Deep Learning based Multimedia Data Mining for Autism Spectrum Disorder (ASD) Diagnosis

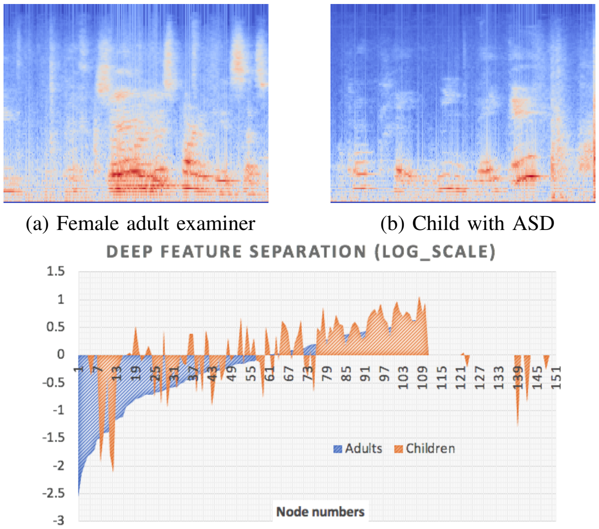
Saad Sadiq *, Micheal Castellanos*, Jacquelyn Moffitt†, Mei-Ling Shyu*, Lynn Perry† and Daniel Messinger†
*_Deptartment of Electrical and Computer Engineering_
†_Department of Psychology_
University of Miami, Coral Gables, FL, USA
Email: {saadsadiq, m.castellanos29, jmoffitt, shyu, lkperry, dmessinger}@miami.edu

***Abstract*—Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by deficits in social communication and restricted and repetitive patterns of behavior. Autism is estimated to affect 1 in 59 children in the United States and costs roughly $35B to the society. Early diagnosis of ASD is vital for promoting early intervention and positive developmental outcomes. Traditional diagnostic procedures for ASD include structured behavioral observation by a trained clinician. Diagnosticians typically rely on the Autism Diagnostic Observation Schedule (ADOS-2) to quantify ASD symptoms. In this paper, we take a parallel approach and investigate language modalities and discover associations between objective measurements of social communication and ASD symptoms. We analyze 33 children with autism and extract their linguistic patterns from their conversations with diagnosticians in a clinical setting. Our methods use Long-Short Term Memory (LSTM) networks to learn Speech Activity Detection (SAD) and speaker diarization patterns to generate the vocal turn-taking metrics. We then use our novel proposed pipeline to predict the ADOS-2 Calibrated Severity Scores (CSS) of Social Affect (SA). The proposed framework achieve state-of-the-art predictive diagnostic estimates of ASD severity compared to industry's leading algorithms. Results compared with the language acquisition system Language ENvironment Analysis (LENA) and other algorithms indicate a significant improvement in the R^2 measure.**

***Keywords*-Multimedia Data mining, Medical Diagnostics, Autism Spectrum Disorder (ASD), Deep Learning**

I. Introduction

Knowledge Discovery and Data Mining (KDD) has been extensively used in interdisciplinary domains such as CRM, education, clinical medicine, fraud detection and genetic data mining [1]. These methods focus on extracting useful knowledge from raw multi-modal data that would be inaccessible by traditional machine learning methods. Recently, there has been a lot of interest to utilize data mining methods in exploring niche behavioral and psychological symptoms such as autism spectrum disorder (ASD) [2]. The challenge of extracting predictive features from highly collaborative domains such as in autism research is that it draws upon research from statistics, databases, pattern recognition, data visualization, and high-performance computing. Thus, until recently, the computation power as well as the penetration of machine learning in behavioral sciences was limited. However, with the advent of ground-breaking deep learning methods such as Convolutional Neural Networks (CNNs) and Long-Short Term Memory (LSTM) networks [3], a myriad of highly predictive deep features are uncovered [4].



(a) Female adult examiner

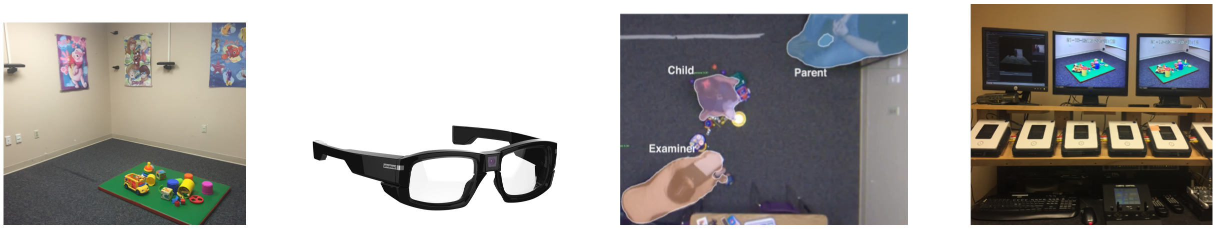
(b) Child with ASD

(c) Deep features from a neural network.

