An Ensemble Learning Method for Early Detection of Autism and Other Developmental Delays using Crowdsourced Video Annotations

Scott Fleming[1], Qandeel Tariq[2], Michael Du[2], Dennis Wall[2]

[1]Biomedical Informatics Training Program, Stanford University

[2]Department of Pediatrics (Systems Medicine), Stanford University

Introduction / Related Work

- Existing methods for diagnosing various forms of developmental delay (DD) in young children are both lengthy and costly
- Diagnosis often requires that a trained professional evaluate ~100 behaviors over several hours of personal interaction with the child[1]
- Detecting DD in the developing world, with scarce resources, is especially difficult, creating an unmet need of scalable diagnostics
- Recent work has demonstrated that a crowd of raters with minimal training can perform comparably to a clinical professional by viewing a short home video of the child and answering ~30 qn's[1].

Methods / Results

- Collected 159 short home videos (~3 minutes each) of Bangladeshi children, many with stunting/high risk for DD
- Six American raters with minimal training gave responses to 30 different questions about each child in a subset of the videos.
- Split the dataset into a training set (114 videos), and test set (45 videos), each with approximately equal balances of children with typical development, Autism Spectrum Disorder (ASD), and speech/language conditions (SLC) distinct from ASD.
- Train set had:
  - 40 children with ASD (mean age = 2 yr 4 mo, SD age = 7 mo)
  - 35 children with SLC (mean age = 2 yr 1 mo, SD age = 6 mo)
  - 39 children with no DD (mean age = 2 yr 6 mo, SD age = 6 mo)
- Test set had:
  - 15 children with ASD (mean age = 2 yr 5 mo, SD age = 7 mo)
  - 15 children with SLC (mean age = 2 yr 1 mo, SD age = 5 mo)
  - 15 children with no DD (mean age = 2 yr 4 mo, SD age = 6 mo)
- For each rater, trained two types of models (Logistic Regression, Random Forest[2]) based on the videos that they rated in two stages
  - Stage 1: Train a classifier to predict no DD vs. DD
  - Stage 2: Train a classifier to predict ASD vs. SLC
- Used innovative weighting scheme to ensemble raters' models:

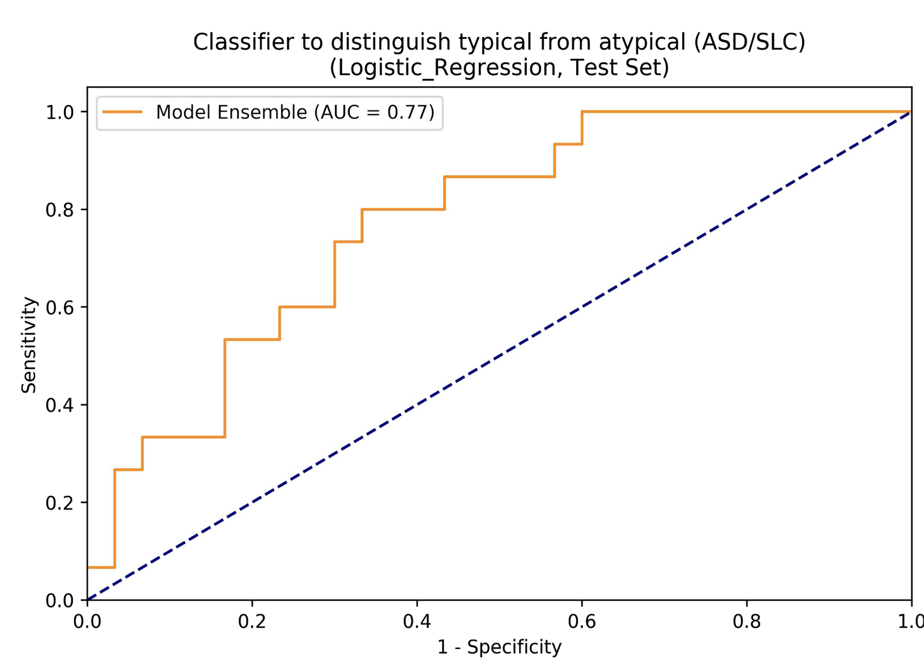
$$\hat{y}^{(i)} = F(\mathbf{x}^{(i)}) \qquad \mathcal{L}(\mathbf{X}, \mathbf{y}, F) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y^{(i)} = F(\mathbf{x}^{(i)})\}$$

