Unsupervised Machine Learning Methods for Diagnosing Autism Spectrum Disorder Using Multimodal Data: A Survey

Antika Roy
Independent Researcher
Email: antika.roy121@gmail.com

Pablo Rivas, *Senior, IEEE*
School of Eng. and Computer Science
Dept. of Computer Science, Baylor University
Email: Pablo_Rivas@Baylor.edu

Mahee Noor Tayba
Department of Computer Science, Baylor University
Email: Mahee_Tayba1@Baylor.edu

***Abstract*—The neurodevelopmental disorder known as autism spectrum disorder (ASD) is becoming increasingly pervasive worldwide. It can be categorized when a person struggles to interact and communicate socially and also exhibits repetitious behaviors and interests. Although early intervention is crucial and may have long-term advantages in the lives of people with ASD, diagnosing ASD is difficult due to its heterogeneous characteristics and a large number of data from various aspects such as genetic, behavioral, electronic health records, and many other domains. This review paper offers an in-depth assessment of papers that use the most commonly used unsupervised machine learning techniques in ASD, such as "k-means clustering" and "Hierarchical clustering." This research aims to identify and highlight the most recent unsupervised machine learning advances in the ASD literature while also attempting to highlight the significant contributions and limits of the selected works to provide insight for future researchers.**

***Index Terms*—neurodevelopmental disorder, heterogeneous characteristics, hierarchical clustering, unsupervised machine learning.**

I. Introduction

Autism spectrum disorder (ASD) is classified by the "Diagnostic and Statistical Manual of Mental Disorders, Fifth edition" as a developmental impairment characterized by clinical guidelines which include difficulties with socio-emotional mutuality, nonverbal communicative behaviors, and creating, sustaining, and comprehending affairs along with limited, recurrent behavioral patterns, preferences, or actions that can last a lifetime [1]. Early developmental periods often show the first indications and symptoms of ASD; however, social deficits and behavioral abnormalities may not be identified as symptoms until a child cannot fulfill social, educational, occupational, or other significant life stage expectations [2]. In 1998, CDC began keeping track of the presence of ASD and the characteristics of kids with ASD in the United States. 1 in 44 eight-year-olds had ASD overall, and males were 4.2 times more likely than girls to have the disorder [3]. In light of recent statistics from the World Health Organization (WHO), one in every 100 children has autism [4]. In terms of the presentation and severity of the symptoms, risk factors, and causation, along with the treatment response, ASD is a complex disorder. Since autism has connections to genetics and neurological development, as well as severe deformities in social interaction and behavior, it provides an opportunity for researchers to investigate the neurobiological origins of social communication abilities and emphasizes the importance of advancing our knowledge of ASD [5].

Although doctors employ standardized diagnostic instruments for ASD diagnoses, one main disadvantage of the procedure is that delivering assessment techniques takes a significant amount of time to complete the evaluation and analyze the results. In detecting ASD, reducing screening duration while increasing correctness is the fundamental goal of machine learning research. Patients with ASD can receive early treatment by shortening the assessment hours. Another goal of the machine learning techniques is to identify the top-rated ASD characteristics by reducing the volume of the corresponding input sample [6].

Machine learning is a field of artificial intelligence that deals with the theory and algorithms for learning from data [7], [8]. We can use machine learning to analyze vast clinical records to enhance patient outcomes, hasten the innovation of treatments and remedies, and computerize repetitive tasks to reduce the chance of human mistakes. Autism researchers are also getting inventive, employing machine-learning technologies to diagnose better, categorize the disorder into subgroups, and provide support to those on the spectrum. There are three major categories for machine learning: supervised learning, where we can work with labeled data; unsupervised learning, where we usually cluster and analyze unlabeled data; and semi-supervised learning, where fewer labels exist in comparison to unlabeled data for training a model. To discover hidden patterns in data, unsupervised learning assesses and groups unlabeled data sets using machine learning techniques [9]. Researchers studying ASD can use unlabeled data from treatment response, linguistic abilities, behavioral and genetic data, and other sources that do not have pre-existing groupings or categories, thanks to unsupervised machine learning.

This systematic review comprehensively discusses the research on the most often utilized unsupervised learning methods: k-Means Clustering and Hierarchical Clustering in ASD studies done since 2016. In addition, this paper examines contemporary research for ASD diagnosis, highlighting its relevance, diverse range of datasets, significant contributions, and limits using k-Means Clustering and Hierarchical Clustering approaches from peer-reviewed publications. This paper's contents are categorized as follows: Section I introduces Autism Spectrum Disorder, the problems and obstacles the subjects confront, and the need to integrate machine learning into this spectrum. Section II describes the methodology that was employed in this study. Section III presents and discusses the acquired results, followed by a conclusion in Section IV.

II. Research Methodology

This literature evaluation has been done to examine how k-Means Clustering and Hierarchical Clustering approaches have been used recently in ASD diagnosis. Using Google Scholar, ScienceDirect, and PubMed databases, an electronic literature search has been performed for pertinent, peer-reviewed papers from 2016 to 2022. Throughout this study, search phrases such as "autism spectrum disorder," "unsupervised machine learning," "k-means clustering in ASD," and "Hierarchical clustering in ASD" were combinedly used. The abstract and methodology portions of the publications were analyzed, and the selection process for an article included the following conditions: (a) peer-reviewed journal publication; (b) contained diagnostic data of people with autism; (c) the k-means clustering & hierarchical clustering method of unsupervised machine learning was used; and (d) the publication's impact factor was more than 4. To reduce the possibility of bias, PRISMA suggestions were taken into account. As a result, we first chose research abstracts, and then he evaluated entire texts to make sure the inclusion requirements were met. The purpose of the study/objective, kind of the study/datasets, number of ASD participants, ML models, discussion, findings, and future work were all retrieved as significant information for the study. The next section provides a detailed discussion of the research's findings.

III. Results and discussion

More than 50 publications were discovered as a consequence of the study's abstract review of these papers. The whole content of the publications was scrutinized in depth, and 18 of them met the aforementioned requirements, while the others were excluded from the survey. Seven of the included publications utilized the k-Means Clustering approach, see Table I, while eleven used the Hierarchical Clustering method, see Table II. Papers were grouped according to the clustering approach employed in each paper. The reviewed papers relevant to that subject are presented after briefly describing the approach at the beginning of each section.

Two types of machine learning models are utilized in ASD detection: supervised and unsupervised. Unsupervised learning learns from untagged data, whereas supervised learning uses input variables to determine a goal classification. The purpose of unsupervised learning is to learn and discover hidden patterns in massive datasets. Cluster analysis is the most prominent example of unsupervised learning. Clustering may be utilized on Electronic Health Records (EHRs) in health care to identify distinct subgroups of patients with ASD and typically developing individuals. In ASD, unsupervised learning clusters are statistically examined to see if levels of various measurable variables (such as cognition, linguistic skills, and sensory behavior) differ between clusters. The most popular and generally used unsupervised approaches are K-mean Clustering, Hierarchical Clustering, Model-based Clustering, and Node Based Resilience Clustering. The following sections discuss 18 studies that employed the most often used unsupervised learning approaches, K-mean Clustering and Hierarchical Clustering, as well as their implementation, dataset, essential contributions, and limitations.

