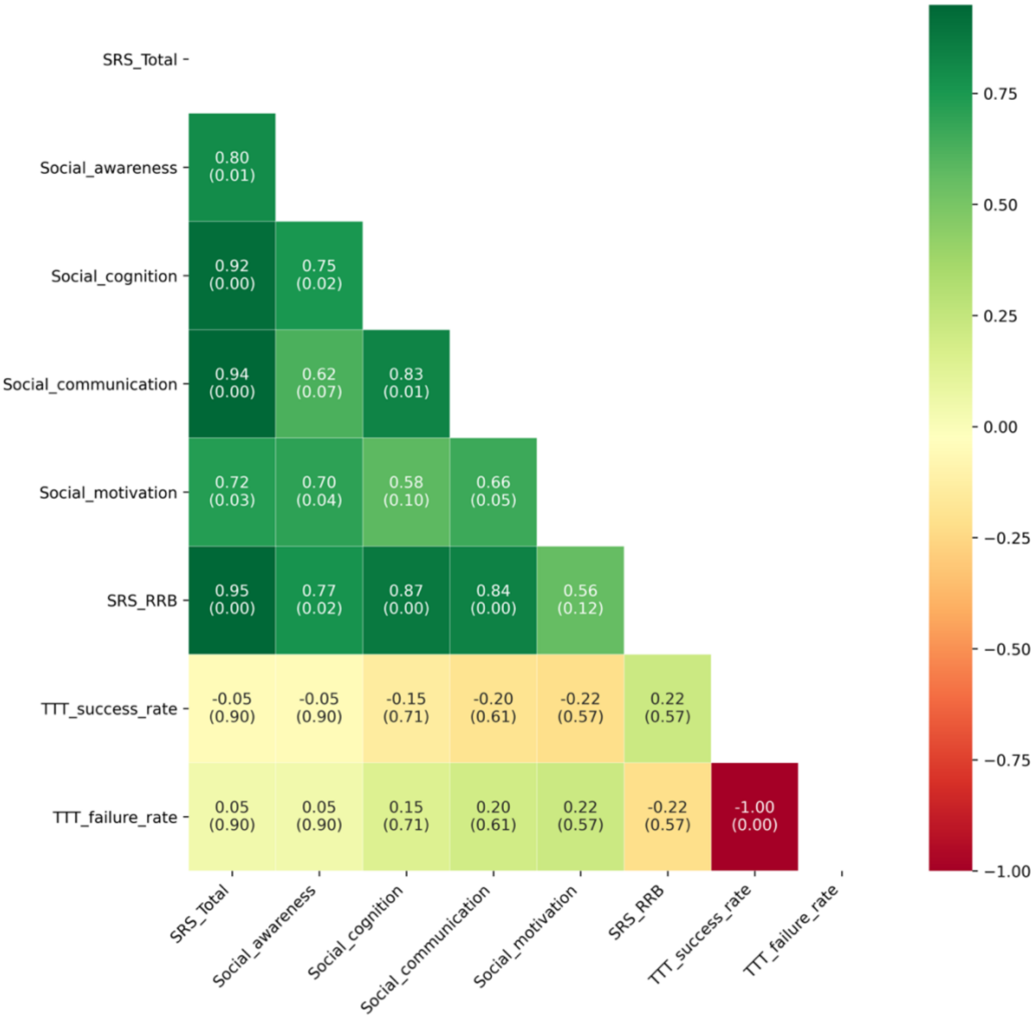
*DCDQ: Developmental Coordination Disorder Questionnaire, TTT: Turn Taking Task*

c. K-CBCL



*K-CBCL: Korean Child Behavior Checklist, PDD: Pervasive Developmental Disorder, ADHD: Attention Deficit Hyperactivity Disorder, ODD: Oppositional Defiant Disorder, TTT: Turn Taking Task*

