Original Paper

Classifying Autism From Crowdsourced Semistructured Speech Recordings: Machine Learning Model Comparison Study

Nathan A Chi[1]; Peter Washington[2], BA, MSc; Aaron Kline[1], BS; Arman Husic[1], BSc; Cathy Hou[3]; Chloe He[4], BS; Kaitlyn Dunlap[1], BA; Dennis P Wall[1,4,5], PhD

[1]Division of Systems Medicine, Department of Pediatrics, Stanford University, Palo Alto, CA, United States

[2]Department of Bioengineering, Stanford University, Stanford, CA, United States

[3]Department of Computer Science, Stanford University, Stanford, CA, United States

[4]Department of Biomedical Data Science, Stanford University, Stanford, CA, United States

[5]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, United States

Corresponding Author:
Dennis P Wall, PhD
Division of Systems Medicine
Department of Pediatrics
Stanford University
3145 Porter Drive
Palo Alto, CA, 94304
United States
Phone: 1 650 666 7676
Email: dpwall@stanford.edu

Abstract

Background: Autism spectrum disorder (ASD) is a neurodevelopmental disorder that results in altered behavior, social development, and communication patterns. In recent years, autism prevalence has tripled, with 1 in 44 children now affected. Given that traditional diagnosis is a lengthy, labor-intensive process that requires the work of trained physicians, significant attention has been given to developing systems that automatically detect autism. We work toward this goal by analyzing audio data, as prosody abnormalities are a signal of autism, with affected children displaying speech idiosyncrasies such as echolalia, monotonous intonation, atypical pitch, and irregular linguistic stress patterns.

Objective: We aimed to test the ability for machine learning approaches to aid in detection of autism in self-recorded speech audio captured from children with ASD and neurotypical (NT) children in their home environments.

Methods: We considered three methods to detect autism in child speech: (1) random forests trained on extracted audio features (including Mel-frequency cepstral coefficients); (2) convolutional neural networks trained on spectrograms; and (3) fine-tuned wav2vec 2.0—a state-of-the-art transformer-based speech recognition model. We trained our classifiers on our novel data set of cellphone-recorded child speech audio curated from the *Guess What?* mobile game, an app designed to crowdsource videos of children with ASD and NT children in a natural home environment.

Results: The random forest classifier achieved 70% accuracy, the fine-tuned wav2vec 2.0 model achieved 77% accuracy, and the convolutional neural network achieved 79% accuracy when classifying children's audio as either ASD or NT. We used 5-fold cross-validation to evaluate model performance.

Conclusions: Our models were able to predict autism status when trained on a varied selection of home audio clips with inconsistent recording qualities, which may be more representative of real-world conditions. The results demonstrate that machine learning methods offer promise in detecting autism automatically from speech without specialized equipment.

(JMIR Pediatr Parent 2022;5(2):e35406) doi: 10.2196/35406

KEYWORDS

autism; mHealth; machine learning; artificial intelligence; speech; audio; child; digital data; mobile app; diagnosis

Introduction

Autism spectrum disorder (ASD, or autism) encompasses a spectrum of disorders characterized by delayed linguistic development, social interaction deficits, and behavioral impairments [1]. Autism prevalence has rapidly increased in recent years: according to the Centers for Disease Control and Prevention, autism rates have tripled since 2000 to 1 in 44 children in 2018 [2]. In the United States alone, over 5 million individuals are affected [3], and nearly 75 million are affected worldwide. Despite the increasing prevalence of autism, access to diagnostic resources continues to be limited, with 83.86% of all American counties not having any [4]. These nationwide inadequacies in autism resources are compounded by the lengthy nature of diagnosis. On average, the delay from the time of first consultations with health care providers to the time of diagnosis is over 2 years. Such extensive delays often cause diagnosis at a later age (usually ≥4 years old) [5], which may result in greater lifelong impacts, including a higher likelihood of psychotropic medication use, lower IQ scores, and reduced language aptitude [6,7]. Given that timely autism identification and intervention has been shown to improve treatment success and social capabilities, research has focused on its early detection [7-11].

Although symptoms vary across individuals, prosody abnormalities are among the most notable signs of autism, with multiple studies suggesting that affected children display peculiarities including echolalia, monotonous intonation, and atypical pitch and linguistic stress patterns [12-14]. Given this, an effective artificial intelligence sound classifier trained to detect speech abnormalities common in children with autism would be a valuable tool to aid autism diagnostic processes.

Prior research [15,16] investigated prosodic disorders in children with autism to varying degrees of success. Cho et al [17] developed models that achieved 76% accuracy on a dataset of recorded interviews between children and unfamiliar adults, trained on data recorded at a consistent location using a specialized biosensor device with 4 directional microphones. Similarly, Li et al [18] achieved high accuracies when training on speech data recorded with multiple wireless microphones, providing high purity recordings at a central recording location (a hospital). However, both used data collected in centralized, unfamiliar locations with high-quality recording equipment. Such research, while promising, does not accelerate the process of autism detection because it requires the use of specialized equipment and centralized recording locations to provide consistent audio quality, posing significant barriers to the widespread availability of automatic diagnosis tools. Additionally, interacting with unknown adults in foreign environments could be stressful and possibly affect the behavior of children with autism, thus leading to observations that are not generalizable to the real world.

In this work, we propose a machine learning–based approach to predict signs of autism directly from self-recorded semistructured home audio clips recording a child's natural behavior. We use random forests, convolutional neural networks (CNNs), and fine-tuned wav2vec 2.0 models to identify differences in speech between children with autism and neurotypical (NT) controls. One strength of our approach is that our models are trained on mobile device audio recordings of varying audio quality. Therefore, unlike other studies, our approach does not necessitate specialized high-fidelity recording equipment. Additionally, we attempt to capture naturalistic speech patterns by recording children playing educational games with their parents in a low-stress home environment. Finally, our approach does not require a trained clinician to converse with the child. To our knowledge, our method is the first to aurally detect symptoms of autism in an unstructured home environment without the use of specialized audio recording devices.