Figure 1. Differences in spectral and deep features between audio of adult and child, which allows us to model predictions based on the type of audio data.

Diagnoses of ASD can be difficult to obtain as they can only be made upon observation by a highly trained examiner [5]. The difficulties in processing data such as child speech in a daily-life environment have been highlighted at the 2017 JSALT Summer Workshop at CMU [6], where it became apparent that unconventional speech containing mumble, cry, overlaps and other artifacts required finer models and motivated the organization of the 2018 DIHARD Challenge [7]. To date, there have not been large-scale investigations that objectively measure social communication disturbances in children with ASD [8].

In this paper we use data from a sizable ASD study and mine useful patterns using efficient deep learning. Spectrogram samples from the study, shown in Figure 1(a,b) illustrates the differences in audio patterns between adult and



(a) Examination room with play objects for Child-Examiner interactions.

(b) Pivothead glasses with embedded video camera worn by ADOS-2 examiner.

(c) Examination setup of child, examiner, and parent during ADOS-2 from a ceiling-mounted camera.

(d) The control room to capture the media.

Figure 2. Resources to study objective measures at the Early Play and Development Laboratory in the Department of Psychology.

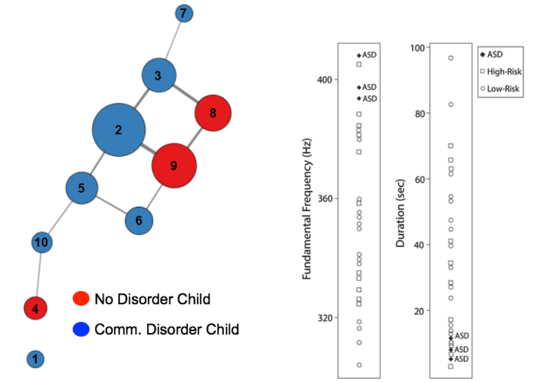
child vocalizations. This allows us model predictions using CNNs and separate speakers based on their deep features as shown in Figure 1(c). Currently the "gold standard" for autism diagnosis includes the Autism Diagnostic Observation Schedule (ADOS-2) [9], a semi-structured play-based assessment. Our data is collected following the ADOS-2 standard in a controlled environment as shown in Figure 2a. Using face mounted Pivothead glasses, worn by examiners and parents (Figure 2b), we observe social communication as well as restricted and repetitive behaviors of interests. The assessment takes between 40 to 60 minutes to complete and includes a variety of materials and activities that are chosen based upon the child's language fluency (Figure 2c). The children are treated on a variety of symptoms and given a calibrated severity score (CSS). The diagnosis is captured from all attendees and processed for multi-modal analysis in the control room (Figure 2-d). This paper uses deep learning methods to investigate the association between objective measurements of social communication and ASD symptoms in 33 children having ASD. Linguistic patterns of, children having ASD, are explored to derive objective measures that directly predict the CSS Social Affect. A data mining pipeline that uses these vocal patterns is developed to model a non-parametric estimator that accurately classifies ADOS-2 symptoms.

The rest of the paper is organized as follows. Section II highlights the methods developed by other research teams. The methodology applied as well as the developed prediction framework are described in Section III. The experiments and results are explained in Section IV. Finally, the conclusion is given in section V with indications of improvements and future work.

II. Previous Work

Autism Spectrum Disorder is defined by restricted, repetitive patterns of behavior, as well as persistent disturbances of social communication and interaction across multiple contexts [10]. These patterns can be subdivided into linguistic, social, facial and movement features that explain ASD symptoms.

Linguistic deficits are a key component both of the diagnostic profile of ASD and the early identification of this disorder [11]. Initial referrals are often based on concerns over language delay, with other social challenges becoming more apparent later on [12]. Cohen et al. [13] found that higher child latency to parent approach was associated with higher ADOS Social Affect severity scores. A social network analysis indicated that children with communication disorders (deaf/hard-of-hearing children) exhibited more heterogeneous social interaction patterns than other children [14] (Figure 3a). The average distance between child and parent was also related to the child's RRB CSS at the level of a trend, $r = .74, p \leq .10$.



(a) Vocal interactions of children (within 1.5m and mutually oriented). Node size shows total vocalizations with peers.

(b) The frequency and duration of vocalizations distinguish children at low and high risk for ASD.

Figure 3. Studies on child communications are indicative of differentiation between children with and without ASD.

Esposito et al. evaluated and distinguished ASD risk and outcome patterns from vocal differences notable early in development [15]. 15-month-old infants at high risk for ASD, who went on to receive a diagnosis of ASD by 36 months, were observed to produce cries that are shorter and of a higher fundamental frequency than their peers (Figure 3b). Young children with ASD can also be differentiated from their typically developing peers (18-37 months) based on the content of language produced (e.g., length of utterance, number of nouns, etc.) [16]. However, research addressing the cross-context stability of social communication disturbances is rare in part because the field has lacked efficient methods for measuring behavior [17]. Further application of these measures may improve diagnostic accuracy as well as elucidate the importance of child relative positioning and approach in social interactions.