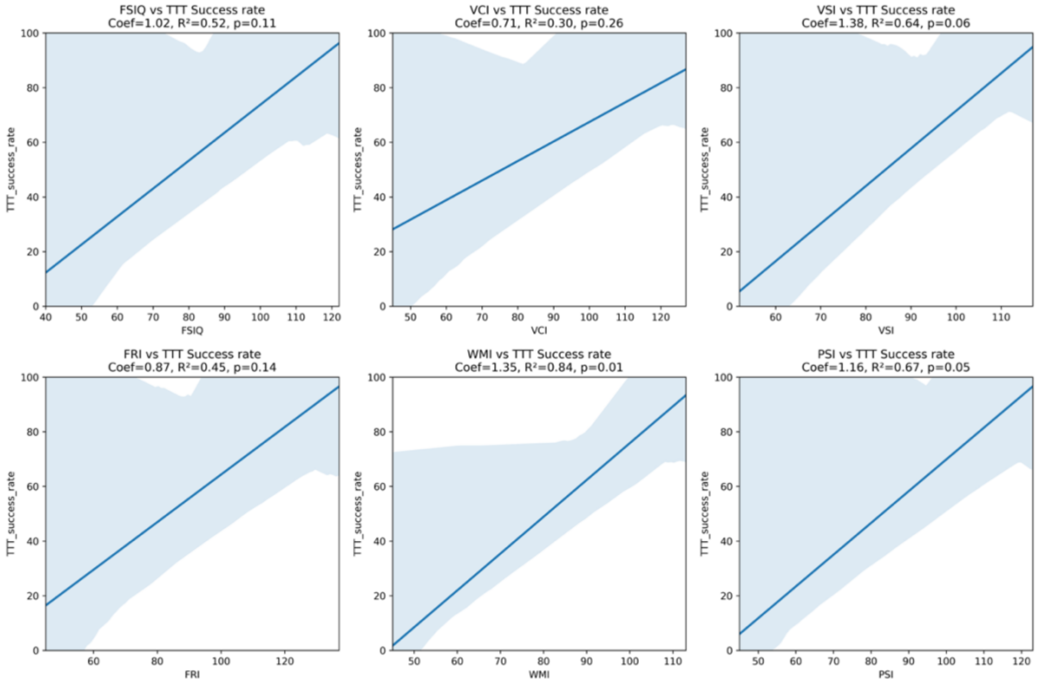
d. SRS-2



*SRS: Social Responsiveness Scale, RRB: Restricted and Repetitive Behaviors, TTT: Turn Taking Task*

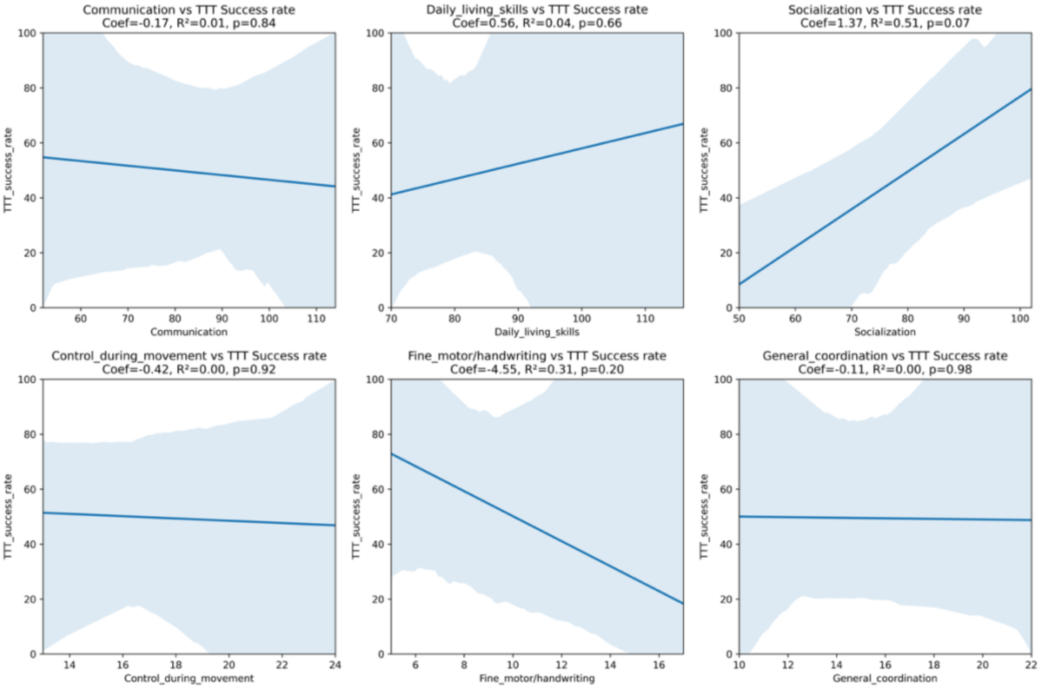
**Figure S2.** Linear regression analysis between turn taking success and cognitive function, NPT scores (a. Cognitive function, b. K-Vineland-II and DCDQ, c. K-CBCL, d. SRS-2)