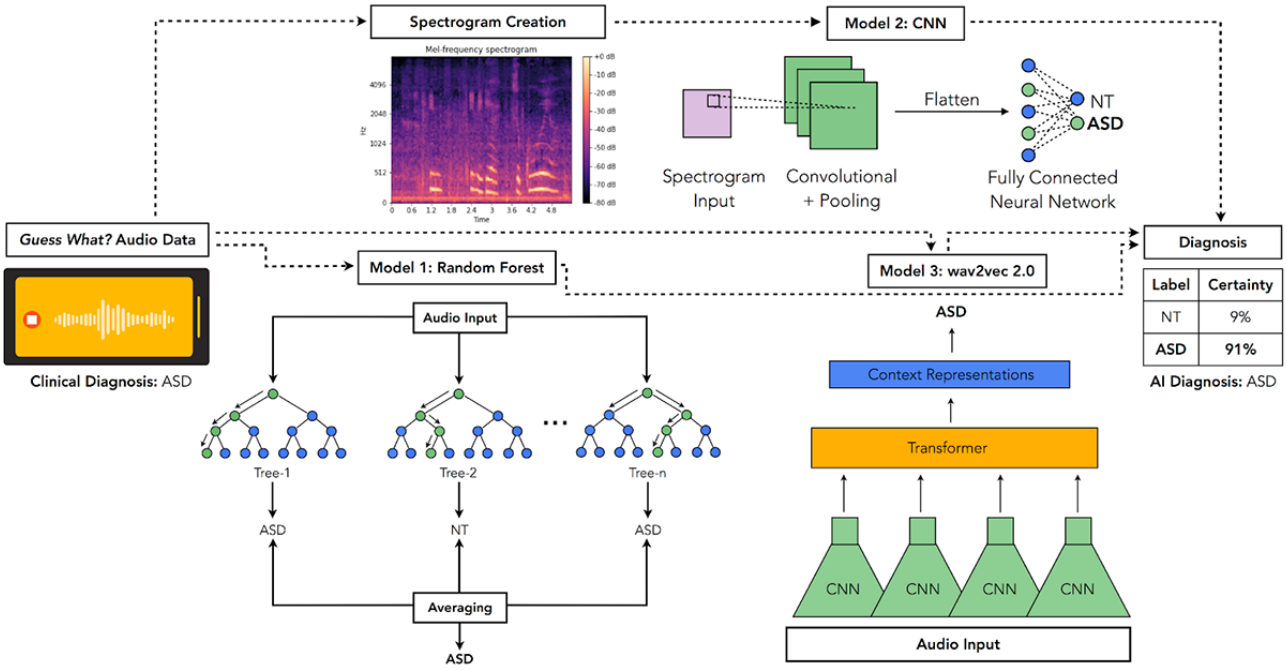
Methods

Data Acquisition

Process

We obtained audio data of NT children and children with autism in a home environment through *Guess What?*, a mobile game designed for prosocial play and interaction at home between 2- to 8-year-old developing children and their parents [19-23] (Figure 1, "*Guess What?* Audio Data"). During a game session, parents and children choose either a charades game (acting out emotions, characters, sports, chores, or objects) or a simple quiz game (identifying colors, shapes, numbers, and word spellings). Children are directed to follow the rules of gameplay, while parents serve as game mediators. Throughout the session, parents record their children by placing their smartphones on their foreheads with the front-facing camera oriented toward the child. After each 90-second session, parents are given the option to view their child's game session video recording and share it with our research team.

Figure 1. Overview of audio-based AI detection pipeline. First, the educational video game *Guess What?* crowdsources the recording of videos of NT children and children with ASD from consenting participants. Audio of children's speech is manually spliced from the videos and 3 models are trained on this audio data. The first is a random forest classifier, which uses an ensemble of independently trained decision trees. The second is a CNN. The third is a fine-tuned wav2vec 2.0 model. Model 1 takes commonly used speech recognition features as input, model 2 learns from spectrograms of the audio, and model 3 takes the raw audio data itself as input. AI: artificial intelligence; ASD: autism spectrum disorder; CNN: convolutional neural network; NT: neurotypical.



Distribution of Demographics

We collected a total of 77 videos of 58 children participating in gameplay, recorded in the span of 4 years from 2018 to 2021. The participants ranged in age from 3-12 years old and included 20 children with ASD (19 male and 1 female) and 38 NT children (15 male, 22 female, and 1 unspecified). The median age of the children with ASD was 5 years; the median age of the NT children was 9.5 years. Parents involved in the study consented to sharing their videos with our research team and provided their child’s age, sex, and diagnosis.

Advantages

This pipeline offers several benefits over traditional diagnostic workflows. Since only a smartphone is necessary, more children can be assessed for autism than through in-lab procedures, with lower costs of time and health care resource use. Through *Guess What?*, a traditionally time-intensive health care process for diagnosis could potentially be transformed into a quick and enjoyable process. Furthermore, children recorded at home may be more likely to behave in a naturalistic manner.

Data Preprocessing

Home videos are naturally variable in quality; their data contains a number of irregularities that must be addressed prior to analysis. In particular, parents or children would often join in gameplay simultaneously, resulting in a variety of voices, sometimes overlapping with one another. This overlap of voices can complicate the isolation and extraction of the child’s voice. In order to remove adult speech, we manually sampled only child speech from each video, ensuring that each resulting clip did not include any voice other than the child’s. Each child contributed a mean of 1.32 videos and 14.7 clips, resulting in a total data set size of 850 audio clips, representing 425 ASD and 425 NT clips. The 850 clips were split into 5 folds, as shown in Table 1, in preparation for 5-fold cross-validation. When creating the folds, we included the restriction that all clips spliced from a given child’s video had to be included in the same fold to prevent models from learning from child-specific recording idiosyncrasies, including environmental background noise and audio quality.

Table 1. Distribution of 850 audio clips across 5 folds. Each of the 3 models was trained on the same distribution of clips with 5-fold cross-validation.

Group	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Neurotypical	87	87	81	83	87
Autism spectrum disorder	87	87	81	83	87

Classifiers

We investigated 3 machine learning methods to predict autism from audio, each represented in Figure 1.

Random Forest

We trained random forests on a set of audio features (Mel-frequency cepstral coefficients, chroma features, root mean square, spectral centroids, spectral bandwidths, spectral rolloff, and zero-crossing rates) typically used in traditional signal processing speech recognition. We also tried training other models (including logistic regression, Gaussian Naive Bayes, and AdaBoosting models), which did not perform as well. We implemented the random forest model in scikit-learn and used the following manually chosen hyperparameters: $\max_{\text{depth}}$=20,000, $n_{\text{estimators}}$=56, $\max_{\text{features}}$=15, $\min_{\text{samples split}}$=10, $\min_{\text{samples leaf}}$=20, $\min_{\text{weight fraction leaf}}$=0.1.