Natural language processing has also been studied to model ASD symptoms. Luo et al. [18] used standard natural language processing methods to digitize and visualize these descriptions. The complex patterns of these descriptive sentences exhibited a difference in semantic space between individuals with ASD and control participants. Studies from neuroscience have highlighted that the corpus callosum and intracranial brain volume holds significant information for detection of ASD [19]. The approach uses multi-source joint analysis scheme for collecting/processing data for ASD classification. Their approach uses mostly linguistic features based on vocalizations and word counts and classifies the tones and content of the conversations to find cues in ASD. Vocalization patterns across interactive and non interactive contexts have shown to indicate a strong predictive correlation for early communication development [20].

Spectrograms have long been used to directly model vocal patterns in speech of infants and children. Authors in [21] use spectograms to model infant crying patterns based on differences in the distribution of energy in the spectrograms. Azizi [22] demonstrated several abnormal features in the spectrographic analysis of children with ASD. Visual assessment investigates the intonation pattern, mean of pitch, amplitude, duration, intensity, and tilt in the ASD. A skill such as prosody that unobtrusively conveys emotional and pragmatic aspects of speech may be particularly vulnerable in children with autism, but those who do not have learning difficulties may be capable of increasing their prosodic awareness and ability [22].

Recently, usage of deep learning based computer vision methods towards objective measurement of of child facial expressions have seen significant increase in interest [23]. These measurements are used to better understand children's emotional reactivity and early interaction [24], and detected patterns of smile [25] and head motion atypicalities [26] in children with ASD. Children with ASD show atypical patterns of attention to internal features of the face, particularly the eyes. In such conditions, computer vision methods can be used to capture both gazes and smiles at adults from children in first-person-video [27]. An issue that can arise with deep learning methods is the semantic gap between the low-level deep features and their high-level semantic meaning [28]. There has been considerable effort towards bridging the connection between deep features and the conceptions formed by representation systems [29]. Other notable efforts include analyzing brain imaging activate patterns to investigate patterns of functional connectivity [30]. Rad et al. [31] studied atypical postural or motor behaviors in social interactions using deep learning methods. They used deep learning to learn the discriminating features from multi-sensor accelerometer signals.

III. The Framework

Movement toward objective quantification of ASD symptoms has the potential to increase the reach of screening, decrease time to diagnosis, and reduce ethnic and sex-based disparities [32]. These objective measurements can produce quantitative indices of ASD symptoms that could inform clinical categorization and referral [33].

In this paper we propose a pipeline framework that mines for meaningful objective measurements of social communication and language across contexts and how they are associated with autism severity indices and language capabilities. The overall framework diagram in Figure 4 shows the complete process from raw audio to ADOS-2 CSS score predictions. The objective is to separate speakers and utilize the vocalizations and vocal duration as predictive objective measures. We start by transforming the captured audio into Mel Frequency Cepstral Coefficients (MFCC) and train a CNN using manually coded samples, to ignore the non-speech segments. The Long Short-Term Memory (LSTM) [3] networks identify speakers and speaker changes on a frame-by-frame level [34]. The speaker activity and change are used to generate diarization metrics to incorporate the number and duration of vocalization regimes, e.g., child speaking versus child yelling. This data is used to train a Synthetic Random Forest with interactions and 1000 replications to predict the Social Affect scores. The Synthetic Random Forest prediction model estimates ADOS-2 Calibrated Severity Scores (CSS) with R^2 of 0.402.

Our methods use hand annotated samples of from 33 ADOS-2 examination interviews. Around 2000, 2-second clips were hand annotated as ”Adult”, ”Child”, ”Both”, and ”Irrelevant”. These annotations were used to generate spectrograms and train a CNN to classify out the irrelevant audio. A batch size of 32 spectrograms were selected. An adaptive learning rate schedule was chosen that increases by 5% when validation error rate decreases by 1%. Furthermore, the learning rate decreases by 20% when the validation error rate decreases by 0.5%. We halt the training process when the validation error rate drops below 0.1%. Since we are dealing with only three closely related classes, regularization and dropout were not used, as they reduce the classification accuracy. We repeat the train/test process several times with different weight initialization. The best classification accuracy for removing irrelevant artifact frames was 84.7% with a feed-forward network, an initialized learning rate of 0.001, and a momentum coefficient of 0.75. The network had 0.65M parameters. Details of computing MFCC, LSTM Noisemes event detection and speaker diarization are provided in the following sections.