a. Cognitive function



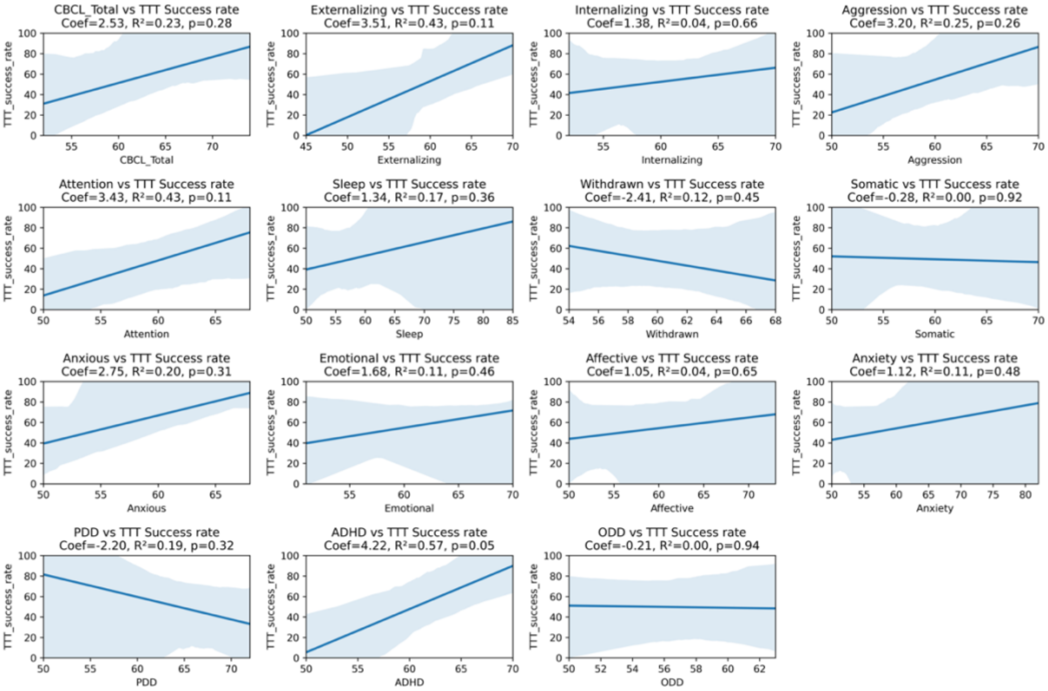
*FSIQ: Full Scale Intelligence quotient, VCI: Visual Comprehension IQ, VSI: Visual Spatial IQ, FRI: Fluid Reasoning IQ, WMI: Working Memory IQ, PSI: Processing Speed IQ, TTT: Turn Taking Task*