CNN Model

We trained a CNN using spectrograms of our data as input [24,25]. Our spectrograms were synthesized via the Python package Librosa. Figure 2 shows an example of the spectrograms used to train the CNN. The CNN, represented in Figure 3, consists of 9 layers each with alternating convolution and max pooling layers, as well as 3 dense layers with a L2 regularization penalty of 0.01. We investigated both training a small CNN (~8 million parameters) from scratch and fine-tuning the image recognition model Inception v3 (with ~33 million parameters) trained on ImageNet [26]. However, our CNN model with 8 million parameters ultimately performed slightly better than the transfer learning approach, likely due to the irrelevance of ImageNet features to spectrograms. Our final CNN model, which we train for 15 epochs (until training performance stopped improving), has 8,724,594 parameters.

Figure 2. Mel-frequency spectrogram for a neurotypical child speech segment, spliced from a *Guess What?* gameplay video. This spectrogram was one of 850 used to train the convolutional neural network model with 8 million parameters, which yielded the highest accuracy of the 3 best-performing models.

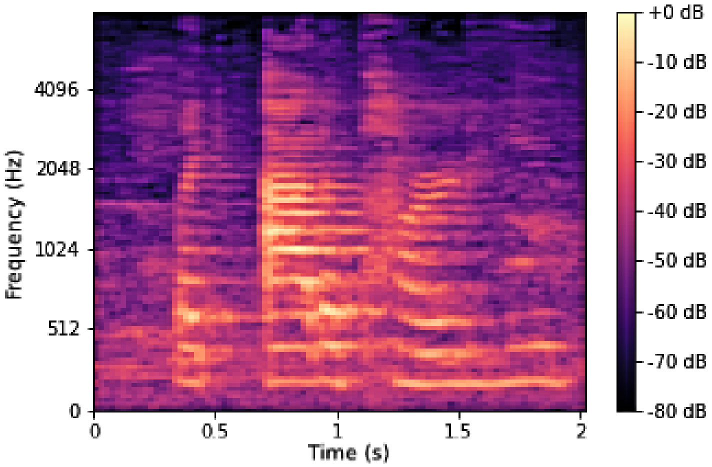
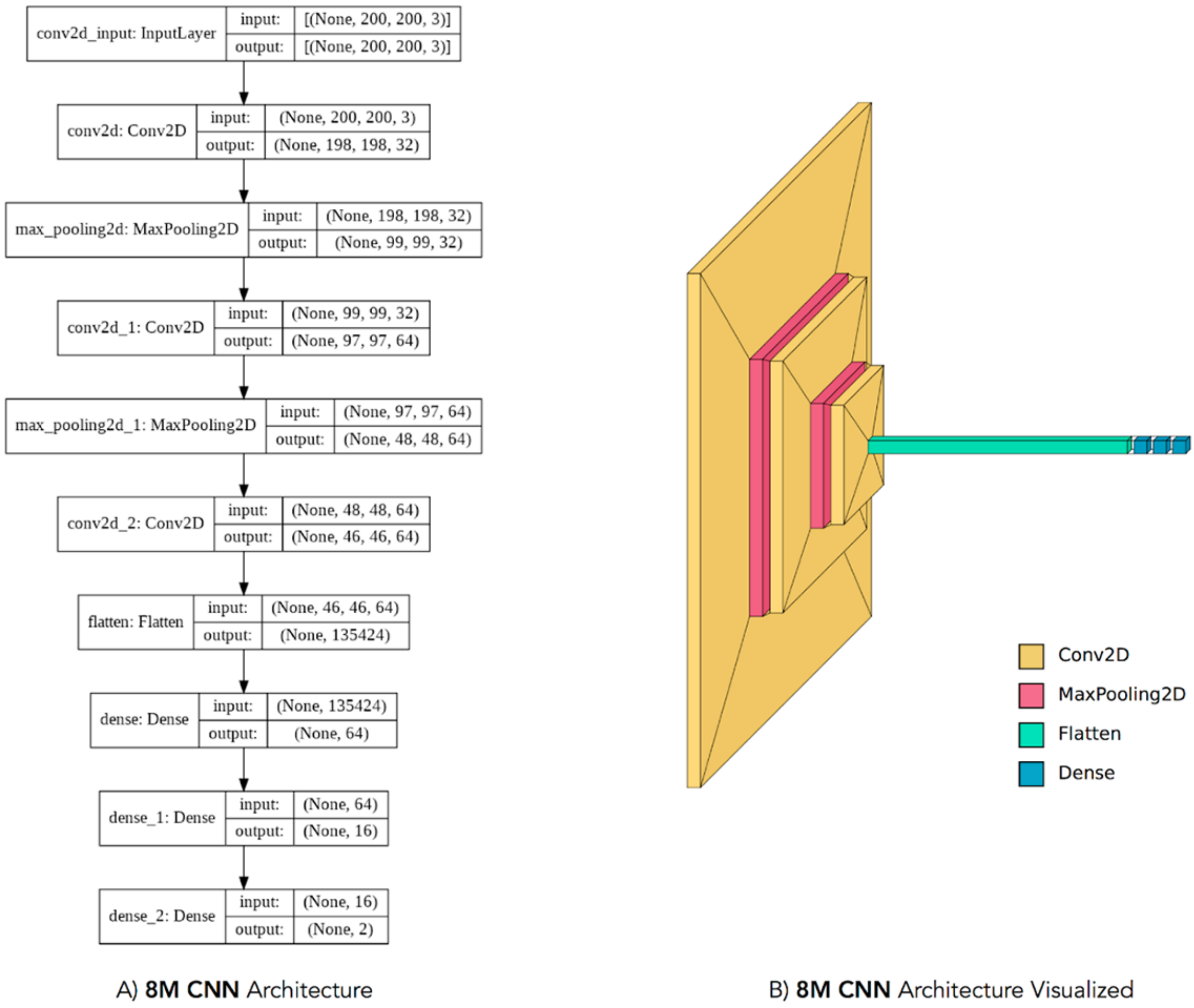


Figure 3. (A) and (B) represent the same 8M CNN model architecture. This architecture performed best out of all of our tested architectures, including a fine-tuned Inception v3 model. (B) was in part created with the Python package Visualkeras. 8M CNN: convolutional neural network with 8 million parameters.


