INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Comparing hand-crafted features to spectrograms for autism severity estimation

M. Eni[1], I. Dinstein[2], and Y. Zigel[1]

[1]Department of Biomedical Engineering, Ben-Gurion University of the Negev, Israel
[2]Departments of Psychology and of Cognitive and Brain Sciences, Ben-Gurion University of the Negev, Israel
marinamu@post.bgu.ac.il, dinshi@bgu.ac.il, yaniv@bgu.ac.il

Abstract

In this work, we compared two different input approaches to estimate autism severity using speech signals. We analyzed 127 audio recordings of young children obtained during the Autism Diagnostic Observation Schedule 2nd edition (ADOS-2) administration. Two different sets of features were extracted from each recording: 1) hand-crafted features, which included acoustic and prosodic features, and 2) log-mel spectrograms, which give the time-frequency representation. We examined two different Convolutional Neural Network (CNN) architectures for each of the two inputs and compared the autism severity estimation performance. We showed that the hand-crafted features yielded lower prediction error (normalized RMSE) in most examined configurations than the log-mel spectrograms. Moreover, fusing the estimated autism severity scores of the two feature extraction methods yielded the best results, where both architectures exhibited similar performance (Pearson R=0.66, normalized RMSE=0.24).

Index Terms: ADOS, audio, autism, CNN, features, severity estimation, spectrogram.

1. Introduction

Many speech signal processing algorithms utilize different feature extraction methods; some features are more task-specific (hand-crafted features [1], [2]), whereas others are more generic (e.g., spectrograms [3]). For example, Eyben et al. presented the openSMILE feature extraction toolkit in 2010 [4], which enables the extraction of audio and video features for signal processing and machine learning, particularly features enabling emotion recognition from speech. In 2016 the eGeMAPS was presented [5], a set of 88 acoustic features that can be used in various automatic voice analyses [6]. These and other hand-crafted features are widely used for emotion recognition [7] and in the quantification of speech disorders such as Autism Spectrum Disorder (ASD) [6], [8]–[11].

Today's speech processing and recognition algorithms also utilize spectrograms [12], [13], the time-frequency representations of the audio signal. Since humans do not perceive frequencies linearly, the mel-scale [14] which approximates the frequency spacing of a human cochlea, was developed to produce mel-spectrograms that reflect the frequency bands perceived by the human ear [7].

In this study, we explore the influence of these two feature extraction methods (i.e., hand-crafted and mel-spectrograms) on the ASD severity estimation system. ASD is a neurodevelopmental disorder characterized by speech disorders, including high pitch frequency, greater pitch inconsistencies, echolalia (speech repetition), and the generation of more distressed vocalizations [15], such as crying and screaming. The severity and manifestation of ASD symptoms can vary widely among individuals, but common characteristics of ASD include difficulties with social interaction, communication, and repetitive and restricted patterns of behavior, interests, or activities [16]. One of the most widely used observational tools for the assessment of ASD in children is the Autism Diagnostic Observation Schedule 2nd edition (ADOS-2) [17], which is based on DSM-5 criteria [18] and is currently considered the gold standard for identifying individuals with ASD. An ADOS session lasts 40-60 min. The total clinician-rated scores range from 0-30, with higher scores indicating more severe symptoms.

In our previous work, we extracted a group of 48 hand-crafted acoustic and conversational features from children's speech [19]. Our findings showed that these features could be used to estimate the severity of core ASD symptoms using a Convolutional Neural Network (CNN) architecture that was trained on 56 audio recordings. L. Nanni et al. [20] compared hand-crafted features and features extracted by deep CNN, and examined their combination for different image classification tasks. Their experiments showed that the fusion of different feature extraction methods outperforms the standard approaches. Similar conclusions were obtained in voice pathology detection [21] and in speech emotion recognition [22]. Based on these findings, we hypothesize that each feature extraction method may capture information that is less expressed by the other method. Therefore, the fusion between the two methods may give a better representation of ASD children's speech.

The main goal of this study was to provide a comparison between two feature extraction methods, and their combination, for estimating autism severity from speech recordings. We developed a DNN-based ASD severity estimation system combining hand-crafted features and spectrograms. We examined two different CNN architectures for each input and compared the autism severity estimation performance. The proposed system provides a robust, effective, and non-invasive method for estimating ASD severity in young children.

2. Dataset

We analyzed speech recordings from ADOS-2 assessments taken from ASD diagnoses. Audio recordings of 127 children, aged 4.09±1.31 years (see data specifications in Table 1), were selected from the Azrieli National Centre for Autism and Neurodevelopment Research (www.autismisrael.org). During the ADOS-2 assessments, the children were recorded using a distant microphone (CHM99, AKG, Vienna, Austria), placed approximately one meter from the child, thus allowing the child to move freely around the room. The language level of the children who participated in this study ranged from toddlers who are not verbally fluent to children with fluent speech. The

10.21437/Interspeech.2023-658

study was approved by the SUMC Ethics committee and parents of all participating toddlers provided written informed consent. All research was performed in accordance with the guidelines and regulations of the Helsinki committee. The audio sampling rate was 44.1 kHz (16 bits per sample), which was then downsampled to 16 kHz. The overall duration of the speech recordings was 41.76±11.49 min ([19.80, 77.36] min).

Table 1: *Children's demographics. Mean (std)*

#Children	ADOS score	Age (y)	Boys (%)
All	12.57 (7.52)	4.09 (1.31)	79
ASD (77%)	15.43 (5.83)	4.17 (1.26)	79
Not-ASD (23%)	2.29 (3.36)	3.82 (1.45)	79

3. Proposed approach

In this study, we propose several feature-extraction methods and system configurations for estimating ASD severity (total ADOS scores). The total ADOS score is composed of the assessment of the social communication impairments and restricted and repetitive patterns of behavior. Our proposed estimation system can be divided into several modules (Figure 1): 1) audio database acquisition, 2) child vocalization detection and segmentation, 3) feature extraction, which includes the hand-crafted features and the log-mel spectrograms, and 4) ASD severity estimation using two different architectures: CNN1D and SLINet. We examined several system configurations of the feature-sets / vocalization-durations / architectures and observed their impact on estimating the ASD severity scores.

3.1. Detection of child vocalizations

The audio dataset was manually segmented and labeled for different speakers using an in-house labeling Graphical User Interface (GUI). For this study, we only focused on child-labeled audio segments. Detection of child vocalizations from the labeled segments, separated by silence, was then applied. The algorithm uses energy and energy thresholds to detect the start and end of each vocalization [19]. These vocalizations may include speech, crying, screaming, and other vocalic events. Since too short vocalizations may not contain enough information, in the current study, we used a vocalization duration threshold; vocalizations shorter than this threshold were excluded. Several duration threshold values were examined: 110/200/500 ms. The maximum duration was not limited.

3.2. Feature extraction

Two feature extraction methods for estimating ASD severity score were implemented and tested. For this purpose, we extracted both the hand-crafted features and spectrograms from each recording.

3.2.1. Hand-crafted features

Here we utilized ASD-related features, similar to [19]. From each detected vocalization, we extracted nine types of features (see Table 2). Next, we selected sub-groups of 10 consecutive vocalizations of the child and computed a variety of statistics from their features (e.g., mean, standard deviation) in order to capture distributional changes. This derived a feature vector of 49 features for the selected sub-group of vocalizations. We performed this procedure 100 times, selecting random sub-groups of sequential vocalizations from each recording. We combined the 100 feature vectors into a single feature matrix, yielding an input matrix of 100×49, one matrix per child. Because of the randomness of vocalization selection in each sub-group, we ran the feature extraction five times for each recording, estimated the severity score for each feature matrix, and calculated the mean value for each recording.

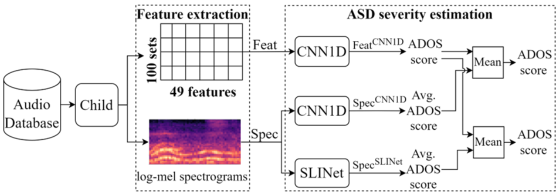

Figure 1: *Block diagram of the ASD severity estimation system. Two feature extraction methods (hand-crafted: Feat, log-mel spectrograms: Spec) were compared along with two CNN architectures (CNN1D and SLINet). In overall, five models were derived: $Feat^{CNN1D}$, $Spec^{CNN1D}$, $Spec^{SLINet}$, $Feat^{CNN1D}$ $+Spec^{CNN1D}$, and $Feat^{CNN1D}$ $+Spec^{SLINet}$.*

Table 2: *List of hand-crafted features.*

	Feature type	Statistics	Size
1	Pitch	μ, σ^2, σ^2/μ, min F_0, max F_0, VC/UV	10
2	Formants & Bandwidths	μ, σ^2, F_1, F_2, $\vert F_1 - F_2\vert$, BW_1, BW_2	10
3	Voicing	μ, σ^2	2
4	Energy	μ, σ^2, E, ΔE. ΔΔE, |ΔE|	8
5	Spectral slope	μ, σ^2, [20,500] Hz, [500,1500] Hz, VC/UV	8
6	ZCR	μ, σ^2, VC/UV	6
7	Jitter	μ, σ^2	2
8	Duration	μ, σ^2	2
9	Quantity	#Vocalizations	1

VC: voiced vocalization (pitch defined in most frames), UV: unvoiced vocalization, F_0: fundamental frequency, BW: bandwidth, ZCR: Zero-Crossing Rate, Size: number of features.

3.2.2. Log-Mel spectrograms

For each detected vocalization, we generated log-mel spectrograms. Since the majority (78%) of the vocalizations ranged from 0.03s to 1s, the child's vocalizations were set to a fixed length of 1 s either by trimming or zero-padding. The remainder of the trimmed vocalization was discarded if its length was below the defined minimum duration threshold (110/200/500 ms). The one-sec vocalizations were converted to mel-spectrograms using the Python Librosa 0.9.2 library. The mel-spectrograms were calculated using 40 mel filters with a Hann window size of 40 ms with an overlap of 20 ms. Next, the mel-spectrograms were dB-scaled to generate the log-mel spectrograms, which were then used as input to the network.

Since the number of the vocalizations could differ, during model training all the spectrograms, of each individual child's recording, were assigned the same ADOS score as the entire audio label, i.e., the score reported by the clinician during the ADOS assessment (the actual ADOS score). The final ADOS score predicted by our approach was thus the average of the predicted score values of all the spectrograms for the audio recording.

3.3. Architectures

We compared two different CNN architectures (see Figure 2) to estimate the ASD severity score, as expressed in the total ADOS score:

The CNN1D: a CNN architecture proposed in [19] and is based on one-dimensional convolutional layers. This network takes as input the feature matrix of a recording, or a spectrogram of a 1 sec vocalization, and returns the predicted/estimated ADOS score. The architecture is built out of seven hidden layers, with ReLU activation function, and one output layer with a linear activation. The two dropout layers randomly set input units to 0 with a frequency of 0.5 at each step during training time.

The SLINet: a CNN architecture proposed by M. Kaushik et al. [23], especially for the detection of Specific Language Impairment (SLI) from audio recordings of children aged 6-12. Children with SLI and those diagnosed with ASD share the common feature of poor spoken communication skills [24]. Delays and deficits in language are generally among the first symptoms of the autistic condition and are the core features of SLI. In addition to similarities in language impairments, this detection algorithm has successfully classified children with SLI. Thus, we retrained the SLINet CNN architecture on our ASD database and used it on the spectrogram feature extraction method alone. The SLINet network is based on two-dimensional convolutional layers with ReLU activation function and includes 3 dropout layers with drop rates of 0.2, 0.4, and 0.5, respectively. Since our goal was to estimate a severity score, we changed the output layer from a dense layer with a Softmax activation function with two output neurons, which was used for the classification task, to a linear activation function with one output neuron that used for the regression task.

In addition to each model, we also examined fusing the two methods, "Feat+Spec", by averaging the two predicted scores of each child. See the network's initialization method and the total number of parameters in Table 3.

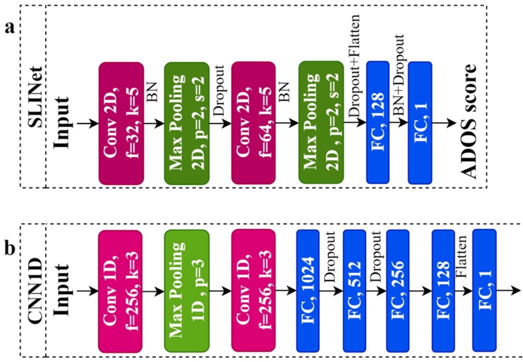

Figure 2: *The examined architectures for ASD severity score (ADOS) estimation: (**a**) SLINet; (**b**) CNN1D. BN stands for Batch Normalization, 'k' for kernel size, 'f' for filters, 'p' for pooling size, and 's' for strides.*

3.4. Training

The two model architectures were trained using the RMSprop (Root Mean Square Propagation) optimizer [25] with mean squared error (MSE) loss function and validated with 5-fold cross-validation for multiple epochs. The models were implemented in Python 3.8 using TensorFlow's Keras API 2.4.1. The training was conducted on a server with NVIDIA A40 GPU. The whole feature extraction-training-testing process took up to three days. Early stopping was applied to prevent overfitting of the model; hence the number of epochs was different for every fold, and the best-performing model with the highest Pearson correlation between the predicted and actual scores for entire recordings, was retained. For each fold, in order to identify the optimal model, we used a random search algorithm with 5-fold cross-validation and optimized three hyperparameters: batch size (2^5 - 2^9), learning rate (10^{-4} - 10^{-1}), and number of epochs ([10, 2000]). We trained the spectrogram models for 500 iterations (to save run time) and applied early stopping if there was no improvement on the validation set for ten epochs. The chosen parameters for each model in each fold are described in the Supplementary material.

Performance was evaluated using the Normalized Root Mean Square Error (NRMSE, (1)) and the Pearson correlation coefficient, R, between the actual and the estimated/predicted ADOS scores.

$$NRMSE = \frac{RMSE}{max(y_{actual}) - min(y_{actual})} \tag{1}$$

where RMSE stands for the Root Mean Square Error, and y_{actual} for the actual/observed severity scores of the children in the training dataset, as derived by the clinician during the ADOS assessment.

Table 3: *Networks characteristics*

Network architecture	Weights initialization	Biases initialization	Total number of parameters
CNN1D	Uniform	Zeros	Feat: 1,190,785 Spec: 1,189,761
SLINet	Xavier	Zeros	Spec: 568,897

4. Results

Five system configurations for ASD severity estimation were tested: 1) Log-mel spectrograms using SLINet (Spec^{SLINet}), 2) Log-mel spectrograms using CNN1D (Spec^{CNN1D}), 3) Hand-crafted features using CNN1D (Feat^{CNN1D}), 4) Fusion between the two extraction methods where the spectrograms were trained using the SLINet (Feat^{CNN1D} + Spec^{SLINet}), and 5) Fusion between the two extraction methods where the spectrograms were trained using the CNN1D (Feat^{CNN1D} + Spec^{CNN1D}). Figure 3 shows the performance measures (Pearson correlation and NRMSE) of the five systems.

Of the three vocalization duration thresholds, 200 ms achieved the best performance in most configurations, as shown in Figure 3. Table 4 shows the ASD severity estimation performances for the five different systems configurations for vocalizations longer than 200ms. Here, the performance measures are shown as the mean values and standard deviation among the five folds. In the case of hand-crafted features, the mean was also calculated across the five random feature matrices extracted for each recording. These results show that the fusion between the predicted severity scores of the hand-crafted feature extraction method and the predicted scores of the log-mel spectrograms achieved the highest performance (R = 0.66, NRMSE = 0.24).

5. Discussion

In this study, we analyzed speech signals of young Hebrew-speaking children recorded during their ADOS assessments. As our primary goal, we compared the performance of handcrafted features and log-mel spectrograms to estimate the ASD severity scores. Our findings showed that the hand-crafted features performed better in most configurations than the log-mel spectrograms, deriving lower prediction error. However, the fusion between the two methods achieved the best performance. Both fused networks evidenced similar performance when comparing the SLINet and the authors proposed CNN1D architecture. These results suggest that the network that was initially proposed to classify children with SLI may also be suitable for diagnosing children with autism at different levels.

Analysis of child vocalizations showed that using vocalizations longer than 200 ms contributed to better performance of the ASD estimation system. Further statistical analysis of the demographic characteristics of children revealed no statistical difference in prediction of the ADOS score between boys and girls using the spectrogram extraction method. However, the analysis demonstrated a superior outcome in R-value for children exceeding the median age as compared to those in the younger cohort using the $\text{Feat}^{\text{CNN1D}}$ model.

In an additional comparison, we extracted a feature vector of 88 eGeMAPS features [5] from each child vocalization using the Python openSMILE toolkit. This set of features was initially proposed as a simple acoustic parameter set for automated voice analysis in paralinguistic or clinical settings. We used a multilayer perceptron neural network with three fully connected hidden layers with 128, 64, and 32 nodes. The highest results using eGeMAPS were derived using the 110 ms minimum vocalization length, resulting in mean R of 0.569±0.100 and NRMSE of 0.271±0.016. In contrast, the fusion of spectrograms and hand-crafted features achieved better results than eGeMAPS.

Finally, we examined the effect of a mel-scale, and replaced the log-mel spectrograms with ordinary linear spectrograms, generated with 512 FFT points. We trained the two described architectures while limiting the min vocalization length to 200 ms. In this experiment, the linear spectrograms had higher prediction error than the log-mel spectrograms, deriving; 1) CNN1D: R = 0.40±0.16, NRMSE = 0.28±0.02, and 2) SLINet: R = 0.56±0.06, NRMSE = 0.26±0.02. These results underscore the advantages of the mel-scale when processing speech signals of children for autism estimation.

The current study has a few strengths and limitations. First, we analyzed a relatively large sample size of young children (age 1-7 years) diagnosed with ASD; some were minimally verbal, and some spoke fluently. As far as we know, this collected data is one of the largest speech databases of children with autism this age, specifically Hebrew-speaking children. One potential limitation of this study is the exclusive testing of our system on Hebrew language. Nevertheless, the utilization of prosodic features offers a potential pathway for overcoming this constraint. Recordings of the children's speech were acquired utilizing a remote microphone, allowing them greater freedom of movement within the recording room; moreover, a wearable microphone can distract the child and interfere with diagnostic process [26]. However, the signal to noise ratio of the acquired signal from a remote microphone is lower than a wearable microphone and may bias ASD severity estimation.

Table 4: *ASD severity estimation performances of the five system configurations - for vocalizations longer than 200ms.*

Feature type	System configuration	R $\mu\pm\sigma$ (median)	NRMSE $\mu\pm\sigma$ (median)
Log-mel spectrograms	$\text{Spec}^{\text{SLINet}}$	0.665 ± 0.117 (0.626)	0.263 ± 0.022 (0.257)
	$\text{Spec}^{\text{CNN1D}}$	0.642 ± 0.093 (0.670)	0.260 ± 0.013 (0.258)
Hand-crafted features	$\text{Feat}^{\text{CNN1D}}$	0.614 ± 0.089 (0.625)	0.240 ± 0.017 (0.236)
Fusion (Feat+Spec)	**$\text{Feat}^{\text{CNN1D}}$ + $\text{Spec}^{\text{SLINet}}$**	**0.663 ± 0.081 (0.656)**	**0.236 ± 0.011 (0.239)**
	$\text{Feat}^{\text{CNN1D}}$ + $\text{Spec}^{\text{CNN1D}}$	**0.657 ± 0.069 (0.672)**	**0.237 ± 0.010 (0.234)**

6. Conclusions

This study provides valuable insights into the use of hand-crafted features and log-mel spectrograms for estimating ASD severity scores in young Hebrew-speaking children. Our findings highlight the potential for a fusion of these two methods to achieve the best performance. This speech analysis algorithm could be highly valuable in evaluating early ASD risk and as a unique outcome measure for quantifying ASD severity, providing beneficial clinical utility.

7. Acknowledgements

This study was supported by the Israeli Ministry of Science and Technology (Grant no. 3-17422).

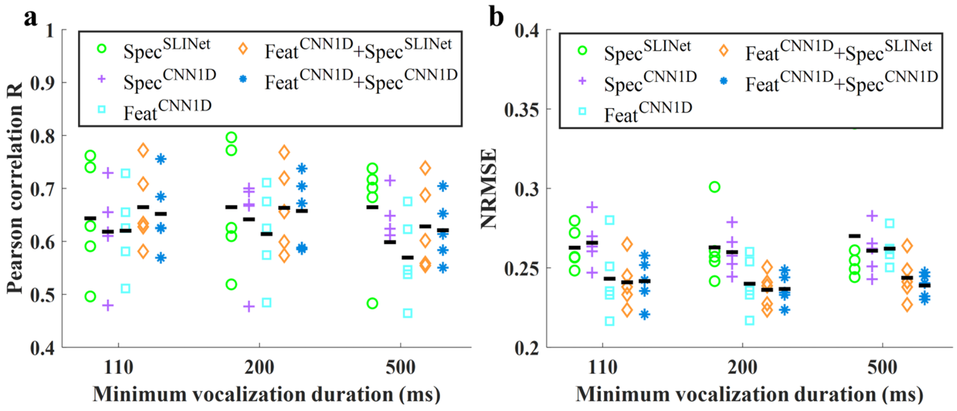

Figure 3: *ASD severity estimation system performance of different system configurations. The metrics: **(a)** Pearson correlation R, and **(b)** Normalized Root Mean Squared Error (NRMSE) were calculated between the predicted and the actual ASD severity scores. Each dot (marker) represents an iteration from 5-folds cross validation. Marker types indicate 5 different system configurations, while horizontal black lines represent the average value across the 5-folds.*

8. References

[1] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011,” *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, 2012.

[2] A. C. Salem *et al.*, “Evaluating atypical language in autism using automated language measures,” *Sci. Rep.*, vol. 11, no. 1, p. 10968, Dec. 2021.

[3] J. Wagner, D. Schiller, A. Seiderer, and E. André, “Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?,” in *Interspeech 2018*, 2018, pp. 147–151.

[4] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in *Proceedings of the international conference on Multimedia - MM ’10*, 2010, pp. 1459–1462.

[5] F. Eyben *et al.*, “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[6] F. B. Pokorny *et al.*, “Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, no. August, pp. 309–313, 2017.

[7] Z. Zhao *et al.*, “Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition,” *IEEE Access*, vol. 7, pp. 97515–97525, 2019.

[8] R. Francese, M. Frasca, and M. Risi, “Automatic creation of a Vowel Dataset for performing Prosody Analysis in ASD screening,” in *2021 25th International Conference Information Visualisation (IV)*, 2021, pp. 29–34.

[9] J. C. Y. Lau *et al.*, “Cross-linguistic patterns of speech prosodic differences in autism: A machine learning study,” *PLoS One*, vol. 17, no. 6, p. e0269637, Jun. 2022.