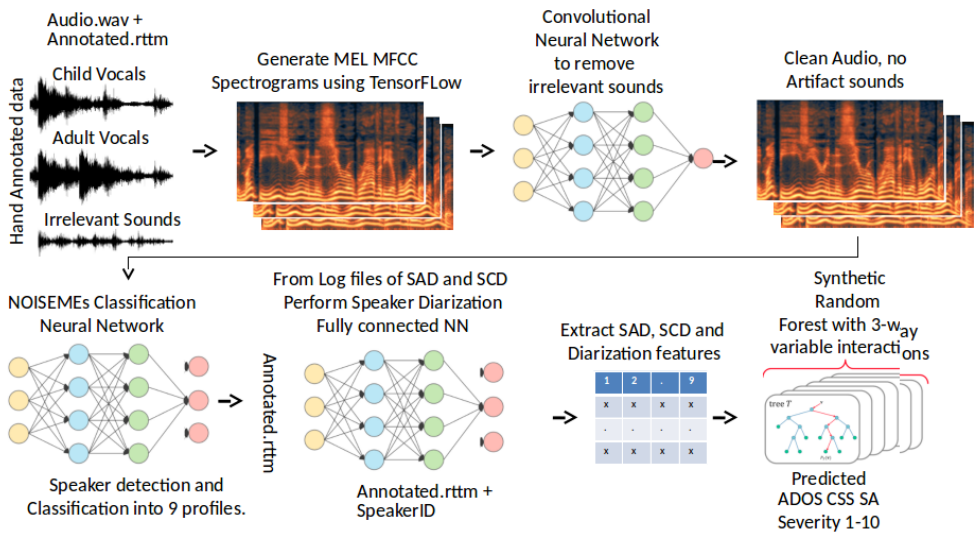


Figure 4. The overall workflow of the proposed framework.

A. Mel-Frequency Cepstral Coefficients (MFCCs)

The audio signal is fed to the pipeline in the form of frequency spectrum representation called Mel-Frequency Cepstral Coefficients spectrograms (MFCC). Spectral representations of audio signal are widely used in speech and speaker recognition because of their low-dimensional and perceptually-relevant representation of the signal. The most common method to calculate spectrograms is using the Short-time Fourier Transform (STFT) to compute the sinusoidal frequency and phase content of the time varying signal. First the energy spectrogram is computed by taking the magnitude of the complex-valued STFT $\sqrt{a^2+b^2}$; where a and b are the real and complex components of the STFT. GPU implementation of TensorFlow Discrete Fourier Transform (DFT) operations are used without any specific kernel optimizations. In the TensorFlow DFT, usually there are 513 unique bins or frequency banks to compute the energy spectrogram. The Mel-spectrograms are calculated from the energy spectrograms using TensorFlow as well. Given a vector x of n input amplitudes such as $x[0], x[1], x[2], ..., x[N-1]$. The Discrete Fourier Transform yields a set of n frequency magnitudes defined as:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi kn}{N}} \tag{1}$$

Here, k is used to denote the frequency domain ordinal, n is used to represent the time-domain ordinal and N is the length of the sequence to be transformed. The signal must be restricted to be of a size of power of 2 i.e. N has to be power of 2 or must be zero-padded otherwise. Real sine waves can be expressed as the sum of complex sine waves using Euler's identity

$$x[n] = Acos\left(2\pi \frac{kn}{N}\right) = \frac{A}{2} e^{j2\pi kn/N} + \frac{A}{2} e^{-j2\pi kn/N} \tag{2}$$

Since the DFT is a linear function, the DFT of sum of sine waves is the sum of the DFT of each sine wave. So for the spectral case, you get 2 DFTs, one for the positive frequencies and one for the negative frequencies, which are symmetric. The Mel-spectrogram magnitudes are further compressed by applying a non-linear logarithmic compression. This helps to balance the data in low and high energy regions of the spectrum that more aptly represents the human auditory behavior.

B. Noisemes Event Detection

The proposed framework uses deep learning based event detection using the Noisemes audio classifier [35]. Noisemes is a neural network that detects speaker activity by evaluating frame-level probabilities of 17 types of sound events (called "noisemes"), including speech, singing, crying, etc. It uses a pre-trained bidirectional Long Short-Term Memory (LSTM) network with 400 hidden units in each direction. The method extracts acoustic features using the OpenSMILE toolkit [36]. Overall 6,669 low-level acoustic features such as MFCC and fundamental frequencies are extracted and then reduced to 50 dimensions using PCA. The output of this tools is the time labels with 'speech' or 'non-speech' tags used to calculate the speaker activity detection (SAD) and speaker change detection (SCD). To attribute each occurrence of speech to a specific speaker, the output of Noisemes is fed into a diarization tool called DiarTK [34].

C. DiarTK Diarization

DiarTK is an open source non-parametric clustering and re-alignment method from the DiViMe library [34]. It contains a set of algorithms which were designed to automatically detect and label speaker turns in naturalistic audio recordings. DiarkTK works on the principle of agglomerative information bottleneck clustering [37] which is a bottom up clustering approach based on conditional distribution of data.