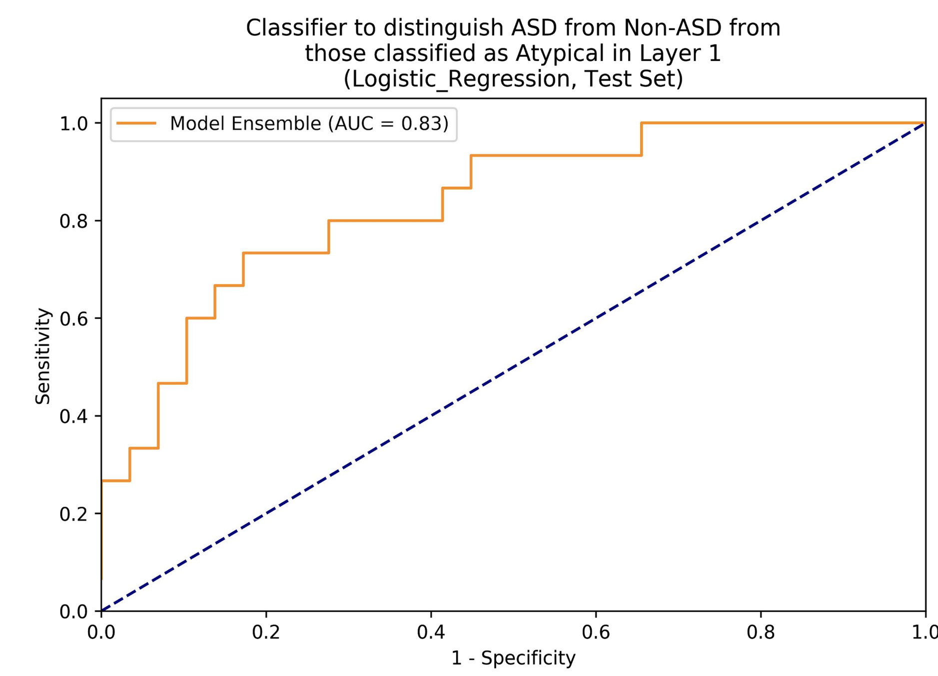
$$\bar{F}(\mathbf{X}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y^{(i)} = mode(\mathbf{y})\} \qquad F^* = \arg\max_{F} \mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\mathcal{L}(\mathbf{X}, \mathbf{y}, F)\right]$$

$$z_j = \mathcal{L}(\mathbf{X}, \mathbf{y}, F^*) - \mathcal{L}(\mathbf{X}, \mathbf{y}, \bar{F})$$

$$\omega_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad \tilde{y}^{(i)} = \sum_{j=1}^{K} \omega_j F_j^*(\mathbf{x}^{(i)})$$