b. K-Vineland-II and DCDQ



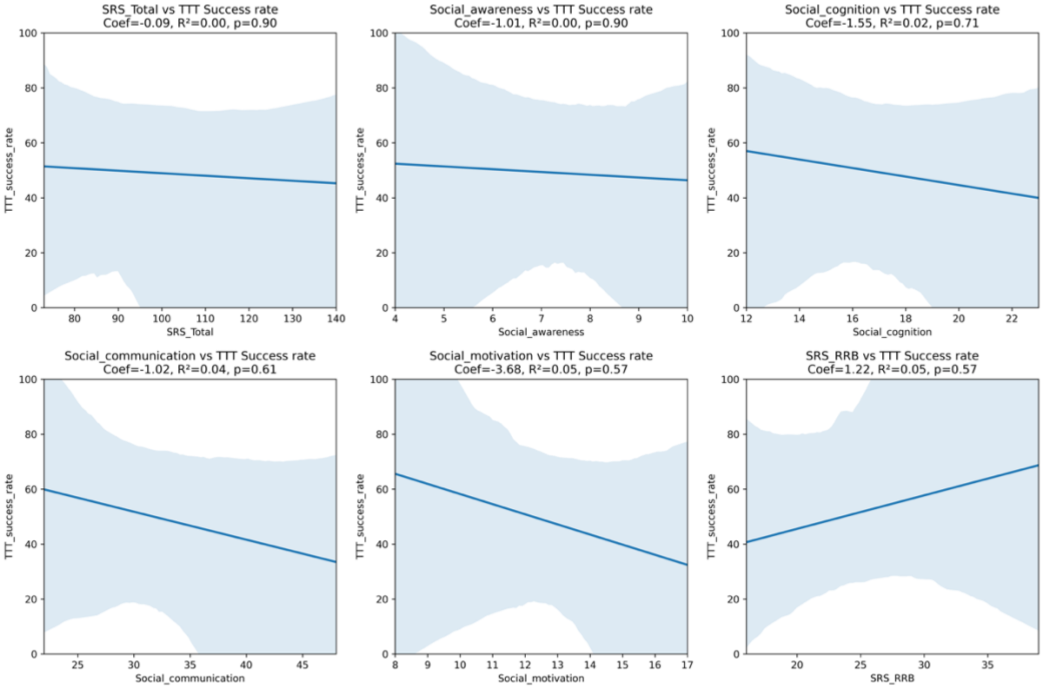
*DCDQ: Developmental Coordination Disorder Questionnaire, TTT: Turn Taking Task*

c. K-CBCL



*K-CBCL: Korean Child Behavior Checklist, PDD: Pervasive Developmental Disorder, ADHD: Attention Deficit Hyperactivity Disorder, ODD: Oppositional Defiant Disorder, TTT: Turn Taking Task*

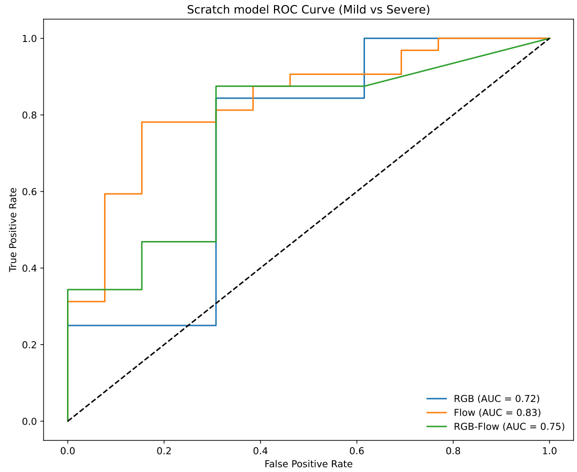
d. SRS-2



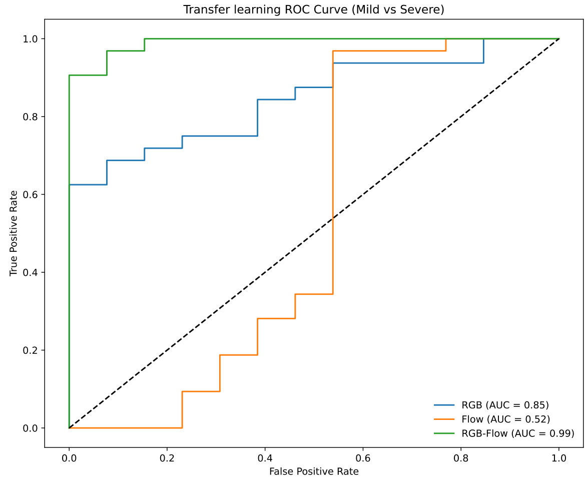
*SRS: Social Responsiveness Scale, RRB: Restricted and Repetitive Behaviors, TTT: Turn Taking Task*

**Figure S3.** Receiver Operating Characteristic (ROC) curve for predicting severity of social skills in ASD (Mild vs. Severe) with and without pretrained weight a. I3D model (without pretrained weight), b. Transfer learning model (I3D model with pretrained weight)

a. I3D model (without pretrained weight)



b. Transfer learning model (I3D model with pretrained weight)



# ABSTRACT IN KOREAN

## 자폐 스펙트럼 장애 아동의 사회성 중증도를 예측하기 위한 사회적 상호작용 기반 딥러닝 시스템 개발 및 검증

연세대학교 일반대학원

의생명시스템정보학교실

이주현

자폐 스펙트럼 장애(ASD)를 가진 아동들은 사회적 상호작용에 어려움을 겪으며, 이러한 사회적 능력은 ASD 진단을 위한 중요한 측정 기준입니다. 그러나 기존의 사회적 능력 측정 방법은 비용이 많이 들고 시간이 오래 걸리며, 검사자의 편견을 포함할 수 있습니다. 따라서 ASD 아동의 사회적 능력을 객관적이고 표준화된 방법으로 측정할 수 있는 도구가 필요합니다. 본 연구의 목적은 1) ASD 아동의 비언어적 사회적 의사소통 기술을 디지털화하는 프로토콜을 개발하고, 2) 개발된 비언어적 사회적 의사소통 기술 측정 프로토콜의 타당성을 평가하며, 3) 개발된 프로토콜을 통해 수집된 비디오 데이터를 사용하여 ASD 아동의 비언어적 사회적 의사소통 기술을 예측할 수 있는 딥러닝 모델을 개발하는 것입니다. 이 연구는 전향적 관찰 연구로, ADOS-2 및 신경심리학적 검사를 사용하여 아동의 사회성을 평가했습니다. 아동의 사회적 상호작용을 측정하고 비디오로 기록하기 위해 ’Turn taking’ 프로토콜을 개발하였습니다. 이 프로토콜을 통해 수집된 데이터는 자폐

아동의 사회성 중증도를 예측하기 위해 설계된 세 가지 다른 딥러닝 모델인 RGB 모델, Optical flow 모델, 그리고 RGB-Optical flow late fusion 모델을 훈련하는 데 사용되었습니다.연구에는 9명의 참가자의 데이터가 포함되었습니다. 개발된 비언어적 사회적 의사소통 기술 측정 프로토콜의 평가는 중증도가 다른 두 그룹(Mild: 중앙값(IQR): 100.0 [87.5 to 100.0], Severe: 중앙값(IQR): 12.5 [0.0 to 66.7], p-value = 0.048) 사이에 turn taking 수행률에서 유의미한 차이를 보였습니다. RGB-Optical flow late fusion 딥러닝 모델은 정확도(93.33%), 정밀도(0.91), 재현율(1.0), F1 점수(0.96), ROC curve (AUC, 0.99)에서 높은 성능을 보여 자폐 아동의 사회성 중증도를 예측하는 데 있어 우수한 성능을 보였습니다. Grad-CAM 알고리즘은 이러한 모델에 적용되었으며, 모델이 예측을 위해 아동의 얼굴과 장난감 상호작용에 주로 초점을 맞추고 있음을 밝혀냈습니다. 이 연구는 컴퓨터 비전 및 딥러닝에 적합한 표준화된 비디오 데이터 수집 프로토콜을 사용하여 행동 바이오마커 데이터셋을 수집하고 비언어적 사회적 의사소통 기술을 측정할 수 있는 가능성을 처음으로 입증하였습니다. 본 연구의 결과에 따르면, 비언어적 사회적 의사소통 기술을 객관적으로 측정하는 것은 ASD 진단을 위한 객관적 정보를 제공하거나 사회성 향상을 위한 치료 프로그램의 효과를 객관적으로 측정하는 데 좋은 대안이 될 수 있습니다.

**핵심 단어:** 자폐스펙트럼, 비언어적 사회적 의사소통 기술, 비언어적 사회적 의사소통 기술 중증도 예측 딥러닝 모델, 전이 학습