[10] K. Ochi, N. Ono, K. Owada, M. Kuroda, S. Sagayama, and H. Yamasue, “Entrainment Analysis for Assessment of Autistic Speech Prosody Using Bottleneck Features of Deep Neural Network,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, vol. 2022-May, pp. 8492–8496.

[11] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, “Learning Converse-Level Multimodal Embedding to Assess Social Deficit Severity for Autism Spectrum Disorder,” in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, vol. 2020-July, pp. 1–6.

[12] R. Doshi *et al.*, “Extending Parrotron: An End-to-End, Speech Conversion and Speech Recognition Model for Atypical Speech,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6988–6992.

[13] N. A. Chi *et al.*, “Classifying Autism From Crowdsourced Semistructured Speech Recordings: Machine Learning Model Comparison Study,” *JMIR Pediatr. Parent.*, vol. 5, no. 2, p. e35406, Apr. 2022.

[14] H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network,” *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[15] A. M. Plumb and A. M. Wetherby, “Vocalization Development in Toddlers With Autism Spectrum Disorder,” *J. Speech, Lang. Hear. Res.*, vol. 56, no. 2, pp. 721–734, Apr. 2013.

[16] J. Baio *et al.*, “Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014,” *MMWR. Surveill. Summ.*, vol. 67, no. 6, pp. 1–23, Apr. 2018.

[17] I. Kamp-Becker *et al.*, “Diagnostic accuracy of the ADOS and ADOS-2 in clinical practice,” *Eur. Child Adolesc. Psychiatry*, vol. 27, no. 9, pp. 1193–1207, Sep. 2018.

[18] American Psychiatric Association, “Cautionary Statement for Forensic Use of DSM-5,” in *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition*, 5th ed., Arlington, VA: American Psychiatric Publishing, Inc, 2013, p. 991.

[19] M. Eni, I. Dinstein, M. Ilan, I. Menashe, G. Meiri, and Y. Zigel, “Estimating Autism Severity in Young Children From Speech Signals Using a Deep Neural Network,” *IEEE Access*, vol. 8, pp. 139489–139500, 2020.

[20] L. Nanni, S. Ghidoni, and S. Brahnam, “Hand-crafted vs. non-handcrafted features for computer vision classification,” *Pattern Recognit.*, vol. 71, pp. 158–172, Nov. 2017.

[21] A. N. Omeroglu, H. M. A. Mohammed, and E. A. Oral, “Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion,” *Eng. Sci. Technol. an Int. J.*, vol. 36, p. 101148, Dec. 2022.

[22] C. Wang, Y. Ren, N. Zhang, F. Cui, and S. Luo, “Speech emotion recognition based on multi-feature and multi-lingual fusion,” *Multimed. Tools Appl.*, vol. 81, no. 4, pp. 4897–4907, Feb. 2022.

[23] M. Kaushik, N. Baghel, R. Burget, C. M. Travieso, and M. K. Dutta, “SLINet: Dysphasia detection in children using deep neural network,” *Biomed. Signal Process. Control*, vol. 68, no. May, p. 102798, Jul. 2021.

[24] S. Durrleman and H. Delage, “Autism Spectrum Disorder and Specific Language Impairment: Overlaps in Syntactic Profiles,” *Lang. Acquis.*, vol. 23, no. 4, pp. 361–386, Oct. 2016.

[25] A. M. Taqi, A. Awad, F. Al-Azzo, and M. Milanova, “The Impact of Multi-Optimizers and Data Augmentation on TensorFlow Convolutional Neural Network Performance,” *Proc. - IEEE 1st Conf. Multimed. Inf. Process. Retrieval, MIPR 2018*, vol. April, pp. 140–145, 2018.

[26] F. E. Z. El Arbaoui, K. El Hari, and R. Saidi, “A Survey on the Application of the Internet of Things in the Diagnosis of Autism Spectrum Disorder,” in *International Conference on Advanced Technologies for Humanity*, 2021, pp. 29–41.