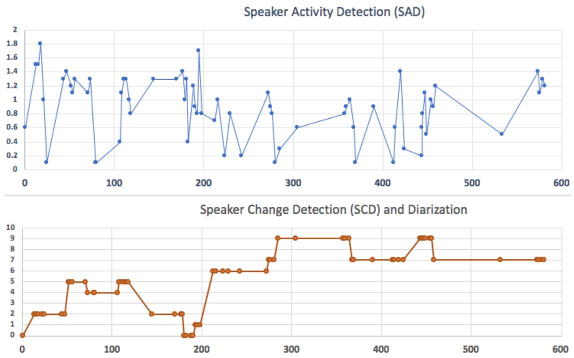


Figure 5. A reference sample case study shows the SAD, SCD and Diarization scores with timing.

The diarization step helps us learn vocal turn taking patterns in child and adult speech. The DiarTK model is imported in the virtual machine as a C++ open source toolkit. At the end of the process, the resulting clusters correspond to identified speakers with their vocal occurrences and durations of their speech. The estimated final diarization output is shown in Figure 5. We extract this speaker information and use in a Synthetic Random Forest to train the final prediction model that estimates ADOS-2 Calibrated Severity Scores (CSS).

D. Synthetic Random Forest

Random forests were chosen as the final classification stage of the framework because of their ability to handle very high dimensional spaces. The metrics from Noisemes and DiarTK are converted into a table of features containing vocalization frequency and duration. We use this data to train a synthetic random forest with three-way interactions and 1000 replications to predict the ADOS-2 CSS score. The random forest is implemented as an aggregation of *ntree* number of trees, usually in thousands, and each tree is grown by bootstrapping a randomly sampled vector *mtry* from the complete dataset. Each tree in the random forest collection is grown non-deterministically with a two-stage method i.e. bootstrapping and random variable selection resulting in substantially de-correlated trees. Each tree is grown to contain *nodesize* samples in the terminal node.

The forest is built by growing the trees based on a random vector θ_k such that the tree predictor $h(\mathbf{x}, \theta_k)$ represents a predicted probability specified by the class, ranging from 0 to 1. Thus, the vector θ_k contains the predicted probabilities of the outcome variable Y. The final predictions are defined as the unweighted average over the collection of predictor trees as shown in Equation (3), where $h(\mathbf{x}; \theta_k), k = 1, ..., ntree$ are the collection of the tree predictors and $\mathbf{x}$ represents the observed input variable vector of length *mtry* with the associated i.i.d random vector θ_k.

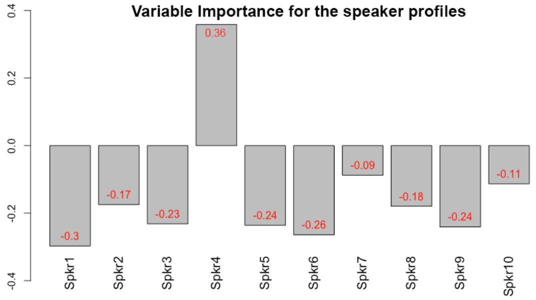


Figure 6. Variable Importance Metric illustrating the reciprocity of correlation between vocalization and SA.

$$\overline{h}(\mathbf{x}) = (1/ntree) \sum_{k=1}^{ntree} h(\mathbf{x}; \theta_k). \tag{3}$$

As $k \to \infty$, the Law of Large Numbers ensures

$$E_{\mathbf{X},Y}(Y - \overline{h}(\mathbf{X}))^2 \to E_{\mathbf{X},Y}(Y - E_\theta(\mathbf{X}; \theta))^2, \tag{4}$$

where θ represents the predicted probabilities of the outcome variable averaged over *ntree* trees. The convergence in Equation (4) implies that the random forests do not overfit.

Synthetic features are calculated using out-of-bag (OOB) data to avoid overfitting. To guarantee that error rates and variable importance are regularized, same sized bootstrap draws are performed on all trees in the construction of the synthetic forest. The variable importance metric of the vocal turn-taking is shown in Figure 6.

IV. Experiment and Results

Many of the symptoms associated with ASD affect a person's social communication, usually with verbal and vocal cues. Easily recognizable vocal symptoms of ASD includes difficulty in conversation, talking at length, and having an unusual tone of voice. In our study, vocal and verbal cues from 33 children having ASD were collected and analyzed using the pipeline proposed in section III.

The study was structured to perform ADOS-2 evaluations by a clinical diagnostician in a 12 ft x 15 ft x 10 ft room equipped with a ceiling mounted microphone (Figure 2a). In addition, the examiner and parent each wore a Pivothead 1080 HD 8MP point of view (PoV) video-recording camera in the form of eyeglasses. The examiners presented the children with a series of toys and activities and created opportunities to engage in social interactions. The information about each child's social behaviors were recorded (e.g., eye contact, language used, gestures, and social reciprocity) as well as restricted and repetitive behaviors and interests (e.g., physical repetitive behaviors such as hand flapping, playing with toys in specific ways, repetitive routines, and insistence on sameness). The children were rated on a variety of symptoms in each of these domains and receive a calibrated severity score (CSS) ranging from 1 indicating little to no symptoms of ASD to 10 i.e. high level of symptoms.