conv2d_input: InputLayer | input: [(None, 200, 200, 3)] | output: [(None, 200, 200, 3)]
conv2d: Conv2D | input: (None, 200, 200, 3) | output: (None, 198, 198, 32)
max_pooling2d: MaxPooling2D | input: (None, 198, 198, 32) | output: (None, 99, 99, 32)
conv2d_1: Conv2D | input: (None, 99, 99, 32) | output: (None, 97, 97, 64)
max_pooling2d_1: MaxPooling2D | input: (None, 97, 97, 64) | output: (None, 48, 48, 64)
conv2d_2: Conv2D | input: (None, 48, 48, 64) | output: (None, 46, 46, 64)
flatten: Flatten | input: (None, 46, 46, 64) | output: (None, 135424)
dense: Dense | input: (None, 135424) | output: (None, 64)
dense_1: Dense | input: (None, 64) | output: (None, 16)
dense_2: Dense | input: (None, 16) | output: (None, 2)
Conv2D
MaxPooling2D
Flatten
Dense

A) **8M CNN** Architecture

B) **8M CNN** Architecture Visualized

Wav2vec 2.0

We fine-tuned wav2vec 2.0, a state-of-the-art transformer model pretrained on a self-supervised audio denoising task [27]. Although wav2vec 2.0 is typically used for speech-to-text decoding, prior work [28] has demonstrated its utility for suprasegmental tasks such as emotion prediction. We used the facebook/wav2vec2-base variant and fine-tuned for 264 steps. The final model has 95 million parameters.

Summary

For each method, we trained and evaluated using 5-fold cross-validation. We ensured that clips from a child are maintained in one fold to prevent the model from artificially performing better by learning user recording idiosyncrasies (eg, background noise). For each fold, we saved the weights for the highest performing model after training and reported mean accuracy (with threshold 0.5), precision, recall, F_1 score, and area under the receiver operating characteristic curve (AUROC), averaged over the 5 folds.

Results

Of our models, the best-performing model was the CNN model with 8 million parameters, achieving 79.3% accuracy, 80.4% precision, 79.3% recall, 79.0% F_1 score, and a mean AUROC score of 0.822 (Table 2). Our wav2vec 2.0 model performed comparably with our best CNN, achieving 76.9% accuracy, 78.2% precision, 74.6% recall, and 76.8% F_1 score, and a mean AUROC score of 0.815. On the other hand, our highest performing lightweight machine learning model (random forest) performed somewhat worse than the other 2 models, with 69.7% accuracy, 68.7% precision, 74.4% recall, 69.4% F_1 score, and a mean AUROC score of 0.740.

Our receiver operating characteristic (ROC) curves for the top 3 highest performing models of each category are included in Figure 4, panels A, C, and E. In each figure, ROC curves for each individual fold and the mean curve are reported. One point of interest is that each figure has variation in area under the curve (AUC) values between folds to some degree. Moreover, these variation trends are similar between models: for instance, each model appears to perform well on fold 2 while performing relatively poorly on fold 3. This suggests that the data in each fold may be too limited, resulting in folds that have differences in content that cause varying model performance from fold to fold. This disparity between AUC values is the greatest in Figure 4A, perhaps explainable by the random forest classifier's small size and lightweight traits. The wav2vec model (Figure 4E) has the most unvarying results, implying that it is better at consistently performing well at classifying unseen data than either of the other two models. This is expected, given that the wav2vec model contains far more parameters than either of the other two models and is more robust.

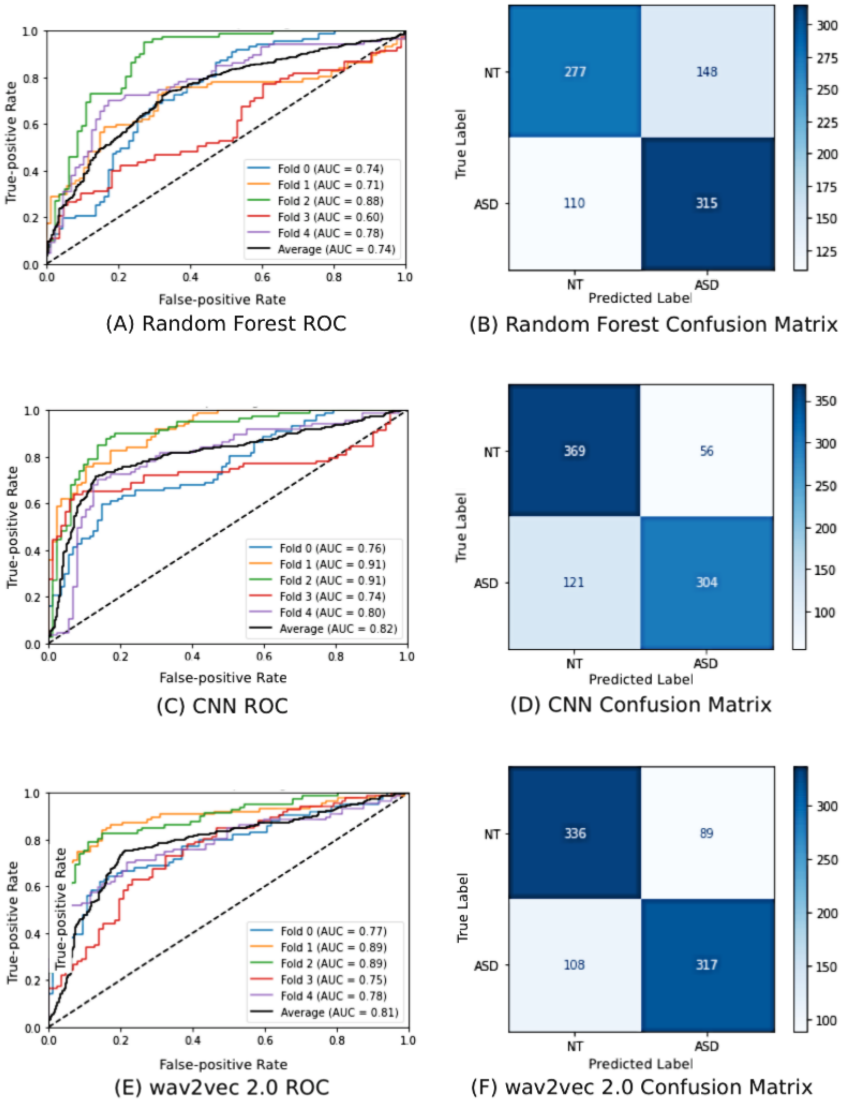
In Figure 4, panels B, D, and F, we provide confusion matrices for all 3 highest performing models. Figure 4D and Figure 4F show that both the CNN and wav2vec models have relatively few false positive predictions, while Figure 4B shows that the random forest classifier has a relatively large number of false positive predictions. All models have similar false negative prediction rates.

