

# Classification Predictive Modeling of Dyslexia

Y-C.YU<sup>1</sup>, K. SHYNTASSOV<sup>1</sup>, A. ZEWGE<sup>2</sup>, \*L.A. GABEL<sup>3</sup>,

<sup>1</sup>Electrical & Computer Engineering, <sup>2</sup>Computer Science, <sup>3</sup>Neuroscience, Lafayette College, Easton, PA

003.12  
B35

## INTRODUCTION

Dyslexia is a reading disability that affects children across language orthographies, despite adequate intelligence and educational opportunity. If learning disabilities remain untreated, a child may experience long-term social and emotional problems, which may influence future success in all aspects of their lives. Early detection and intervention will help to close the gap between typically developing and reading impaired children in acquiring reading skills. We have demonstrated that animal models of dyslexia, genetic models based on candidate dyslexia susceptibility genes, and children with specific reading impairment show a common deficit on a virtual Hebb-Williams maze task. Since virtual maze task does not require oral reporting (rapid access to phonological processing) or rely on text, performance is not influenced by a potential difference in reading experience between groups. Although the correlation between dyslexia and the performance in the virtual Hebb-Williams maze task has been demonstrated, classification of atypical participants (i.e., dyslexic participants) through real-time observation of their performance on the virtual Hebb-Williams maze task is not feasible at this time. A computational model that can predict reading ability based on maze learning performance, would enable real-time feedback of the performance in the form of at-risk percentages for reading. Reading data and maze learning outcomes were analyzed from 227 school-aged children (8-14 years of age). Applying multiple variables (e.g. biological sex, time to complete the task, errors committed, and deviation from the true path) into machine-learning based computational models resulted in the prediction accuracy above 70%. Successful development of this predictive model would allow for early detection of risk for reading impairment, which can lead to early interventions to close the gap between typically developing and reading impaired children in acquiring reading skills.

## METHODS

### Participants

227 school-aged children, 8-14 years of age. Woodcock-Johnson IV Tests of Achievement subtests: Letter-Word Identification, Word Attack, Reading Recall, and Passage Comprehension. Reading cluster scores for Reading, Basic Reading, and Reading Comprehension were calculated from the four subtests. Reading impairment (atypical) was defined as at least 1 standard deviation below the mean (Mean: 100; SD: 15). Reliability coefficients for cluster scores, which consist of two or more subtests, range from 0.85 to 0.94.

### Virtual Hebb-Williams Maze

The virtual maze environment as shown in Fig. 1 was created using the Unreal Development Kit. Data collection and analysis were implemented in Java. All software packages were installed on laptop PCs for experiments. Mazes were displayed at a resolution of 1600 × 900 in full-screen mode. The interior of a maze was divided into 6 × 6 cells. Each cell consists of 256 × 256 units, with the goal and start boxes 300 units deep, and a viewing height (eye level) of 85 units from the ground. A Logitech Attack 3 joystick was provided for participants to navigate through the virtual environment at a constant velocity of 175 units/sec and a maximum turn rate of 96°/sec. More detailed information of the virtual Hebb-Williams maze for this study is available in Gabel et al., 2016.