Methods / Results





Pred. (Top) vs. Actual (Left) for LR, Stage 1

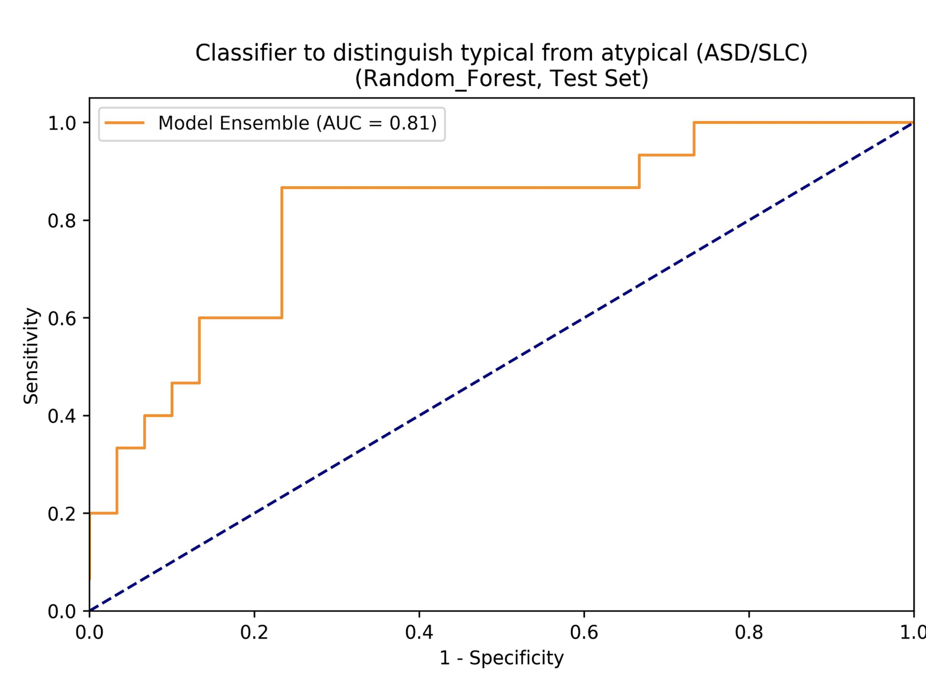
	ASD/SLC	Normal
ASD/SLC	30	0
Normal	14	1

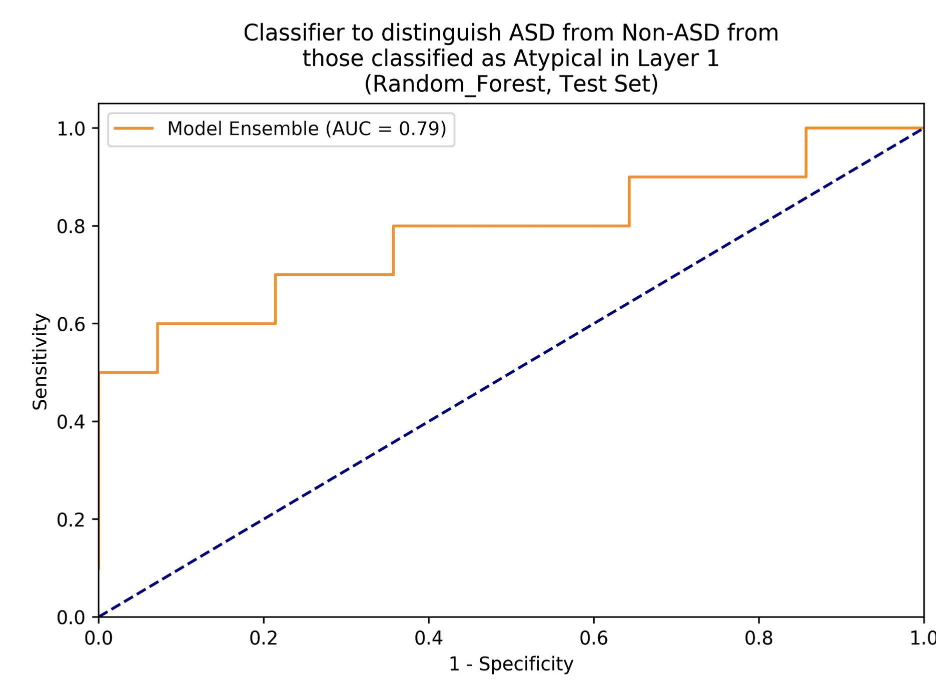
Predicted (Top) vs. Actual (Left) for LR