Table 2. Performances on *Guess What?* data set. Results are reported with standard deviation over 5 different runs for each model.

Model	Accuracy, mean (SD)	Precision, mean (SD)	Recall, mean (SD)	F_1 score, mean (SD)	AUROC[a], mean (SD)
Random forest	0.697 (0.013)	0.687 (0.010)	0.744 (0.247)	0.694 (0.013)	0.740 (0.09)
Convolutional neural network	0.793 (0.013)	0.804 (0.014)	0.793 (0.014)	0.790 (0.014)	0.822 (0.010)
Wav2vec 2.0	0.769 (0.005)	0.782 (0.021)	0.746 (0.031)	0.768 (0.006)	0.815 (0.077)

[a] AUROC: area under the receiver operating characteristic curve.

Figure 4. (A) ROC curve for random forest model. (B) Confusion matrix for random forest model. (C) ROC curve for 8M CNN. (D) Confusion matrix for CNN. (E) ROC curve for wav2vec 2.0 model. (F) Confusion matrix for wav2vec 2.0 model. All models were tested and trained on the *Guess What?* audio data set, composed of child speech segments taken from educational gameplay videos. 8M CNN: convolutional neural network with 8 million parameters; ASD: autism spectrum disorder; AUC: area under the curve; NT: neurotypical; ROC: receiver operating characteristic.



(A) Random Forest ROC

(B) Random Forest Confusion Matrix

(C) CNN ROC

(D) CNN Confusion Matrix

(E) wav2vec 2.0 ROC

(F) wav2vec 2.0 Confusion Matrix

Discussion

Principal Results

We trained multiple models to detect autism from our novel data set of audio recordings curated from the educational video game *Guess What?* We presented a set of systems that classify audio recordings by autism status and demonstrated that both CNNs and state-of-the-art speech recognition models are capable of attaining robust performance on this task, with lightweight statistical classifiers still achieving reasonable results.

Privacy

One consideration for any recorded audio medical diagnosis is privacy [29-31], which is particularly important for studies involving commonly stigmatized disorders like autism [32]. We note that since our proposed models are relatively lightweight, they could feasibly be deployed at home on mobile devices, allowing for private offline symptom detection as well as privacy-preserving federated learning approaches [33]. Prior work investigated using federated learning techniques to preserve privacy while boosting model performance on a functional magnetic resonance imaging classifier task; a similar framework might be feasible for autism diagnosis, affording a greater degree of privacy for parents who wish for a diagnostic signal but hesitate to share videos with strangers [34].

Limitations

One limitation of our approach is the relative imbalances in the gender distribution of children who comprised our speech data set. Our data set included a split between 95% males with ASD and 5% females with ASD for autistic speech segments, as well as a 39% NT male, 58% NT female, and 3% NT unknown gender split for NT speech segments. Our data set had a sizable imbalance in terms of the relative proportion of males and females with ASD represented. Although some imbalance is to be expected due to the naturally skewed autism sex ratio, our imbalance was larger than the observed real-world 4:1 to 3:1 male-to-female incidence ratio, which would result in a data set containing an 80%-75% male and 20%-25% female split for ASD segments [35,36]. Therefore, despite being closer to replicating actual conditions than prior work, our data set may still not be completely representative of real-world conditions. Additionally, while we require parents to disclose their child's clinical diagnosis by choosing from options not widely known to those who have not received a clinical evaluation, these labels are self-reported and thus unverified.

Another limitation of our work is that we evaluated on a relatively small data set. Additionally, manually splicing videos to isolate child voices is a time-intensive process that may not be scalable to larger data sets. The alternative—automatically isolating voices through blind signal separation—is an exceptionally challenging task [37,38]. However, it poses a potential area of interest and is possibly a necessary hurdle to overcome to develop widely available and consistently effective autism machine learning diagnosis resources.

Future Work

One strength of our approach is the relatively small amount of data required to train the model. Our models were trained on clips spliced from a total of 115.5 minutes of audio yet still yielded relatively accurate results, implying that training on more data may improve performance.

Therefore, future directions include testing our models' performance with additional data from a wider selection of both children with autism and NT children. One particular area of interest may be wearable devices such as Google Glass [39,40]; previous work [41-44] investigated delivering actionable, unobtrusive social cues through wearables. Such approaches have been demonstrated to improve socialization among children with ASD [10,45], suggesting that they could also be used to collect naturalistic data similar to this experiment in an unobtrusive way.

Another area of interest for future work may be examining the possibility of leveraging a distributed workforce of humans for extracting audio-related features to bolster detection accuracy. Previous work examined the use of crowdsourced annotations for autism, indicating that similar approaches could perhaps be applied through audio [31,46-51]. Audio feature extraction combined with other autism classifiers could be used to create an explainable diagnostic system [52-64] fit for mobile devices [60]. Previous work investigated using such classifiers to detect autism or approach autism-related tasks like identifying emotion to improve socialization skills; combining computer vision–based quantification of relevant areas of interest, including hand stimming [58], upper limb movement [63], and eye contact [62,64], could possibly result in interpretable models.

Conclusions

Use of automatic audio classification could help to accelerate and improve the accuracy and objectivity of the lengthy diagnosis process for autism. Our models were able to predict autism status by training on a varied selection of home audio clips with inconsistent recording quality, which may be more representative of real-world conditions. Overall, our work suggests a promising future for at-home detection of ASD.