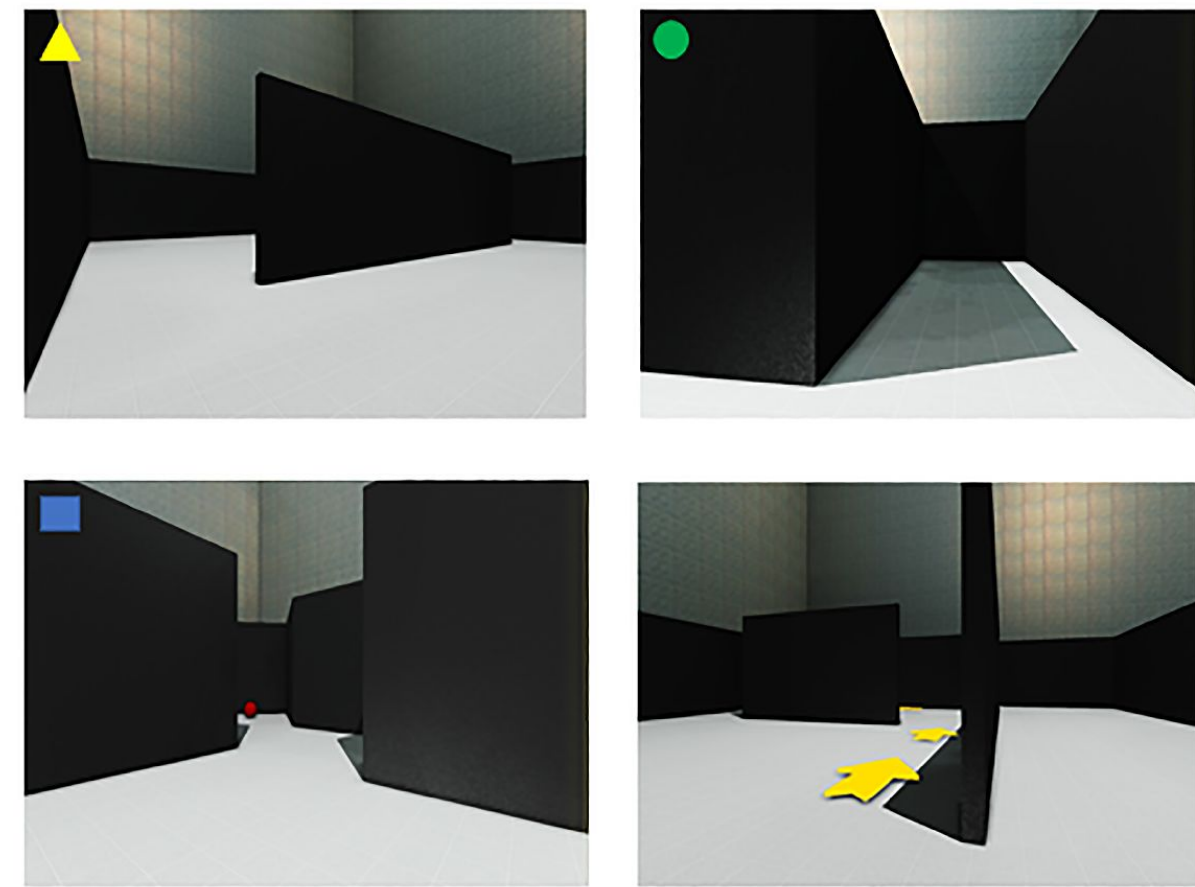
Maze 1 was used as a training maze for participants to get acquainted to the virtual by finding the red ball in the maze. Once the target was acquired by the participant, a congratulatory statement appeared on the screen. Participants completed each trial consecutively until all six trials were complete, with a 120-second maximum completion time per trial. When the maximum allowable time was reached, the interior walls were designed to come up, guiding the participant to the goal box. Participants were randomly oriented within the start box at the beginning of each trial. All participants completed Hebb-Williams mazes 1, 5, 6, 8, 11, and 12 in the same day. The raw data generated by the virtual maze program were saved in text files that include x and y coordinates of the participant's position in a 1536 (or 6 × 256) by 1536 maze grid and a corresponding time stamp.

### Data Processing

Due to the large size of the raw datasets, the original 1536 by 1536 maze grid was scaled down to a 24 by 24 grid for efficiency purposes, which results in a scaling coefficient of 64. Therefore, every x and y coordinate was divided by this scaling coefficient. Cells visited by a participant (numbered 0 to 23) were stored in an array list. Every participant's array list was then compared to the ground-truth list of the same maze for the trial. The ground truth list of each maze was generated by moving through the virtual maze in the shortest path while the data were recorded. However, since the distance travelled by the participants varies from each other, the sizes of their lists differ, which meant that direct comparisons with the ground-truth list was impossible. Thus, every cell visited by the participant was first matched to the nearest cell from the ground-truth list to determine the performance score.

