DEEP NEURAL NETWORK APPROACHES FOR AUTISM DETECTION IN CHILDREN USING VOCAL BIOMARKERS: A SURVEY

Qurrat-ul-ain[1, 2], Aznul Qalid Md Sabri [1*], Erma Rahayu Binti Mohd Faizal Abdullah[1], Nurul Binti Japar[1], Nazia Perwaiz[3], Aisha Shabbir[2], Manjeevan Seera[4]

[1]Department of Artificial Intelligence,
Faculty of Computer Science and Information Technology,
University of Malaya, 50603, Kuala Lumpur, Malaysia

[2]NUST Institute of Civil Engineering,
School of Civil and Environmental Engineering,
National University of Sciences and Technology,
Islamabad, Pakistan

[3]School of Electrical Engineering and Computer Science,
National University of Sciences and Technology,
Islamabad, Pakistan

[4]School of Business,
Monash University Malaysia,
Selangor, Malaysia

Emails: s2036785@siswa.um.edu.my[1,2], aznulqalid@um.edu.my[1]* (Corresponding Author), erma@um.edu.my[1], nuruljapar@um.edu.my[1], nazia.perwaiz@seecs.edu.pk[3], aisha.shabbir@nice.nust.edu.pk[2], mseera@gmail.com[4]

ABSTRACT

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by diverse social and communication challenges, often reflected in atypical speech patterns. Vocal biomarkers have thus emerged as a promising, non-invasive avenue for early detection. This survey analyzes approximately 90 peer-reviewed studies published between 2004 and 2024, evaluating deep neural network (DNN)-based methods, particularly Siamese Neural Networks (SNNs), for ASD detection through vocal data. The studies collectively involved sample sizes ranging from 15 to over 1,700 participants, across various age groups from infants to adults. Performance metrics from these studies reported diagnostic accuracies up to 98%, sensitivity reaching 96.7%, and specificities up to 94.2%. The review highlights the effectiveness of SNNs even in limited-data scenarios and outlines challenges such as the lack of standardized vocal features and dataset diversity. It concludes with recommendations for future research to support the development of scalable, real-world solutions for early ASD diagnosis.

Keywords: Autism Spectrum Disorder; Vocal Biomarkers; Deep Neural Networks; Machine Learning; Siamese Neural Network.

1.0 INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition marked by behavioral deficits, including social and communication challenges, and repetitive behaviors [1]. Communication difficulties encompass struggles with nonverbal cues, inappropriate responses in conversations, and expressing needs (DSM-5) [2]. Individuals with ASD also exhibit a wide range of cognitive abilities, from low functioning to high levels of reasoning [3], [4].

The number of children diagnosed with autism has significantly increased in recent years, from one in 59 [5] to one in 44 [6]. While this growing prevalence reflects an actual increase in cases, it can also be partly attributed to improved diagnostic methods, earlier detection in younger children, and heightened public awareness [3].

Current research identifies two primary forms of autism: early-onset and regressive. Children with early-onset autism display distinct behaviors, such as difficulty responding to their name, making eye contact, paying attention to others, and sharing behaviors that differ from those of typically developing peers [7]. These variances become particularly evident during the second year [8], although signs of ASD have been identified in some studies as early as before the first birthday [7], [8], [9]. Autism has a higher prevalence of early onset in its population.

Children who had seemed to be developing normally in the regressive phase start to lose their social and communicative abilities between the ages of 16 and 20 months [10]. As per the information adapted from DSM-5, Fig. 1 shows potential indicators and symptoms of communication difficulties in autistic children. Delays in developing expressive or functional language skills by the age of five are frequently cited as one of the main indicators of autism [11] although this is not a diagnostic criterion for ASD. Parents often identify language delays in their children at around 18 months of age [12].

The type of intervention applicable for language development requires a thorough understanding of the children's skills to recognize the differential language trajectories in autistic children. Studies indicate that after the age of five, only a small percentage of children acquire functional language [13][14]. In a review of 535 children, 47% developed fluent speech by the age 8, while 70% achieved only phrase speech [15]. Another review found that children aged 5 to 7 typically began developing language skills, with about one-third reaching the stage of phrase speech [16]. Additionally, 61 minimally vocal children showed significant gains in spontaneous communication when intervention strategies using speech-generating devices and naturalistic settings were employed. Currently, researchers developing automated detection methods for minimally verbal autistic children face challenges due to the lack of objective techniques for evaluating their linguistic abilities [17].

Fig 1. Potential indicators and symptoms of communication impairment in children with autism, based on DSM-5 criteria

1.1 Speech Assessments of Autistic Children

Perceptual differences are prominent in the speech of autistic individuals, with early accounts noting atypical speech habits [18][19]. Blinded listeners have also found autistic individuals' speech distinct from neurotypical peers [20][21]. However, most research relies on structured tasks (e.g., word lists or sentence reading) or autism experts for evaluation. Studies on prosodic impairments in autistic children show varying success. For instance, [22] achieved 76% accuracy using data from a single site and specialized biosensors with directional microphones, while [23] reported high accuracy with speech data from wireless microphones in a hospital setting. Both studies relied on centralized, unfamiliar environments and high-fidelity recording devices, limiting practicality for automated diagnostic tools. Stress from interactions with unfamiliar adults in such settings may further reduce real-world applicability.

With the current advancement in machine learning methodologies, the automated diagnostic measures are of utmost benefit to both the clinicians and researchers for devising measures, which help with the low-cost, reliability, and early detection of expressive language developmental delays in autistic children. For the precise and prompt diagnosis of ASD in children, numerous researchers are working to develop early-detection tools based on machine learning and artificial intelligence algorithms [18]. The utilization of computer vision to train machine learning models has led researchers to a successful detection of ASD using face images of children with sound precision [19]. However, this methodology invades the privacy of children and requires continual updates of images throughout childhood. Retinal and brain imaging have also been used as objective screening methods for ASD diagnosis. In contrast, speech data is increasingly used to detect language disorders and ASD while preserving participant confidentiality. Each child's linguistic patterns are unique, like fingerprints, and repetitive phrases and words are core characteristics of autistic children. Such speech abnormalities can serve as effective measures for early detection and diagnosis of ASD using linguistics and machine learning.

Despite these advancements, current diagnostic procedures remain constrained by several limitations. Traditional assessments are time-consuming, costly, and reliant on subjective observations by clinicians and caregivers, often leading to inconsistent or delayed diagnoses. These challenges emphasize the urgent need for objective, scalable, and efficient tools for early autism detection and classification, particularly among minimally verbal children. Advanced diagnostic tools, such as neuroimaging, often incur high costs, making them less feasible for widespread implementation in publicly funded healthcare systems [20].

Deep learning surpasses traditional models by enabling automatic feature learning. While deep neural networks (DNNs) underpin modern speech systems, their reliance on large datasets poses challenges for limited data

scenarios like impaired speech modeling. To tackle this, we explore Siamese neural networks for autism detection, which perform well even with smaller datasets.

Detecting autism using vocal biomarkers poses several challenges, primarily due to the absence of standardized biomarkers for diagnosis. This research provides a comprehensive, up-to-date review of existing methods using Siamese neural networks for autism detection through vocal biomarkers, filling a gap due to the limited number of such surveys available. This study offers the first extensive analysis of deep neural network approaches, particularly focusing on Siamese neural networks, for detecting autism using speech data.

SNNs are particularly well-suited for ASD diagnosis due to their ability to learn similarity functions between paired inputs, making them effective even with limited and imbalanced datasets—a common issue in autism research. These networks have demonstrated success in identifying distinguishing features in neuroimaging and behavioral data, enhancing classification accuracy and generalizability across patient populations [21], [22]. By reducing dependence on large, labeled datasets and emphasizing contrastive learning, SNNs offer a more robust and data-efficient framework for identifying minimally verbal or atypically presenting children with ASD.

1.2 Main Contributions

Recently, the computer vision and artificial intelligence community has increasingly focused on using speech patterns for the automatic detection of autism, resulting in various proposals for developing precise and quantitative methods for this purpose. Table 1 presents an overview of the existing surveys related to Siamese Neural Networks (SNNs), vocal biomarkers (VB), autism detection, and associated studies, using a checkmark (✓) to indicate studies within specific categories. This survey marks the first thorough analysis of recent advancements in Siamese Neural Networks for vocal biomarker-based autism detection, focusing on articles published over the last two decades (since 2004). Key contributions of this survey include:

1. This paper systematically reviews the advancements in SNN techniques for detecting Autism Spectrum Disorder (ASD) using vocal biomarkers, providing a detailed assessment of how these methods have evolved. It serves as a valuable resource for researchers seeking to understand the current landscape of voice-based autism detection.

2. This survey provides a comprehensive review of existing SNN-based approaches, comparing them with state-of-the-art deep learning models. It highlights the strengths and weaknesses of these methods while identifying key challenges, such as the lack of standardized vocal biomarkers and scalability issues in data collection. Additionally, it offers clear recommendations for selecting the most effective techniques and outlines promising directions for future research, guiding the development of more robust and scalable solutions.

3. By focusing on the role of SNNs in vocal biomarker-based detection, this survey underscores the transformative potential of these techniques in early ASD diagnosis. The insights gained could significantly influence both research and clinical practices, paving the way for more accurate, non-invasive diagnostic tools that can be widely adopted.

These contributions establish this survey as an important resource for advancing the field of autism detection, offering a foundation for future innovations and practical implementations in the use of deep neural networks for speech-based diagnosis.

1.3 Survey Organization

This study comprises nine sections. Section 2.0 outlines the search strategy and eligibility criteria. Section 3.0 reviews vocal biomarkers for autism detection. Section 4.0 presents the clinical background of speech-related issues in ASD. Section 5.0 discusses experimental procedures using DNNs for data collection. Section 6.0 summarizes publicly accessible speech datasets. Section 7.0 reviews methods for automatically detecting autism using deep neural networks. Section 8.0 explains and highlights the survey's findings.

2.0 SEARCH STRATEGY AND ELIGIBILITY CRITERIA

This review adhered to the methodology proposed by Kitchenham [23] for conducting systematic reviews in software engineering, which emphasizes thorough planning, careful study selection, and rigorous data synthesis. The eligibility criteria for this survey included the suitability of deep neural network-based methods, particularly Siamese Neural Networks (SNN), for automated autism detection through vocal biomarkers.

To define the scope of the review, clear eligibility criteria were established. Studies were included if they

(i) provided a well-defined algorithmic approach for ASD detection using vocal or speech-based features, and

(ii) presented empirical findings derived from experiments, datasets, or model evaluations. Only studies that demonstrated tangible results were retained.

Table 1. Summary of existing surveys and reviews related to Siamese Neural Networks (SNN), Vocal Markers (VM) and Autism Spectrum Disorder Detection (ASD)

Publication	Year	Topic	Scope		
			SNN	VM	ASD
[51]	2022	A survey on SNN	✓	x	x
[52]	2022	A survey on SNN	✓	x	x
[53]	2021	A survey on SNN	✓	x	x
[54]	2021	A survey on SNN	✓	x	x
[55]	2020	A survey on SNN	✓	x	x
[56]	2022	A survey on deep neural networks and speech processing	✓	✓	x
[57]	2003	A survey on speech prosody for ASD	x	✓	✓
[58]	2017	A survey on VM for ASD	x	✓	✓
[59]	2022	A study on VM for ASD	x	✓	✓
[60]	2020	A survey on neural network models for ASD	x	x	✓
[61]	2022	A survey on ASD detection using machine learning	x	x	✓
[62]	2022	A survey on ASD detection using deep learning	✓	x	✓
[63]	2023	A survey on language in ASD	x	✓	✓
[64]	2024	A survey on ASD language assessments	x	✓	✓
Our work	2024	A survey on deep neural network approaches for autism detection in children using vocal biomarkers	✓	✓	✓

Research articles focusing on unrelated modalities such as visual data, functional magnetic resonance imaging (fMRI), electroencephalography (EEG), or any other brain imaging techniques were excluded, as this review specifically concentrates on voice and speech-based ASD detection.

A comprehensive literature search was carried out across multiple digital libraries and academic databases. These included IEEE Xplore, Elsevier, ScienceDirect, Web of Science, Google Scholar, ProQuest, and Frontiers. To ensure broad coverage of relevant studies, a combination of search terms was used with logical operators such as AND and OR. Keywords included: deep learning, neural networks, Siamese neural networks, machine learning, autism, autism spectrum disorder, ASD, speech, voice, vocal biomarkers, acoustics, prosody, and related terms.

Following this multi-database search, duplicate entries were removed. Articles were initially screened by reviewing their titles and abstracts. Full-text reviews were then conducted on the shortlisted studies to confirm alignment with the eligibility criteria. As a result of this rigorous selection process, a total of approximately 90 relevant studies were identified, spanning a publication timeframe from 2004 to February 2024.

The distribution of publications over the years reveals a noticeable upward trend. Between 2004 and 2015, the number of relevant articles remained relatively low, reflecting the nascent stage of deep learning applications in ASD research. However, from 2016 onwards, and especially between 2018 and 2023, there was a marked increase in publications. This aligns with the broader surge in deep learning research and increased accessibility to speech datasets and computational tools. A visualization of this trend is provided in Fig. 2, illustrating the growing academic interest in voice-based autism detection.

