Selection of Eye-Tracking Stimuli for Prediction by Sparsely Grouped Input Variables for Neural Networks: towards Biomarker Refinement for Autism

Beibin Li
University of Washington
Seattle Children's Research Institute
beibin@cs.washington.edu

Erin Barney
Seattle Children's Research Institute
erin.barney@seattlechildrens.org

Caitlin Hudac
University of Alabama
cmhudac@ua.edu

Nicholas Nuechterlein
University of Washington
nknuecht@cs.washington.edu

Pamela Ventola
Yale University
pamela.ventola@yale.edu

Linda Shapiro
University of Washington
shapiro@cs.washington.edu

Frederick Shic
University of Washington
Seattle Children's Research Institute
fshic@uw.edu

ABSTRACT

Eye tracking has become a powerful tool in the study of autism spectrum disorder (ASD). Current, large-scale efforts aim to identify specific eye-tracking stimuli to be used as biomarkers for ASD, with the intention of informing the diagnostic process, monitoring therapeutic response, predicting outcomes, or identifying subgroups with the spectrum. However, there are hundreds of candidate experimental paradigms, each of which contains dozens or even hundreds of individual stimuli. Each stimuli is associated with an array of potential derived outcome variables, thus the number of variables to consider can be enormous. Standard variable selection techniques are not applicable to this problem, because selection must be done at the level of stimuli and not individual variables. In other words, this is a grouped variable selection problem. In this work, we apply lasso, group lasso, and a new technique, Sparsely Grouped Input Variables for Neural Network (SGIN), to select experimental stimuli for group discrimination and regression with clinical variables. Using a dataset obtained from children with and without ASD who were administered a battery containing 109 different stimuli presentations involving 9647 features, we are able to retain strong group separation even with only 11 out of the 109 stimuli. This work sets the stage for concerted techniques designed around engines to iteratively refine and define next-generation biomarkers using eye tracking for psychiatric conditions. http://github.com/beibinli/SGIN

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; **Neural networks**; • **Applied computing** → *Health informatics.*

KEYWORDS

eye-tracking, feature selection, group selection, stimuli selection, paradigm selection, neural network, autism

ACM Reference Format:
Beibin Li, Erin Barney, Caitlin Hudac, Nicholas Nuechterlein, Pamela Ventola, Linda Shapiro, and Frederick Shic. 2020. Selection of Eye-Tracking Stimuli for Prediction by Sparsely Grouped Input Variables for Neural Networks: towards Biomarker Refinement for Autism. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers), June 2–5, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3379155.3391334

1 INTRODUCTION

Eye tracking (ET) has become well-regarded as a powerful marker of cognition and human variability [Shic 2016], with several large-scale ongoing efforts in progress to develop next generation biomarker applications that use eye tracking for mental health applications [Loth et al. 2017; Ness et al. 2019; Webb et al. 2019]. Key among these efforts are validation studies aimed at studying the usability of eye-tracking paradigms, typically designed for psychological studies, to serve as markers of treatment response, predictors of latter outcomes, and subgrouping to identify subgroups most likely to benefit from pharmacological or behavioral intervention [FDA-NIH Biomarker Working Group 2016].

Standard psychological investigations using eye tracking in the study of neuropsychiatric conditions rely on collection of eye-tracking data during specific probes designed around psychological or psychiatrically-relevant constructs, for example psychometric investigations of attentional disengagement processes in attention deficit hyperactivity disorder (ADHD) or free viewing of social interactions in autism spectrum disorder (ASD) [Karatekin 2007]. Hundreds of different experimental paradigms have been developed

ETRA '20 Full Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7133-9/20/06…$15.00
https://doi.org/10.1145/3379155.3391334

to study such psychological phenomena, with many paradigms focused around specific areas of research enquiry related to, but not defining of, psychiatric conditions of relevance. Each paradigm, in turn, is composed of dozens or even hundreds of individual stimuli, with each stimulus heterogeneous in its contribution to aggregate statistical results.

A key question in the field is: given the myriad collection of diverse paradigms available as potential markers of a condition, with each paradigm associated with multiple stimuli, and each stimulus associated with multiple outcome measures, how can a parsimonious subset of stimuli be selected to address a specific need? This question is important both theoretically and practically. From the theoretical side, identification of eye-tracking paradigms aligned with specific targets (e.g. social interaction deficits in ASD) provides illumination regarding the fundamental nature of the condition and hints as to ultimate mechanism. From the practical side, use of an eye-tracking paradigm for a specific purpose necessitates trade-offs in usability: the longer the session associated with a biomarker, the more burdensome that biomarker becomes, and the less likely it will be to see practical implementation.

Adding to the complexity is the sheer volume of possible outcome measures available within even a single eye tracking stimulus presentation: while psychological studies typically define a small number of primary outcome measures relevant to the question under investigation, the flexibility of eye tracking is such that a large number of possible measures can be derived, ranging from properties of saccades and fixations, to rates of blinking, to traversal patterns defined by their information theoretic properties. This leads to a complex selection process, and, given the limited sample sizes and data available to inform this process, necessitates specialized techniques appropriate for large *p* small *n*.

In this paper, we aim to (1) apply and validate lasso, group lasso, and Sparsely Grouped Input Variables for Neural Networks (SGIN) algorithms to select eye-tracking stimuli; (2) find a subset of useful eye-tracking stimuli to classify children with and without autism; (3) examine the possibility to select eye-tracking stimuli for predicting directly-observed behavioral ratings of autism severity (ADOS) and IQ, as well as parent-directed ratings associated with degree of autistic symptomatology (SRS) and adaptive functioning (Vineland), in children.

We review related literature on eye tracking and feature selection in Section 2; introduce our eye-tracking paradigms, stimuli, and outcome measures, the experiment design, the feature extraction methods, and machine learning models in Section 3; and show our results to reduce eye-tracking paradigms in Section 4. We discuss the limitations of our work and future directions in Section 5, and conclude our work in Section 6.