## RESULTS

### 1 Hebb-Williams maze configurations and virtual platform

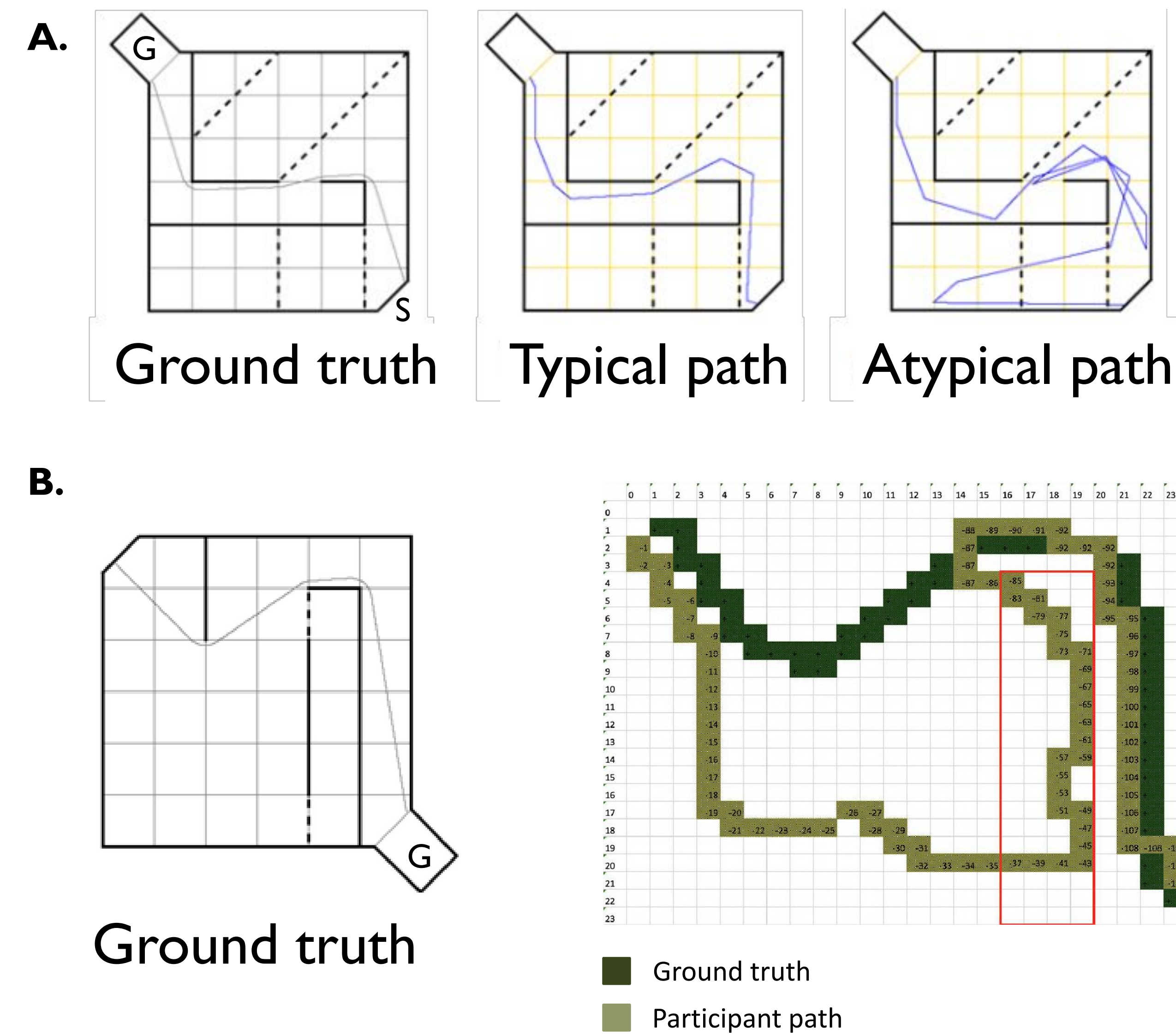


Line drawings of the mazes used for training (mazes 1 and 5) and testing (mazes 6, 8, 11, and 12). The goal box containing the target (i.e., red ball) is identified by the letter "G," and the start position is identified by the letter "S." The number listed below the drawing indicates the Hebb-Williams maze configuration. The solid black lines represent walls in the maze; the black dotted lines represent error zones, which are not visible to participants; and the gray trace is the true path between the start and the goal box. The specific position of the avatar in the scenes from maze 5 are represented by specific points in the schematic of maze 5 (triangle, entering the maze; circle, dead end; square, identifying the target; arrows, true path revealed once 120 s has elapsed).

### 3 Identify training algorithms that can classify both typical and atypical readers with good accuracy

Algorithm	With age		Without age	
	Atypical	Typical	Atypical	Typical
Fine Tree	2	7	2	7
Medium Tree	2	7	2	7
Coarse Tree	1	10	1	10
Linear	1	9	1	9
Discriminant				
Logistic	1	9	1	9
Regression				
Gaussian Naive	2	10	2	10
Bias				
Kernel Naive Bias	1	9	1	9
Linear SVM	1	10	1	10
Quadratic SVM	0	10	0	10
Cubic SVM	0	10	0	10
Fine Gaussian	0	10	0	10
SVM				
Medium Gaussian	0	10	0	10
SVM				
Coarse Gaussian	0	10	0	10
SVM				
Fine KNN	1	9	0	8
Medium KNN	0	10	0	10
Coarse KNN	0	10	0	10
Cosine KNN	0	10	0	10
Cubic KNN	0	10	0	10
Weighted KNN	0	10	0	10
Boosted Trees	1	9	1	9
Bagged Trees	0	10	0	10
Subspace	1	10	1	10
Discriminant				
Subspace KNN	1	10	1	10
RUSBoosted	7	7	6	7
Trees				
Narrow NN	1	9	0	10
Medium NN	2	10	2	9
Wide NN	1	10	1	9
Bilayered NN	1	10	0	9
Trilayered NN	2	10	2	10