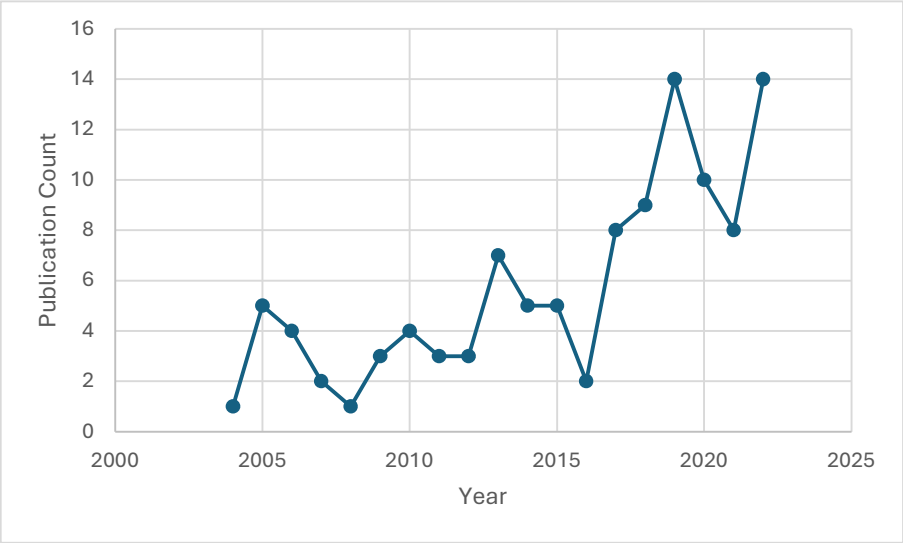

Fig 2. Year-wise publication trend (2004–2022) on deep learning approaches for ASD detection using vocal biomarkers. A noticeable rise in research is seen after 2017, highlighting growing interest in speech-based diagnostic methods.

In analyzing the extracted studies, several key patterns were observed. A wide variety of deep learning models were employed, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), Autoencoders, and hybrid architectures. Siamese Neural Networks (SNNs) were specifically explored in cases involving smaller or imbalanced datasets due to their robustness in low-data environments. Over time, researchers have transitioned from handcrafted feature extraction techniques toward end-to-end deep learning pipelines capable of automatically learning salient vocal features from raw audio.

The datasets used in these studies also varied in size and type. Publicly available resources such as CHILDES, TalkBank, DE-ENIGMA, and ASDBank were among those utilized. These datasets offer annotated child speech data across various languages and contexts. In many cases, custom datasets were also developed by individual research teams to suit specific tasks such as echolalia recognition, prosodic analysis, or speech emotion classification.

Performance evaluation was a consistent theme across studies, with most reporting metrics such as accuracy, precision, recall, F1-score, sensitivity, specificity, and area under the ROC curve (AUC). Reported accuracies ranged from 68% to 98%, depending on model architecture, dataset quality, and task complexity.

A sample timeline of publication frequency is shown in Fig. 2, while future sections of this paper include detailed tables summarizing the techniques, datasets, and results across reviewed studies.

3.0 VOCAL MARKERS FOR AUTISM SPECTRUM DISORDER DETECTION

The speech behavior of children with autism is highly variable, making it one of the biggest challenges in detecting ASD using speech assessment methods. This heterogeneity also contributes to the underdevelopment of standardized vocal biomarkers for diagnosing ASD in children. Various acoustic features have been assessed in combination for automatic autism detection using machine learning models. This multivariate approach enables the selection of multiple features for speech analysis, rather than relying on a single acoustic feature to differentiate between ASD and typically developing children.

Pitch, loudness, length, and voice quality are the four primary areas of speech production that comprise research on prosody in people with ASD, as defined in [24] and [25]. Limited variation in pitch, inappropriate pitch and pitch modulation, and oscillations between excessive loudness and quietness, frequently with abrupt and inappropriate shifts between the two, have all been noticed in the speech patterns of individuals with ASD [26], [27], [28], [29] and [30]. Additionally, people with ASD can speak slowly or quickly [31], [28], [26]. They also have a unique voice that has been described as "hoarse," "harsh," and "hyper-nasal" [31], [30], and they frequently use more sounds like squeals, growls, and yells [32].

According to many studies, autistic children make fewer speech-like sounds and consonants as well as less complex syllable forms throughout the prelinguistic phase of development [33], [34] and [35]. The diagnosis is particularly difficult because of other language development disorders that might also be presenting themselves at an early age in children [36]. The assessment of the speech of autistic children generally results in marking them

as sounding atypical. Subjectively, they have been identified as monotone, too fast, or too slow, too loud, etc. [37], [28], [38] and [39]. Although early research in this field presented different findings, when measured using objective acoustic parameters like pitch, studies have shown inconsistent results. Some studies have reported pitch to be of higher frequency in autistic children as contrast to the neuro-typical children [40], [41] while others have reported opposite findings [42], [43], [44]. The studies based on intensity of speech have also been inconsistent in their findings [45], [46], [47].

There are discrepancies in speech rate investigations [20], [64], [65]. A systematic review of 45 studies by [34] on acoustic speech characteristics found that the cause of abnormal perceptual quality remains unclear, and the studied features don't fully explain it. The study in [55] suggests that autistic children with enhanced perceptual abilities may "tune in" to speech models, improving their speech but lacking precision due to motivation or ability issues. Empirical research is divided, with some finding normal articulatory skills [66], [67], while others note nonconforming prosody patterns [68], [69]. These contradictions may stem from population heterogeneity [70], varying sample sizes, and differences in study criteria and perceptual measurements of articulatory precision.

In [48] the authors found differing vocal characteristics between ASD and TD children. [49] analyzed infant vocalizations for early ASD screening, focusing on 18-month-olds. Twenty TD and twenty ASD children were studied, yielding a 97% accuracy rate. K-fold cross-validation was employed, ensuring unseen subjects in the test fold. In [50], the researchers observed in their analysis of autistic people's speech articulation that the studies have been using inconsistent perceptual procedures to measure the articulation and hence suggested that fine grained techniques may be introduced for better understanding of such deficits since categorizing the speech as binary (correct or incorrect) may not be as suitable measure for detecting autism.

In [74], a study compared the articulatory accuracy of autistic and neurotypical adults and children, using manual analysis of speech recordings, which required intensive data preprocessing. A recent meta-analysis by [35] examined acoustic markers of autism in vocalizations of American and Danish autistic children. The study identified several key features, including pitch, intensity, and voice breaks, as potential markers of autism across languages and cultures. The authors call for further research to validate these markers and explore their use in acoustic diagnosis.

4.0 CLINICAL DIAGNOSTIC MEASURES OF SPEECH ASSESSMENT FOR AUTISM

Current ASD diagnostic processes involve several key components, including diagnostic checklists, physical screening, developmental monitoring, and supplementary testing. Developmental monitoring involves observing and assessing a child's progress during regular health checkups. Subsequently, specialists conduct routine physical examinations to rule out any physical conditions by examining factors such as the patient's vision, hearing, and basic motor skills. The aim is to ensure that any lack of response from the patient, for instance, is not due to hearing impairment, rather than assuming it is related to ASD. The patient's behavior is then assessed using standardized checklists, including the Modified Checklist for Autism in Toddlers, the Autism Diagnostic Observation Schedule, and the Autism Diagnostic Interview-Revised. Lastly, additional testing is performed by specialists, frequently including in-person assessments. There are four main phases in the current clinical techniques for diagnosing ASD, as shown in Fig. 3. The clinical diagnostic techniques that focus on speech evaluation for the detection of ASD are listed in this section. A key characteristic of ASD is shortfalls in language and social communication. The heterogeneity and symptoms of the disease may also be increased by additional factors that interact with the core symptoms, such as sensory processing issues and attention issues.

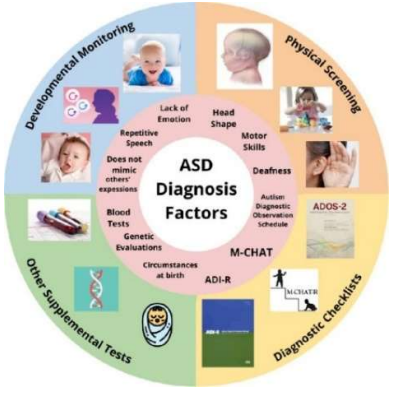

Fig 3. A four-stage process illustrating the current clinical practices for diagnosing ASD, according to [126], (1) Developmental Monitoring; (2) Physical Screening to rule out sensory or motor deficits; (3) Use of standardized Diagnostic Checklists; and (4) Additional testing such as in-person evaluations and parent-reported measures.

Research has demonstrated that prosodic abnormalities—such as irregularities in rhythm, tone, and stress—are common in families with a genetic predisposition to ASD [37]. Recent reviews of literature show studies investigating prosody in ASD, such as [65] and [66]. Table 2 provides a list of different acoustic features that have been utilized in different machine learning based approaches for autism detection.

These findings highlight the significance of accurately evaluating prosodic elements for the detection, diagnosis, and monitoring of ASD. Despite efforts to identify acoustic and prosodic irregularities for objective evaluations, measuring them in clinical settings remains notably difficult.

Table 2. List of Acoustic Features that have been assessed throughout different ML-based studies.

S. No	Acoustic Features	S. No	Acoustic Features
1.	Canonical Transitions	**9.**	Pronunciation Quality
2.	Dominant Frequencies	**10.**	Signal Energy
3.	Duration / Speech Rate	**11.**	Spectral Harmonicity / Entropy
4.	Formants Frequencies	**12.**	Strength of Excitation
5.	Harmonic-to-Noise Ratio (HNR)	**13.**	Voice Quality / Jitter / Shimmer
6.	Mel-frequency Cepstral Coefficients	**14.**	Volume / Loudness
7.	Linear Prediction Cepstrum Coefficients	**15.**	Zero-Crossing Rate
8.	Pitch / Fundamental Frequency		

4.1 Manual Diagnostic Methods of Vocal Analysis for Autism

Speech-language pathologists (SLPs) use various clinical measures to assess speech and language in individuals with ASD. In this section, we examine these clinical measures used to identify speech deficits related to ASD. Fig. 4 presents a list of these manual diagnostic measures.

The Autism Diagnostic Observation Schedule (ADOS-2; [67]) uses a semi-structured, standardized approach to assess individuals for ASD symptoms outlined in the DSM-5 through guided activities. In ADOS-2 Module 3, only one item (A2) assesses "Speech abnormalities related to autism" (rate/intonation/rhythm/volume), rated on a scale from 0 to 2. Achieving a score of 2 indicates a crude grouping of unusual traits. Similarly, The Social Responsiveness Scale (SRS-2) [68], an extension of the SRS [69], assesses social impairment and autistic traits in children aged 4 to 18. Administered by a familiar informant, it takes about 15 to 20 minutes to complete and contains 65 items rated on a Likert scale from 1 to 4, with one reversed item. Constantino & Gruber's SRS includes an item evaluating limited voice quality (item 53), such as speaking in an unusual tone.

Meanwhile, the Diagnostic and Statistical Manual of Mental Disorders [70], with its seven diagnostic criteria, does not include considerations of speech or voice quality. The Children Communication Checklist (CCC-2; [71]) is another popular tool with 70 standardized items to evaluate logical and social communication in children with the age range of 4 to 17. Caregivers rate behaviors on a 4 – point scale (0 for less than once a week to 3 for several times a day). The checklist comprises 10 subscales, each with seven items, five addressing weaknesses and two strengths. These evaluate various aspects, including speech patterns, syntax, semantics, coherence, language initiation, repetitive language, use of context, non-verbal communication, social interactions, and interests. Age-adjusted scores (mean = 10; SD = 3) are calculated from raw scores for each subscale. A General Communication Composite (GCC) score (mean = 100; SD = 15) sums scores from scales A to H. The Social Interaction Difference Index (SIDI) computes the difference between four pragmatic sub-scores (E through H) and four structural language sub-scores (A through D), which suggests the presence of autism if the number of negative values is greater. The CCC-2 is a parent completed checklist and requires time and accurate responses from parents, which are then processed by an expert to provide the diagnosis.

The Social Communication Checklist (SCC) is another tool, consisting of a 40-item checklist that assesses social communication skills in children aged 4 to 14 years. It includes questions associated with social interactions, language abilities, and pragmatic skills. Additionally, the Social Communication Questionnaire, SCQ [72], is a parent-reported tool for evaluating communication and social deficits in ASD individuals. It covers language development, social interaction, and repetitive behaviors.

For language assessment, Clinical Evaluation of Language Fundamentals (CELF-5; [73]) is a standardized test that measures various aspects of language, including receptive and expressive language abilities, as well as semantics and pragmatic language use. Preschool Language Scale (PLS; [74]) is a standardized test designed to assess language abilities in children from birth to 7 years. It includes measures of expressive and receptive and language, semantics, and syntax. It offers a variety of tasks and stimuli, such as pointing to pictures, answering

questions, and completing sentences, which allow for the assessment of different language domains and offer a comprehensive view of a child's language skills.

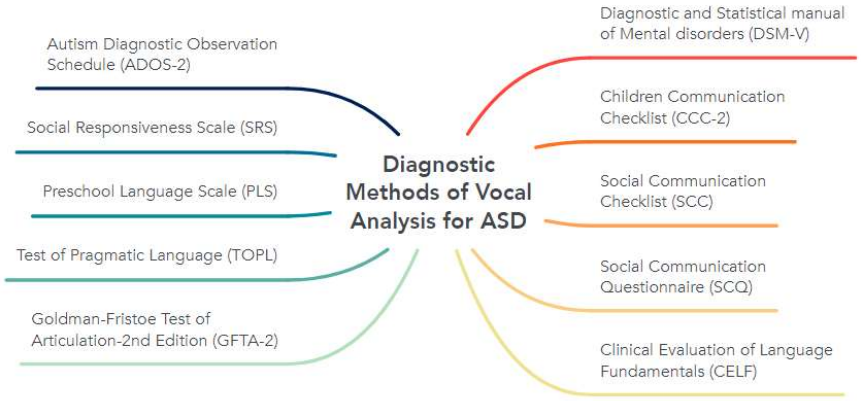

Fig 4. Clinical tools and diagnostic practices used by speech-language pathologists for analyzing vocal and communication behaviors in children with ASD

Another key tool, the Test of Pragmatic Language (TOPL-2) is a standardized assessment evaluating pragmatic language skills, such as conversation initiation and maintenance, gesture use, and comprehension of nonliteral language. In terms of speech articulation, the Goldman-Fristoe Test of Articulation-2nd Edition (GFTA-2; [75]) is a standardized evaluation tool used to assess the articulation skills of children and adults. It is designed to measure the ability to produce individual sounds, as well as the overall intelligibility of speech. The GFTA-2 assesses the production of 43 consonant sounds and provides a standardized score for each sound, as well as an overall score for articulation ability. The test can be used to identify speech sound disorders, plan treatment, and monitor progress over time. The GFTA-2, widely used by speech-language pathologists, is suitable for individuals aged 2 to 21 and typically takes 15 to 20 minutes to administer. Published by Pearson Assessments, it's one of the most popular articulation assessments in speech-language pathology.

All these speech assessment methods are manual processes and dependent on assessment by a SLP, which is both time consuming and cost intensive. Clinicians typically rely on longer samples of spoken language across various semantic, syntactic, and grammatical contexts for assessments. Machine learning-based speech processing methods enhance efficiency, which can aid in the early detection of autism and, consequently, early interventions for autistic children. For example, the Language ENvironment Analysis (LENA) system offers an automated approach to assess vocal development in preschoolers with ASD. This system enables the identification and analysis of a child's vocalizations from recordings made throughout an entire day in natural settings, with durations of up to 16 hours. LENA software, available for both research and commercial use, can derive several indices reflecting a child's vocal development, as demonstrated in studies [76], [77], [78].

5.0 EXPERIMENTAL PROTOCOLS FOR AUTISM DETECTION USING SPEECH DATA

The experimental protocol for autism detection through speech includes standardized practices such as participant numbers, selection criteria, tasks, recording device specifications, placement, and duration. This section examines voice-based data acquisition protocols used in autism detection studies and explores deep neural network approaches.

Participants: These studies vary significantly in participant numbers. Table 3 summarizes existing research, including subject count, gender, age, and vocal feature details. Many studies involve relatively few participants, making it challenging to determine the validity and significance of results for establishing a universal vocal biomarker. Gender-wise, most studies include both male and female participants.

Selection Criteria of Participants and Tasks: The crucial stage of identifying participants for vocal biomarkers-based autism detection involves utilizing both quantitative (software-based speech analysis) and qualitative (clinical questionnaire-based) methods to gather data. For voice recording in speech disorder studies, diverse clinical questionnaire-based pre-testing techniques, detailed in Section IV, are used to choose appropriate participants. In studies that utilize public datasets for speech-based analysis using deep learning techniques, the participants are usually pre-identified as ASD children through the database.

Table 3. Participants information in voice-based experiments for ASD detection where F for female and M for male is used

Study	Subjects	Gender	Age Group (Mean, SD)	Other Information	Type of Detection
[85]	81 participants (ASD: 30, TD: 51)	-	-	Comparison between machine learning-based voice analysis and evaluations by 10 speech therapists.Single-word utterances used as stimuli.	Machine learning-based voice analysis compared to auditory evaluations by speech therapists.
[86]	116 ASD videos, 46 TD videos	-	ASD (mean=4.83, SD=2.25), TD (mean=2.92, SD=1.16)	Behavioural features assessed using a mobile web portal and measured by three blinded raters.	Mobile detection of ASD using home videos
[87]	1876 instances	625 M, 477 F	Child Dataset (mean=6.35, SD=2.36), Adolescents Dataset (mean=14.1, SD=1.57), Adult Dataset (mean=29.7, SD=16.5)	ASD screening using an app (ASDTests)	ASD Screening
[88]	24 (RAVDESS), 2 (TESS) and 91 (CREMA-D)	12 M, 12 F (RAVDESS), 0 M, 2 F (TESS) and 48 M, 43 F (CREMA-D)	21 -33 years old (RAVDESS), 26 & 64 years old (TESS) and 20 - 74 years old (CREMA-D)	-	Emotion Recognition in ASD children
[89]	15	11 M, 4 F	7 - 12 years	Introduced a new dataset featuring 15 children with ASC in a setting where child interacts with the robot	Echolalic Vocalisations Recognition in Autistic children
[90]	1500	-	3-6 years	CUChild127 database of Cantonese-speaking preschool children	Speech Sound Disorder
[91]	191 children (126 ASD, 65 TD)	ASD:78 M, 25 F; TD: 48 M, 40 F	9 - 42 months (mean age = 27.39 months, SD = 9.11 months)	Audio data from ASD diagnoses at SNUBH (2016-2018) and Living Lab (2019-2021). ASD detection from voices of the children without directly getting the specific characteristics.	ASD Detection

Table 3. Continued

Study	Subjects	Gender	Age Group (Mean, SD)	Other Information	Type of Detection
[92]	58 children 77 videos (ASD: 20, TD: 38)	19 M, 1 F (ASD) & 15 M, 22F (TD) & 1 unspecified	3-12 years	A new dataset of child speech audio recorded via cell phones, collected from a mobile game “Guess What?" developed by Stanford	ASD Detection
[93]	72 Hebrew speaking children (ASD: 56, TD:06, DD:10)	63 M, 9 F	ASD (mean age = 50.3, SD = 14.8), TD (mean age = 38.3, SD = 14.9), DD (mean age =56.2, SD =15.9)	-	ASD Severity
[94]	120 subjects	57 M, 63 F	-	The corpora were extracted from 60 phone calls. This is the INTERSPEECH 2013 Computational Paralinguistics Challenge dataset	Classification of Emotions and ASD
[95]	70 children (ASD: 58)	60 M, 10 F	26 - 125 months (mean age = 56.31)	Audio database gathered from actual ADOS Module 2 screening sessions.	Severity Level of Atypical Prosody in children
[96]	118 children (ADOS Module 2) & 71 children (ADOS Module 3)	105 M, 13 F (ADOS Module 2), 64 M, 7 F (ADOS Module 3)	Age: 26 - 142 months, mean age = 57.77 (ADOS Module 2) & Age: 24 - 166 months, mean age = 91.95 (ADOS Module 3)	-	Speech and Language abnormalities of ASD Children
[97]	199 patients (ASD: 56, TD: 143)	-	-	Data was taken from semi-structured and unstructured medical forms	ASD Detection using NLP
[98]	1758 subjects	-	-	Dataset taken from Fadi Thabtah based on AQ-10, Recommendations of the features contributing the most towards ASD were made	Early Prediction of ASD
[99]	33 ASD children	-	-	Vocal and verbal cues were collected	ASD Severity Estimation
[100]	39 Infants	23 M, 16 F	6 - 24 months	eGeMAPS Speech Feature Dataset	ASD Detection in Infants

When collecting data from participants for deep learning techniques aimed at diagnosing autism based on speech patterns, some studies have used sensors and devices instead of traditional diagnostic methods to identify children with autism [79], [80]. Past studies have mainly utilized machine learning or deep learning on static autism datasets, assessing their effectiveness primarily through accuracy, sensitivity, and specificity metrics [81].

6.0 PUBLIC SPEECH DATASETS FOR AUTISM ASSESSMENTS

After a thorough investigation, it was found that publicly available speech datasets for detecting autism are limited, with many research groups developing and using their own datasets due to the scarcity of accessible data. Table 4 details the publicly available speech datasets used for ASD detection.

The DE-ENIGMA Database (Riva et al., 2020, [82]): In the DE-ENIGMA studies on teaching recognition of emotions, 128 children aged 5 to 12 (including 19 females) from Britain and Serbia participated. They were randomly allocated to either robot-assisted or adult-assisted activities according to steps 1 through 4 of the emotion training programs "Teaching Children with Autism to Mind Read" [83]. Each child attended 4-5 sessions, recorded with various audio, video, and depth recording devices. Parental consent for inclusion in the database was obtained for 121 children. The DE-ENIGMA database is poised to become the largest dataset for studying autistic interactions, offering valuable audio and video recordings for exploring various aspects of autism-related behavior, such as child–robot interactions, emotion recognition, and cross–cultural comparisons. Additionally, it serves as essential training data for machine learning research in autism, streamlining participation in this field and potentially saving significant time and resources.

The TalkBank Database: The TalkBank is a digital repository of language data that includes recordings and transcripts of conversations, interviews, and other spoken interactions. The TalkBank database is an interdisciplinary resource that is used by researchers from fields such as linguistics, communication sciences, psychology, and learning to study various aspects of language use, acquisition, and processing. The TalkBank database includes a wide range of language data, such as recordings of child language development, bilingual conversations, second language learning, aphasia and other language disorders, and typical adult language use. The data is annotated with linguistic and behavioral codes to facilitate analysis and is available to researchers for free through the TalkBank website. TalkBank also includes tools for analyzing language data, such as programs for creating and aligning transcripts, automated analysis of prosody and speech acoustics, and tools for conducting statistical analyses on linguistic data.

CHILDES (Child Language Data Exchange System): The TalkBank database houses CHILDES (Child Language Data Exchange System), encompassing children's speech across various conditions (e.g., Autism, Down's syndrome, hearing impairment) and languages (e.g., English, Dutch, Greek, Mandarin**)** [84]**.**

ASDBank: The TalkBank database features ASDBank, containing data related to language development both from children and adolescents with autism, using similar methods and analyses as the CHILDES database [101].

UCI (University of Irvine, California) Machine Learning Repository: The UCI Machine Learning Repository provides a range of resources, including databases, domain theories, and data generators, which the machine learning community uses for empirical algorithm analysis [102]. Among its datasets are three related to autism, each containing various instances. These datasets, collected globally via the 'ASDTests' mobile app by Dr. Fadi Fayez, cover different age groups: children (4–11), adolescents (12–17), and adults (18+). While these datasets offer valuable insights, some data may be incomplete, which could potentially affect clinical decisions.

Bergelson SEEDLingS Homebank Database: SEEDLingS investigate how environmental factors and early linguistic input influence infants' language development, focusing on word learning in babies aged 6 to 18 months. By analyzing their surroundings, researchers aim to understand how infants process and organize words and objects **[103]**. The dataset combines lab studies with home audio and video recordings to track language growth over time for 44 infants. Currently, available data includes audio recordings from visits at 6 and 7 months, with plans to add data from 8–17-month visits in the future. The sample primarily consists of middle-class families from English-speaking environments with no known vision or hearing issues. Collected at the University of Rochester, this dataset has potential applications for autism detection in infants.

Multimodal Dyadic Behavior Database (MMDB): The Multimodal Dyadic Behavior (MMDB) dataset is a valuable resource for studying toddlers' social and communicative behavior, featuring video, audio, and physiological data from 160 sessions of semi-structured play interactions between trained adults and children aged 15 to 30 months. Designed to elicit social attention and non-verbal communication, this dataset focuses on

milestones relevant to ASD [104]. The audio data, in particular, can be utilized for ASD detection using vocal biomarkers.

Table 4. Publicly available speech datasets used for autism detection

Dataset	Data Type	Subjects	Other Information
De-ENIGMA Database	Multi-Modal Database	121	Data was collected through studies conducted on emotion learning of autistic children with a humanoid robot
CHILDES Database - TalkBank Project	Speech Recordings	54	Contains spoken language multimedia data of children and adolescents
ASDBank – TalkBank Project	Speech Recordings	86	Contains spoken language multimedia data of children and adolescents
UCI Machine Learning Repository	Multivariate	396	Results from the AQ-10 child test were used as the datasets. This dataset is related to classification and predictive tasks.
Bergelson SEEDLingS Homebank Database	Speech Recordings	44 subjects, 87 instances	These recordings were collected monthly at home for 44 infants aged 6-18 months, along with some pilot data. The collection comprises audio recordings of a whole day, hour-long video recordings of one hour, and in-lab eye tracking data.
MMDB (Multimodal Dyadic Behavior) Database	Multimodal	160 instances	This collection offers unique multimodal recordings (video, audio, and physiological) capturing toddlers' social and communicative behaviour.
Infant Brain Imaging Study (IBIS)	MRIs, Home Audio Recordings	503	This study, directed by researchers across four clinical sites and a center at the Montreal Neurological Institute (MNI), seeking to perform a MRI/DTI and communicative analysis of infants at high risk for autism—specifically, their siblings—at 6, 12, 24, and 36 months of age.

Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS): The RAVDESS database consists of validated speech and song recordings from twenty-four professional actors (12 females, 12 males), each delivering lexically coordinated statements in a neutral North American accent. The recordings include expressions of happiness, calmness, surprise, sadness, disgust, anger, and fear for speech, and calmness, anger, happiness, sadness, and fear for song [105]. Each emotion is presented at two intensity levels, plus a neutral expression, across face-voice, face-only, and voice-only formats. Ratings from 247 evaluators (7,356 recordings) and test-retest data from 72 participants confirm high emotional validity, intensity, and reliability. Corrected accuracy and composite "goodness" measures are provided to aid stimulus selection. All recordings are freely available under a Creative Commons license.

Infant Brain Imaging Study (IBIS): The Infant Brain Imaging Study (IBIS) is a collaborative research project investigating early brain development in infants and children with autism to identify and understand early brain changes that could support early diagnosis and treatment [106]. Led by investigators at four clinical sites—

University of North Carolina (UNC), University of Washington (UW), Washington University (WU), and Yale University—and coordinated at the Montreal Neurological Institute, the study provided developmental assessments and reimbursed expenses for participating families. It focused on high-risk infants, primarily siblings of individuals with autism, and included MRIs at 6, 12, and 24 months, along with home audio recordings at 12- and 15-months using LENA devices.

7.0 INVESTIGATING AUTOMATIC AUTISM DETECTION METHODS

Fig. 5 shows the general deep learning framework used to deal with automatic autism assessment. It includes data pre-processing, feature extraction and classification or prediction task. In the following section, we will review each step of this general pipeline and study the different proposed approaches in each step.

7.1 Pre-Processing Audio Data

Pre-processing audio data is a critical measure in detecting autism in individuals. By utilizing these pre-processing techniques, it is possible to enhance the efficiency of speech-based autism detection methods [107], [108]. The first step in pre-processing audio data is to remove any background noise that may be present. This part is known as filtering. This can be achieved by using noise reduction techniques such as spectral subtraction or Wiener filtering. Studies have shown that eliminating silent pauses and noise can enhance the performance of ML/DL models [109]. Text processing for speech/audio synthesis is even more intricate and includes tasks such as text normalization, tokenization, and sentence segmentation [110]. Therefore, improving the system's performance is just as critical in speech synthesis as it is in speech recognition.

7.2 Feature Extraction

To reduce computational complexity and analyze specific features, the data can be segmented into smaller sections or frames. Feature extraction techniques can then be applied to the segmented data to obtain relevant information such as pitch, energy, and spectral features. As illustrated in [111], this standard pre-processing and feature extraction process includes techniques like MFCC. These features serve as inputs for machine learning models to classify autism. Additionally, normalization and scaling methods can be used to standardize the data and prepare it for analysis.

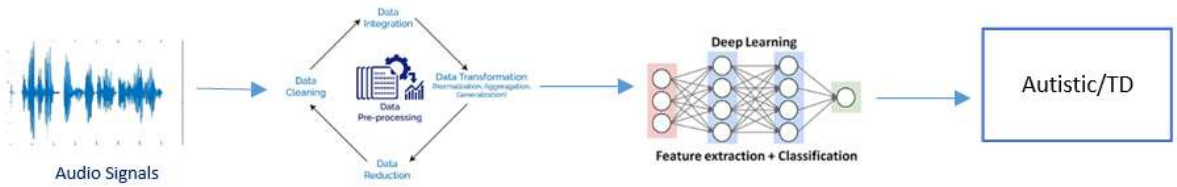

Fig 5. A generalized deep learning pipeline for speech-based ASD detection, consisting of data preprocessing (e.g., noise removal), feature extraction (e.g., MFCC, prosody analysis), and classification using neural network models.

Overall, pre-processing audio data plays a crucial role in accurately detecting autism in individuals. Some of these feature extraction techniques are given below:

Prosody analysis: This involves analyzing features of voice such as intonation, frequency, and rhythm to identify abnormalities that may be associated with autism.

Formant analysis: This technique involves analyzing the resonance frequencies in the speech signal to identify potential indicators of autism.

Mel-frequency cepstral coefficients (MFCCs): This is a common technique used in speech recognition and involves extracting spectral features of the speech signal that are sensitive to changes in the human voice.

Pitch tracking: This involves analyzing the fundamental frequency of the speech signal, which can be used to identify potential indicators of autism.

Speech rate analysis: This technique involves analyzing the rate at which a person speaks, as individuals with autism may exhibit abnormalities in their speaking rate. Overall, feature extraction techniques for autism detection from audio/speech can vary with respect to the specific features of the speech signal being analyzed. It is important to use a combination of techniques to achieve higher accuracy in detecting autism.

7.3 Machine Learning-Based Speech Assessment Methodologies for Autistic Children

Many researchers are working to introduce early-detection technologies using machine learning algorithms and artificial intelligence for accurate and timely diagnosis of ASD in children [61]. However, research on speech assessment using machine learning approaches for ASD diagnosis remains limited. The study in [2] delves into machine learning applications for ASD diagnosis, aiming to enhance screening efficiency and accuracy. The authors weigh the merits and drawbacks of these methods, noting a flaw in current ASD screening tools tied to DSM-IV in comparison with DSM-5 criteria. They advocate for renewing testing tools to align with DSM-5 classification, emphasizing the need for revised diagnostic algorithms.

In [85], ML-based voice analysis was compared to speech therapists' judgments in distinguishing ASD in adolescents through one-word utterances. With 81 participants (51 typically developing, 30 ASD), aged 3-10 and without comorbid issues, SVM was applied on 24 features. SVM showed superior accuracy (76%) in ASD identification compared to speech therapists. This study highlights the potential of ML algorithms in voice prosody analysis as a valuable screening tool for ASD. A recent study at [91] introduced a voice-based method for detecting ASD in children, employing an end–to–end neural network model. This approach, unlike traditional methods, doesn't explicitly extract deterministic features. Rather, it merges two feature-extraction models with a bidirectional long short-term memory (BLSTM) classifier to determine whether voices indicate ASD or typical development (TD), offering a probability score in the process.

Early diagnosis can also be facilitated through assistive technologies and the Internet of Things (IoT), which leverage machine learning algorithms and deep learning to enhance diagnosis and patient care. The paper at [112] systematically reviews IoT-based approaches for ASD, highlighting key research trends and presenting a technical taxonomy to categorize existing methods and algorithms. It also includes a statistical and functional analysis of these approaches, focusing on evaluation metrics like accuracy and sensitivity. Authors in [86] used machine learning (ML) to diagnose ASD by classifying home videos of children. They employed eight ML algorithms on 62 two-minute-long home videos of American children with and without ASD. The study aimed to assess the accuracy of ML algorithms in identifying ASD in a digital setting, achieving a 92% accuracy rate. Additionally, they attained an 85% accuracy (AUC) and 76% sensitivity for diagnosing ASD in children compared to those with other developmental delays, as well as distinguishing atypical from developmentally delayed children. These findings advocate for the adoption of a mobile video-based and ML-driven methodology for early and remote autism detection in children, including those in Bangladesh.

In [113], the authors investigated the speech patterns of autistic children in contrast to typically developing (TD) children. The focus was on analyzing a specific set of indicative and interrogative sentences spoken by both groups and comparing various aspects such as intonation pattern, pitch, amplitude, duration, intensity, and tilt. The primary goal of the research was to assess how the intonation pattern, pitch average, amplitude, duration, intensity, and tilt varied between children with ASD and TD children. The collected data revealed that while the amplitude, duration, intensity, mean pitch, and tilt showed similarities between the ASD and TD groups, there were significant differences in the intonation pattern, mean pitch (linked to the severity of autism), and amplitude, specifically in indicative sentences.

To elaborate further, the findings indicated that autistic children exhibited a monotonous intonation pattern and difficulties in producing meaningful indicative and interrogative sentences, particularly in relation to these three mentioned features.

7.4 Deep Learning Automatic Autism Detection

To conduct a literature review on automatic autism detection using deep learning, it is essential to first identify the different categories of deep learning approaches, such as classification, regression, or generation. Based on [114] and [115], deep learning methods can be broadly classified into four categories: discriminative models, generative models, representative models, and hybrid models. This section also reviews studies employing various neural network architectures, either discriminative or generative, for automatic autism detection.

The primary architectures considered are Convolutional Neural Networks (CNNs), with a preference for multimodal approaches due to their higher accuracy compared to single-method approaches, as shown in Fig. 6. Additionally, this section examines studies that utilize Siamese neural network models for automatic autism detection using speech data. The influence of deep learning has been expanding in the field of research, including speech analysis [116] and classification of ASD [117], [118]. Deep learning techniques, specifically CNNs and RNNs, as well as the BLSTM model have been applied or recommended as methods for detecting autism in children, according to sources [96], [119], [120], [121], [122].

Traditional ASD screening methods are time-consuming, hindering early detection and raising costs. In a recent study in [87], researchers proposed an automated approach using deep neural networks (DNNs) to detect autistic traits objectively, replacing subjective scoring. The DNNs were trained on labeled cases and controls, achieving promising results in root mean square error (RMSE), specificity, sensitivity, and accuracy (refer to (1), (2), and (3)) through ten-fold cross-validation.

$$Accuracy = \frac{True\ Positives + Tru\quad Negatives}{n} \tag{1}$$

$$Sensitivity = \frac{True\ Positives}{True\ Positives + Fal\quad Negatives} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Target\ Output_i - Actual\ Output_i)^2}{n}} \tag{3}$$

Comparative analysis with other machine learning algorithms revealed the superiority of DNNs. Integrating deep learning into ASD screening methods could enhance early identification, facilitating approach to vital support for affected role and families and promoting their overall health.
Here are some of the deep learning techniques that have been used:

7.4.1 Convolutional Neural Networks (CNNS)

Convolutional Neural Networks (CNNs) are a widely acknowledged deep learning architecture known for their exceptional performance across diverse applications. Comprising convolutional layers followed by pooling layers, CNNs enable robust learning of input data representations at various levels of abstraction in the feature hierarchy. Recent research has demonstrated CNNs' effectiveness in autism detection through vocal data analysis, where they automatically learn features from raw audio signals and predict autism status using one or multiple classifiers [96], [123], [124].

CNNs and their variations perform exceptionally well in speech-related tasks such as language classification, narrator verification, and sentiment identification. Traditionally, CNNs for audio utilize fixed-duration segments, but [123] advocate for using features from the entire speech signal. They propose integrating a spatial pyramid pooling (SPP) layer into the CNN, removing the limitation of having fixed length segments and allowing the CNN to incorporate features from signals having varying lengths during training. This method yields diverse feature maps from the convolution layer. Additionally, they introduce a CNN-based segment-level pyramid match kernel (CNN-SLPMK) for classification using support vector machines (SVMs), achieving similar results to state-of-the-art techniques in emotion recognition.

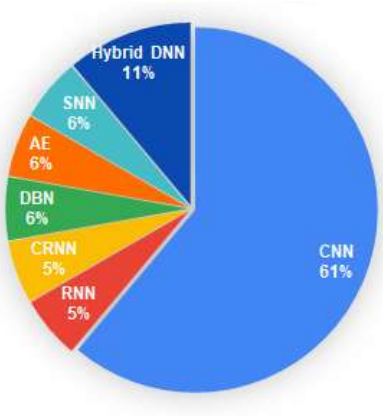

Fig 6. Overview of neural network architectures employed across reviewed studies for automatic autism detection. This includes Convolutional Neural Networks, Recurrent Neural Networks, Siamese Networks, and hybrid models, with a breakdown of their respective use cases and performance in vocal biomarker analysis.

In [96], the authors propose methods for detecting stereotyped idiosyncratic phrases and unusual rhythmic pattern in speech related to ASD. For atypical prosody, both hand-crafted feature-based methods and end-to-end deep learning frameworks are suggested, along with cross-validation and score-level fusion techniques for enhanced performance. Unweighted average recall (UAR) was used as a performance evaluation measure (Refer to (4)).

$$UAR = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i'}{N_i} \tag{4}$$

where n, Ni, N'i denote the number of classes, the number of samples that belong to class i and the number of correctly classified samples that belong to class i, respectively.

To detect stereotyped idiosyncratic phrases in speech transcripts, the method involves extracting text features using language models, dependency treebanks, TF-IDF, and LIWC software, followed by an SVM backend. The authors assembled a database of spontaneous Mandarin speech captured during Autism Diagnostic Observation Schedule (ADOS) Module 2 and Module 3 sessions, comprising 118 and 71 children, respectively. Experimental results on this database demonstrate efficient prediction of unusual rhythmic patterns and labelled characteristic phrases code for young children at risk of ASD, with unweighted accuracy of 88.1% and 77.8%, respectively.