2 RELATED WORK

2.1 Eye-Tracking for Children with ASD

In the last two decades, eye-tracking technology has been widely applied in the autism research community to study joint [Navab et al. 2012] and social attention [Guillon et al. 2014; Liberati et al. 2017], visual preference [Pierce et al. 2016], responses to dyadic bids [Campbell et al. 2014], theory of mind capabilities [von dem Hagen et al. 2013], facial expression recognition [Król and Król 2019; Sterling et al. 2008], attention preferences at both semantic [Chen and Zhao 2019] and perceptual [Shic et al. 2007; Wang et al. 2015] levels, among other important abilities for individuals with ASD [Frazier et al. 2017].

Several studies have applied feature selection to analyze patterns of visual attention from eye-tracking data. For instance, Wang et al. [2015] grouped the same eye-tracking features across different stimuli together in order to characterize factors influencing visual attention in individuals with ASD; Coutrot and Guyader [2017] applied lasso and Expectation-Maximization to select features for saliency maps. These studies provide insight for interpreting eye-tracking behaviour and for identification of important features impacting visual attention, but the development of practical tools for application in neurodevelopmental and neuropsychiatric conditions requires different foci, including the need to balance parsimony and efficiency with effective operation.

The length of an eye-tracking session is almost solely attributable to the number and length of stimuli trials presented to participants. A traditional eye-tracking battery can contain hundreds of individual stimuli spread over multiple eye-tracking paradigms to acquire data from children with ASD. However, such long experimental batteries are often stressful for young children and infants. Many children are unable to pay full attention throughout long eye-tracking sessions. Shorter sessions are more tolerable, reduce participant burden, improve data quality, and require less staff time to administer.

Current methods for selecting stimuli and paradigms for eye-tracking outcome measures use more straightforward winnowing of experimental batteries based on independent trial discriminability [Frazier et al. 2018]. While already showing high promise, such approaches do not leverage the full potential of machine learning advances, which can implicitly consider interrelationships between trials. However, care is warranted because machine learning and statistical learning models can suffer from overfitting when the number of features are too large, especially relative to the number of subjects. Machine learning practitioners often apply regularization to avoid overfitting. In this work, our goal is to apply this idea further to reduce the number of stimuli for future eye-tracking studies, leading to more parsimonious batteries tuned to specific outcome goals.

2.2 Group Sparsity in Machine Learning

As discussed, removing highly-correlated paradigms and stimuli can reduce tangible and intangible data acquisition costs and avoid overfitting. Selecting eye-tracking stimuli can be considered as a special case of feature selection in machine learning, which is usually solved by imposing regularization to achieve sparsity. In this paper, we say a group is sparse when all model parameter weightings associated with the input variables in that group are zero. Sparse groups are said to have been removed by the model. We will say that a model has achieved sparsity when it has removed groups of variables during training.

Sparsity has been studied in machine learning for decades, and the lasso is perhaps the most famous example. The standard lasso is a linear model that utilizes a λ-weighted l-1 regularization to enforce sparsity in the model's input variables [Tibshirani 1996].

The group lasso (GL) introduces the notion of groups to extend the standard lasso. Let $X \in \mathbb{R}^{n \times p}$ be the matrix of n samples and p features, $y \in \mathbb{R}^n$ be the vector of n labels, and $\beta^* \in \mathbb{R}^p$ be the vector of optimal model parameters.

$$\beta^* = \arg\min_{\beta} ||y - \sum_{i=1}^{k} X_i \beta_i||_2^2 + \lambda \sum_{i=1}^{k} \sqrt{p_i} ||\beta_i||_2$$

As shown in the equation above, the group lasso partitions the p features into k sets, each of which we will refer to as a group X_i for $i \in \{1, \ldots, k\}$. Denote the cardinality of each X_i as p_i so that we can rewrite the matrix $X = \{X_1, X_2, ..., X_k\}$ and parameter vector $\beta = \{\beta_1, \beta_2, ..., \beta_k\}$, where $X_i \in \mathbb{R}^{n \times p_i}$ represents p_i features and $\beta_i \in \mathbb{R}^{p_i}$ represents the associated parameters for group X_i. The norm for regularization is not squared in the group lasso; otherwise, it becomes weighted ridge regression.

In this paper, we will apply lasso and group lasso as the traditional machine learning methods to reduce the number of eye-tracking stimuli for children with autism.

3 METHODS

3.1 Eye-Tracking Paradigms and Experiment

There are six types of eye-tracking paradigms in the experiment with a total of 109 distinct stimuli. These stimuli were designed by clinical psychologists and researchers to characterize features known to be atypical or impaired in individuals with ASD. Participants in this work watched these 109 stimuli during the eye-tracking experiment with different counterbalancing orders. In this study, our main goal is to test whether we need all 109 stimuli to achieve satisfactory classification and regression performance. The associated stimulus ID and number of features for each paradigm are show in Table 1, and the paradigms are as follows:

- **Activity Monitoring (AM)** - Two actors engage in a play-based activity while speaking to one another. This paradigm taps into social motivation and understanding of shared play.
- **Biomotion (BM)** - Biological motion preference paradigm, with a point-light-display of a human figure engaged in an activity shown on one side of the screen, and a control set of moving dots (either rotating or phase-scrambled) shown on the other side of the screen. This paradigm taps into primitive social preferences.
- **Dyadic Bids (DB)** - Three actors sit around a table while having a conversation. Periodically, an actor turns and asks a question directly to the camera, emulating a direct conversational exchange with the participant. This task examines sensitivity to overtures for social engagement by others.
- **Dynamic Scenes (DS)** - These are clips drawn from professional, child-appropriate movies, edited to be 60 seconds while preserving an overarching cohesive narrative. This paradigm examines more complex, naturalistic scene viewing involving people engaged in a wide variety of tasks.
- **Social Referencing (SR)** - Scenes in which an actor engaged in stressful activities (e.g. stacking a tall thin block tower, inflating a balloon till near-burst). These episodes depict the activities' escalation (e.g. balloon continues to expand), critical event (e.g., balloon pops and the pieces begin flapping wildly), and resolution (e.g. actors sigh with relief). This paradigm examines automaticity of information-seeking from others.
- **Theory of Mind (ToM)** - A protagonist continuously engages in search/play activities with specific objects while a second, antagonist actor, unbeknown to the protagonist, constantly and randomly interferes with the protagonist's goals. This paradigm gauges spontaneous interest in others' intentions as well as comprehension of the mental frame and perspective of others and is relevant for perspective taking and associated skills.