### 2 Comparison between ground-truth and participant performance



Ground truth

**A.** Line drawings of Maze 8 depicting the ground truth, the completion path for a participant with a typical reading profile, and an atypical reader. **B.** Line drawing of Maze 12 depicting the ground truth (left). Performance score calculation algorithm visualization for maze 12 (right). Light green cells demonstrate the participant's path from the start box (S) to the goal box (G). Dark green cells represent the ground truth path. The red border indicates the error zone.

### 4 Sensitivity and accuracy of classification

Case description	Atypical	Typical
Without gender and age	5	7
With gender	6	7
With age	6	8
With both gender and age	7	7

RUSBoosted Trees Algorithm test with data from Mazes 5, 6, 8, 11, 12

Case description	Atypical	Typical
Without gender and age	5	6
With gender	5	7
With age	5	7
With both gender and age	5	8

RUSBoosted Trees Algorithm test with data from Mazes 5, 6, 8, 11

Case description	Atypical	Typical
Without gender and age	6	8
With gender	5	8
With age	7	8
With both gender and age	7	8

RUSBoosted Trees Algorithm test with data from Mazes 6, 8, 11, 12

Case description	Atypical	Typical
Without gender and age	7	6
With gender	5	8
With age	5	8
With both gender and age	6	8

RUSBoosted Trees Algorithm test with data from Mazes 5, 6, 8, 11, 12, Trials 1-4

## DISCUSSION

- A computational model has been developed to classify human participants with dyslexia by using experimental data from 227 participants (8-14 years of age) in response to the virtual Hebb-Williams maze task.
- Data obtained from each participant include the participant's biological sex, age, and performance scores from 36 trials (6 trials for each of the 6 mazes).
- Classification Learner App (MATLAB) was used to train the model with 29 machine learning algorithms
- Classification accuracy for atypical participants was only achieved with the RUSBoosted (random under sampling boosting) Trees algorithm.
- A systematic approach was used to evaluate the impact of training data reduction on accuracy of the classification model.
- The data demonstrate that the model performed equally well without considering the training maze (Maze 1), but the contribution of biological sex and age was significantly greater when the size of the data was reduced.
- Performance on some mazes (i.e. Maze 5 & Maze 12) had a greater impact on the accuracy of the model than others
- Consistent with previous findings, removing the last two trials (Trials 5 & 6) from the model didn't influence the accuracy of the model

## CONCLUSIONS

- In future studies, training data should include performance scores from trials 1-4 for Mazes 5, 6, 8 11, and 12, and include biological sex and age
- Early identification, prior to the time that children at risk for RD experiencing learning difficulties, allows for early intervention
- If a predictive model that can classify children at risk for reading impairment before they can read, with a strong degree of accuracy, the virtual maze may be a low-cost, time efficient, easy-to-use tool for the early detection of reading impairment

## REFERENCES

- Gabel LA, Battison A\*, Truong DT, Lindström ER, Voss K\*, Yu Y-C, Roongruengratanakul S\*, Shyntassov K\*, Riebesell S\*, Tournanos N\*, Nielsen-Pheiffer C, Paniagua S, Gruen, JR (2022) Orthographic depth may influence the degree of severity of maze learning performance in children at risk for reading disorder. *Developmental Neuroscience*, 44, 651-670 [Epub ahead of print 2022 Oct. 12].
- Gabel, LA, Voss, K\*, Johnson, E, Lindström, ER, Truong, DT, Murray, EM\*, Cariño, K\*, Nielsen, CM, Paniagua, S., Gruen, JR (2021). Identifying dyslexia: Link between maze learning and dyslexia susceptibility gene, *DCDC2*, in young children. *Developmental Neuroscience*, 43, 116-133. [Epub ahead of print 2021 Jun 29].
- Yu Y-C, Shyntassov K\*, Zewge A\*, and Gabel LA, (2022) *Classification Predictive Modeling of Dyslexia*, 2022 56th Annual Conference on Information Sciences and Systems (CISS), 2022, pp. 177-181. doi: 10.1109/CISS53076.2022.9751182

## ACKNOWLEDGEMENTS

- This work was supported by:
- Alan & Wendy Pesky Foundation Research Grant to LAG.
  - Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R15HD087937 (LAG, EL, Y-CY and JRG). *The content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH.*
  - Alexander von Humboldt Foundation (LAG)