Acknowledgments

The work was supported in part by funds to DPW from the National Institutes of Health (1R01EB025025-01, 1R01LM013364-01, 1R21HD091500-01, 1R01LM013083), the National Science Foundation (Award 2014232), The Hartwell Foundation, Bill and Melinda Gates Foundation, Coulter Foundation, Lucile Packard Foundation, Auxiliaries Endowment, the ISDB Transform Fund, the Weston Havens Foundation, and program grants from Stanford's Human Centered Artificial Intelligence Program, Precision Health and Integrated Diagnostics Center, Beckman Center, Bio-X Center, Predictives and Diagnostics Accelerator, Spectrum, Spark Program in Translational Research, and MediaX, and from the Wu Tsai Neurosciences Institute's Neuroscience:Translate Program. We also acknowledge generous support from David Orr, Imma Calvo, Bobby Dekesyer, and Peter Sullivan. PW would like to acknowledge support from Mr Schroeder and the Stanford Interdisciplinary Graduate Fellowship (SIGF) as the Schroeder Family Goldman Sachs Graduate Fellow.

Conflicts of Interest

DPW is the founder of Cognoa.com. This company is developing digital health solutions for pediatric health care. All other authors declare no competing interests.

References

1. Hodges H, Fealko C, Soares N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. Transl Pediatr 2020 Mar;9(Suppl 1):S55-S65 [FREE Full text] [doi: 10.21037/tp.2019.09.09] [Medline: 32206584]
2. Maenner MJ, Shaw KA, Bakian AV, Bilder DA, Durkin MS, Esler A, et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. MMWR Surveill Summ 2021 Dec 03;70(11):1-16 [FREE Full text] [doi: 10.15585/mmwr.ss7011a1] [Medline: 34855725]
3. Hyman SL, Levy SE, Myers SM, Council on Children with Disabilities, Section on Developmental and Behavioral Pediatrics. Identification, Evaluation, and Management of Children With Autism Spectrum Disorder. Pediatrics 2020 Jan 16;145(1):e20193447. [doi: 10.1542/peds.2019-3447] [Medline: 31843864]
4. Ning M, Daniels J, Schwartz J, Dunlap K, Washington P, Kalantarian H, et al. Identification and Quantification of Gaps in Access to Autism Resources in the United States: An Infodemiological Study. J Med Internet Res 2019 Jul 10;21(7):e13094 [FREE Full text] [doi: 10.2196/13094] [Medline: 31293243]
5. Zuckerman K, Lindly OJ, Chavez AE. Timeliness of Autism Spectrum Disorder Diagnosis and Use of Services Among U.S. Elementary School-Aged Children. Psychiatr Serv 2017 Jan 01;68(1):33-40 [FREE Full text] [doi: 10.1176/appi.ps.201500549] [Medline: 27476809]
6. Bargiela S, Steward R, Mandy W. The Experiences of Late-diagnosed Women with Autism Spectrum Conditions: An Investigation of the Female Autism Phenotype. J Autism Dev Disord 2016 Oct 25;46(10):3281-3294 [FREE Full text] [doi: 10.1007/s10803-016-2872-8] [Medline: 27457364]
7. Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J, et al. Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. Pediatrics 2010 Jan;125(1):e17-e23 [FREE Full text] [doi: 10.1542/peds.2009-0958] [Medline: 19948568]
8. Fenske EC, Zalenski S, Krantz PJ, McClannahan LE. Age at intervention and treatment outcome for autistic children in a comprehensive intervention program. Analysis and Intervention in Developmental Disabilities 1985 Jan;5(1-2):49-58. [doi: 10.1016/s0270-4684(85)80005-7]
9. Dawson G, Jones EJ, Merkle K, Venema K, Lowy R, Faja S, et al. Early behavioral intervention is associated with normalized brain activity in young children with autism. J Am Acad Child Adolesc Psychiatry 2012 Nov;51(11):1150-1159 [FREE Full text] [doi: 10.1016/j.jaac.2012.08.018] [Medline: 23101741]
10. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, et al. Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder: A Randomized Clinical Trial. JAMA Pediatr 2019 May 01;173(5):446-454 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.0285] [Medline: 30907929]
11. Sandbank M, Bottema-Beutel K, Crowley S, Cassidy M, Dunham K, Feldman JI, et al. Project AIM: Autism intervention meta-analysis for studies of young children. Psychol Bull 2020 Jan;146(1):1-29 [FREE Full text] [doi: 10.1037/bul0000215] [Medline: 31763860]
12. Grossi D, Marcone R, Cinquegrana T, Gallucci M. On the differential nature of induced and incidental echolalia in autism. J Intellect Disabil Res 2013 Oct;57(10):903-912. [doi: 10.1111/j.1365-2788.2012.01579.x] [Medline: 22676294]
13. Nakai Y, Takashima R, Takiguchi T, Takada S. Speech intonation in children with autism spectrum disorder. Brain Dev 2014 Jun;36(6):516-522. [doi: 10.1016/j.braindev.2013.07.006] [Medline: 23973369]
14. Paul R, Augustyn A, Klin A, Volkmar FR. Perception and production of prosody by speakers with autism spectrum disorders. J Autism Dev Disord 2005 Apr;35(2):205-220. [doi: 10.1007/s10803-004-1999-1] [Medline: 15909407]
15. Xu D, Gilkerson J, Richards J, Yapanel U, Gray S. Child vocalization composition as discriminant information for automatic autism detection. 2009 Presented at: Annual International Conference of the IEEE Engineering in Medicine and Biology Society; September 3-6, 2009; Minneapolis, MN p. 2518-2522. [doi: 10.1109/IEMBS.2009.5334846]
16. Lee JH, Lee GW, Bong G, Yoo HJ, Kim HK. Deep-Learning-Based Detection of Infants with Autism Spectrum Disorder Using Auto-Encoder Feature Representation. Sensors (Basel) 2020 Nov 26;20(23):6762 [FREE Full text] [doi: 10.3390/s20236762] [Medline: 33256061]
17. Cho S, Liberman M, Ryant N, Cola M, Schultz R, Parish-Morris J. Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations. Interspeech. 2019. URL: https://www.isca-speech.org/archive/interspeech_2019/cho19_interspeech.html [accessed 2022-03-29]