A. *k-means clustering*

Among the most fundamental and extensively applied unsupervised machine learning algorithms is k-means clustering. In this technique, each data point is assigned to the nearest cluster applying Euclidean distance to keep the centroids as compact as feasible after locating k centroids [10]. The selected publications in which k-means clustering is most recently employed in autism spectrum disorder are discussed in the following paragraphs. Table 1 presents a summary of papers that employed the k-means clustering approach. In addition, we listed the types of data used in the research, the study's overall objective, key contributions, and specific constraints.

In 2016, Lingren and colleagues [11] conducted a study utilizing EHRs from several institutions, using a computerized technique for deriving cohorts and analyzing comorbidity trends in patients with ASD. The comorbidity clustering analyses were performed on a massive group of 20,658 patients with ASD from three sites that revealed three clusters: psychiatric, developmental, and seizure disorder clusters. The data was pre-processed by transforming the patients' "ICD-9 (International Classification of Diseases, Ninth Edition) codes into Phenotype Wide Association Study (PheWAS) categories" while excluding a few categories where patients constituted below 0.5 percent of the total [11]. Then, using the k-means clustering method, they clustered the patients characterized by the generated "PheWAS code vectors". They used k-means clustering to determine the clustering using the "highest silhouette coefficient" by varying the number of clusters from 2 to 20. Each institution has a similar clustering tendency, according to the researchers. Three to four minor clusters (5–20%) and one major cluster were found in each collection. One of many comorbidity types dominates each of the smaller clusters: 1) psychological issues such as depression, hyperkinetic syndrome, anxiety disorder, and OCD; 2) developmental issues such as lack of coordination, dyslexia, and numerous ear, skin, and other physical system diseases; and 3) epilepsy and recurring seizures. A significant incidence of comorbidities or category of comorbidities was not found in the more significant cluster, which comprised about 60–80 percent of the sample.

TABLE I: List of papers reviewed on k-means clustering.

List of the papers				
Author, Date of Publication	Type of Study	Objective	Key Contributions	Limitations & Future Study
Lingren et al. [11], July 29,2016	EHR databases from three institutions	This paper used several institutions' EHRs to forward the research on ASD by creating an automated system for cohort extraction and studying the co-occurrence characteristics of comorbidities in ASD patients.	1) The three locations' combined cohort of 20,658 ASD patients is the largest ASD cohort ever. 2) Suggested automated cohort selection techniques open up possibilities for further large-scale EHR research and personalized ASD cure. 3) Psychiatric, developmental, and seizure problems- three clusters were discovered from this three-site cohort.	1) The combined cohort doesn't represent a gold standard because evaluating all 20K+ patients' charts wasn't possible. 2) In the future, the methodology can be used on cohorts without ICD-9 code in the ASD diagnoses 3) Using the clustering algorithms of this study in measuring morbidity in other complicated illnesses
Easson et al. [12], February 01,2019	Resting-state fMRI (rs-fMRI) data (145 autistic males and 121 TD males) collected from ABIDE dataset	This study discovered distinct clusters of functional connectivity patterns in a population of ASD and TD patients.	1) Emphasized the significance of addressing FC-based subgroups while studying brain-behavior interactions in ASD and controls. 2) Implementation of multivariate statistical analysis to discover the best association between brain activity measurements and trial design or group membership.	1) The subtypes were defined using a single preprocessing method. Different preprocessing techniques can be applied across FC-based subtypes. 2) As age was not adjusted in the FC data, the k-means clustered subtypes differ. 3) Future research will need to include people with low-functioning ASD to see if they have the same subgroups.
Stevens et al. [13], December, 2017	Records of challenging behaviors collected from 2,116 verified ASD patients	The focus of this research is to present the first cluster analysis-based machine learning assessment of challenging behaviors that discovers prevalent challenging behavior profiles and co-occurrences of behavior.	1) First study of its kind, with 2,116 patients, examines challenging behaviors. 2) While the occurrence of many challenging behaviors is typical, the results show that a dominating behavior develops in the majority of cases. 3) According to the findings, gender disparities in problematic behaviors should be considered while treating them.	Only considers the existence of challenging behaviors, not their purpose. A functional component in challenging behaviors will be investigated in the future.
Obara et al. [14], October 04,2018	ADI-R data of 17 persons (13 boys and 4 girls) with ASD	This study's objective was to see if machine learning approaches could identify a subset of ASD people with vitamin B6 responsiveness based on phenotypic factors.	1) Using specified phenotypic factors, focus on finding a subgroup of people with ASD who are vitamin B6 sensitive. 2) Hypothesis testing among specified variables and their combinations 3) Analyzed the data using two alternative techniques, affinity propagation and k-medoids	1) Predictive value necessitates additional studies using various algorithms, such as support vector machine, as well as an evaluation of the results' correctness, probably utilizing a greater number of participants and validation data sets. 2) More information may be required to evaluate vitamin B6 response and determine how common it is.
Hyde et al. [15], December 19, 2019	Policies, procedures, and views of people with HFASD from an online survey of 285 employers	This study employs unsupervised ML to investigate the employers' policies and practices regarding ASDs, focusing on understanding flaws in recruitment procedures that can be resolved through a combo of educational initiatives and alterations in the corporate world.	The clustered data allows researchers to examine the recruiting procedures of organizations who hire people with ASD vs. those that do not.	Not available

Table I continued				
Author, Date of Publication	Type of Study	Objective	Key Contributions	Limitations & Future Study
Silleresi et al. [16], January 27, 2020	Language capacities and nonverbal tasks of 51 (age range:6-12 years) verbal children having ASD	This study advocated a comprehensive evaluation of both the language and intellectual capacities of children with ASD, putting heterogeneity at the forefront of the investigation by expanding previous work on language in autism across the full spectrum.	1) Found improvements in recognizing language characteristics in ASD is based on investigating children throughout the spectrum and using strong structural language and NV assessments. 2) According to the study, every one of the profiles ”end up” with bilingual kids, suggests that regardless of whether an ID is present, a bilingual language context for an autistic kid does not impede adequate structural language functioning.	1) Future research on the phenotypic similarities between autism with impaired language (ASD-LI) and Developmental Language Disorder (DLD) and autism with normal language (ASD-LN) and TD. 2) To address the quantity and relative prevalence of profiles requires a bigger dataset of children with poor nonverbal IQ and language impairment. 3) Additional study is required to explore how linguistic and intellectual ability profiles vary across time.
Narita et al. [17], August 17, 2020	Genotypic data, phenotypic variables, and history of treatment from the Simons Simplex Collection	The researchers investigated whether grouping people with ASD using a clustering algorithm on the basis of phenotypic and treatment response factors may be utilized to differentiate more genetically homogenous people with ASD.	1) If indeed the set of data contains numerous heterogeneous subgroups, only a subgroup with a considerably lower number of homogenous people might uncover high-impact genetic variables. 2) The study’s most noteworthy conclusion was that lowering the sample size might boost statistical power.	1) Significant discrepancies between the two genotyping systems could have influenced the replication study outcomes. 2) A bigger sample size is required for future research on cluster validation. 3) Additional study is required to explore how linguistic and intellectual ability profiles vary across time.

TABLE II: List of papers reviewed on hierarchical clustering.

List of the papers				
Author, Date of Publication	Type of Study	Objective	Key Contributions	Limitations & Future Study
Lombardo et al. [18], October 18, 2016	Genomics and systems biology on discovery and replication datasets of individuals with ASC (n=694) and without ASC (n=249)	Their discovery of identifiable, reproducible, and sturdy ASC subgroups with varying specific mentalizing abilities as evaluated by the RMET in adolescence constitutes a significant step forward in the accuracy of knowledge of mentalizing issues in ASC people.	1) The findings have the ability to advance personalized medicine goals in such manners that can hasten therapeutic effects for patients. 2) The technique of subgrouping is a significant breakthrough in this study.	1) The complete spectrum of autism heterogeneity is unlikely to be reflected by the RMET test alone. Subsets of RMET items might be studied to a greater extent. 2) Specificity and sensitivity in deciphering mentalizing heterogeneity might be improved. 3) Multi-group item response theory (IRT) analysis might be used to larger sample sizes across subgroups.
Elwin et al. [19], December 05, 2016	Data from a previous validation study of SR-AS	The goal was to discover subgroups of ASC persons that had comparable sensory attributes.	1) Discovered groupings of people with varying levels of sensory symptom frequency 2) Sought to investigate the prevalence of psychiatric comorbidity in the ASC population and any relationships between cluster membership and comorbidity	1) Absence of more comprehensive validation of the SR-AS 2) Both samples lacked a measure of ASC features, and the population sample lacked information on mental diseases, including ASC. 3) Further evaluation of psychometric features is required. 4) In comparison to regular growth, ASC needs concentration on developmental features of sensory function.

Table II continued				
Author, Date of Publication	Type of Study	Objective	Key Contributions	Limitations & Future Study
Pichitpunpong et al. [20], March 28, 2019	ADI-R data of 85 male individuals with ASD	In this work, they aimed to analyze the proteome profiles of ASD patients by minimizing the variability of the people with regard to ASD through the use of a phenotypic subgrouping technique.	1) Proteomic research identifies unique differentially expressed proteins in an ASD subgroup 2) Functional and network prediction analyses have implicated genes and pathways.	1) To eliminate bias, a full meta-analysis and/or a comprehensive evaluation of existing proteome profiles could be performed. 2) in the future, the study's findings should be validated in a wider sample with gender, age, and ethnicity-matched controls.
Kurochkin et al. [21], June 21, 2019	1366 metabolites were found in gray matter in the prefrontal cortex of 32 people with autism and 40 people with control	The results suggest that metabolic alterations identified in blood could indicate metabolite intensity changes in the brain.	1) Paved the way for universal metabolic biomarker-based healthcare practices 2) Uncovered a number of previously unknown metabolic pathways related to ASD 3) Measured the brain metabolome in nonhuman primates	1) Thorough validation is required 2) Small sample size due to the scarcity of ASD brain tissue samples
Smith et al. [22], February 15, 2019	Metabolites from children having ASD from the project - Children's Autism Metabolome Project (CAMP)	The purpose was to study data from the CAMP population to find metabotypes linked with ASD that might allow for categorization based on common metabolic features.	1) Identification and use of ASD metabotypes can contribute to implementable metabolic tests that can assist in early detection and classification for targeted therapy initiatives. 2) Metabolite correlations within ASD demonstrate unique groups of amine metabolites. 3) ASD-Related AA:BCAA imbalance metabotypes have been identified	1) Since blood plasma levels of amine metabolites really aren't significantly correlated with brain levels, it is challenging to figure out the link between the changes in plasma levels and brain levels. 2) Uncertainty exists about the specificity of AADMs for ASD in comparison to other neurodevelopmental diseases. 3) There aren't enough tissue samples or animal models to explore the molecular pathways underlying amino acid dysregulation metabotype (AADMs).
Zheng et al. [23], January 08, 2020	Measures of children's development (such as cognitive and linguistic ability)	This study used HCPC to see whether there are identifiable subgroups among preschoolers with ASD on the basis of numerous behavioral and developmental parameters in addition to autistic symptom assessments.	1) First to discover subgroups of preschoolers with ASD using HCPC. 2) Persistent IQ disparity across subgroups may be a possible phenotypic characteristic of preschoolers with ASD.	1) Sample size (N = 188) is inadequate. 2) There was no valid investigation to reproduce the subgroup outcomes with a separate group of preschool-aged kids with ASD. 3) More research is needed to discover latent component structures and linkages collected across many measures.
Dwyer et al. [24], June 15, 2020	Auditory electrophysiological data were collected from 96 typically developing and 243 autistic children.	This study intends to investigate brain diversity in autism and conventional sensory processing by utilizing responses to identify clusters of children with comparable patterns of intensity-dependent auditory processing.	The study emphasizes loudness-dependent normalized responses, enabling it to avoid individual distinctions in biophysical aspects like the thickness of the skull, while using GFP provides a comprehensive measure of neural response strength, which eliminates the necessity for predetermined decisions regarding assessing specific elements or electrode sites.	1) To validate exploratory results, replication may be required. 2) Does not resolve the variability associated with traditional ERP latencies and topographies 3) The sensory questionnaire used was not created especially for autistic children. 4) Further study is needed since the variability in sensory reactions irrelevant to loudness is not considered.

Table II continued				
Author, Date of Publication	Type of Study	Objective	Key Contributions	Limitations & Future Study
McDougal et al. [25], November 2020	Reading, mathematics, and attention abilities of twenty-two autistic children between ages 6 to 16 years and fifty-nine TD children between ages 6 to 11 years	To explore reading and arithmetic accomplishment profiles for children with and without ASD, focusing on the function of attention in these profiles, to get a clear idea of individual variations.	1) The capacity to divide attention characterizes reading and math profiles in autistic students. 2) Transdiagnostic subgroups defined by attention and accomplishment. 3) Children who are less attentive and achieve less show relative deficits in math.	Conducting cluster analysis with a limited number of participants from both groups, especially the ASD group, is not optimal.
Paakki et al. [26], May 16, 2021	Resting-state (RS) fMRI data from 28 individuals with ASD and 27 normally developing (TD) controls.	To investigate nonsequential volume-wise techniques to characterize the time-varying functional connectivity of resting state (RS) fMRI brain networks (RSNs).	The comparison of voxelwise signal changes and nonsequential volume collecting into CAPs gives a supplementary perspective on connectivity and a different approach to sliding window analysis.	1) Censorship lowers freedom and can also erase the signal of interest. 2) RS was not compared to task data, and the link between the CAPs of the rest and task data should be investigated further.
Benabdallah et al. [27], March 18, 2022	Resting-state fMRI data from ABIDE database	The goal was to validate ASD ideas while also improving autism detection.	The suggested strategy is novel in that it uses elimination as a tool to identify autistic brain connection changes and demonstrate how they contribute to the differentiation of ASD from controls.	Not available
Bullen et al. [28], April 2022	Completed assessments of arithmetic and reading achievement, Theory of Mind (ToM), working memory, IQ, and inferential reasoning of 78 individuals	The goal of this study is to expand on past research on achievement profiles in autistic children by studying variations in a wider range of cognitive skills.	1) Reading fluency differentiated subgroups better than other factors. 2) The findings have significant implications for general education inclusion.	1) This study's narrow sample composition (White 65 percent and male 80 percent) can only give information on a tiny fraction of the autistic community. 2) Variation in measurement may contribute to disparities between studies.

Using Resting-state fMRI (rs-fMRI) data of 145 autistic males and 110 typically developing (TD) males, Easson et al. [12] applied an analytical data-driven, dimensional method to describe subtypes in all participants based on discrete clusters of functional connectivity (FC), as well as to link FC patterns to distinct behavioral traits in these subgroups. Two FC-based subtypes (85 ASD and 54 TD in Subtype 1, 60 ASD, and 67 TD in Subtype 2) were established using k-means clustering. Compared to Subtype 1, Subtype 2 was distinguished by higher FC within networks and lesser FC between networks, particularly between the default mode network (DMN) and other resting-state networks (RSNs). According to connectivity between FC and behavior observed in and between RSNs for IQ, ADOS (Autism Diagnostic Observation Schedule), RRB (Restricted and Repetitive Behaviors) scores for Subtype 1, and all behavioral measures for Subtype 2, identical behavioral profiles can be connected to various functional correlations of the brain. Individuals in each subtype had similar IQ and SRS scores, while ASD individuals in both categories had similar ADOS scores.

Stevens et al. [13] published the first assessment of challenging behaviors in a wide group of 2,116 patients based on machine learning. According to the statistics, the overall problematic behavior profiles of the population will probably feature a variety of challenging behaviors. They discovered seven clusters using k-means cluster analysis (calculating the k value with the "elbow method"), with patients from the sample providing good representation in each cluster. Meaningful behavior profiles were determined by clustering, indicating the presence of a dominating single challenging behavior in most clusters. After evaluating clusters of behaviors for the full sample, the study was done individually on the male and female sample groups. The findings imply that the existence of gender differences regarding challenging behaviors on the autism spectrum should be considered in therapy.

In 2018, Obara et al. [14] stated that they were the first to try to identify a subset of people with ASD who are vitamin B6 responsive using specified phenotypic factors. First, they concentrated on indications and biomarkers known to be possible vitamin B6 response indicators. They used k-medoids clustering analysis and affinity propagation (AP) to evaluate these factors' capacity to identify a subgroup with ASD. AP was initially used using the preprocessed dataset to split the patients into two groups based on possible vitamin B6 responsiveness. To test the resilience of the AP clustering, k-medoids clustering was performed. The AP study accurately classified probable vitamin B6-responsive people with ASD (cluster 1). All of the individuals were clearly divided into five groups, with clusters 2–5 consisting of people who had a poor reaction to vitamin B6. The k-medoids study also demonstrated good categorization. There were five clusters, and the outcome was identical to that of the AP except for one individual categorized in Cluster 3 by the k-medoids approach but in Cluster 2 by the AP.

The study conducted by Hyde et al. [15] used k-means clustering to investigate the policies, practices, and views of 285 employers about persons with "higher functioning" ASD (HFASD). To generate the data matrix, 41 survey questions and subquestions were used, each with a Boolean variable. The four categories of hiring, training, accommodation, and retention were then applied to the variables. The average score for each respondent was then computed in each of the four categories. Using the k-means clustering, the data was divided into five clusters. Cluster 5 has the greatest HFASD hiring rate over the last five years (86%) and the maximum cluster members. The two most prominent clusters, 3 and 5, had the most dramatic average scores in each category. This might imply that employers have no policies and practices in the workplace or that they have some that are exceptionally beneficial. Cluster 2 and cluster 5 only had employment rates higher than the survey average of 58%. For entry-level occupations, more than half of the employers in every one of these clusters required a college degree. Clusters 1 and 3 had the lowest percentages of employing HFASD (24% and 26%, respectively) and also had percentages of needing a college degree below 50% (32% and 29%, respectively) separately).

Recently, Silleresi et al. [16] used explicitly motivated assessments to carry out a thorough evaluation of 51 verbal children ranging in age from 6 to 12 years old with ASD in terms of language (particularly structural language abilities) and intellectual abilities (namely, nonverbal (NV) cognitive abilities). They investigated potential connections between structural language and NV skills, which were assessed using different LITMUS (Language Impairment Testing in a Multilingual Setting) repetition tests for language capacities and NV activities such as the Block Design, RPM, and Matrix Reasoning tasks for cognitive capabilities in ASD children. They employed an integrative strategy based on cluster analysis, which identified five unique profiles. Each of the four rationally possible matches of nonverbal abilities and structural language indicated in the ICD-11 was discovered in these five profiles. Three profiles appeared in children with typical language ability, while two developed in language-impaired youngsters.

Narita et al. [17] investigated whether grouping ASD patients using k-means clustering algorithm with phenotypic characteristics and treatment history might be utilized to differentiate more genetically homogenous ASD patients. Using phenotypic data from the Simons Simplex Collection (SSC), they performed cluster-based genome-wide association analysis (GWASs) using k-means clustering with a cluster size of 15. First, they did a standard genome-wide association analysis (GWAS) using 597 ASD patients and 370 controls. Then, based on the clustering results, they separated the patients and performed GWAS in every subgroup versus controls (cluster-based GWAS). In the replication step, they also performed cluster-based GWAS on a different SSC data set that contained 712 probands and 354 controls.In the second stage of cluster-based GWASs, they discovered 65 chromosomal loci, including 30 intragenic loci situated in 21 genes and 35 intergenic loci that met the criterion. Different ASD subgroups might be linked to statistically relevant loci. Since these disorders are thought to have a common etiology, at least in part, with ASD, the findings imply that the statistically relevant SNPs may describe the symptoms of autism.

B. Hierarchical clustering in ASD Research

Hierarchical clustering (HC) is an unsupervised clustering approach that includes forming clusters with dominant ordering from top to bottom, which is often shown as a tree diagram known as a dendrogram. The dendrogram ends with a set of clusters, where each cluster holds similar items within itself but differs from the others. We usually see two kinds of hierarchical clustering strategies. The first one is Agglomerative Hierarchical Clustering which can be described as "hierarchical clustering algorithms that begin with each entity as its own (singleton) cluster and then iteratively merge entities and clusters into a single cluster, constituting the entire group of entities" [29] and the other is Divisive Hierarchical Clustering which is "the reverse of agglomerative algorithms; begins with the entire set of entities as one cluster, which is then iteratively divided (usually bifurcated) until each entity is its own (singleton) cluster" [29]. In the following parts, we will look at some of the most recent studies in which Hierarchical clustering has been employed in autism spectrum disorder. Table 2 contains an overview of chosen studies that used the hierarchical clustering method, including the types of data analyzed in the study, along with its main contributions and specific shortcomings.

In 2016, Lombardo et al. [18] investigated variability in mentalizing ability as judged by the Reading the Mind in the Eyes Test; RMET in people with and without ASC using genomes and systems biology data from two huge distinct cognitive sample sets (n=694; n=249). RMET item-level data from all participants were combined to create a two-dimensional matrix, which was then translated into a distance matrix across individuals. Each dataset's distance matrices were turned into a topological overlap (TO) matrix for grouping into subgroups. Then the TO matrices were fed into agglomerative hierarchical clustering with Ward's method. A dynamic hybrid tree-cutting technique (deepSplit = 1) was used to subdivide the clustered dendrograms. Their clustering method identified five unique ASC subgroups and four separate TD subgroups that are present in both the Discovery and Replication datasets. Three subgroups of ASC adults (45-62%) show indications of significant impairments, whereas other subgroups are effectively unimpaired.

Elwin et al. [19] discovered sensory subgroups of people with ASC in a psychiatric cohort. Hierarchical clustering was used to evaluate the concept of groupings using sensory symptoms that occur at varying frequencies. The clustering analysis's agglomeration coefficients and dendrogram in the ASC sample recommended a three-cluster solution: low, intermediate, and high. The low-frequency group's (n=37) readings were all lower than the ASC sample mean, with sensory motor reactivity being especially low. When compared to cluster one, the intermediate group (n=17) had significantly greater levels of sensory interests, high awareness/hyper-reactivity, and sensory/motor challenges, but not low awareness/hyporeactivity. All measures were elevated in the high-frequency group, and the co-occurrence of Low awareness/Hyporeactivity and High awareness/Hyperreactivity was noticeable. The significant variation between clusters was the frequency of sensory complaints. The idea of an overall frequency/severity variance between clusters is supported by the cluster solution.

Pichitpunpong et al. [20] first identified subgroups using clustering analysis on the Autism Diagnostic Interview-Revised (ADIR) scores of 85 individuals with ASD, and then they reexamined the transcriptome profiles of those who have and haven't ASD to identify dysfunctional genes. Cluster analysis of ADI-R data showed 4 ASD clusters or groups, termed G1 (24 people), G2 (11 people), G3 (30 people), and G4 (20 people), having ASD with significant linguistic impairment and with transcriptome profiling uncovering dysregulated genes in each category. The proteome study on the ASD subgroup with significant language impairment found eighty-two changed proteins. The ASD subgroup with acute language difficulties exhibited diazepam-binding inhibitor (DBI) protein at significantly lower levels, and DBI expression levels were associated with numerous ADI-R item scores.

Kurochkin and colleagues [21] correlated metabolite intensity differences observed in the brain's prefrontal cortex (PFC) to variations detected in urine and blood in 2019. They studied variations in the intensities of 1366 metabolites found in the gray matter of the prefrontal cortex of 40 control and 32 people with autism. As the distance metric, they used hierarchical clustering using the 1-Pearson correlation coefficient to uncover patterns of age-related intensity differences for metabolites linked to ASD in both ASD and control samples. They divided the tree into four groups using the complete-linkage approach of hierarchical clustering. The separation to a high number of clusters revealed no unique patterns and produced clusters with few metabolites (n¡20). Fifteen percent of these metabolites had substantially altered concentrations in ASD and were grouped in sixteen metabolic pathways. Ten pathways in ASD patients' urine and blood were changed, allowing for the development of novel diagnostic tools. Moreover, metabolic analyses in forty chimps and forty macaques revealed an abundance of different metabolite intensities exclusive to humans, corroborating the hypothesis that ASD is caused by the breakdown of new evolutionary cortical mechanisms.

To discover if an imbalance of amino acids (AAs) was a more prevalent phenomenon in individuals who had ASD, Smith et al. [22] tried to compare plasma metabolites of 516 autistic children to those of 164 age-matched normally developing children enrolled in the Children's Autism Metabolome Project (CAMP). They used hierarchical clustering and pairwise Pearson correlation analysis to discover amine metabolites with coregulated metabolism. The Pearson correlation coefficient's 1-absolute was the value used as a measure of dissimilarity for calculating the distances for clustering. For hierarchical clustering, the Wards' approach was utilized. The glycine cluster is located in Cluster 1. BCAAs and phenylalanine are found in Cluster 2. Glutamate and aspartate are present in Cluster 3. Groups of identified AAs with positive associations were inversely linked with BCAA levels in ASD. Three amino acid dysregulation metabotypes linked with ASD were discovered as a result of imbalances between these two groups of AAs.

In 2020, Zheng et al. [23] divided the 188 preschoolers having ASD into 3 separate categories according to a variety of behavioral and developmental dimensions by using Hierarchical Clustering on the nine Principal Components (HCPC) of the dataset. For clustering, the first set of PCs that accounts for more than 85% of the variation was chosen, and it revealed latent patterns among the variables that influenced the clustering. They then performed HCA on the PCs by determining the Euclidean distance with Ward's minimum variance approach. On the basis of the results from the dendrogram and inertia graph, they produced a three-cluster solution. Children in Cluster 1 exhibited higher cognitive, linguistic, and adaptive skills levels than the rest of the sample and less severe sensory issues, repetitive behaviors, and social symptoms. Kids in Cluster 2 exhibited roughly the same levels of adaptive, linguistic, and cognitive abilities as those in Cluster 1, but their social deficiencies and sensory and repetitive behaviors were more severe. Children in Cluster 3 had the worst social, repetitive, and sensory symptoms as well as the lowest levels of cognitive, linguistic, and adaptive skills.

Dwyer et al. [24] employed hierarchical clustering to separate early autistic and generally developing children into categories depending on the normalized global field power (GFP) of their event-related potentials (ERPs) in response to auditory stimuli at four distinct loudness levels (50, 60, 70, and 80 dB SPL). The relative intensities of their brain reactions at various noise levels are measured by this GFP. In this investigation, the normalized GFP levels of each participant from each time point in each loudness condition were clustered using Ward's agglomerative hierarchical approach and displayed using heatmaps. Ward's method attempts to locate clusters in multivariate Euclidean space based on the distance by reducing variance within each cluster. Participants were divided into four groups: C1 includes 18 typically developing and 53 autistic individuals, C2 includes 17 typically developing and 24 autistic individuals, C3 includes 31 typically developing and 32 autistic individuals, and C4 includes 15 typically developing and 23 autistic individuals. There were significant differences across clusters in the normalized GFP reactions to sounds of various intensities. Although autistic and typically developing individuals' cluster assignments overlapped quite a bit, autistic participants were inclined to exhibit a pattern of reasonably linear rises in reaction intensity followed by an excessively strong response to 70 dB stimuli. Individuals with autism who had this pattern tended to score higher on cognitive ability tests. Additionally, there was a tendency for participants who were normally developing to fall disproportionately into a cluster with disproportionately/nonlinearly strong 60 dB responses. Auditory distractibility was shown to be greater in a cluster of ASD patients who responded disproportionately significantly to the loudest (80 dB) noises, and auditory distractibility was likewise linked to similarly robust reactions to loud stimuli. This seems to demonstrate the co-occurrence of behavioral and neural sensory abnormalities.

Another research [25] looked into the effect of attention on academic accomplishment in children who are either with or without ASD. Three transdiagnostic subgroups, each containing children with good, average, and poorer divided attention and academic ability, were identified by a hierarchical cluster analysis of 81 children based on their reading, mathematics, and divided attention scores. In this analytical method, instances are successively grouped into homogeneous clusters, and in every stage, the squared Euclidean distance between two clusters is measured; the clusters with the smallest distance are then joined to form a single cluster. Profiles A ("good-attention-higher-achieving"), B ("average-attention-average-achieving"), and C ("poor-attention-lower-achieving"), respectively, represented 6.2 percent, 70.4 percent, and 23.5 percent of the sample. When compared to their IQ, children with average or above-average attention and achievement scores indicated a relative strength in mathematics. Children with lower divided attention and accomplishment scores, on the other hand, exhibited a disparity in math achievement compared to their IQ and, at the same time, reading achievement.

Paakki [26] and his colleagues used hierarchical clustering (HC) on the RS-fMRI data of twenty-eight teenagers with autism spectrum disorder (ASD) and their twenty-seven typically developing (TD) control to categorize the image volumes. Each participant's fMRI signal was temporally adjusted voxelwise by dividing by the temporal SD after removing the mean. A data matrix created by reshaping and concatenating 11,930 resting state (RS) fMRI volumes from fifty-five teenage subjects was then imported into the R environment. All the BOLD fMRI volumes that had escaped censoring were then subjected to clustering. As they used the Ward method to perform hierarchical clustering, a cosine similarity matrix was transformed into a distance matrix. They discovered comparable brain state proportions in 58 co-activation patterns (CAPs) with clustering intervals from 2 to 30. Each cluster's respective fMRI volumes were merged. They used a group-independent component technique to select fourteen key RSNs for simplification. The average z-scores of the RSNs allowed them to significantly reorganize the RSNs and calculate the proportion of voxels inside every RSN with a noteworthy group difference. These findings were combined to uncover global group-specific characteristics.

Recently, Benabdallah et al. [27] proposed a methodology that focuses on ASD adaptive methods to test autism detection ideas concerning brain connectivity. Their strategy involved inhibiting brain connections associated with a certain idea. They extracted such linkages by combining specialized techniques. To identify the weak connections and the local/long-range connections, they used a minimum spanning tree and hierarchical clustering, respectively. These two methodologies were employed to validate the long-range underconnectivity to create connectivity matrices. They used the minimum spanning tree in order to streamline a complicated network depending on weights. They employed hierarchical clustering to separate the brain areas into groups and then confirmed the inter- and intra-connection of the brain in this way. The application of clustering enabled the detection of impairments in ASD brain connectivity. They could identify a long-range connectivity impairment and demonstrate that it is unaffected by gender, age, and other factors like handedness.

Bullen et al. [28] attempted to expand on earlier research of achievement profiles in autistic kids by studying disparities over a wider spectrum of cognitive skills. Firstly, the research investigated four achievement variables using hierarchical cluster analysis employing Euclidean distance and full linkage as cluster analysis parameters. The clustering yielded two distinct achievement groups with substantial differences in every one of the four academic success measures. The first subgroup, which included 55 people (70 percent of the dataset), had a below-average distribution of reading comprehension and reading fluency, an average distribution of calculation abilities, and a low average distribution of problem-solving abilities. The second group included 23 individuals (30 percent of the sample) with above-average calculation and problem-solving skills and average reading comprehension and reading fluency. The groups differed considerably in IQ, working memory, and reading fluency ability. Theory of Mind (ToM), inferential reasoning, and symptomatology was not different across groups. The variations in performance on cognitive and academic characteristics were then analyzed in order to determine what differentiates education in the cluster-defined academic groupings.

IV. Analysis, Challenges and Future Work

In this paper, we intend to perform a literature review on the application of k-means clustering and hierarchical clustering methods in ASD research. A total of 18 articles were discovered and analyzed. The most popular approach was hierarchical clustering, which was followed by k-means clustering.

Hierarchical clustering is the most used clustering algorithm of the two methods studied in the articles because of its interpretable and informative structure. So, by examining the dendrogram, researchers and physicians can more easily determine how many clusters there are. Researchers looked for subgroups in several sorts of data related to ASD studies. Such as behavioral data, genetic data, sensory abilities, linguistic and intellectual abilities, brain imaging information, math and reading skills, phenotypic characteristics, and treatment response information.

K mean clustering has been used successfully to find co-occurrence trends of comorbidities in patients with ASD. [11] When a fundamental disease or disorder coexists with one or more other diseases or disorders, this is referred to as comorbidity. Autistic children commonly have medical comorbidities. Comorbid medical illnesses profoundly influence a child's growth and behavior. Early diagnosis and treatment of these comorbidities will benefit the child's learning capability and his or her personal and family conditions.

Recent neuroscience and brain imaging advancements have opened the door for more detailed knowledge of the brain's function and structure. The strength of functional connections between pairs of regions is a notion that is extensively used to derive features from fMRI data. We highlighted how k-means clustering and hierarchical clustering approaches were utilized to detect unique groups of FC patterns in ASD individuals and controls using resting state fMRI data [12], [26], [27]. These subgroups will lead to a better knowledge of neurological functioning patterns, which will help us identify these disorders and understand the variables leading to their morphology.

The characteristics of autism spectrum disorder (ASD) are genetically and phenotypically diverse and frequently considered a barrier to understanding its origin, diagnosis, treatment, and prognosis. The cluster analyses [14], [17] contained in this review study can assist in establishing clinically significant ASD phenotypic clusters, which will aid in a better understanding of autism's heterogeneity. It would also provide significant information for the research of the disorder's origin, diagnosis, therapy, and prognosis. Cluster analysis in genetic and metabolic biomarkers [18], [21], [22], together with genetic discoveries and clinical observations in ASD, can outline a scenario in which frequent and unusual variants combine to establish a diagnosis.

ASD researchers have also made strides in using machine learning to understand the behavioral and social elements of ASD. In this review, we covered the benefits and drawbacks of several behavioral characteristics of ASD patients, such as challenging behaviors, linguistic and intellectual abilities, sensory abilities, reading, arithmetic, and attention abilities. To put it briefly and generally, grouping ASD data can help doctors learn more about the processes supporting treatment plans by enabling them to discover which way is the best suited by various types and intensities of support and therapy. Because early diagnosis is critical in autism research, these groups of ASD patients and features can help researchers discover novel approaches to treat ASD while improving the accuracy and timeliness of the current diagnostic process. The results have significant implications for inclusion in general education. It is unclear how to facilitate these children's learning in the classroom the most effectively.

Much of the literature concentrates on findings of the significance of social and behavioral support for children with ASD. The research study [28] revealed that autistic children without intellectual disabilities might be more prone to struggle with learning, which should be taken into account in mainstream education. Present research indicates that a considerable proportion of autistic youngsters face challenges with reading and math achievement. To examine the accomplishment profiles of autistic children without an intellectual handicap, hierarchical cluster analysis was performed in this study. The results have significant implications for inclusion in general education. Unfortunately, there is a shortage of knowledge regarding the best ways to support these kids' classroom learning.

Using k-means clustering, the researchers investigated the practices, policies, and perceptions of two hundred and eighty-five employers regarding individuals having "higher functioning" ASD (HFASD) [15]. Researchers can compare the employment procedures of companies that employ people with ASD to those that do not by using the data examined in the results. Suppose scholars comprehend the division of hiring practices in this direction. In that case, they will be more likely to understand the elements that enable effective recruiting and retention of workers with disabilities and how to encourage these practices in a broader range. In addition, researchers can use this information to design employer training programs that emphasize successful results and give examples of what other firms have done to establish productive workplaces for their impaired employees.

We identified the following challenges:

1) The majority of research papers share a problem with smaller datasets. Therefore, bigger-scale research studies are also required to determine if cluster organization stays consistent across more extensive groups of ASD patients.
2) Age and gender are two of the most important factors to consider in ASD studies because results may vary depending on the patient's age or gender. For example, gender disparities in problematic behaviors on the spectrum should be considered when treating them. Also, k-means clustered sub-types differed dramatically because age wasn't adjusted for the FC data.
3) Some aspects of ASD research require validation and future development. Such as, in terms of sensory attributes, further psychometric feature evaluation is needed.

In conclusion, clustering in unsupervised machine learning offers clinicians and researchers a practical way to leverage individual variations among persons with ASD to advance diagnosis and better understand patients at different levels on the spectrum. In this review, we studied different data types that may reveal various underlying patterns of ASD data, all of which may offer distinct and complementary information and clinical value. To consolidate the studies on unsupervised machine learning in ASD, additional research is required on other clustering techniques besides hierarchical clustering and k-means clustering. The main contributions, limits, and overall challenges of the studies discussed in this paper can serve as a model for future researchers who wish to forecast ASD meltdown. The information displayed and categorized in the tables can provide a comprehensive picture of the most recent usage of k-means and hierarchical clustering algorithms.

V. Conclusion

The primary goal of our research was to review and analyze the latest findings on two of the most extensively used unsupervised machine learning methods, "k-Means Clustering" and "Hierarchical Clustering" for detecting groups or patterns in people with Autism Spectrum Disorder (ASD) syndromes across various data types. Unsupervised ML methods have often yielded satisfactory outcomes in diagnosing ASD. In this work, different data sources may reveal various hidden structures within ASD data, all of which may offer distinct and complementary information and clinical significance. We think there is a good chance that unsupervised machine learning researchers will band together and encourage other researchers to join in the effort to recognize and address this unique issue.

Acknowledgment

P. Rivas performed part of this work while funded by the National Science Foundation under grants CNS-2136961 and CNS-2210091.

References

[1] "Diagnostic criteria," Nov 2022. [Online]. Available: https://www.cdc.gov/ncbddd/autism/hcp-dsm.html 1

[2] G. Arbanas, "Diagnostic and statistical manual of mental disorders (dsm-5)," *Alcoholism and psychiatry research*, vol. 51, pp. 61–64, 2015. 1

[3] M. J. Maenner, K. A. Shaw, A. V. Bakian, D. A. Bilder, M. S. Durkin, A. N. Esler, S. M. Furnier, L. Hallas, J. A. Hall-Lande, A. Hudson, M. M. Hughes, M. Patrick, K. J. Pierce, J. N. Poynter, A. Salinas, J. Shenouda, A. C. Vehorn, Z. Warren, J. N. Constantino, M. Dirienzo, R. T. Fitzgerald, A. Grzybowski, M. H. Spivey, S. Pettygrove, W. M. Zahorodny, A. Ali, J. G. Andrews, T. Baroud, J. Gutierrez, A. S. Hewitt, L. C. Lee, M. L. C. Lopez, K. C. Mancilla, D. McArthur, Y. D. Schwenk, A. Washington, S. Williams, and M. E. Cogswell, "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, pp. 1 – 16, 2021. 1

[4] [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders 1

[5] C. Lord, E. H. Cook, B. L. Leventhal, and D. G. Amaral, "Autism spectrum disorders," *Neuron*, vol. 28, pp. 355–363, 2000. 1

[6] M. M. Rahman, O. L. Usman, R. C. Muniyandi, S. Sahran, S. Mohamed, and R. A. Razak, "A review of machine learning methods of feature selection and classification for autism spectrum disorder," *Brain Sciences*, vol. 10, 2020. 1

[7] P. Rivas, *Deep Learning for Beginners: A beginner's guide to getting up and running with deep learning from scratch using Python*. Packt Publishing Ltd, 2020. 1

[8] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data*. AMLBook New York, 2012, vol. 4. 1

[9] "Supervised vs. unsupervised learning: What's the difference?" [Online]. Available: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning 1

[10] M. James, "Some methods for classification and analysis of multivariate observations.[in] proceedings of 5-th berkeley symposium on mathematical statistics and probability," 1967. 2

[11] T. Lingren, P. Chen, J. Bochenek, F. Doshi-Velez, P. Manning-Courtney, J. Bickel, L. W. Welchons, J. Reinhold, N. M. Bing, Y. Ni, W. J. Barbaresi, F. D. Mentch, M. A. Basford, J. C. Denny, L. Vazquez, C. L. Perry, B. Namjou, H. Qiu, J. Connolly, D. J. Abrams, I. A. Holm, B. A. Cobb, N. Lingren, I. Solti, H. Hakonarson, I. S. Kohane, J. B. Harley, and G. K. Savova, "Electronic health record based algorithm to identify patients with autism spectrum disorder," *PLoS ONE*, vol. 11, 2016. 2, 3, 10

[12] A. K. Easson, Z. Fatima, and A. R. McIntosh, "Functional connectivity-based subtypes of individuals with and without autism spectrum disorder," *Network Neuroscience*, vol. 3, no. 2, pp. 344–362, 02 2019. [Online]. Available: https://doi.org/10.1162/netn_a_00067 3, 6, 10

[13] E. Stevens, A. Atchison, L. Stevens, E. Hong, D. Granpeesheh, D. R. Dixon, and E. J. Linstead, "A cluster analysis of challenging behaviors in autism spectrum disorder," *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 661–666, 2017. 3, 6

[14] T. Obara, M. Ishikuro, G. Tamiya, M. Ueki, C. Yamanaka, S. Mizuno, M. Kikuya, H. Metoki, H. Matsubara, M. Nagai, T. Kobayashi, M. Kamiyama, M. Watanabe, K. Kakuta, M. Ouchi, A. Kurihara, N. Fukuchi, A. Yasuhara, M. Inagaki, M. Kaga, S. Kure, and S. Kuriyama, "Potential identification of vitamin b6 responsiveness in autism spectrum disorder utilizing phenotype variables and machine learning methods," *Scientific Reports*, vol. 8, 2018. 3, 6, 10

[15] K. Hyde, A. J. Griffiths, C. Giannantonio, A. E. HURLEY-HANSON, S. K. Mathur, and E. J. Linstead, "Exploring the landscape of employers for individuals with autism spectrum disorder," *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1961–1967, 2019. 3, 7, 10

[16] S. Silleresi, P. Prévost, R. Zebib, F. Bonnet-Brilhault, D. Conte, and L. Tuller, "Identifying language and cognitive profiles in children with asd via a cluster analysis exploration: Implications for the new icd-11," *Autism Research*, vol. 13, 2020. 4, 7

[17] A. Narita, M. Nagai, S. Mizuno, S. Ogishima, G. Tamiya, M. Ueki, R. Sakurai, S. Makino, T. Obara, M. Ishikuro, C. Yamanaka, H. Matsubara, Y. Kuniyoshi, K. Murakami, F. Ueno, A. Noda, T. Kobayashi, M. Kobayashi, T. Usuzaki, H. Ohseto, A. Hozawa, M. Kikuya, H. Metoki, S. Kure, and S. Kuriyama, "Clustering by phenotype and genome-wide association study in autism," *Translational Psychiatry*, vol. 10, 2019. 4, 7, 10

[18] M. V. Lombardo, M.-C. Lai, B. Auyeung, R. J. Holt, C. Allison, P. L. Smith, B. Chakrabarti, A. N. V. Ruigrok, J. Suckling, E. T. Bullmore, A. J. S. P. F. E. T. S. M. B. B. B.-C. B. B. C. Cata, A. J. Bailey, S. Baron-Cohen, P. F. Bolton, E. T. Bullmore, S. J. Carrington, M. Catani, B. Chakrabarti, M. C. Craig, E. M. Daly, S. C. L. Deoni, C. Ecker, F. Happé, J. Henty, P. Jezzard, P. J. Johnston, D. K. Jones, M.-C. Lai, M. V. Lombardo, A. Madden, D. Mullins, C. M. Murphy, D. G. M. Murphy, G. Pasco, A. N. V. Ruigrok, S. A. Sadek, D. Spain, R. Stewart, J. Suckling, S. J. Wheelwright, S. C. R. Williams, C. E. Wilson, C. Ecker, M. C. Craig, D. G. M. Murphy, F. Happé, and S. Baron-Cohen, "Unsupervised data-driven stratification of mentalizing heterogeneity in autism," *Scientific Reports*, vol. 6, 2016. 4, 7, 10

[19] M. Elwin, A. Schröder, L. Ek, T. Wallsten, and L. Kjellin, "Sensory clusters of adults with and without autism spectrum conditions," *Journal of Autism and Developmental Disorders*, vol. 47, pp. 579 – 589, 2017. 4, 8

[20] C. Pichitpunpong, S. Thongkorn, S. Kanlayaprasit, W. Yuwattana, W. Plaingam, S. Sangsuthum, W. M. Aizat, S. N. Baharum, T. Tencomnao, V. W. Hu, and T. Sarachana, "Phenotypic subgrouping and multi-omics analyses reveal reduced diazepam-binding inhibitor (dbi) protein levels in autism spectrum disorder with severe language impairment," *PLoS ONE*, vol. 14, 2019. 5, 8

[21] I. Kurochkin, E. E. Khrameeva, A. Tkachev, V. Stepanova, A. A. Vanyushkina, E. Stekolshchikova, Q. Li, D. Zubkov, P. Shichkova, T. Halene, L. Willmitzer, P. Giavalisco, S. Akbarian, and P. Khaitovich, "Metabolome signature of autism in the human prefrontal cortex." *Communications biology*, vol. 2 1, p. 234, 2019. 5, 8, 10

[22] A. M. Smith, J. J. King, P. R. West, M. A. Ludwig, E. L. R. Donley, R. E. Burrier, and D. G. Amaral, "Amino acid dysregulation metabotypes: Potential biomarkers for diagnosis and individualized treatment for subtypes of autism spectrum disorder," *Biological Psychiatry*, vol. 85, pp. 345–354, 2019. 5, 8, 10

[23] S. Zheng, K. Hume, H. Able, S. L. Bishop, and B. A. Boyd, "Exploring developmental and behavioral heterogeneity among preschoolers with

asd: A cluster analysis on principal components," *Autism Research*, vol. 13, 2020. 5, 8

[24] P. Dwyer, X. Wang, R. D. Meo-Monteil, F. Hsieh, C. D. Saron, and S. M. Rivera, "Defining clusters of young autistic and typically developing children based on loudness-dependent auditory electrophysiological responses," *Molecular Autism*, vol. 11, 2020. 5, 8

[25] E. McDougal, D. M. Riby, and M. Hanley, "Profiles of academic achievement and attention in children with and without autism spectrum disorder." *Research in developmental disabilities*, vol. 106, p. 103749, 2020. 6, 9

[26] J.-J. Paakki, J. S. Rahko, A. Kotila, M.-L. Mattila, H. Miettunen, T. M. Hurtig, K. Jussila, S. Kuusikko-Gauffin, I. K. Moilanen, O. Tervonen, and V. J. Kiviniemi, "Co-activation pattern alterations in autism spectrum disorder–a volume-wise hierarchical clustering fmri study," *Brain and Behavior*, vol. 11, 2021. 6, 9, 10

[27] F. zahra Benabdallah, A. D. E. Maliani, D. Lotfi, R. Jennane, and M. E. Hassouni, "An autism spectrum disorder adaptive identification based on the elimination of brain connections: a proof of long-range underconnectivity," *Soft Comput.*, vol. 26, pp. 4701–4711, 2022. 6, 9, 10

[28] J. C. Bullen, M. C. Zajic, N. McIntyre, E. J. Solari, and P. C. Mundy, "Patterns of math and reading achievement in children and adolescents with autism spectrum disorder," *Research in Autism Spectrum Disorders*, 2022. 6, 9, 10

[29] P. Arabie, L. J. Hubert, and J. D. Carroll, "Clustering," *Encyclopedia of Social Measurement*, p. 317–320, 2005. [Online]. Available: http://dx.doi.org/10.1016/b0-12-369398-5/00069-4 7