Table 1: Paradigm types and the corresponding stimulus ID. In the third column, we provide the average (and standard deviation) number of features extracted from a stimulus in the corresponding paradigm. The last column shows the length of each stimulus in seconds.

Paradigm Type	Stimulus ID	# features	Len
Activity Monitoring	1 - 12	50 (0.00)	20s
Biomotion	13 - 52	58 (2.41)	15s
Dyadic Bids	53 - 76	177 (3.89)	15s
Dynamic Scenes	77	41 (0.00)	60s
Social Referencing	78 - 93	48 (0.00)	15s
Theory of Mind	94 - 109	106 (8.64)	15s

We will first ignore the nature of these different eye-tracking paradigms when applying machine learning algorithms, and then we will re-evaluate these six paradigms after finding suitable subsets of stimuli to classify children with ASD and without ASD.

3.2 Data Acquisition

We recruited 64 children for this eye-tracking experiment (mean age = 76.32 months, SD age = 16.3 months; 43 males, 21 females). Thirty-two participants were diagnosed with ASD, and the other thirty-two participants have different diagnostic labels including typical development, language delay, and other developmental concerns (non-ASD). There is no significant difference for age and sex between the ASD and non-ASD population. Besides the diagnostic label of ASD or not, ADOS severity score, SRS total score, and Vineland ABC Standard Score were obtained by clinicians and trained psychometrists for 30 of the participants with ASD. Stanford-Binet Intelligence Scale IQ test scores were obtained for all participants.

The ADOS (Autism Diagnostic Observation Schedule) severity score is a standardized measure of autism severity, administered as a direct assessment of a child by a trained administrator, that is held as a gold standard for clinical, genetic, and neurobiological research [Gotham et al. 2009]. The SRS-2 (Social Responsiveness Scale, second edition) aims to aid in diagnosis and treatment planning for autism, with a high SRS score indicating deficiencies in reciprocal social behavior, which might lead to severe interference with everyday social interactions [Bruni 2014]. This assessment is questionnaire-based, and typically directed towards caregivers or educators to answer about a specific child. The Vineland-3 ABC

(Adaptive Behavior Scale) Standard Score, with a mean of 100 and standard deviation 15, is based on "Communication", "Daily Living Skills", and "Socialization" adaptive behavior domains for children [Sparrow et al. 2016; Yang et al. 2016]. It is delivered by a trained administrator to a caregiver as a structured interview. Several studies suggest IQ can be used as a predictor for young children with autism [Charman et al. 2011; Harris and Handleman 2000; Mayes and Calhoun 2003]. The Stanford-Binet Intelligence Scale is one such validated measure of IQ, administered directly to a child by a trained administrator. These four measurements provide different yet important clinical and psychological information for participants with ASD.

In the eye-tracking experiments, an SR Eyelink 1000 Plus sampling binocularly at 500 Hz was used for data collection. Experiments were shown on a Samsung SyncMaster 2233RZ 22-inch monitor running at 60hz (1680 x 1050 pixels). Participants were positioned 650 mm from the eye-tracking camera at the start of each session. Experiments are programmed in Neurobehavioral Stimulus Presentation in a comprehensive delivery system enabling both eye-tracking quality control (multiple embedded 5-point calibrations) and behavioral management (ability to spontaneously create breaks using videos to manage non-compliant behaviors by participants). Protocols for data acquisition were aligned with those described in [Webb et al. 2020]. 95% trimmed mean calibration error was 0.751 visual degrees. No specific instructions were given to participants except to watch the screen. Thus the interpretation of administered trials relied on the characterization of passive (or implicit) attentional process inherent to the individual during stimulus display. Individuals with ASD are differentially impacted by the nature of verbal instruction, e.g. as observed in tests of intelligence [Dawson et al. 2007], and as such passive viewing studies without explicit verbal direction have some of the greatest applicability across levels of function as well as age in the study of neurodevelopment.

3.3 Feature Extraction for Eye-Tracking Data

In the post-hoc analysis, we apply the Dynamic Programming on Distance Dispersion fixation identification [Li et al. 2016] algorithm, using partial fixations weighted by the number of gaze points within each region, and region-of-interest (ROI) analysis to obtain clean features. Quality control measures include calibration accuracy, calibration stability, and data collection rates. Participants' mean fixation time, number of fixations, saccade duration, time of first saccade, valid ET duration (samples collected within a ROI multiplied by the sampling rate), percentage of gaze on screen, percentage of missing data, gaze data quality, and other oculomotor features are extracted for each paradigm. These oculomotor statistics are also evaluated for multiple ROIs and within specific time windows corresponding to events of interest in stimuli. The list of derived results to be considered was not exhaustive, but rather reflected specific, high-level perspectives on variables of note that were expected to provide separation of ASD/non-ASD groups and utility in the monitoring and tracking of intervention effects in children with ASD.

For example, ROIs are annotated by experts for each frame of video stimuli, with no composite regions overlapping but with the inclusion of combined variables (e.g. head = eyes + mouth + hair + skin). For the Activity Monitoring, Dynamic Scenes, and Social Referencing paradigms, oculomotor features on each ROI are extracted and include regions such as bodies, heads, eyes, mouths, activities, and distractors. For the Dyadic Bids and Theory of Mind paradigms, visual response latency for events that occur in the paradigms are additionally included. For the Biomotion paradigms, the ROI is defined by the screen location (i.e. left, right) and the nature of the region (biological or control condition). Visual latency to orient to biological motion is also included.