18. Li M, Tang D, Zeng J, Zhou T, Zhu H, Chen B, et al. An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. Computer Speech & Language 2019 Jul;56:80-94. [doi: 10.1016/j.csl.2018.11.002]
19. Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall DP. Guess What?: Towards Understanding Autism from Structured Video Using Facial Affect. J Healthc Inform Res 2019 Oct 2;3(1):43-66 [FREE Full text] [doi: 10.1007/s41666-018-0034-9] [Medline: 33313475]
20. Kalantarian H, Jedoui K, Washington P, Wall DP. A Mobile Game for Automatic Emotion-Labeling of Images. IEEE Trans Games 2020 Jun;12(2):213-218 [FREE Full text] [doi: 10.1109/tg.2018.2877325] [Medline: 32551410]
21. Kalantarian H, Washington P, Schwartz J, Daniels J, Haber N, Wall D. A gamified mobile system for crowdsourcing video for autism research. 2018 Jul 4 Presented at: IEEE International Conference on Healthcare Informatics (ICHI); June 4-7, 2018; New York, NY. [doi: 10.1109/ICHI.2018.00052]
22. Kalantarian H, Jedoui K, Washington P, Tariq Q, Dunlap K, Schwartz J, et al. Labeling images with facial emotion and the potential for pediatric healthcare. Artif Intell Med 2019 Jul;98:77-86 [FREE Full text] [doi: 10.1016/j.artmed.2019.06.004] [Medline: 31521254]
23. Kalantarian H, Jedoui K, Dunlap K, Schwartz J, Washington P, Husic A, et al. The Performance of Emotion Classifiers for Children With Parent-Reported Autism: Quantitative Feasibility Study. JMIR Ment Health 2020 Apr 01;7(4):e13174 [FREE Full text] [doi: 10.2196/13174] [Medline: 32234701]
24. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
25. Hershey S, Chaudhuri S, Ellis D. CNN architectures for large-scale audio classification. 2017 Mar 5 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); March 5-9, 2017; New Orleans, LA. [doi: 10.1109/ICASSP.2017.7952132]
26. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2016 Jun 27 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.308]
27. Baevski A, Auli M, Mohamed A, Zhou H. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. 2020 Presented at: 34th Conference on Neural Information Processing Systems; December 6-12, 2020; Virtual URL: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf
28. Pepino L, Riera P, Ferrer L. Emotion recognition from speech using WAV2VEC 2.0 embeddings. Interspeech. 2021. URL: https://www.isca-speech.org/archive/interspeech_2021/pepino21_interspeech.html [accessed 2022-03-29]
29. Kotz D, Gunter CA, Kumar S, Weiner JP. Privacy and Security in Mobile Health: A Research Agenda. Computer (Long Beach Calif) 2016 Jun;49(6):22-30 [FREE Full text] [doi: 10.1109/MC.2016.185] [Medline: 28344359]
30. Washington P, Yeung S, Percha B, Tatonetti N, Liphardt J, Wall D. Achieving trustworthy biomedical data solutions. Biocomputing 2021 2020:1-13. [doi: 10.1142/9789811232701_0001]
31. Washington P, Tariq Q, Leblanc E, Chrisman B, Dunlap K, Kline A, et al. Crowdsourced privacy-preserved feature tagging of short home videos for machine learning ASD detection. Sci Rep 2021 Apr 07;11(1):7620 [FREE Full text] [doi: 10.1038/s41598-021-87059-4] [Medline: 33828118]
32. Alshaigi K, Albraheem R, Alsaleem K, Zakaria M, Jobeir A, Aldhalaan H. Stigmatization among parents of autism spectrum disorder children in Riyadh, Saudi Arabia. Int J Pediatr Adolesc Med 2020 Sep;7(3):140-146 [FREE Full text] [doi: 10.1016/j.ijpam.2019.06.003] [Medline: 33094144]
33. Yang Q, Liu Y, Chen T, Tong Y. Federated Machine Learning. ACM Trans Intell Syst Technol 2019 Mar 31;10(2):1-19. [doi: 10.1145/3298981]
34. Li X, Gu Y, Dvornek N, Staib LH, Ventola P, Duncan JS. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med Image Anal 2020 Oct;65:101765 [FREE Full text] [doi: 10.1016/j.media.2020.101765] [Medline: 32679533]
35. Fombonne E. Epidemiology of Pervasive Developmental Disorders. Pediatr Res 2009 Jun;65(6):591-598. [doi: 10.1203/pdr.0b013e31819e7203]
36. Loomes R, Hull L, Mandy WPL. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. J Am Acad Child Adolesc Psychiatry 2017 Jun;56(6):466-474. [doi: 10.1016/j.jaac.2017.03.013] [Medline: 28545751]
37. Wang D, Chen J. Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Trans Audio Speech Lang Process 2018 Oct;26(10):1702-1726. [doi: 10.1109/taslp.2018.2842159]
38. Shreedhar Bhat G, Shankar N, Panahi I. A computationally efficient blind source separation for hearing aid applications and its real-time implementation on smartphone. The Journal of the Acoustical Society of America 2019 Oct;146(4):2959-2959. [doi: 10.1121/1.5137282]
39. Kline A, Voss C, Washington P, Haber N, Schwartz H, Tariq Q, et al. Superpower Glass. GetMobile: Mobile Comp and Comm 2019 Nov 14;23(2):35-38. [doi: 10.1145/3372300.3372308]
40. Voss C, Washington P, Haber N. Superpower glass. 2016 Presented at: 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing; 2016; Heidelberg, Germany. [doi: 10.1145/2968219.2968310]