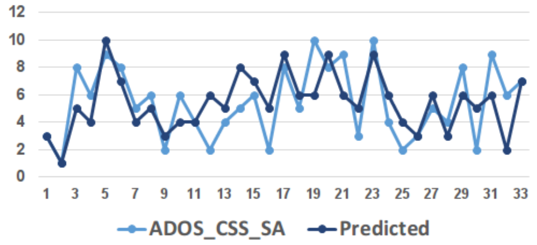


Figure 7. Prediction of ADOS-2 social affect symptom severity based on vocal turn-taking produced by deep learning models indicates a high correlation with R^2=0.402.

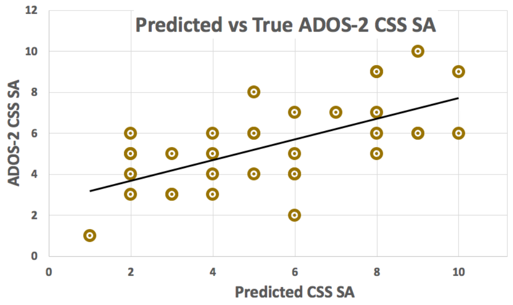


Figure 8. Linear model fitted on the predicted vs True values indicates an R^2 of .402 and a p-value of 0.6×10^{-5}.

In this paper, we expanded the objective measurements of social communication by using raw audio data from the ADOS-2 sessions and perform automatic diagnosis of ASD severity. Audio signal of the assessments were processed using the proposed pipeline providing automated vocalization detection and yielding child and adult vocal initiation counts and turn counts within conversational vocal blocks. The 10 detected profiles are 'Background', 'Speech', 'Speech Non-English', 'Mumble', 'Singing', 'Singing with Music', 'Music', 'Human non-speech', 'Cheer', and 'Crowd'.

The framework extracts 20-dimensional Mel-frequency Cepstral Coefficients (MFCC) with Cepstral Mean Normalization (CMN). Since we only utilize MFCC features, frames containing brief bursts of vocalizations such as "shouts", "blips" are classified by considering the surrounding signal patterns in the frame. The Noisemes classifier produces top-5 hypotheses for classification and considers a frame for speech only if the corresponding hypotheses appear in the target 17 Noiseme classes. This allows for running the entire examination audio without much pre-processing. The Noisemes class labels are fed to the diarization system for profile detection. Few of the diarized classes were often confused due to the close proximity of these utterances. Vocalizations such as "mumble" was confused as "speech" and "singing", "speech non-English" was confused with "crowd" etc. However some of the confusions were unexpected and difficult to explain such as "Cheer" and "Human non-speech" was sometimes confused with "background" etc. "Background" accounts for sounds such as children playing with toys, door closing and sounds from other physical objects.

The classified profiles were tabulated for their vocal occurrence counts and duration of speech and subsequently fed to the Synthetic Random Forest. The final estimated ADOS-2 SA scores versus the ground truth are shown in Figure 7. The proposed framework achieves highly accurate predictions with a high coefficient of determination with the ground truth. Figure 8 shows a strong linear relationship between the predicted and true CSS SA class labels.

A. *Comparison with Other Methods*

As a comparative baseline, we contrast the proposed framework with other common machine learning methods. The compared methods were provided with the same objective measures i.e. vocal turn-taking and vocalization durations to predict the ADOS-2 SA scores. It was observed that even when identical objective measures were used, the modeled fit lacked objectivity and produced imprecise results. The predicted $\hat{SA}$ scores of each model were compared to true SA scores using a linear regression fit in a leave-one-out cross validation fashion. Table I presents the fit summary for each of the compared estimators. The general trend is that the prediction accuracy increases with more complicated estimators. It was observed that Nearest Neighbor and Naive Bayes did not perform much better than plain linear regression. Introducing Random Forests produced only a slight improvement over simple linear regression whereas using the synthetic random forest, in the proposed pipeline, with 3-way interactions and parameter tuning performs much better. Support vector machines (SVM) using the linear kernel achieved a test frame R^2 of 0.21 which was among the best but nearly half of the proposed framework.

An indication of significant association between $\hat{SA}$ and SA is the p-value. The p-value for each model tests the null hypothesis and a low p-value (< 0.05) indicates that the two variables are significantly related. While some methods were close to α the p-value was significant in only SVM and the proposed method. F-statistic is another predictor of correlation between the predicted and true labels with values further from 1 indicate stronger relationship. In the proposed framework, F-statistic was 20.84 which is relatively larger than 1 given the size of our data. AIC and BIC are both penalized-likelihood criterion used to compare non-nested models, which ordinary statistical tests cannot do. A lower value of AIC and BIC indicates closeness to the true model.