	ASD	Normal	SLC
ASD	4	0	11
Normal	0	1	14
SLC	1	0	14

Pred. (Top) vs. Actual (Left) for LR, Stage 2

	SLC	ASD
SLC	28	1
ASD	11	4





Pred. (Top) vs. Actual (Left) for RF, Stage 1

	ASD/SLC	Normal
ASD/SLC	22	8
Normal	2	13

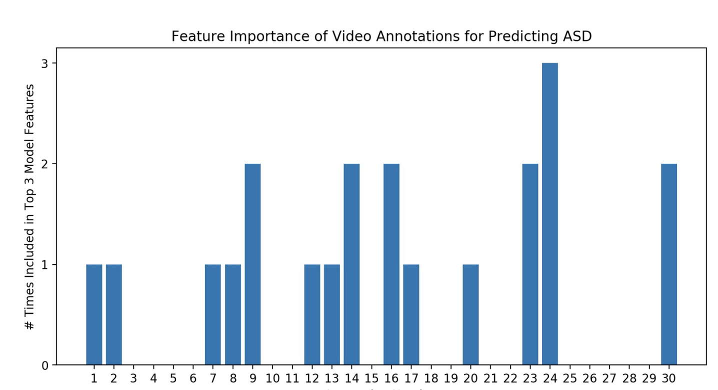
Predicted (Top) vs. Actual (Left) for RF

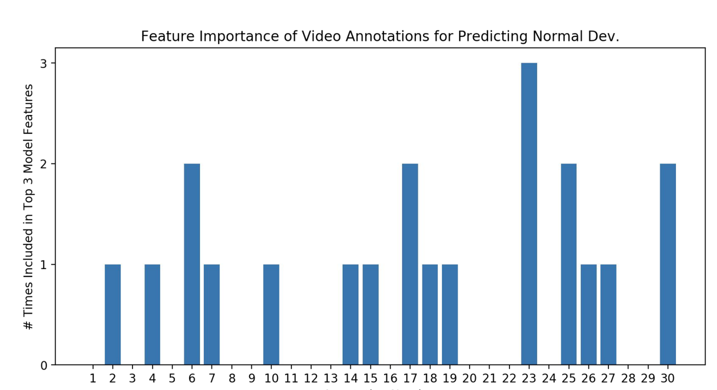
	ASD	Normal	SLC
ASD	5	5	5
Normal	0	13	2
SLC	0	3	12

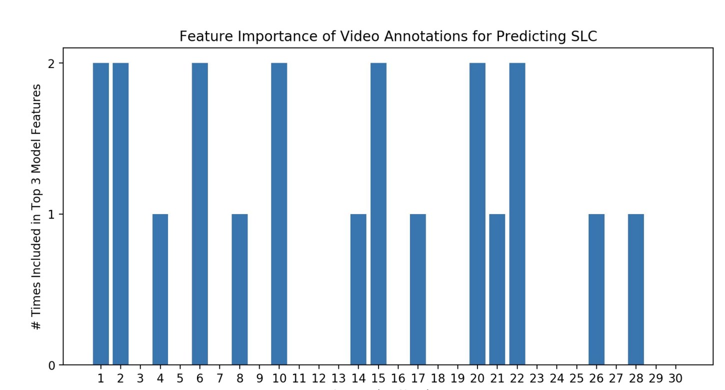
Pred. (Top) vs. Actual (Left) for RF, Stage 2

	SLC	ASD
SLC	14	0
ASD	5	5

- Logistic regression ensemble achieved stage 1 accuracy of 69%, stage 2 accuracy of 73%, overall accuracy 42%
- Random forest ensemble achieved stage 1 accuracy of 78%, stage 2 accuracy of 79%, overall accuracy of 67%
- Calculating feature importance ranks for random forest ensemble reveals underlying focus of the successful model:

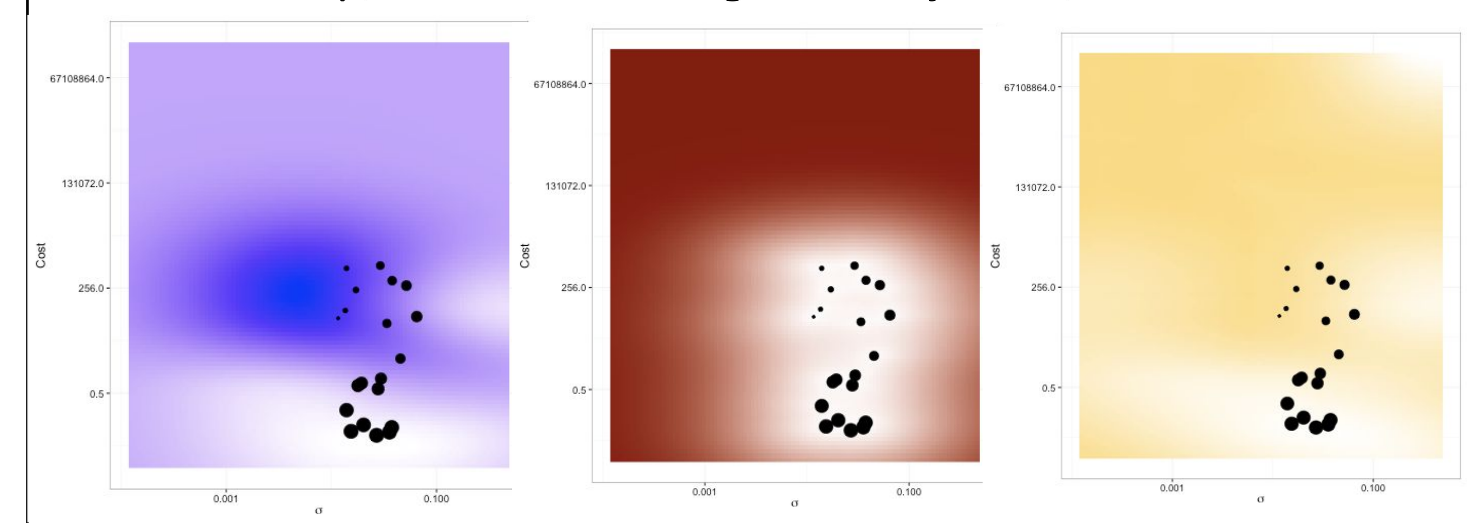






Methods / Results

- We used hyperopt, a python library for hyperparameter optimization, to automate the model tuning process[3]
- Hyperopt utilizes Sequential Model-Based Optimization (SMBO) with a Tree-structured Parzen Estimator (TPE) to maximize Expected Improvement (EI)[4]
- At each step, constructs surrogate of objective, illustrated here[5]:



Conclusion

- These findings demonstrate the potential utility of cross-cultural and cross-linguistic video annotation for scalable early detection and diagnosis of developmental delays in the developing world
- Future work includes:
- Instead of imputing before training the model, use a model that can implicitly handle missing values (i.e. tree models, XGBoost[6])
- Try first inferring the "true" annotations for each video, then build a single model off of that.
- Try different weighting schemes, i.e. instead of "boost" in accuracy, try using RMSE from a gold-standard rater

References

[1] Tariq, Qandeel, et al. "Use of artificial intelligence and manual feature tagging for rapid and mobile detection autism spectrum disorder". *Under review* (2018).

[2] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12.Oct (2011): 2825-2830.

[3] Bergstra, James, Dan Yamins, and David D. Cox. "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms." *Proceedings of the 12th Python in Science Conference*. 2013.

[4] Bergstra, James S., et al. "Algorithms for hyper-parameter optimization." *Advances in neural information processing systems*. 2011.

[5] Kuhn, Max. "Bayesian Optimization of Machine Learning Models." Revolution Analytics, 19 June 2016, 9:35, blog.revolutionanalytics.com/2016/06/bayesian-optimization-of-machine-learning-models.html.

[6] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016.

Acknowledgements

I would like to thank Dennis Wall for supporting me during this rotation, the Biomedical Informatics Training Program for the retreat, and the NIH for its support through a BD2K Fellowship.