In [95], the authors devised an automated system to quantify atypical prosody in young children with ASD. They utilized the openSMILE toolkit to extract acoustic features and employed a SVM as a reference. Additionally, they proposed various deep neural network setups to model atypical prosody directly from speech spectrograms. Combining deep learning with the baseline enhanced system performance. Using speech recordings from ADOS tasks, they achieved up to 90% accuracy. Table 5 shows the results from this study. Their study included 70 children, 58 diagnosed with ASD, showing effective prediction of atypical prosody scores for at-risk young children.

The study at [92] investigated machine learning's potential in autism detection from children's speech, considering the prevalent prosody abnormalities in autistic children. The study employed RFs, CNNs, and fine-tuned wav2vec 2.0 on a unique dataset of child speech recorded through a cellphone using a mobile game app capturing autism and neurotypical children's natural home interactions. Results showed 70% to 79% accuracy rates in classifying ASD or NT audio, demonstrating machine learning's efficacy in autism detection without specialized equipment, despite inconsistent recordings. The study used 5-fold cross-validation and collected 77 videos of 58 children (20 ASD, 38 NT) aged 3 to 12 over four years (2018-2021). Parents provided the child's details and consented to video sharing for research purposes. The performance of the Random Forest, 8M CNN, and wav2vec 2.0 models, each trained and evaluated on the Guess What? audio dataset, and was assessed using ROC curves and confusion matrices (Fig. 7 (A–F)).

Table 5: Three categories classification results on testing set from [95] (UAR(seg) stands for calculating UAR with respect to segment)

Model	Inputs	UAR (seg)
SVM	OpenSmile features	50.5%
CNN + RNN	STFT spectrogram	34.62%
	CQT spectrogram	35.48%
RNN	STFT spectrogram	45.62%
	CQT spectrogram	36.56%

In [97], the workers developed a machine learning (ML) model employing Natural Language Processing (NLP) to detect ASD. They converted hand-written medical forms to digital formats, pre-processed the data, and conducted classification, significantly streamlining the ASD detection process. Out of 199 digital forms examined, 56 were confirmed ASD cases (positive samples), while 143 were non-ASD cases (negative samples).

Previous research indicates that speech difficulties in children with ASD may offer insights into the severity of their condition. Authors in [93] examined speech recordings of Hebrew-speaking children who underwent ADOS assessments, extracting prosodic, acoustic, and conversational features. They identified 60 features from recordings of 72 children, finding correlations between certain features (e.g., pitch variability, Zero Crossing Rate) and ADOS scores. Utilizing DNN algorithms, with a CNN yielding the best results, they achieved a mean RMSE of 4.65 and a mean correlation of 0.72 with true ADOS scores. Additionally, [98] introduced a hybrid deep learning and Explainable Artificial Intelligence (XAI) approach for early and precise ASD prediction. The explainable model g(x') with an input x' for an original model f(x) and input variables for a model x = (x1, x2, ...xn) was used in this study (refer to (5)).

$$f(x) = g(x') = \varphi_0 + \sum_{i=1}^{n} \emptyset_i x_i' \tag{5}$$

where n is the number of input features and φ_0 is the constant value when all input values are missing. Their framework not only enhances prediction accuracy but also provides recommendations for clinicians, aiding in earlier identification of ASD traits in toddlers.

The work in [99] examined language patterns in children with autism using Long Short-Term Memory (LSTM) networks, correlating them with ASD symptoms. They analyzed conversations between children with autism and diagnosticians, predicting Calibrated Severity Scores (CSS) of Social Affect (SA) in the ADOS-2 assessment. Annotated samples from 33 ADOS-2 interviews were utilized to label 2000 2-second audio clips as "Adult", "Child", "Both", or "Irrelevant". Spectrograms were generated and CNN was trained to filter out unrelated acoustics. A batch size of 32 spectrograms was chosen. The predictions in this study were defined as the unweighted average over the collection of predictor trees as shown in the equation (6) below, where $h(x; \theta_k)$, *k = 1, ..., ntree* are the collection of the tree predictors and x represents the observed input variable vector of length *mtry* with the associated i.i.d random vector θ_k.

$$\bar{h}(x) = \left(\frac{1}{ntree}\right) \sum_{k=1}^{ntree} h(x; \theta_k) \tag{6}$$

This pipeline framework achieved superior predictive indicative evaluations of ASD severity in comparison to other algorithms like Language Environment Analysis (LENA), as measured by R^2.

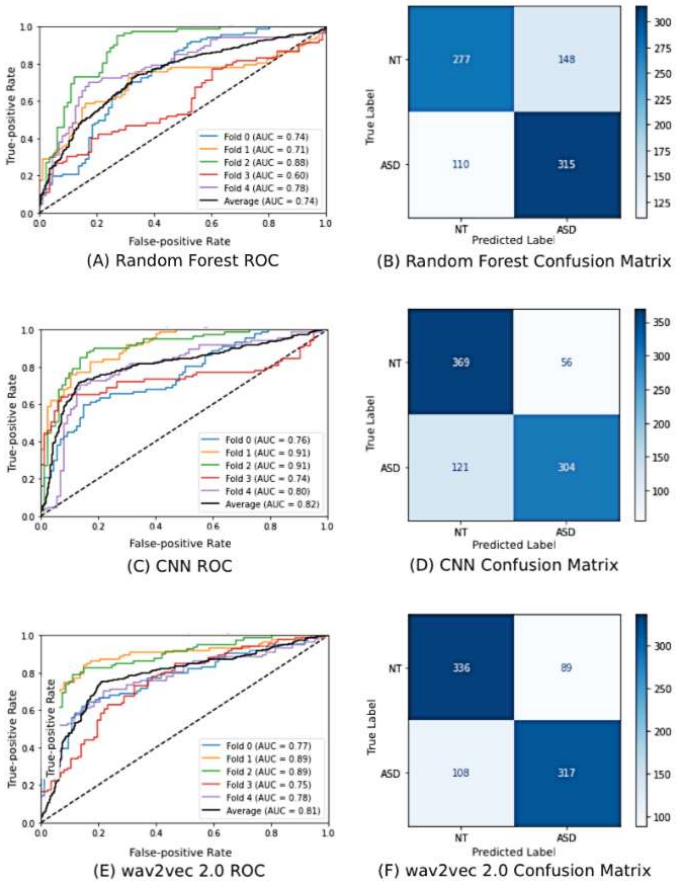

Fig 7. (A) Receiver Operating Characteristic (ROC) curve for the Random Forest model; (B) Corresponding confusion matrix for Random Forest; (C) ROC curve for the 8M Convolutional Neural Network (CNN); (D) Confusion matrix for the 8M CNN; (E) ROC curve for the wav2vec 2.0 model; and (F) Confusion matrix for wav2vec 2.0.

7.4.2 Recurrent Neural Networks (RNNS)

Recurrent neural networks (RNNs) are a type of deep learning algorithm specialized in handling sequential data. Unlike standard models like multilayer perceptron (MLPs) or CNNs, RNNs consider past information during training, crucial for sequences where input order matters. They excel in remembering information from previous samples and learning sequential patterns, making them ideal for tasks like analyzing vocal biomarkers in speech recordings to predict autism status.

In [88], an emotion recognition system based on speech was developed to assist autistic children in understanding human sentiments during social exchanges. This system employs machine learning and deep learning techniques, utilizing an ensemble learning approach combining multiple algorithms for real-time speech recognition. Models including support SVM, MLP, and RNN were trained on datasets like RAVDESS, TESS, CREMA-D, and a custom dataset. Two sets of audio features were compared, and a facial expression recognition model was integrated for multimodal processing, enhancing emotion prediction accuracy. This study's speech processing aspect offers a foundation for extending ASD diagnosis and identifying vocal biomarkers from audio data.

7.4.3 Convolutional Recurrent Neural Networks (CRNNS)

CRNNs are employed to analyze vocal biomarkers from speech recordings, capturing both spectral and temporal characteristics. Utilizing CNNs, spectral features are extracted, followed by RNNs to process temporal dynamics and a classifier for autism prediction.

In [89], a 2D CNN integrated with a bidirectional LSTM network was used to recognize echolalia, with log Mel-spectrograms as input for spectral feature extraction. Recurrent layers learned long-term temporal context, and a feedforward neural network classified the dataset. Data from The DE-ENIGMA, Horizon 2020 initiative, was utilized, and leave one subject out cross-validation achieved 83.5% classification accuracy. Fig 8. presents the confusion matrix of the best classification result of this study.

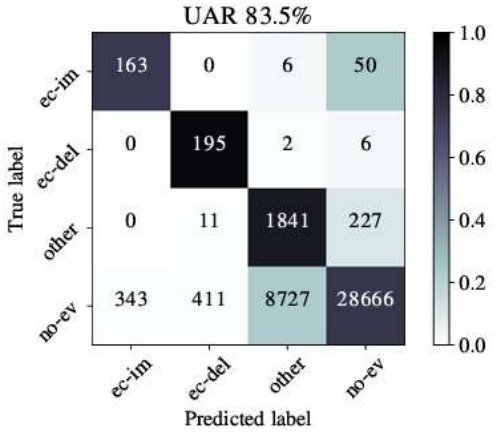

Fig 8. Confusion matrix of the best classification result obtained by CRNN3 [89]

7.4.4 Deep Belief Networks (DBNS)

Deep Belief Networks (DBNs) are utilized to learn hierarchical representations of vocal biomarkers extracted from speech recordings. This involves training a DBN to capture increasingly abstract representations of the speech signal, followed by a classifier for predicting autism status.

In a recent study at [125], an XAI-based ASD diagnosis (XAI-ASD) model was introduced, aiming for accurate ASD detection and classification. The proposed technique incorporates Bacterial Foraging Optimization (BFO) for feature selection and the Whale Optimization Algorithm (WOA) with the DBN model for classification. Hyperparameters of the DBN model are finely tuned using WOA to improve classification performance. Extensive simulations on an ASD dataset were conducted to evaluate diagnostic effectiveness.

7.4.5 Autoencoders (AE)

Autoencoders have been used to learn compressed representations of vocal biomarkers extracted from speech recordings. This approach involves training an autoencoder to learn a compressed representation of the speech signal and then using a classifier to predict autism status based on this representation.

The study in [100] presented a deep learning approach for detecting ASD, utilizing various models. They employed a auto-encoder for feature extraction which was a pre-trained model and a joint optimization strategy to enhance strength for broadly dispersed and raw data. When applied to the eGeMAPS speech feature dataset, this method resulted in better ASD detection performance in children in contrast to using the raw dataset on its own. Fig 9. shows each data distribution as a two-dimensional scatter plot.

Audio data were collected from ASD diagnoses at SNUBH between 2016 and 2018, with IRB approval for analysis. The dataset included evaluations of 39 infants aged 6 to 24 months, diagnosed by a psychiatrist for children using DSM-5 criteria. Diagnoses were based on various assessment tools such as ADI-R, BeDevel-I, ADOS-2, BeDevel-P, K-CARS, SCQ, and SRS, with an average age of 19.20 months (SD = 2.52).

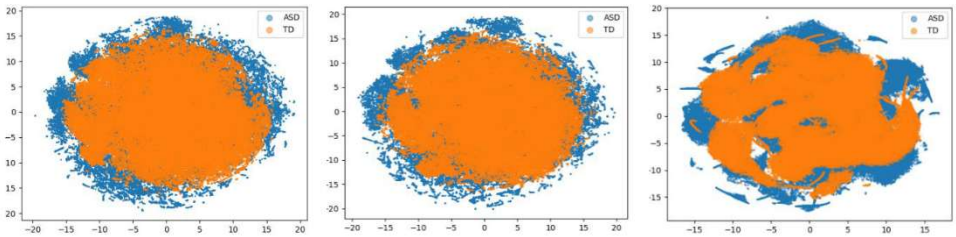

Fig 9. Two-dimensional scatter plot for (a) eGeMAPS-88, (b) eGeMAPS-54, and (c) the AE processed by t-stochastic neighbor embedding (t-SNE) [100]

7.4.6 Recurrent Siamese Neural Network (RSNN)

Automatic detection of speech sound disorders (SSD) in children with or without ASD using acoustics-based assessment is highly desirable. However, the accessibility of a properly annotated disordered speech dataset is critical for developing an accurate automatic speech assessment system, which is particularly challenging for autistic children's speech.

The authors in [90] introduced a child speech disorder detection system trained solely on normal speech. Using a Siamese recurrent network, the system learns pronunciation resemblance and difference between phone pairs in an embedding space. To detect speech sound disorders, the network measures the distance between test and desired phones, training a binary classifier. Incorporating speech attribute features, it assesses pronunciation quality and provides diagnostic feedback. Results show the Siamese recurrent network, combining speech attributes and phone posterior features, achieves 0.941 detection accuracy. Focusing on CUChild127, a database of child speech, it aims to aid research on Cantonese-speaking preschoolers' speech sound disorders, comprising recordings from 1,500 children (aged 3-6), including 310 identified with SSD. Each child read 127 Cantonese words during recording, covering consonants and vowels. This work holds potential for extension to other languages.

7.4.7 Hybrid Deep Neural Network Approaches (HDNN)

Hybrid Deep Neural Network (HDNN) methods enhance autism detection by combining diverse neural networks to capitalize on their strengths and mitigate weaknesses, aiming for a more precise and resilient detection model.