Table I
Performance Comparison with other common machine learning regression methods.

Methods	R-sq	F-statistic	P-value	Log-Likelihood	AIC	BIC	Conf. Int. Lo	Conf. Int. Hi
Linear Regression	0.061	1.993	0.161	-77.262	158.5	161.5	-0.011	0.621
Nearest Neighbor	0.087	2.960	0.095	-76.786	157.6	160.6	-0.058	0.672
Naive Bayes	0.111	3.90	0.057	-72.221	156.7	159.7	-0.011	0.721
Random Forest	0.162	9.7	0.062	-74.239	152.5	155.5	0.141	0.766
Support Vector Machines	0.215	8.518	0.006	-74.285	152.6	155.6	0.169	0.955
Proposed Pipeline	0.402	20.84	0.6e-5	-69.80	143.6	146.6	0.442	1.154

Since the sample size is relatively small, the proposed pipeline achieves only slightly lower values of AIC and BIC but still lower than the rest.

V. Conclusion

Autism is increasingly prevalent in the United States with increased concentration in ethnic minority and low-resource populations. Early diagnosis of ASD is the key to control, treat, and mitigate the symptoms. However, due to myriad of reasons children are not diagnosed until 4 years of age. In this paper, an automatic machine learning based approach is proposed to detect audio regimes that directly estimate ASD Severity Social Affect scores. This approach can help improve diagnostic accuracy and mitigate ethnic and economic disparaites in ASD diagnosis. The speaker activity and speaker change are used on a frame-by-frame level to diarize the vocal turn-taking in the audio signal. These features are used to train a synthetic random forest and predict the outcome scores. The proposed model achieves state-of-the-art performance in predicting ADOS-2 symptoms. Comparative analyses with other machine learning methods indicate that the proposed method not only achieves better R^2 but also sharper confidence intervals. Future research directions include a unified recording device and software toolkit for automatic speech processing developed by the LENA Foundation.

Acknowledgment

The research reported in this paper was supported by grants from the National Science Foundation (1620294) and the National Center for Special Education Research, Institute of Education Sciences Award (R324A180203).

References

[1] E. W. Ngai, L. Xiu, and D. C. Chau, “Application of data mining techniques in customer relationship management: A literature review and classification,” *Expert systems with applications*, vol. 36, no. 2, pp. 2592–2602, 2009.

[2] D. S. Murray, J. S. Anixt, D. L. Coury, K. A. Kuhlthau, J. Seide, A. Kelly, A. Fedele, D. Eskra, and C. Lannon, “Transforming an autism pediatric research network into a learning health system: Lessons learned,” *Pediatric Quality & Safety*, vol. 4, no. 2, p. e152, 2019.

[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, p. 436, 2015.

[4] N. Takahashi, M. Gygli, and L. Van Gool, “Aenet: Learning deep audio features for video analysis,” *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 513–524, 2017.

[5] R. C. Sheldrick, M. P. Maye, and A. S. Carter, “Age at first identification of autism spectrum disorder: an analysis of two us surveys,” *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 56, no. 4, pp. 313–320, 2017.

[6] N. Ryanta, E. Bergelson, K. Church, A. Cristia, J. Du, S. Ganapathy, S. Khudanpur, D. Kowalski, M. Krishnamoorthy, R. Kulshreshta *et al.*, “Enhancement and analysis of conversational speech: Jsalt 2017,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2018, pp. 5154–5158.

[7] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe *et al.*, “Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge,” in *Proc. Interspeech*, 2018, pp. 2808–2812.

[8] I. Chin, M. S. Goodwin, S. Vosoughi, D. Roy, and L. R. Naigles, “Dense home-based recordings reveal typical and atypical development of tense/aspect in a child with delayed language development,” *Journal of child language*, vol. 45, no. 1, pp. 1–34, 2018.

[9] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, “The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism,” *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[10] S. M. Myers, C. P. Johnson *et al.*, “Management of children with autism spectrum disorders,” *Pediatrics*, vol. 120, no. 5, pp. 1162–1182, 2007.

[11] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®).* American Psychiatric Pub, 2013.

[12] I.-M. Eigsti, A. B. de Marchena, J. M. Schuh, and E. Kelley, “Language acquisition in autism spectrum disorders: A developmental review,” *Research in Autism Spectrum Disorders*, vol. 5, no. 2, pp. 681–691, 2011.

[13] I. L. Cohen, J. M. Gardner, B. Z. Karmel, and S.-Y. Kim, “Rating scale measures are associated with noldus ethovision-xt video tracking of behaviors of children on the autism spectrum,” *Molecular autism*, vol. 5, no. 1, p. 15, 2014.

[14] L. K. Perry, E. B. Prince, A. M. Valtierra, C. Rivero-Fernandez, M. A. Ullery, L. F. Katz, B. Laursen, and D. S. Messinger, "A year in words: The dynamics and consequences of language experiences in an intervention classroom," *PloS one*, vol. 13, no. 7, p. e0199893, 2018.