41. Washington P, Voss C, Haber N. A wearable social interaction aid for children with autism. 2016 Presented at: Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems; 2016; San Jose, CA. [doi: 10.1145/2851581.2892282]
42. Daniels J, Haber N, Voss C, Schwartz J, Tamura S, Fazel A, et al. Feasibility Testing of a Wearable Behavioral Aid for Social Learning in Children with Autism. Appl Clin Inform 2018 Jan 21;9(1):129-140 [FREE Full text] [doi: 10.1055/s-0038-1626727] [Medline: 29466819]
43. Washington P, Voss C, Kline A, Haber N, Daniels J, Fazel A, et al. SuperpowerGlass. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017 Sep 11;1(3):1-22. [doi: 10.1145/3130977]
44. Daniels J, Schwartz JN, Voss C, Haber N, Fazel A, Kline A, et al. Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism. NPJ Digit Med 2018 Aug 2;1(1):32 [FREE Full text] [doi: 10.1038/s41746-018-0035-3] [Medline: 31304314]
45. Daniels J, Schwartz J, Haber N, Voss C, Kline A, Fazel A, et al. 5.13 Design and Efficacy of a Wearable Device for Social Affective Learning in Children With Autism. Journal of the American Academy of Child & Adolescent Psychiatry 2017 Oct;56(10):S257. [doi: 10.1016/j.jaac.2017.09.296]
46. Washington P, Leblanc E, Dunlap K, Penev Y, Kline A, Paskov K, et al. Precision Telemedicine through Crowdsourced Machine Learning: Testing Variability of Crowd Workers for Video-Based Autism Feature Recognition. J Pers Med 2020 Aug 13;10(3):86 [FREE Full text] [doi: 10.3390/jpm10030086] [Medline: 32823538]
47. Washington P, Paskov K, Kalantarian H, Stockham N, Voss C, Kline A, et al. Feature Selection and Dimension Reduction of Social Autism Data. Pac Symp Biocomput 2020;25:707-718 [FREE Full text] [Medline: 31797640]
48. Washington P, Kalantarian H, Tariq Q, Schwartz J, Dunlap K, Chrisman B, et al. Validity of Online Screening for Autism: Crowdsourcing Study Comparing Paid and Unpaid Diagnostic Tasks. J Med Internet Res 2019 May 23;21(5):e13668 [FREE Full text] [doi: 10.2196/13668] [Medline: 31124463]
49. Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, et al. Detecting Developmental Delay and Autism Through Machine Learning Models Using Home Videos of Bangladeshi Children: Development and Validation Study. J Med Internet Res 2019 Apr 24;21(4):e13822 [FREE Full text] [doi: 10.2196/13822] [Medline: 31017583]
50. Washington P, Leblanc E, Dunlap K. Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder. Biocomputing 2021 2020:14-25. [doi: 10.1142/9789811232701_0002]
51. Washington P, Leblanc E, Dunlap K. Crowd Annotations Can Approximate Clinical Autism Impressions from Short Home Videos with Privacy Protections. medRxiv Preprint published online on July 6, 2021. [doi: 10.1101/2021.07.01.21259683]
52. Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP. Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS Med 2018 Nov 27;15(11):e1002705 [FREE Full text] [doi: 10.1371/journal.pmed.1002705] [Medline: 30481180]
53. Leblanc E, Washington P, Varma M, Dunlap K, Penev Y, Kline A, et al. Feature replacement methods enable reliable home video analysis for machine learning detection of autism. Sci Rep 2020 Dec 04;10(1):21245 [FREE Full text] [doi: 10.1038/s41598-020-76874-w] [Medline: 33277527]
54. Haber N, Voss C, Fazel A, Winograd T, Wall D. A practical approach to real-time neutral feature subtraction for facial expression recognition. 2016 Mar 07 Presented at: IEEE Winter Conference on Applications of Computer Vision (WACV); 2016; Lake Placid, NY. [doi: 10.1109/wacv.2016.7477675]
55. Washington P, Kalantarian H, Kent J, Husic A, Kline A, Leblanc E, et al. Training Affective Computer Vision Models by Crowdsourcing Soft-Target Labels. Cogn Comput 2021 Sep 27;13(5):1363-1373. [doi: 10.1007/s12559-021-09936-4]
56. Washington P, Park N, Srivastava P, Voss C, Kline A, Varma M, et al. Data-Driven Diagnostics and the Potential of Mobile Artificial Intelligence for Digital Therapeutic Phenotyping in Computational Psychiatry. Biol Psychiatry Cogn Neurosci Neuroimaging 2020 Aug;5(8):759-769 [FREE Full text] [doi: 10.1016/j.bpsc.2019.11.015] [Medline: 32085921]
57. Washington P, Kline A, Mutlu O. Activity recognition with moving cameras and few training examples: Applications for detection of autism-related headbanging. 2021 May Presented at: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems; May 2021; Yokohama, Japan. [doi: 10.1145/3411763.3451701]
58. Lakkapragada A, Kline A, Mutlu O. Classification of Abnormal Hand Movement for Aiding in Autism Detection: Machine Learning Study. Arxiv. Preprint published on August 18, 2021 URL: https://arxiv.org/abs/2108.07917 [accessed 2022-03-30]
59. Washington P, Kalantarian H, Kent J. Improved Digital Therapy for Developmental Pediatrics using Domain-Specific Artificial Intelligence: Machine Learning Study. JMIR Pediatrics and Parenting 2020 Dec 23. [doi: 10.2196/26760]
60. Banerjee A, Washington P, Mutlu C, Kline A, Wall D. Training and profiling a pediatric emotion recognition classifier on mobile devices. Arxiv. Preprint published on August 22, 2021 URL: https://arxiv.org/abs/2108.11754 [accessed 2022-11-05]
61. Hou C, Kalantarian H, Washington P, Dunlap K, Wall D. Leveraging video data from a digital smartphone autism therapy to train an emotion detection classifier. medRxiv Preprint published online on August 1, 2021. [doi: 10.1101/2021.07.28.21260646]
62. Chong E, Clark-Whitney E, Southerland A, Stubbs E, Miller C, Ajodan EL, et al. Detection of eye contact with deep neural networks is as accurate as human experts. Nat Commun 2020 Dec 14;11(1):6386 [FREE Full text] [doi: 10.1038/s41467-020-19712-x] [Medline: 33318484]

63. Crippa A, Salvatore C, Perego P, Forti S, Nobile M, Molteni M, et al. Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities. J Autism Dev Disord 2015 Jul 5;45(7):2146-2156. [doi: 10.1007/s10803-015-2379-8] [Medline: 25652603]
64. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Applying Machine Learning to Kinematic and Eye Movement Features of a Movement Imitation Task to Predict Autism Diagnosis. Sci Rep 2020 May 20;10(1):8346 [FREE Full text] [doi: 10.1038/s41598-020-65384-4] [Medline: 32433501]

Abbreviations

ASD: autism spectrum disorder
AUC: area under the curve
AUROC: area under the receiver operating characteristic curve
CNN: convolutional neural network
NT: neurotypical
ROC: receiver operating characteristic

Edited by S Badawy; submitted 09.12.21; peer-reviewed by A Benis, S Taraman; comments to author 04.01.22; revised version received 18.01.22; accepted 25.01.22; published 14.04.22

Please cite as:
Chi NA, Washington P, Kline A, Husic A, Hou C, He C, Dunlap K, Wall DP
Classifying Autism From Crowdsourced Semistructured Speech Recordings: Machine Learning Model Comparison Study
JMIR Pediatr Parent 2022;5(2):e35406
URL: https://pediatrics.jmir.org/2022/2/e35406
doi: 10.2196/35406
PMID: