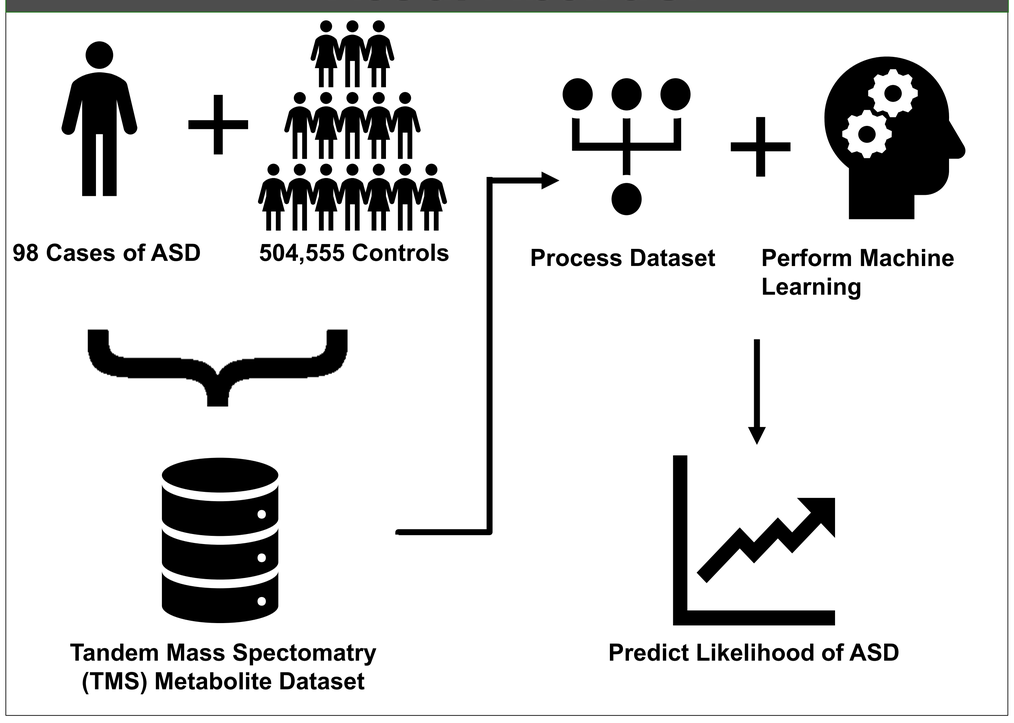
Predicting Autism Spectrum Disorder (ASD) at Birth Using Machine Learning

Brady Hoskins[1], Leo Brueggeman[1], Jacob Michaelson[1]
[1]Department of Psychiatry, The University of Iowa

Introduction

- Incidence of autism spectrum disorder (ASD) has recently reached 1/59 children
- Diagnoses of ASD are done with time consuming, behavioral tests
- Early diagnosis and intervention consistently leads to better outcomes
- There is great need for a biomarker of ASD that could improve diagnostic efforts
- Using newborn metabolite data, we trained machine learning models to detect ASD risk factors

Visual Abstract



Methods

- During preprocessing we normalized the data to account for time
- We also accounted for the ratio of males to females in the ASD category
- Controls are unlabeled and likely have ASD at population levels
- Table 1. shows an overview of the dataset used during the preprocessing stage

Population Distribution		
Cases = 98	Males = 258,054	Male Cases = 76
Controls = 504,555	Females = 246,501	Female Cases = 22

Table 1. Distribution of Data from DevGenes database.

- We tested several classification algorithms to predict ASD status using the 'caret' package in R
- Model evaluation is done by finding the Area Under the Receiving Operator Curve (AUROC)
- The AUROC is found by plotting the sensitivity and specificity of a model
- Sensitivity and specificity calculations are show in Figure 1.

2.1 SENSITIVITY, SPECIFICITY AND ACCURACY

Outcome of the diagnostic test	Condition (e.g. Disease) As determined by the Standard of Truth		
	Positive	Negative	Row Total
Positive	TP	FP	TP+FP (Total number of subjects with positive test)
Negative	FN	TN	FN + TN (Total number of subjects with negative test)
Column total	TP+FN (Total number of subjects with given condition)	FP+TN (Total number of subjects without given condition)	N = TP+TN+FP+FN (Total number of subjects in study)

Table 1. Terms used to define sensitivity, specificity and accuracy

Figure 1. Chart showing how to find sensitivity and specificity [1].

Results

- Figure 2. shows the Receiving Operator Characteristics values from a 5-fold repeated cross validation
- Hyperparameter optimization was done via random search
- Random forest models show superior classification performance, with a mean AUROC of 0.5513 in cross validation

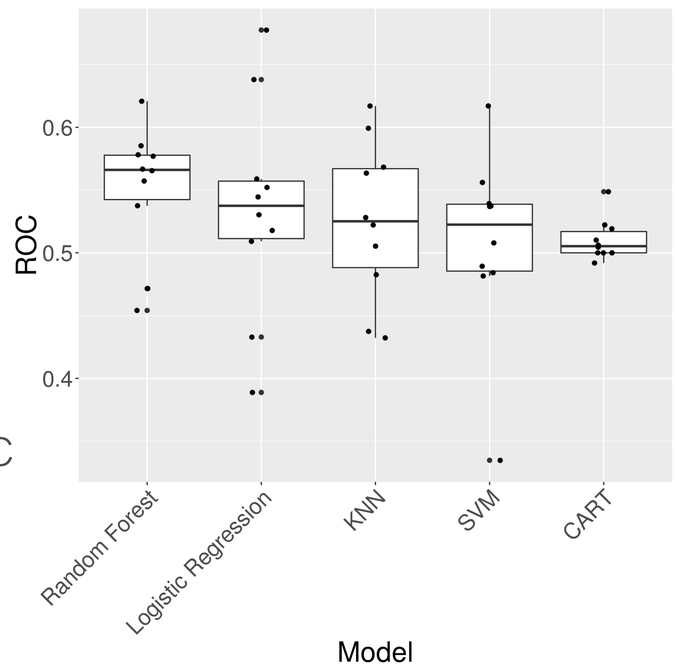


Figure 2. Boxplot comparison of cross-validated models ROC values.

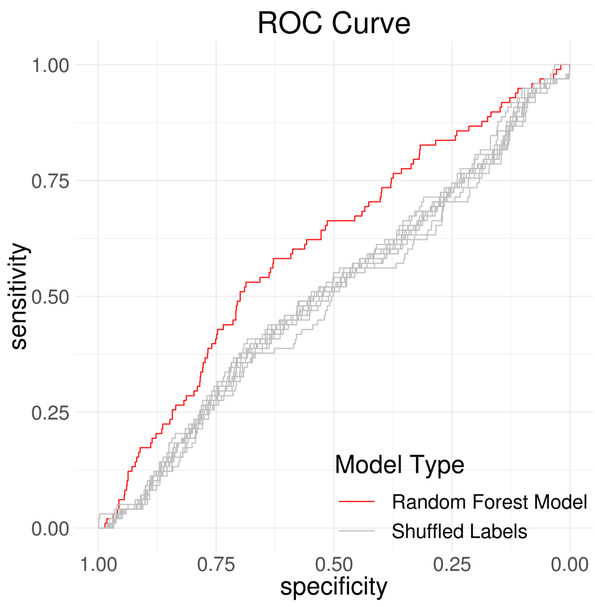


Figure 3. Plot of Random Forest ROC Curve with ROC Curve of assigning cases at random.

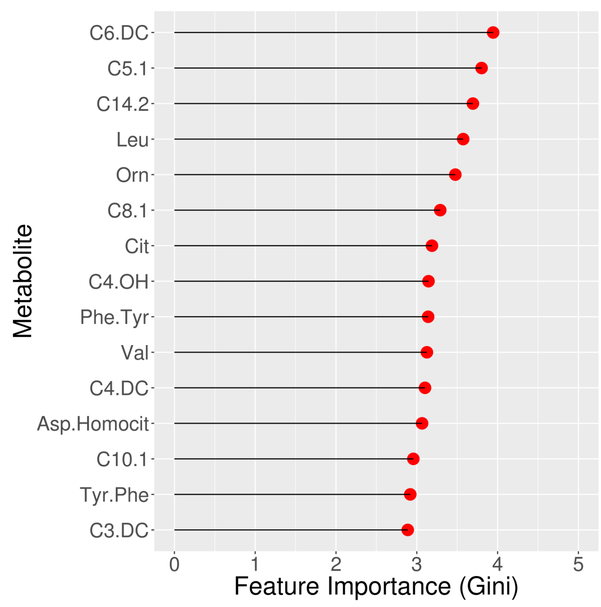


Figure 4. Plot of feature importance based on Mean Decrease Gini.

- Figure 3. a random forest model was fit on the full data (OOB predictions shown) achieving an AUROC of 0.6053
- As a control, 10 iterations of model fitting with shuffled cases are shown as grey lines on ROC plot
- Figure 4. feature importance of random forest model shows several Acylcarnitine are important
- Interestingly, C6-DC or Methylglutarylcarnitine has been found to be associated with the metabolic disease (HMG-CoA lyase deficiency) that causes hypoglycemia and seizures [2]
- Also, Leucine (Leu) and Citrulline (Cit) have been significantly connected to children with ASD [3]

Conclusion

- This research is ongoing but our work thus far has shown that machine learning is able to detect some metabolic risk factors for ASD at birth
- Moving forward, we aim to utilize deep learning methods such as the use of auto encoders
- We will continue to receive diagnoses of ASD from our DevGenes database that should enhance the predictability of our models

Acknowledgements & References

I would like to thank Dr. Michaelson for affording me the opportunity to conduct research in his lab and Leo Brueggeman for the assistance throughout my project in helping me learn more every day.

[1] Zhu, Wen & Zeng, Nancy & Wang, Ning. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS ® implementations. NorthEast SAS users group, health care and life sciences.

[2] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {date}. World Wide Web URL: https://omim.org/

[2] Tirouvanziam R, Obukhanych TV, Laval J, Aronov PA, Libove R, Banerjee AG, Parker KJ, O'Hara R, Herzenberg LA, Herzenberg LA, Hardan AY. Distinct plasma profile of polar neutral amino acids, leucine, and glutamate in children with Autism Spectrum Disorders. J Autism Dev Disord. 2012 May;42(5):827-36.