[15] G. Esposito, M. del Carmen Rostagno, P. Venuti, J. D. Haltigan, and D. S. Messinger, "Brief report: Atypical expression of distress during the separation phase of the strange situation procedure in infant siblings at high risk for asd," *Journal of Autism and Developmental Disorders*, vol. 44, no. 4, pp. 975–980, 2014.

[16] S. Tek, L. Mesite, D. Fein, and L. Naigles, "Longitudinal analyses of expressive language development reveal two distinct language profiles among young children with autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 44, no. 1, pp. 75–89, 2014.

[17] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3414–3421.

[18] S. X. Luo, J. A. Shinall, B. S. Peterson, and A. J. Gerber, "Semantic mapping reveals distinct patterns in descriptions of social relations in adults with autism spectrum disorder," *Autism Research*, vol. 9, no. 8, pp. 846–853, 2016.

[19] H. Sharif and R. A. Khan, "A novel framework for automatic detection of autism: A study on corpus callosum and intracranial brain volume," *arXiv preprint arXiv:1903.11323*, 2019.

[20] B. Franklin, A. S. Warlaumont, D. Messinger, E. Bene, S. Nathani Iyer, C.-C. Lee, B. Lambert, and D. K. Oller, "Effects of parental interaction on infant vocalization rate, variability and vocal type," *Language Learning and Development*, vol. 10, no. 3, pp. 279–296, 2014.

[21] A. Chittora and H. A. Patil, "Spectral analysis of infant cries and adult speech," *International Journal of Speech Technology*, vol. 19, no. 4, pp. 841–856, 2016.

[22] Z. Azizi, "The acoustic survey of intonation in autism spectrum disorder," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2207–2207, 2015.

[23] W.-S. Chu, F. De la Torre, J. F. Cohn, and D. S. Messinger, "A branch-and-bound framework for unsupervised common event discovery," *International journal of computer vision*, vol. 123, no. 3, pp. 372–391, 2017.

[24] W. I. Mattson, J. F. Cohn, M. H. Mahoor, D. N. Gangi, and D. S. Messinger, "Darwin's duchenne: Eye constriction during infant joy and distress," *PloS one*, vol. 8, no. 11, p. e80161, 2013.

[25] B. L. Lambert-Brown, N. M. McDonald, W. I. Mattson, K. B. Martin, L. V. Ibañez, W. L. Stone, and D. S. Messinger, "Positive emotional engagement and autism risk." *Developmental psychology*, vol. 51, no. 6, p. 848, 2015.

[26] K. B. Martin, Z. Hammal, G. Ren, J. F. Cohn, J. Cassell, M. Ogihara, J. C. Britton, A. Gutierrez, and D. S. Messinger, "Objective measurement of head movement differences in children with and without autism spectrum disorder," *Molecular autism*, vol. 9, no. 1, p. 14, 2018.

[27] S. R. Edmunds, A. Rozga, I. E, E. Karp, J. Rehg, W. L. Stone, and Y. Li, "A novel method for quantifying eye-to-eye gaze during naturalistic social interactions finds preliminary differences between asd and td toddlers," *Interational Meeting for Austism Research.*, 2016.

[28] S. Sadiq and M.-L. Shyu, "Cascaded propensity matched fraud miner: Detecting anomalies in medicare big data," *Journal of Innovative Technology*, vol. 1, no. 1, pp. 51–61, 2019.

[29] S. Sadiq, M.-L. Shyu, and D. J. Feaster, "Counterfactual autoencoder for unsupervised semantic learning," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 9, no. 4, pp. 1–20, 2018.

[30] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018.

[31] N. M. Rad and C. Furlanello, "Applying deep learning to stereotypical motor movement detection in autism spectrum disorders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 1235–1242.

[32] P. Howlin, I. Magiati, and T. Charman, "Systematic review of early intensive behavioral interventions for children with autism," *American journal on intellectual and developmental disabilities*, vol. 114, no. 1, pp. 23–41, 2009.

[33] E. Courchesne, T. Pramparo, V. H. Gazestani, M. V. Lombardo, K. Pierce, and N. E. Lewis, "The asd living biology: from cell proliferation to clinical phenotype," *Molecular psychiatry*, p. 1, 2018.

[34] A. Le Franc, E. Riebling, J. Karadayi, Y. Wang, C. Scaff, F. Metze, and A. Cristia, "The aclew divime: An easy-to-use diarization tool," in *Proc. INTERSPEECH*, 2018.

[35] S. Burger, Q. Jin, P. F. Schulam, and F. Metze, "Noisemes: Manual annotation of environmental noise in audio streams," School of Computer Science, Carnegie Mellon University, Tech. Rep., 12 2012. [Online]. Available: https://kilthub.cmu.edu/ndownloader/files/11903183

[36] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[37] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 250–255.