Researchers in [94] investigated using speech utterances to classify emotions and ASD, employing ensemble classification techniques. The study evaluated the performance of SVM, DNN, and k-nearest neighbors (KNN), alongside auditory characteristics from openSMILE. They introduced the Acoustic Segment Model (ASM), incorporating temporal information for classification by learning ASMs for each emotion and ASD category. Combining machine learning and ASM techniques in an ensemble system, evaluations were performed on datasets from the INTERSPEECH 2013 Computational Paralinguistics Challenge organizer.

These results highlight the potential of voice acoustic parameters as a specific diagnostic tool for ASD. However, further research is needed to develop reliable and ethical systems for clinical use, addressing privacy and potential stigmatization concerns.

8.0 DISCUSSION

The identification of vocal biomarkers for the automatic detection of autism can play a crucial role in the early diagnosis of autism. Numerous deep learning approaches have been suggested and extensively explored in existing literature, as discussed in this survey paper. However, despite these advancements, several challenges and limitations persist, impeding the practical implementation of these techniques.

To improve autism detection, studies in [102] and [103] suggest using advanced machine learning models like deep neural networks and transformers. By moving beyond support vector machines and basic regression, they can explore ensemble learning or algorithm combinations for better robustness. Additionally, improving

generalization to broader, undiagnosed populations through transfer learning and training on diverse, larger datasets is crucial.

Automation and transparency need improvement in studies like [104] and [108], which lack detailed model documentation and require manual data segmentation. Addressing this by improving transparency in reporting architectures, training processes, and hyperparameters is crucial. Automating tasks like speech segmentation using unsupervised or self-supervised learning could reduce bias and enhance scalability, making models more efficient and suitable for real-world applications without manual intervention.

Some studies, such as [89] and [92], face challenges with limited datasets and the specific focus of their detection tasks. Addressing the dataset limitations could involve collecting larger and more diverse samples, particularly from different demographic groups, to balance out representation issues like gender imbalances. Moreover, expanding the feature set to capture more nuanced speech patterns, in addition to echolalia or basic speech emotion recognition, would broaden the scope of detection systems. One recommended direction is the development of cross-linguistic vocal biomarker benchmarks, enabling models to learn invariant features across languages and accents, thus improving generalizability.

Additionally, integrating real-time ASD screening tools into telehealth platforms could dramatically expand accessibility, particularly in under-resourced regions. This could be achieved by deploying lightweight versions of trained models on mobile devices, leveraging on-device processing and cloud synchronization for data analysis.

Finally, some studies, such as [97] and [98], highlight the need for external validation and clinical trials. These models, while promising, need to be validated against real-world clinical datasets to ensure their reliability and effectiveness in medical practice. Establishing partnerships with hospitals or medical institutions would provide access to clinical data that could be used to rigorously test these models. It is also essential to build shared clinical data pipelines that allow real-time annotation, consent tracking, and model feedback integration, thereby accelerating iterative improvement in deployed diagnostic tools. Additionally, incorporating multi-modal data, such as video and audio, alongside text, could improve accuracy by leveraging complementary information from different types of data. This approach would not only ensure external validation but also enhance the overall capability of ASD detection models.

The comparative performance of deep learning models in ASD detection using vocal biomarkers reveals several consistent patterns and model-specific strengths. Notably, CNNs and their extensions, such as CNN-SLPMK have demonstrated strong performance in emotion recognition and autism classification tasks, particularly when entire speech signals were used rather than fixed-length segments. The addition of spatial pyramid pooling enhanced feature diversity, leading to more accurate predictions compared to traditional CNNs.

SNNs emerged as particularly effective in scenarios with limited or imbalanced datasets, leveraging contrastive learning to model similarity between input pairs. Studies such as [90] achieved a 94.1% accuracy rate in detecting speech sound disorders in children using only normal speech data, demonstrating that SNNs can generalize well even without extensive ASD-specific training samples.

Recurrent architectures like BLSTM and CRNN also performed well in tasks involving temporal dynamics, such as echolalia recognition and atypical prosody detection. For example, [89] reported an 83.5% UAR using a CNN-BLSTM hybrid on child-robot interaction data. These models excel in capturing long-term dependencies and contextual shifts in speech, which are often critical in identifying ASD-related anomalies.

Several factors influenced model performance across studies:

Data quality and preprocessing played a critical role. Manual segmentation and non-standardized annotation pipelines reduced model scalability and reliability in some cases [91], [92].

Demographic diversity of datasets impacted generalizability. Many models struggled with gender or language bias due to skewed datasets, limiting their deployment in multicultural settings.

The choice of acoustic features also affected the outcomes. Models that incorporated broader prosodic and spectral features, such as pitch variability, Zero Crossing Rate, or intensity contours, often achieved higher correlations with clinical ASD severity scores [93].

Despite promising accuracy scores, several models lacked external validation, with few tested against real-world clinical data. This gap highlights the need for robust evaluation frameworks, standardized protocols, and collaborative data sharing initiatives to improve reproducibility and deployment.

Other than the improvements suggested above, some limitations persist for the task of autism detection based on voice data. In the following sections, we outline these limitations and propose potential solutions to address them in general.

Variability in vocal characteristics: Vocal biomarkers may be influenced by various factors such as age, gender, language, cultural background, and individual differences. This variability makes it challenging to establish universal standards for detecting autism based solely on vocal features. To address this limitation, collecting a large and diverse dataset that includes individuals from different age groups, genders, cultural backgrounds, and language abilities can help capture the variability in vocal characteristics. This enables the deep learning model to learn more robust representations of vocal features across different populations.

Lack of standardized protocols: The lack of standardized protocols for data collection, annotation, and analysis hinders the comparison of results across studies, making it difficult to establish reliable vocal biomarkers for autism detection. To address this, consensus on data collection protocols is needed, including recording conditions, participant instructions, and task prompts. Standardized guidelines for annotating vocal biomarkers, with clear definitions and criteria, should also be developed to ensure consistency. Creating benchmark datasets, well-documented and representative of the target population, will aid in comparing algorithms. Involving clinicians and experts in ASD can help refine protocols and ensure clinical relevance.

Limited sample sizes and diversity: Many studies in this field have small sample sizes, which may not adequately represent the diverse population of individuals with autism. A lack of diversity in the collected data can limit the generalizability and applicability of vocal biomarkers. One of the factors behind this is the heterogeneity of this disorder, which leads to diversity in the collected data. An effort can be made to develop inclusive criteria for participant recruitment that promote diversity and representation. Consider factors such as age, gender, cultural background, socioeconomic status, and comorbid conditions when defining the eligibility criteria and implement recruitment strategies that target underrepresented populations to ensure a more diverse sample. This can also be addressed to some extent by encouraging data sharing and collaboration among researchers in the field. This can involve the creation of data repositories or platforms where researchers can deposit and access anonymized vocal biomarker datasets. By sharing data, researchers can collectively analyze larger and more diverse datasets, leading to a better understanding of the variability in vocal characteristics across different populations.

Co-occurring conditions and comorbidities: Autism often co-occurs with other conditions, such as attention deficit hyperactivity disorder (ADHD) or language impairments. These comorbidities can affect vocal characteristics and confound the accuracy of autism detection using vocal biomarkers.

Limited Understanding of Vocal Biomarkers: While vocal biomarkers show promise in autism detection, there is still much to learn about the specific vocal characteristics associated with the disorder. The field is continuously evolving, and the understanding of vocal biomarkers in autism is still developing. This limited understanding may restrict the accuracy and reliability of deep neural network models trained on current knowledge.

Sensitivity and specificity trade-off: Balancing high sensitivity (detecting true positives) and specificity (identifying true negatives) poses a challenge, with a trade-off between them. Improving one may compromise the other, affecting autism detection accuracy.

Ethical considerations and privacy concerns: The use of vocal biomarkers raises ethical considerations regarding privacy and data protection. Collecting and analyzing individuals' voice recordings for autism detection may raise concerns related to consent, data storage, and potential misuse of personal information. To overcome this, rigorous data anonymization techniques can be implemented to protect the privacy of participants, also removing, or encrypting personally identifiable information from the collected data, such as names, addresses, or other identifying details and employing secure data storage protocols, and maintaining data confidentiality to prevent unauthorized access or breaches.

Data Availability and Quality: Deep neural networks require large amounts of high-quality data for training. However, obtaining diverse and well-annotated datasets specifically for autism detection using vocal biomarkers can be challenging. Limited availability of such datasets may restrict the performance and generalizability of the models.

Addressing these limitations requires further research and development, including larger and more diverse datasets, standardized protocols, robust validation techniques, and careful consideration of ethical and privacy concerns.

9.0 CONCLUSION

This systematic review underscores the promising role of deep neural network techniques in the non-invasive, real-time detection of autism through analysis of vocal biomarkers. The reviewed literature, comprising approximately 90 studies spanning two decades, demonstrates that DNN-based models can achieve high

performance, with diagnostic accuracies reported up to 98%, alongside high sensitivity and specificity. These results highlight the considerable potential of machine learning to transform traditional diagnostic practices in ASD by enabling scalable, speech-based screening tools.

However, the review also identifies several persistent challenges. One major concern is the generalizability of current models, largely due to the variability in vocal characteristics among individuals with ASD and the limited availability of diverse, large-scale datasets. Most existing models have been trained on relatively small and homogeneous samples, which restricts their applicability across broader populations. Moreover, many of the reviewed methods still rely heavily on manual speech data segmentation, which reduces efficiency and limits practical deployment in real-world, clinical settings.

To enhance the effectiveness and real-world applicability of these technologies, the paper recommends several key future directions. First, the development of large, demographically diverse, and standardized vocal datasets is essential for improving model training and validation. Second, integrating synthetic data augmentation strategies could help mitigate the limitations of small sample sizes. Third, the adoption of more advanced deep learning architectures – such as transformers, hybrid models, or self-supervised learning – offers new pathways to improve both accuracy and scalability. Lastly, clinical validation through collaboration with medical institutions is critical to ensure robustness and reliability before these models can be deployed as diagnostic aids.

In conclusion, while the use of DNNs for ASD detection via speech analysis is still emerging, this review provides compelling evidence of its potential. Continued advancements in data collection, model development, and clinical validation will be necessary to fully realize the promise of these tools in supporting early, accessible, and accurate autism diagnosis. Table 6 highlights the main takeaways from the reviewed literature.

Table 6: Summary table highlighting main takeaways from the reviewed literature

Study	Deep Learning Approach	Dataset/ Participants	Performance Metrics	Key Findings and Contributions	Gap/Limitations
[85]	SVM on 24 features	81 participants (51 TD, 30 ASD)	Accuracy: 76%	SVM outperformed speech therapists in ASD identification.	Limited exploration of other ML methods for comparison.
[86]	8 ML algorithms (ADTree8, ADTree7, SVM12, SVM10, SVM5, LR10, LR5, LR9,)	116 ASD videos, 46 TD videos	Accuracy: 92% AUC: 85% LR5: 89%	ML models utilizing 30 behavioural features achieved high accuracy. Sparse 5-feature LR classifier (LR5) outperformed others.	Does not specifically address voice-based detection. Limited generalization to undiagnosed populations. Further confirmation of scalability needed.
[87]	Deep Neural Networks	ASD screening app (ASDTests)	Mean RMSE: 17.94% Mean Specificity: 94.22% Mean Sensitivity: 96.72 Mean Accuracy: 95.58%	Proposed deep learning approach improves ASD screening.	Limited detail on network architecture and training.
[88]	Ensemble Learning system combining SVM, MLP & RNN models	RAVDESS, TESS and CREMA-D	Accuracy (SVM): 66.1% Accuracy (MLP): 65.7% Accuracy (RNN): 63.7% Accuracy (Ensemble): 66.5%	Developed an ensemble learning system for real-time speech emotion recognition in children with ASD. Integrated facial expression recognition for improved emotion predictions. Potential for ASD diagnosis from audio data.	Specific performance metrics and dataset details are not provided. Focuses on emotion recognition and not ASD detection.
[89]	CNN & BLSTM RNN	Introduced a new dataset of 15 Serbian ASC children in a human-robot interaction scenario	Unweighted Average Recall: 83.5%	Achieved efficient echolalia recognition using deep learning.	Specific to echolalia recognition.
[90]	Siamese RNN	CUChild127 database of Cantonese-speaking preschool children	Detection Accuracy: 94.1%	Proposed a Siamese recurrent network detects speech sound disorders using normal speech. Utilized speech attribute features.	Validation of results by applying the same approach to datasets of other languages.