A total 9647 features are extracted from the 109 stimuli. We arrange these 9647 features into 109 groups, where each group of features corresponds to one eye-tracking stimulus. As shown in Table 1, most groups have less than 50 features, and 34 groups have more than 100 features.

3.4 Machine Learning Details

After obtaining eye-tracking features from the post-hoc fixation and ROI analyses, we impute missing values by using the unsupervised Expectation-Maximization algorithm [Dempster et al. 1977].

Our eye-tracking paradigm selection problem suffers from "curse of dimensionality" with $p >> k > n$, where the number of features ($p = 9647$) is much larger than the number of samples ($n = 64$), and the number of groups ($k = 109$) is also larger than the number of samples. Because of the large group size, we expect that the machine learning models can learn to remove most of the groups.

We use $1_{\hat{y}>0.5}$ to represent positive predictions in binary classification problems. In regression tasks, we normalize the labels to the range $[0, 1]$ so that all the measurements can be compared on the same scale.

We use 10-fold cross-validation to validate the machine learning models. We acknowledge that the validation performance might be over-optimistic and does not represent real-world performance for ASD/non-ASD classification. ASD is a diverse and complex disorder that is diagnosed behaviorally by clinicians, a process that requires significant effort and professional training. We use the same training/validation split for all models in the experiment so that the comparisons between different machine learning methods are fair. We also normalize all features based on the training set before each cross-validation.

Lasso is agnostic to the definition of groups; so, we say a group is identified as sparse by lasso if lasso identifies all features as sparse in the given group.

We use accuracy, precision, recall, and Area Under the receiver operating characteristic Curve (AUC) score to evaluate classification results, where participants with ASD are considered positive samples. We use Pearson correlations, which range from -1 to 1, to evaluate regression tasks (with random guessing producing r=0). We use the average number of sparse groups in the 10-fold cross validations as the result for one regularization term λ.

3.5 Sparsify Grouped Variables in Neural Network

In the last decade, neural networks (NNs) have gained popularity and tend to outperform linear models when data is not linearly separable. Feng and Simon [2017] and Scardapne et al. Scardapane et al. [2017] applied group l-1 regularization for neural networks to reduce input features individually without considering the group nature of features. In this work, we extend the regularization to multi-layer, non-linear NNs and ensure sparsity for grouped input variables, imposing group $l-1$ regularization on both the input layer and the first hidden layer. This model is named Sparsely Grouped Variables in Neural Network (SGIN). Different from [Meier et al. 2008; Yuan and Lin 2006] and related studies whose primary goal is regularization, SGIN focuses on removing groups of input variables in order to reduce the information a machine learning system requires during inference and to ease the expense of future data collection.

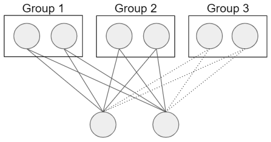


Figure 1: Sparsity for grouped input variables in neural networks. The top row is the input layer; the bottom row is the first hidden layer; dashed lines stand for connections that can be removed to satisfy the sparsity constraint if group 3 is redundant (and thus group 3 could subsequently be removed in the final machine learning pipeline).

Denote J as the element-wise loss and ϕ as an arbitrary loss function in a machine learning system. We define the optimization goal for β, the vector of all parameters in the NN model f, and the regularization term τ as follows.

$$\beta^* = \arg\min_{\beta} J(\beta) + \lambda\tau = \arg\min_{\beta} \phi(f(X), y) + \lambda\tau \tag{1}$$

$$\tau = \sum_{i=1}^{k} \tau_i = \sum_{i=1}^{k} \sqrt{p_i}||\theta_i||_2 \tag{2}$$

Here $\theta \in \mathbb{R}^{p \times q}$ is the matrix of parameters in the first layer of the NN, q is the number of neurons in the first hidden layer, and $\theta \subset \beta$. We let $\theta_i \in \mathbb{R}^{p_i \times q}$ be the parameters between the input variables in group i and all neurons in the first hidden layer. The τ_i is the Frobenius norm of the θ_i matrix rather than the norm of the β_i vector. When k and p are equivalent, each group is composed of one q-dimensional feature vector, and τ reduces to the regularization of the first hidden layer in [Scardapane et al. 2017], an effective method to prune neurons from NNs.

Different from other methods in l-1 regularization for neural network, SGIN provides an efficient algorithm, Stochastic Blockwise Coordinated Gradient Descent, to optimize their loss function by leveraging existing techniques in coordinate descent, blockwise coordinate descent, and stochastic gradient descent. The new optimization algorithm converges faster and finds more sparse groups than the traditional stochastic gradient descent algorithm in empirical experiments.

In this study, we will compare these linear and non-linear machine learning models in their ability to reduce eye-tracking stimuli needed for targeting prediction of clinical variables relevant to studies of individuals with autism.

We use the same hyperparameters (e.g. neural network structure, learning rate scheduling, number of epochs, etc.) for all experiments involving neural networks. For simplicity, we try 1-hidden layer, 2-hidden layer, and 3-hidden layer neural network before the experiments, and then choose the 2-hidden layer neural network (3000 and 500 neurons in the 2 hidden layers respectively) because it has the fastest training convergence.

We do not use bias in neurons in the first hidden layer for implementation convenience, but we use bias for all other layers. Scardapane et al. [2017] also claim that removing the bias term would not affect prediction results. We use ReLU activation for all hidden layers in all experiments for consistency, and the results with Sigmoid activation are similar. Based on convergence of training loss $\phi(f(X), y)$, we use initial learning rate $\eta = 0.1$, train SGINs for 5 epochs, and decrease the learning rate by 50% after each epoch. We use a batch size of 64 for the ASD/non-ASD classification and a batch size of 1 for the regression tasks.

The Institutional Review Board (IRB) from Yale University and the IRB from Seattle Children's Research Institute approved our eye-tracking experiments, data acquisition and data analyses protocols. The eye-tracking data processing pipeline, which is implemented in C++, Matlab, R, and SPSS, along with the data used in this study, is available upon request.

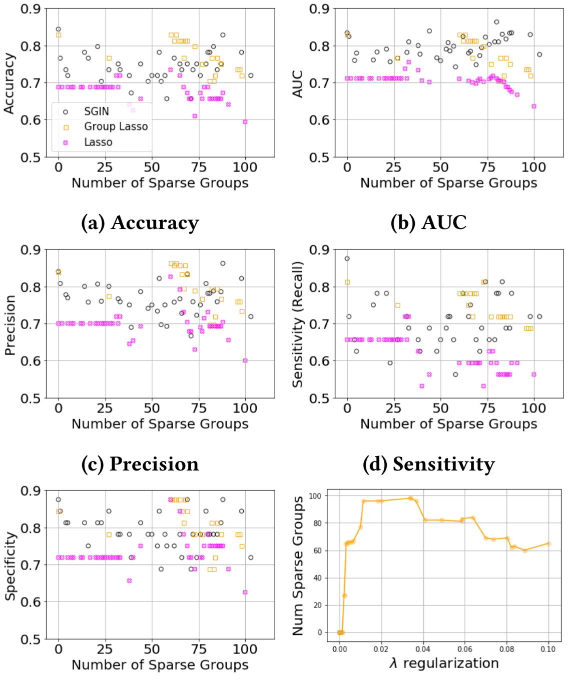
Code for the machine learning experiments are open-sourced in http://github.com/beibinli/SGIN. We use PyTorch [Paszke et al. 2019] to implement the neural networks and SGIN. We also use sklearn for the lasso and pyglmnet [Jas et al. 2020] for the group lasso. All these packages are used in Python 3.6. Details of this implementation, optimization algorithm, and additional applications of SGIN are presented in [Li et al. 2019].

4 RESULTS

4.1 Classification of ASD versus Non-ASD

Our result in Figure 2 suggests that even when 89.9% of the groups are removed (i.e. 99 paradigms are removed), SGIN can still achieve a satisfactory classification result (78.13% accuracy and 82.91% AUC score). The lasso is able to find group sparsity, but its performance is about 15% worse than SGIN (65.6% accuracy and 70.8% AUC score) with 71 stimuli removed. The group lasso performs similarly to SGIN (79.71% accuracy and 79.72% AUC score) on average, though it struggles to learn fewer than 60 groups, even after tuning λ (as shown in Figure 2f), which makes it difficult to interpret the least important groups of features: that is, the group lasso either removes no groups or many, and rarely anything in-between. In the ablation study, using Stochastic Gradient Descent on Loss 1 directly performs worse than SGIN and cannot find redundant groups for this problem.

Even with only 11 groups of variables (out of the original 109 groups), SGIN still achieves good accuracy and AUC on the validation set. This result suggests that substantial redundancy exists in the 109 paradigms: even using a small subset of these paradigms, the eye-tracking experiment can still be effective in between-group classification. Though these methods remain to be validated on larger, unseen test cases, our result suggests that SGIN has the potential to help design more streamlined eye-tracking batteries and select better eye-tracking paradigms in the future.



(a) Accuracy

(b) AUC

(c) Precision

(d) Sensitivity

(e) Specificity

(f) λ tuning for GL

Figure 2: Validation Results for ASD/non-ASD Classification

To compare with other popular machine learning methods in the autism research community, we run Support Vector Machine (with both linear and RBF kernel), decision tree, random forest, and other models for 100 runs each. The SGIN can achieve the best validation accuracy among all methods in all runs, while other linear methods (linear regression, logistic regression, and SVM with linear kernel) also have similarly high performance (around 84% accuracy), possibly because the stimuli were already carefully selected and designed to study children with ASD. However it is important to note that these methods would result in no sparsification. Group lasso has similar distribution with SGIN, but SGIN is the only method that can effectively select stimuli and obtain high accuracy simultaneously.

4.2 Paradigms Selected by SGIN

In our experiment, all machine learning models could not guarantee the selected paradigms are the same across runs with different training data, because this is a $p > n$ problem such that the optimal solution may not be unique. Nevertheless, we run lasso, group lasso, and SGIN for 100 runs each with bootstrap. Then, we count the number of times a paradigm is discarded (sparsified) in all runs. The group lasso is unstable with the bootstrap runs, and most of its runs fail to identify redundant groups. On the other hand, lasso and SGIN agree with each other for most stimuli, as shown in Figure 3,

We compare the probability that a given paradigm would be sparsified in Table 2. Among the six eye-tracking paradigms, Dyadic Bids is the most useful one identified by all three models. Surprisingly, Social Referencing has the largest possibility to be discarded by the machine learning algorithms.

Table 2: Probability a paradigm would be sparsified by the machine learning algorithms. The table is sorted by the probability from SGIN.

Paradigm Type	P(discard)		
	Lasso	Group lasso	SGIN
Dyadic Bids	38.17%	3.08%	24.96%
Theory of Mind	58.19%	5.50%	49.62%
Activity Monitoring	73.17%	6.00%	65.25%
Biomotion	77.50%	5.88%	66.72%
Dynamic Scenes	89.00%	6.00%	71.00%
Social Referencing	86.44%	5.94%	72.00%

Both Dyadic Bids and Theory of Mind contain features for visual latency, and coincidentally these two stimulus types are the two most used paradigms identified by the machine learning models. Biomotion also contains visual latency features, but it does not contain features for gaze on more discrete ROIs like heads, faces, eyes, and bodies. The features extracted in Dyadic Bids and Theory of Mind may contain features extracted from other paradigms, and perform slightly better than those other paradigms, such that the machine learning models preferentially select stimuli from Dyadic Bids and Theory of Mind. More studies are needed to test the robustness of these paradigms.

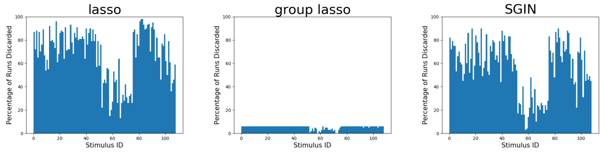


Figure 3: Stimuli discarded in 100 bootstrap runs: *(Left)*: Results from lasso; *(Middle)*: Results from group lasso; *(Right)*: Results from SGIN. The x-axis is the Stimulus ID, and y-axis is the times the corresponding paradigm has been discarded.

4.3 Importance of Eye-Tracking Features

Examination of important eye-tracking features is useful for understanding the specific visual attention properties that characterize children with autism versus those without. We adopt SHAP (SHapley Additive exPlanations) value [Lundberg and Lee 2017] to interpret the importance of features in our trained neural networks. The importance is defined as $V_p = \frac{1}{n}\sum_{i=0}^{n} |s_p(X_i)|$ for feature p, where $s_p(X_i)$ is the SHAP value for feature p with participant i.

Based on this approach, the top 10 important features are shown below, where the corresponding paradigm name is included in parenthesis:

(1) gaze duration toward activities in Stimulus 9 (AM);
(2) gaze duration on biological movements in Stimulus 37 (BM);
(3) proportion of events meeting quality requirements (i.e. good data) in Stimulus 58 (DB);
(4) response latency to non-biological motion in Stimulus 40 (BM);
(5) start time of the 4th fixation (post pre-attentive exploration) in Stimulus 66 (DB);

(6) start time of the 9th fixation (late exploration) in Stimulus 67 (DB);
(7) gaze duration toward eyes in Stimulus 59 (DB);
(8) start time of the 10th region explored in Stimulus 73 (DB);
(9) latency to first orient to the body in Stimulus 105 (ToM)";
(10) number of fixations toward the eyes of the actor with direct gaze in Stimulus 58 (DB).

From this analysis, we find that quality control, ROI looking duration, fixation, and exploration statistics are all important for ASD classification. Overall, validity of gaze is important because children with ASD can provide poorer quality data due to non-compliance and attention issues during presentation of socially-oriented scene content (e.g. see [Chawarska et al. 2012]). Similarly, latency in the Biomotion paradigm provides useful information for understanding ASD phenotypes (e.g. see [Umbricht et al. 2017]). Stimuli indexed by SGIN may be reflective of individual trials comprising experiments which may be particularly good examples of the experimental class.

4.4 Predicting ADOS, SRS, Vineland, and IQ

The eye-tracking paradigms were originally designed to provide separation between children with ASD versus non-ASD children. Thus, the machine learning performance in the previous section is relatively high. We want to see whether these paradigms are as effective when used to predict IQ, ADOS, SRS, and Vineland scores.

SGIN sparsifies the most groups for all the regression tasks. However, the performance of both of these methods was quite poor, with neither SGIN or lasso fitting above chance for predicting ADOS scores, Vineland scores, or SRS scores. However, SGIN ($r = 0.54$, $p < 0.0001^{***}$) successfully predicted IQ scores. This suggests that predicting relative levels of social and adaptive impairment (as indexed by ADOS, SRS, and Vineland) is a challenging problem. Future studies may need to consider stimuli specifically designed for this problem, and/or new techniques to more appropriately model these outcomes. IQ levels are an important outcome measure for understanding the clinical phenotypes of children with ASD, and though the effects are modest, they suggest that eye tracking has the potential, when combined with appropriate sparsification and regression techniques, to lead to efficient methods for obtaining clinically-relevant features across an array of clinical considerations.

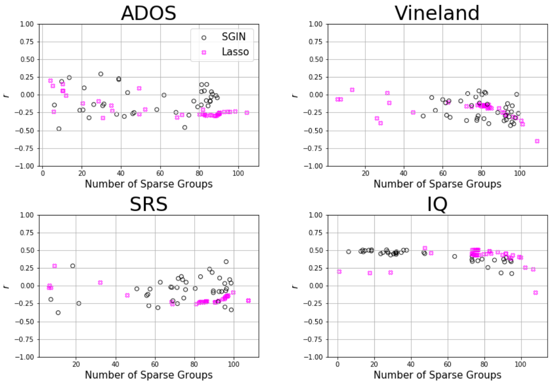


Figure 4: Regression Results for 40 cross-validation runs: the x-axis is the number of sparse groups, and the y-axis is the corresponding correlation coefficient r.

5 DISCUSSION AND FUTURE WORK

In all the experiments, SGIN achieves the top or comparable-to-the-top performance. By using a multi-layer non-linear neural network structure as the base inference model, SGIN can achieve high performance while excluding redundant groups. Even if the training data is linearly distributed, SGIN still has advantages over lasso and group lasso in terms of parsimonious representation and selection of eye-tracking stimuli to be used in prediction. When considering the problem of using eye-tracking data to distinguish between ASD and non-ASD groups, it is important to note that selecting a particular feature for a paradigm in an eye-tracking experiment is considered a pivotal task to inform our understanding of the underlying mechanism. Further organizing paradigms into smaller subgroups can help clinicians and researchers to study these potential biomarkers more efficiently. These explorations require sparsity in bi-level, hierarchical, tree-structure, and overlapping groups. In addition, while temporal information was implicitly included in features through expert organization (i.e. extraction of variables during specific events in stimuli presentation, such as the moment an individual began speaking), data-driven approaches towards isolation of critically informative timing periods is an important avenue for future work.