Table 6: Continued

[91]	End-to-End (E2E) Neural Network	Audio data from ASD diagnoses at SNUBH (2016-2018) and Living Lab (2019-2021)	Accuracy: 71.66% Unweighted Average Recall: 70.81%	Proposed an E2E neural network for ASD identification in children's voices without deterministic feature extraction. Used two feature-extraction models and achieved improved performance.	Manual segmentation of all the prepared data. Automated methods for speech separation to be explored
[92]	Random Forest, CNN, wav2vec 2.0	A new dataset of child speech audio recorded via cell phones, collected from Stanford's "Guess What?" mobile game—an app created to crowdsource videos of both autistic and neurotypical children in their natural home settings	Accuracy (RF): 70% Accuracy (fine-tuned wav2vec 2.0): 77% Accuracy (CNN): 79%	Machine learning effectively detects autism in children's speech. Children's audio classification	Small dataset, Manual splicing of videos, Relative imbalances in the gender distribution of children in dataset
[93]	Several DNN (Multiple Linear Regression, Support Vector Regression SVR, CNN)	72 Hebrew speaking children (ASD: 56, TD:06, DD:10)	RMSE: 4.65 Correlation: 72%	Certain features such as pitch variability and Zero Crossing Rate, were positively correlated with ADOS scores, while others, such as vocal response speed and number of vocalizations, were negatively correlated.	Limited exploration of other regression models. Manual labelling of the audio data
[94]	SVM, DNN, KNN, ASM	INTERSPEECH 2013 Computational Paralinguistics Challenge dataset	UAR (Ensemble): 88.2%	Investigated classification of emotions and ASD using SVM, DNN, KNN, and ASM techniques along with acoustic features. Proposed the Acoustic Segment Model (ASM).	Detection based on emotion. Language models can be utilized for ASM for diagnosis
[95]	SVM, RNN, CNN+RNN, fuse SVM&RNN	Database of 70 children, including 58 with ASD.	Accuracy (fuse SVM &RNN): up to 90%	Developed an automated system for quantifying atypical prosody in young children with ASD.	Audio database is collected in the ADOS module 2 screening environment. Further refinement and validation of the automated system.

Table 6: Continued

[96]	Hand-crafted and deep learning methods for atypical prosody detection. SVM, CNN, RNN, CNN+RNN	Database with 118 children for ADOS Module 2 and 71 children for ADOS Module 3	Unweighted Accuracy: 88.1% for prosody, 77.8% for phrases	Efficiently predicted unusual rhythmic pattern and characteristic phrases in young children at risk of ASD. Achieved high unweighted accuracy for two classification tasks.	Further exploration of feature selection and external validation
[97]	Utilized Natural Language Processing (NLP) machine learning techniques for ASD detection from medical data.	Semi-structured and unstructured medical data of 199 patients.	Accuracy: 83.4% Recall: 91.1%	Model simplified and condensed the process of detecting ASD	Model is only dependent on the textual data and does not verify it through other means like video or audio data
[98]	Explainable deep learning approach for ASD prediction.	AQ-10	Accuracy: 98%	Developed an explainable deep learning approach for predicting ASD along with the recommendation of the features contributing the most in accuracy.	Methodology needs validation through clinical dataset
[99]	Long Short-Term Memory (LSTM) networks	33 children with autism	Predictive Diagnostic Estimates (ADOS-2 CSS of SA). Results show significant improvement in the R2 measure.	Achieved state-of-the-art ASD calibrated severity prediction using LSTM-based speech analysis. Compared favourably with existing algorithms.	Limited information provided about specific metrics and data characteristics in the excerpt. No software toolkit was used for automatic speech processing
[100]	Pre-trained Auto-encoder Model, Joint Optimization	39 Infants (Ages 6 to 24 months), eGeMAPS Speech Feature Dataset	Accuracy (AE): 68.18%	Presented a deep learning approach for ASD detection in infants using pre-trained auto-encoder model as a feature extractor and joint optimization. Improved performance compared to raw data. Utilized audio content from ASD diagnostic videos.	The reliability of the proposed method can be enhanced by incorporating more speech data from infants, refining the audio characteristics, using an auto-encoder, and employing more advanced, deeper, and contemporary model architectures.

REFERENCES

[1] S. E. Levy, “DS Man dell, and R. T Schultz,",” *Autism," The Lancet*, vol. 374, no. 9701, pp. 1627–1638, 2009.

[2] F. Thabtah, “Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment,” in *Proceedings of the 1st International Conference on Medical and health Informatics 2017*, 2017, pp. 1–6.

[3] M. J. Maenner, “Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2016,” *MMWR. Surveillance Summaries*, vol. 69, 2020.

[4] B. Rimland, “Savant capabilities of autistic children and their cognitive implications.,” 1978.

[5] J. Baio, “Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014,” *MMWR. Surveillance Summaries*, vol. 67, 2018.

[6] M. J. Maenner, “Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2018,” *MMWR. Surveillance Summaries*, vol. 70, 2021.

[7] G. T. Baranek, “Autism during infancy: A retrospective video analysis of sensory-motor and social behaviors at 9--12 months of age,” *J Autism Dev Disord*, vol. 29, pp. 213–224, 1999.

[8] R. Palomo, M. Belinchón, and S. Ozonoff, “Autism and family home movies: A comprehensive review,” *Journal of Developmental \& Behavioral Pediatrics*, vol. 27, no. 2, pp. S59--S68, 2006.

[9] E. Werner, G. Dawson, J. Osterling, and N. Dinno, “Brief report: Recognition of autism spectrum disorder before one year of age: A retrospective study based on home videotapes,” *J Autism Dev Disord*, vol. 30, no. 2, p. 157, 2000.

[10] W. A. Goldberg *et al.*, “Language and other regression: assessment and timing,” *J Autism Dev Disord*, vol. 33, pp. 607–616, 2003.

[11] P. Howlin, “Outcomes in autism spectrum disorders,” *Handbook of autism and pervasive developmental disorders*, vol. 1, pp. 201–220, 2005.

[12] S. Mitchell *et al.*, “Early language and communication development of infants later diagnosed with autism spectrum disorder,” *Journal of Developmental \& Behavioral Pediatrics*, vol. 27, no. 2, pp. S69--S78, 2006.

[13] M. K. DeMyer, S. Barton, W. E. DeMyer, J. A. Norton, J. Allen, and R. Steele, “Prognosis in autism: A follow-up study,” *J Autism Child Schizophr*, vol. 3, no. 3, pp. 199–246, 1973.

[14] M. Rutter, D. Greenfeld, and L. Lockyer, “A five to fifteen year follow-up study of infantile psychosis: II. Social and behavioural outcome,” *The British Journal of Psychiatry*, vol. 113, no. 504, pp. 1183–1199, 1967.

[15] E. L. Wodka, P. Mathy, and L. Kalb, “Predictors of phrase and fluent speech in children with autism and severe language delay,” *Pediatrics*, vol. 131, no. 4, pp. e1128--e1134, 2013.

[16] C. Kasari, N. Brady, C. Lord, and H. Tager-Flusberg, “Assessing the minimally verbal school-aged child with autism spectrum disorder,” *Autism Research*, vol. 6, no. 6, pp. 479–493, 2013.

[17] H. McConachie *et al.*, “Systematic review of tools to measure outcomes for young children with autism spectrum disorder,” 2015.

[18] D.-Y. Song, S. Y. Kim, G. Bong, J. M. Kim, and H. J. Yoo, “The use of artificial intelligence in screening and diagnosis of autism spectrum disorder: a literature review,” *Journal of the Korean Academy of Child and Adolescent Psychiatry*, vol. 30, no. 4, p. 145, 2019.

[19] C. Wu *et al.*, “Machine learning based autism spectrum disorder detection from videos,” in *2020 IEEE International Conference on E-health Networking, Application \& Services (HEALTHCOM)*, 2021, pp. 1–6.

[20] S. Cortese *et al.*, “Latest clinical frontiers related to autism diagnostic strategies,” *Cell Rep Med*, 2025.

[21] K. Gao, Y. Sun, S. Niu, and L. Wang, “Informative Feature-Guided Siamese Network for Early Diagnosis of Autism,” in *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11*, Springer, 2020, pp. 674–682.

[22] X. Zhang *et al.*, "Siamese verification framework for autism identification during infancy using cortical path signature features," in *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, IEEE, 2020, pp. 1–4.
[23] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.
[24] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun*, vol. 71, pp. 10–49, 2015.
[25] I. R. Titze, "Toward standards in acoustic analysis of voice," *Journal of Voice*, vol. 8, no. 1, pp. 1–7, 1994.
[26] C. A. M. Baltaxe and J. Q. Simmons III, "Prosodic development in normal and autistic children," *Communication problems in autism*, pp. 95–125, 1985.
[27] W. H. Fay and A. L. Schuler, "Emerging language in autistic children," *(No Title)*, 1980.
[28] W. Goldfarb, N. Goldfarb, P. Braunstein, and H. Scholl, "Speech and language faults of schizophrenic children," *J Autism Child Schizophr*, vol. 2, no. 3, pp. 219–233, 1972.
[29] J. M. Paccia and F. Curcio, "Language processing and forms of immediate echolalia in autistic children," *Journal of Speech, Language, and Hearing Research*, vol. 25, no. 1, pp. 42–47, 1982.
[30] W. Pronovost, M. P. Wakstein, and D. J. Wakstein, "A longitudinal study of the speech behavior and language comprehension of fourteen children diagnosed atypical or autistic," *Except Child*, vol. 33, no. 1, pp. 19–26, 1966.
[31] C. Baltaxe, "Acoustic characteristics of prosody in autism," *Frontier of knowledge in mental retardation*, pp. 223–233, 1981.
[32] S. J. Sheinkopf, P. Mundy, D. K. Oller, and M. Steffens, "Vocal atypicalities of preverbal autistic children," *J Autism Dev Disord*, vol. 30, pp. 345–354, 2000.
[33] R. Paul, A. Augustyn, A. Klin, and F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders," *J Autism Dev Disord*, vol. 35, pp. 205–220, 2005.
[34] E. Werner and G. Dawson, "Validation of the phenomenon of autistic regression using home videotapes," *Arch Gen Psychiatry*, vol. 62, no. 8, pp. 889–895, 2005.
[35] R. Paul, Y. Fuerst, G. Ramsay, K. Chawarska, and A. Klin, "Out of the mouths of babes: Vocal production in infant siblings of children with ASD," *Journal of Child Psychology and Psychiatry*, vol. 52, no. 5, pp. 588–598, 2011.
[36] H. Tager-Flusberg, R. Paul, and C. Lord, "Language and communication in autism," *Handbook of autism and pervasive developmental disorders*, vol. 1, pp. 335–364, 2005.
[37] S. P. Patel *et al.*, "An acoustic characterization of prosodic differences in autism spectrum disorder and first-degree relatives," *J Autism Dev Disord*, vol. 50, pp. 3032–3045, 2020.
[38] L. D. Shriberg, R. Paul, L. M. Black, and J. P. Van Santen, "The hypothesis of apraxia of speech in children with autism spectrum disorder," *J Autism Dev Disord*, vol. 41, pp. 405–426, 2011.
[39] A. Nadig and H. Shaw, "Acoustic and perceptual measurement of expressive prosody in high-functioning autism: Increased pitch range and what it means to listeners," *J Autism Dev Disord*, vol. 42, pp. 499–511, 2012.
[40] M. Kissine and P. Geelhand, "Brief report: Acoustic evidence for increased articulatory stability in the speech of adults with autism spectrum disorder," *J Autism Dev Disord*, vol. 49, pp. 2572–2580, 2019.
[41] M. Sharda *et al.*, "Sounds of melody—Pitch patterns of speech in autism," *Neurosci Lett*, vol. 478, no. 1, pp. 42–45, 2010.
[42] J. J. Diehl, D. Watson, L. Bennetto, J. McDonough, and C. Gunlogson, "An acoustic analysis of prosody in high-functioning autism," *Appl Psycholinguist*, vol. 30, no. 3, pp. 385–404, 2009.
[43] R. B. Grossman, R. H. Bemis, D. P. Skwerer, and H. Tager-Flusberg, "Lexical and affective prosody in children with high-functioning autism," 2010.
[44] A.-M. R. DePape, A. Chen, G. B. C. Hall, and L. J. Trainor, "Use of prosody and information structure in high functioning adults with autism in relation to language ability," *Front Psychol*, vol. 3, p. 72, 2012.
[45] K. Hubbard and D. A. Trauner, "Intonation and emotion in autistic spectrum disorders," *J Psycholinguist Res*, vol. 36, pp. 159–173, 2007.
[46] J. J. Diehl and R. Paul, "Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders," *Res Autism Spectr Disord*, vol. 6, no. 1, pp. 123–134, 2012.