We only recruited 64 participants for our experiments, and only 30 participants had full ADOS, Vineland, and SRS measurements. Even though this proof-of-concept study shows promising results from SGIN, more studies are needed to fully investigate eye-tracking features and meaningful biomarkers for children with ASD. Specifically, the limited power provided by the small sample of children with ASD to predict autism symptomatology may need to be expanded to characterize well-known heterogeneity effects in ASD [Campbell et al. 2014; Chawarska et al. 2014; Happé et al. 2006]. It is important to note that results presented in this manuscript may reflect developmental properties of ASD specific to the school-age period. Much work still remains before these techniques, and eye tracking in general, can be used in a practical fashion for optimizing treatments and for prediction of longer term outcomes [Shic 2016].

6 CONCLUSION

In this study, we compare lasso, group lasso, and SGIN to select eye-tracking stimuli in an experiment for children with autism. The results of our experiment suggest that SGIN can achieve sparsity in grouped input variables for neural networks while maintaining high accuracy in group prediction. In particular, SGIN can classify children with and without ASD with reasonable accuracy with 89.9% of the eye-tracking stimuli removed. SGIN also outperforms lasso in regression of IQ, but more studies are needed to support this result. Future work can design additional tests and structures to examine the robustness of selected stimuli and adaptations to identify more successful methods for prediction of ADOS, SRS, and Vineland scores.

ACKNOWLEDGMENTS

This work is supported by NIH awards K01 MH104739, R21 MH103550, R03 MH092618; the NSF Expedition in Socially Assistive Robotics #1139078; and Simons Award #383661.

REFERENCES

Teryn P Bruni. 2014. Test review: Social responsiveness scale–second edition (SRS-2).

Daniel J Campbell, Frederick Shic, Suzanne Macari, and Katarzyna Chawarska. 2014. Gaze response to dyadic bids at 2 years related to outcomes at 3 years in autism spectrum disorders: a subtyping analysis. *Journal of autism and developmental disorders* 44, 2 (2014), 431–442.

Tony Charman, Andrew Pickles, Emily Simonoff, Susie Chandler, Tom Loucas, and Gillian Baird. 2011. IQ in children with autism spectrum disorders: data from the Special Needs and Autism Project (SNAP). *Psychological medicine* 41, 3 (2011), 619–627.

Katarzyna Chawarska, Suzanne Macari, and Frederick Shic. 2012. Context modulates attention to social scenes in toddlers with autism. *Journal of Child Psychology and Psychiatry* 53, 8 (2012), 903–913.

Katarzyna Chawarska, Frederick Shic, Suzanne Macari, Daniel J Campbell, Jessica Brian, Rebecca Landa, Ted Hutman, Charles A Nelson, Sally Ozonoff, Helen Tager-Flusberg, et al. 2014. 18-month predictors of later outcomes in younger siblings of children with autism spectrum disorder: a baby siblings research consortium study. *Journal of the American Academy of Child & Adolescent Psychiatry* 53, 12 (2014), 1317–1327.

Shi Chen and Qi Zhao. 2019. Attention-Based Autism Spectrum Disorder Screening With Privileged Modality. In *Proceedings of the IEEE International Conference on Computer Vision.* 1181–1190.

Antoine Coutrot and Nathalie Guyader. 2017. Learning a time-dependent master saliency map from eye-tracking data in videos. *arXiv preprint arXiv:1702.00714* (2017).

Michelle Dawson, Isabelle Soulières, Morton Ann Gernsbacher, and Laurent Mottron. 2007. The level and nature of autistic intelligence. *Psychological science* 18, 8 (2007), 657–662.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 1 (1977), 1–22.

FDA-NIH Biomarker Working Group. 2016. *BEST (Biomarkers, EndpointS, and other Tools) Resource.* http://www.ncbi.nlm.nih.gov/books/NBK326791/

Jean Feng and Noah Simon. 2017. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592* (2017).

Thomas W. Frazier, Eric W. Klingemier, Sumit Parikh, Leslie Speer, Mark S. Strauss, Charis Eng, Antonio Y. Hardan, and Eric A. Youngstrom. 2018. Development and Validation of Objective and Quantitative Eye Tracking-Based Measures of Autism Risk and Symptom Levels. *Journal of the American Academy of Child and Adolescent Psychiatry* (Sep 2018). https://doi.org/10.1016/j.jaac.2018.06.023

Thomas W. Frazier, Mark Strauss, Eric W. Klingemier, Emily E. Zetzer, Antonio Y. Hardan, Charis Eng, and Eric A. Youngstrom. 2017. A Meta-Analysis of Gaze Differences to Social and Nonsocial Information Between Individuals With and Without Autism. *Journal of the American Academy of Child and Adolescent Psychiatry* 0, 0 (May 2017). https://doi.org/10.1016/j.jaac.2017.05.005

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

Quentin Guillon, Nouchine Hadjikhani, Sophie Baduel, and Bernadette Rogé. 2014. Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews* 42 (2014), 279–297.

Francesca Happé, Angelica Ronald, and Robert Plomin. 2006. Time to give up on a single explanation for autism. *Nature neuroscience* 9, 10 (2006), 1218–1220.

Sandra L Harris and Jan S Handleman. 2000. Age and IQ at intake as predictors of placement for young children with autism: A four-to six-year follow-up. *Journal of autism and developmental disorders* 30, 2 (2000), 137–142.

Mainak Jas, Titipat Achakulvisut, Aid Idrizović, Daniel Acuna, Matthew Antalek, Vinicius Marques, Tommy Odland, Ravi Prakash Garg, Mayank Agrawal, Yu Umegaki, et al. 2020. Pyglmnet: Python implementation of elastic-net regularized generalized linear models. *Journal of Open Source Software* 5, 47 (2020).

Canan Karatekin. 2007. Eye tracking studies of normative and atypical development. *Developmental Review* 27, 3 (Sep 2007), 283–348. https://doi.org/10.1016/j.dr.2007.06.006

Magdalena Ewa Król and Michał Król. 2019. A novel machine learning analysis of eye-tracking data reveals suboptimal visual information extraction from facial stimuli in individuals with autism. *Neuropsychologia* 129 (2019), 397–406.

Beibin Li, Nicholas Nuechterlein, Erin Barney, Caitlin Hudac, Pamela Ventola, Linda Shapiro, and Frederick Shic. 2019. Sparsely Grouped Input Variables for Neural Networks. arXiv:arXiv:1911.13068

Beibin Li, Quan Wang, Laura Boccanfuso, and Frederick Shic. 2016. Optimality of the distance dispersion fixation identification algorithm. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications.* 339–340.

Alessio Liberati, Roberta Fadda, Giuseppe Doneddu, Sara Congiu, Marco A Javarone, Tricia Striano, and Alessandro Chessa. 2017. A statistical physics perspective to understand social visual attention in autism spectrum disorder. *Perception* 46, 8 (2017), 889–913.

Eva Loth, Tony Charman, Luke Mason, Julian Tillmann, Emily J. H. Jones, Caroline Wooldridge, Jumana Ahmad, Bonnie Auyeung, Claudia Brogna, Sara Ambrosino, and et al. 2017. The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Molecular Autism* 8 (2017), 24. https://doi.org/10.1186/s13229-017-0146-8

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems.* 4765–4774.

Susan Dickerson Mayes and Susan L Calhoun. 2003. Ability profiles in children with autism: Influence of age and IQ. *Autism* 7, 1 (2003), 65–80.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 1 (2008), 53–71.

Anahita Navab, Kristen Gillespie-Lynch, Scott P Johnson, Marian Sigman, and Ted Hutman. 2012. Eye-tracking as a measure of responsiveness to joint attention in infants at risk for autism. *Infancy* 17, 4 (2012), 416–431.

Seth L. Ness, Abigail Bangerter, Nikolay V. Manyakov, David Lewin, Matthew Boice, Andrew Skalkin, Shyla Jagannatha, Meenakshi Chatterjee, Geraldine Dawson, Matthew S. Goodwin, and et al. 2019. An Observational Study With the Janssen Autism Knowledge Engine (JAKE®) in Individuals With Autism Spectrum Disorder. *Frontiers in Neuroscience* 13 (2019). https://doi.org/10.3389/fnins.2019.00111

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems.* 8024–8035.

Karen Pierce, Steven Marinero, Roxana Hazin, Benjamin McKenna, Cynthia Carter Barnes, and Ajith Malige. 2016. Eye tracking reveals abnormal visual preference for geometric images as an early biomarker of an autism spectrum disorder subtype associated with increased symptom severity. *Biological psychiatry* 79, 8 (2016), 657–666.

Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. 2017. Group sparse regularization for deep neural networks. *Neurocomputing* 241 (2017), 81–89.

Frederick Shic. 2016. Eye tracking as a behavioral biomarker for psychiatric conditions: the road ahead. *Journal of the American Academy of Child & Adolescent Psychiatry* 4, 55 (2016), 267–268.

Frederick Shic, Brian Scassellati, David Lin, and Katarzyna Chawarska. 2007. Measuring context: The gaze patterns of children with autism evaluated from the bottom-up. In *2007 IEEE 6th International Conference on Development and Learning.* IEEE, 70–75.

SS Sparrow, Domenic V Cicchetti, and Celline A Saulnier. 2016. Vineland adaptive behavior scales, (Vineland-3). *Antonio: Psychological Corporation* (2016).

Lindsey Sterling, Geraldine Dawson, Sara Webb, Michael Murias, Jeffrey Munson, Heracles Panagiotides, and Elizabeth Aylward. 2008. The role of face familiarity in eye tracking of faces by individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders* 38, 9 (2008), 1666–1675.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

Daniel Umbricht, Marta del Valle Rubido, Eric Hollander, James T McCracken, Frederick Shic, Lawrence Scahill, Jana Noeldeke, Lauren Boak, Omar Khwaja, Lisa Squassante, et al. 2017. A single dose, randomized, controlled proof-of-mechanism study of a novel vasopressin 1a receptor antagonist (RG7713) in high-functioning adults with autism spectrum disorder. *Neuropsychopharmacology* 42, 9 (2017), 1914–1923.

Elisabeth AH von dem Hagen, Raliza S Stoyanova, James B Rowe, Simon Baron-Cohen, and Andrew J Calder. 2013. Direct gaze elicits atypical activation of the theory-of-mind network in autism spectrum conditions. *Cerebral cortex* 24, 6 (2013), 1485–1492.

Shuo Wang, Ming Jiang, Xavier Morin Duchesne, Elizabeth A Laugeson, Daniel P Kennedy, Ralph Adolphs, and Qi Zhao. 2015. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron* 88, 3 (2015), 604–616.

Sara Jane Webb, Frederick Shic, Michael Murias, Catherine A. Sugar, Adam John Naples, Erin Barney, Heather Borland, Gerhard Helleman, Scott P. Johnson, Minah Kim, and et al. 2019. Biomarker Acquisition and Quality Control for Multisite Studies: The Autism Biomarkers Consortium for Clinical Trials. *Frontiers in Integrative Neuroscience* 13 (2019). https://doi.org/10.3389/fnint.2019.00071

Sara Jane Webb, Frederick Shic, Michael Murias, Catherine A Sugar, Adam J Naples, Erin Barney, Heather Borland, Gerhard Hellemann, Scott Johnson, Minah Kim, et al. 2020. Biomarker Acquisition and Quality Control for Multi-Site Studies: The Autism Biomarkers Consortium for Clinical Trials. *Frontiers in Integrative Neuroscience* (2020).

Sabrina Yang, Jessica M Paynter, and Linda Gilmore. 2016. Vineland adaptive behavior scales: II profile of young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders* 46, 1 (2016), 64–73.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 1 (2006), 49–67.