[47] K. Ochi *et al.*, "Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder," *PLoS One*, vol. 14, no. 12, p. e0225377, 2019.
[48] J. Brisson, K. Martel, J. Serres, S. Sirois, and J.-L. Adrien, "Acoustic analysis of oral productions of infants later diagnosed with autism and their mother," *Infant Ment Health J*, vol. 35, no. 3, pp. 285–295, 2014.
[49] J. F. Santos *et al.*, "Very early detection of autism spectrum disorders based on acoustic analysis of pre-verbal vocalizations of 18-month old toddlers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7567–7571.
[50] L. McKeever, J. Cleland, and J. Delafield-Butt, "Aetiology of speech sound errors in autism," *Speech production and perception: Learning and memory*, pp. 109–138, 2019.
[51] Y. Li, C. L. P. Chen, and T. Zhang, "A survey on siamese network: Methodologies, applications, and opportunities," *IEEE Transactions on artificial intelligence*, vol. 3, no. 6, pp. 994–1014, 2022.
[52] O. Ilina, V. Ziyadinov, N. Klenov, and M. Tereshonok, "A survey on symmetrical neural network architectures and applications," *Symmetry (Basel)*, vol. 14, no. 7, p. 1391, 2022.
[53] M. Ondrašovič and P. Tarábek, "Siamese visual object tracking: A survey," *IEEE Access*, vol. 9, pp. 110149–110172, 2021.
[54] D. Chicco, "Siamese neural networks: An overview," *Artificial neural networks*, pp. 73–94, 2021.
[55] A. Nandy, S. Haldar, S. Banerjee, and S. Mitra, "A survey on applications of siamese neural networks in computer vision," in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1–5.
[56] V. Roger, J. Farinas, and J. Pinquier, "Deep neural networks for automatic speech processing: a survey from large corpora to limited data," *EURASIP J Audio Speech Music Process*, vol. 2022, no. 1, p. 19, 2022.
[57] J. McCann and S. Peppé, "Prosody in autism spectrum disorders: a critical review," *Int J Lang Commun Disord*, vol. 38, no. 4, pp. 325–350, 2003.
[58] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, and S. B. Gaigg, "Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis," *Autism Research*, vol. 10, no. 3, pp. 384–407, 2017.
[59] R. Fusaroli, R. Grossman, N. Bilenberg, C. Cantio, J. R. M. Jepsen, and E. Weed, "Toward a cumulative science of vocal markers of autism: A cross-linguistic meta-analysis-based investigation of acoustic markers in American and Danish autistic children," *Autism Research*, vol. 15, no. 4, pp. 653–664, 2022.
[60] P. Lanillos, D. Oliva, A. Philippsen, Y. Yamashita, Y. Nagai, and G. Cheng, "A review on neural network models of schizophrenia and autism spectrum disorder," *Neural Networks*, vol. 122, pp. 338–363, 2020.
[61] V. Zope, T. Shetty, M. Dandekar, A. Devnani, and P. Meghrajani, "ML based approaches for detection and development of autism spectrum disorder: A review," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2022, pp. 79–84.
[62] S. Lylath and L. B. Rananavare, "Efficient Approach for Autism Detection using deep learning techniques: A Survey," in *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, 2022, pp. 1–6.
[63] J. Schaeffer *et al.*, "Language in autism: domains, profiles and co-occurring conditions," *J Neural Transm*, vol. 130, no. 3, pp. 433–457, 2023, doi: 10.1007/s00702-023-02592-y.
[64] T. Girolamo, I.-M. Eigsti, L. Shen, A. Monroe Gulick, and M. L. Rice, "Studies assessing domains pertaining to structural language in autism vary in reporting practices and approaches to assessment: A systematic review," *Autism*, vol. 28, no. 7, pp. 1602–1621, 2024, [Online]. Available: https://research.ebsco.com/linkprocessor/plink?id=4f3204cb-197b-31e1-97cf-2aa1697cd467
[65] S. J. Loveall, K. Hawthorne, and M. Gaines, "A meta-analysis of prosody in autism, Williams syndrome, and Down syndrome," *J Commun Disord*, vol. 89, p. 106055, 2021.
[66] M. Zhang, S. Xu, Y. Chen, Y. Lin, H. Ding, and Y. Zhang, "Recognition of affective prosody in autism spectrum conditions: A systematic review and meta-analysis," *Autism*, vol. 26, no. 4, pp. 798–813, 2022.
[67] C. Lord, "Commentary: Achievements and future directions for intervention research in communication and autism spectrum disorders," *J Autism Dev Disord*, vol. 30, no. 5, p. 393, 2000.
[68] J. N. Constantino, "Social responsiveness scale," in *Encyclopedia of autism spectrum disorders*, Springer, 2021, pp. 4457–4467.
[69] J. N. Constantino and C. P. Gruber, "Social responsive scale (SRS) manual," *Los Angeles, CA: Western Psychological Services*, 2005.

[70] D. American Psychiatric Association, D. S. American Psychiatric Association, and others, *Diagnostic and statistical manual of mental disorders: DSM-5*, vol. 5, no. 5. American psychiatric association Washington, DC, 2013.
[71] D. Bishop, "Children's communication checklist (CCC-2)," *Encyclopedia of autism spectrum disorders*, pp. 915–920, 2021.
[72] L. C. Eaves, H. D. Wingert, H. H. Ho, and E. C. R. Mickelson, "Screening for autism spectrum disorders with the social communication questionnaire," *Journal of Developmental \& Behavioral Pediatrics*, vol. 27, no. 2, pp. S95--S103, 2006.
[73] E. H. Wiig, W. A. Secord, and E. Semel, *Clinical evaluation of language fundamentals: CELF-5*. Pearson, 2013.
[74] I. L. Zimmerman, V. G. Steiner, and R. E. Pond, "Preschool language scale," 1979.
[75] R. Goldman and M. Fristoe, "Goldman-Fristoe test of articulation," 1969.
[76] D. K. Oller *et al.*, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13354–13359, 2010.
[77] J. A. Richards, J. Gilkerson, T. D. Paul, and D. Xu, "The LENA TM automatic vocalization assessment (Technical Report LTR-08-1)," *LENA Foundation*, 2008.
[78] D. Xu, J. A. Richards, and J. Gilkerson, "Automated analysis of child phonetic production using naturalistic recordings," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 5, pp. 1638–1650, 2014.
[79] S. Badotra and S. N. Panda, "A review on software-defined networking enabled iot cloud computing," *IIUM Engineering Journal*, vol. 20, no. 2, pp. 105–126, 2019.
[80] A. Sundas and S. Panda, "IoT and WSN based smart surveillance system for patients with closed-loop alarm," *International Journal of Scientific \& Technology Research*, vol. 8, pp. 508–511, 2019.
[81] F. Thabtah, F. Kamalov, and K. Rajab, "A new computational intelligence approach to detect autistic features for autism screening," *Int J Med Inform*, vol. 117, pp. 112–124, 2018.
[82] G. Riva and E. Riva, "DE-ENIGMA: Multimodal Human--Robot Interaction for Teaching and Expanding Social Imagination in Autistic Children.," *Cyberpsychol Behav Soc Netw*, vol. 23, no. 11, 2020.
[83] P. Howlin, S. Baron-Cohen, and J. A. Hadwin, *Teaching children with autism to mind-read: A practical guide for teachers and parents*. John Wiley \& Sons, 1999.
[84] M. Brian, "The CHILDES Project: Tools for analyzing talk," *Mahwah, NJ \& London: Lawrence Erlbaum*, 2000.
[85] Y. Nakai, T. Takiguchi, G. Matsui, N. Yamaoka, and S. Takada, "Detecting abnormal word utterances in children with autism spectrum disorders: machine-learning-based voice analysis versus speech therapists," *Percept Mot Skills*, vol. 124, no. 5, pp. 961–973, 2017.
[86] Q. Tariq, J. Daniels, J. N. Schwartz, P. Washington, H. Kalantarian, and D. P. Wall, "Mobile detection of autism through machine learning on home video: A development and prospective validation study," *PLoS Med*, vol. 15, no. 11, p. e1002705, 2018.
[87] S. R. Shahamiri, F. Thabtah, and N. Abdelhamid, "A new classification system for autism based on machine learning of artificial intelligence," *Technology and Health Care*, vol. 30, no. 3, pp. 605–622, 2022.
[88] R. Matin, "Developing a speech emotion recognition solution using ensemble learning for children with autism spectrum disorder to help identify human emotions," 2020.
[89] S. Amiriparian *et al.*, "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," 2018.
[90] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child Speech Disorder Detection with Siamese Recurrent Network Using Speech Attribute Features.," in *INTERSPEECH*, 2019, pp. 3885–3889.
[91] J. H. Lee, G. W. Lee, G. Bong, H. J. Yoo, and H. K. Kim, "End-to-end model-based detection of infants with autism spectrum disorder using a pretrained model," *Sensors*, vol. 23, no. 1, p. 202, 2022.
[92] 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, 2022.

[93] 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.
[94] H. Lee *et al.*, "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition.," in *INTERSPEECH*, 2013, pp. 215–219.
[95] T. Zhou, Y. Xie, X. Zou, and M. Li, "An automated assessment framework for speech abnormalities related to autism spectrum disorder," in *3rd International Workshop on Affective Social Multimedia Computing (ASMMC)*, 2017.
[96] M. Li *et al.*, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Comput Speech Lang*, vol. 56, pp. 80–94, 2019.
[97] J. Yuan, C. Holtz, T. Smith, and J. Luo, "Autism spectrum disorder detection from semi-structured and unstructured medical data," *EURASIP J Bioinform Syst Biol*, vol. 2017, pp. 1–9, 2016.
[98] A. Garg *et al.*, "Autism spectrum disorder prediction by an explainable deep learning approach," *Computers, Materials \& Continua*, vol. 71, no. 1, pp. 1459–1471, 2022.
[99] S. Sadiq, M. Castellanos, J. Moffitt, M.-L. Shyu, L. Perry, and D. Messinger, "Deep learning based multimedia data mining for autism spectrum disorder (ASD) diagnosis," in *2019 international conference on data mining workshops (ICDMW)*, 2019, pp. 847–854.
[100] J. H. Lee, G. W. Lee, G. Bong, H. J. Yoo, and H. K. Kim, "Deep-learning-based detection of infants with autism spectrum disorder using auto-encoder feature representation," *Sensors*, vol. 20, no. 23, p. 6762, 2020.
[101] B. MacWhinney, "Understanding spoken language through TalkBank," *Behav Res Methods*, vol. 51, pp. 1919–1927, 2019.
[102] F. F. Thabtah, "Autistic spectrum disorder screening data for children data set," *UCI machine learning repository*, 2017.
[103] E. Bergelson, "Bergelson seedlings homebank corpus," *doi*, vol. 10, p. T5PK6D, 2016.
[104] J. Rehg *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3414–3421.
[105] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS One*, vol. 13, no. 5, p. e0196391, 2018.
[106] J. J. Wolff *et al.*, "Longitudinal patterns of repetitive behavior in toddlers with autism," *Journal of Child Psychology and Psychiatry*, vol. 55, no. 8, pp. 945–953, 2014.
[107] A. L. Higgins, S. F. Boll, and J. E. Porter, "U.S. Patent No. 6,266,633," 2001.
[108] A. Keerio, B. K. Mitra, P. Birch, R. Young, and C. Chatwin, "On preprocessing of speech signals," *International Journal of Signal Processing*, vol. 5, no. 3, pp. 216–222, 2009.
[109] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4470–4474.
[110] U. D. Reichel and H. R. Pfitzinger, "Text preprocessing for speech synthesis," 2006.
[111] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for health: the use of vocal biomarkers from research to clinical practice," *Digit Biomark*, vol. 5, no. 1, pp. 78–88, 2021.
[112] M. Hosseinzadeh *et al.*, "A review on diagnostic autism spectrum disorder approaches based on the Internet of Things and Machine Learning.," *Journal of Supercomputing*, vol. 77, no. 3, pp. 2590–2608, Mar. 2021, doi: 10.1007/s11227-020-03357-0.
[113] Z. Azizi, "The acoustic survey of intonation in Autism Spectrum Disorder," *Journal of the Acoustical Society of America*, vol. 137, no. 4_Supplement, p. 2207, 2015.
[114] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, "Deep learning algorithms for bearing fault diagnostics—A comprehensive review," *IEEE Access*, vol. 8, pp. 29857–29881, 2020.
[115] L. Deng, D. Yu, and others, "Deep learning: methods and applications," *Foundations and trends® in signal processing*, vol. 7, no. 3--4, pp. 197–387, 2014.
[116] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[117] 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 Clin*, vol. 17, pp. 16–23, 2018.
[118] Y. Kong, J. Gao, Y. Xu, Y. Pan, J. Wang, and J. Liu, “Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier,” *Neurocomputing*, vol. 324, pp. 63–68, 2019.
[119] F. B. Pokorny *et al.*, “Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach,” 2017.
[120] 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, 2015.
[121] S. R. Shahamiri and F. Thabtah, “Autism AI: a new autism screening system based on artificial intelligence,” *Cognit Comput*, vol. 12, no. 4, pp. 766–777, 2020.
[122] A. Wawer and I. Chojnicka, “Detecting autism from picture book narratives using deep neural utterance embeddings,” *Int J Lang Commun Disord*, vol. 57, no. 5, pp. 948–962, 2022.
[123] S. Gupta, K. De, D. A. Dinesh, and V. Thenkanidiyoor, “Emotion recognition from varying length patterns of speech using CNN-based segment-level pyramid match kernel based SVMs,” in *2019 National Conference on Communications (NCC)*, 2019, pp. 1–6.
[124] Mustaqeem and S. Kwon, “A CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors*, vol. 20, no. 1, p. 183, 2019.
[125] A. M. Hilal *et al.*, “Modeling of Explainable Artificial Intelligence for Biomedical Mental Disorder Diagnosis.,” *Computers, Materials \& Continua*, vol. 71, no. 2, 2022.
[126] R. Liu and S. He, “Proposing a System Level Machine Learning Hybrid Architecture and Approach for a Comprehensive Autism Spectrum Disorder Diagnosis,” *arXiv preprint arXiv:2110.03775*, 2021.